Protegrity Discover Guide 3.0.2.0

Created on: Aug 8, 2024

Notice

Copyright

Table of Contents

Chapter 1

Background

Data is the most important asset of any organization or business. Data can include critical information and insider secrets, including, but not limited to: user information, business data, operational data, financial information, personally identifiable information, pricing information, and intellectual property.

With the increased use of mobile and cloud applications, organizations have become susceptible to information extending beyond their network's perimeter. Customers can use such applications to bypass the organization firewalls for accessing data. The organization, its employees, its customers, and its external partners expect the data to be safeguarded. If an unauthorized user accesses the data, then it could lead to legal issues, arising out of violations of personal privacy or proprietary rights. Therefore, organizations must secure their data from ever-rising incidents of data security threats.

The challenge that revolves around data management is also observed with legacy systems, where the original implementors are no longer associated with such systems and the supporting documentation is inadequate. As a result, the present owners are not completely aware about the sensitive data that exists within such systems, which increases the overall vulnerability quotient.

To ensure that confidential data remains within the secure perimeter of the enterprise, a data discovery solution must be used to identify enterprise-wide sensitive data. Data discovery is a crucial aspect to achieve the following organizational data compliance and data regulation requirements:

- Payment Card Industry Data Security Standard (PCI DSS)
- General Data Protection Regulation (GDPR)
- Health Insurance Portability and Accountability Act (HIPAA)
- Sarbanes Oxley (SOX)
- Gramm-Leach-Bliley (GLBA)
- Protection of Personal Information Act (POPI), etc.

The sensitive data must be identified, protected, audited, and monitored for risk assessment and management.

Chapter 2

Protegrity Discover Overview

As identifying sensitive data becomes important for any business or organization, Protegrity Discover aims to provide just the right solution to your data discovery requirements. Protegrity Discover is a data discovery system that inspects, locates, and tracks sensitive information in a data-dynamic world.

> **Note:** Before you find out about how Protegrity Discover works, a reading of the commonly used terms with Protegrity Discover is recommended for better understanding. For more information about term definitions, refer to the section *Appendix A: Protegrity Discover-specific Term Definitions*.

Protegrity Discover continuously scans the configured datastores to locate any sensitive data instance residing in the cleartext format. It involves the following step-by-step process:

1. Performing one of the following steps, based on whether the data is structured or unstructured:
   - *Sampling the data* - In case of *structured* data, a chunk of data from the actual data container, which is referred to as sample data, is configured for the data discovery scan.
   - *Reading the data* - In case of *unstructured* data, Protegrity Discover scans or reads the data from the unstructured files.

   For more information about the structured and unstructured data supported by Protegrity Discover, refer to the section *Capabilities and Support*.
2. *Classifying the data* - The data classifiers use evaluation rules to identify the sensitive data.
3. *Catalog the data with additional details* - The data is cataloged with supporting information, such as, when, where, and what data, and is identified as sensitive data along with the observed and estimated count of sensitive data values. The probability of data being classified as sensitive is reported using a confidence scoring mechanism. The catalog does not include the actual sensitive data, but only the data value count.
4. *Obtaining analysis results from the performed scan* - After all classifiers generate findings, the finding with the highest score is selected as the analysis result.

The scan results help you in assessing the risk and raising awareness regarding the sensitive data that is stored in the cleartext format. The identification of the sensitive data helps you in taking measures to protect the data using any data protection method, such as, encryption, tokenization, masking, and monitoring.

With adoption of such a data discovery system, any business or organization can strive to maintain its competitive edge without compromising or risking the confidentiality of its data.

Protegrity Discover can identify the following categories of sensitive data:

- Data categorized as Personally Identifiable Information (PII)
- Data that falls under the scope of the Payment Card Industry Data Security Standard (PCI DSS) compliance

You can also create custom classifiers depending on the type of data that you want to identify.

Chapter 3

System Architecture

This section provides an overview of the architectural components of the Protegrity Discover system and describes how workload is distributed to address multiple processing requirements.

The Protegrity Discover system architecture is divided into two main components: *Orchestrator* and the *Worker*.

The following diagram illustrates a high-level overview of the Protegrity Discover system architecture with its different sub-components.

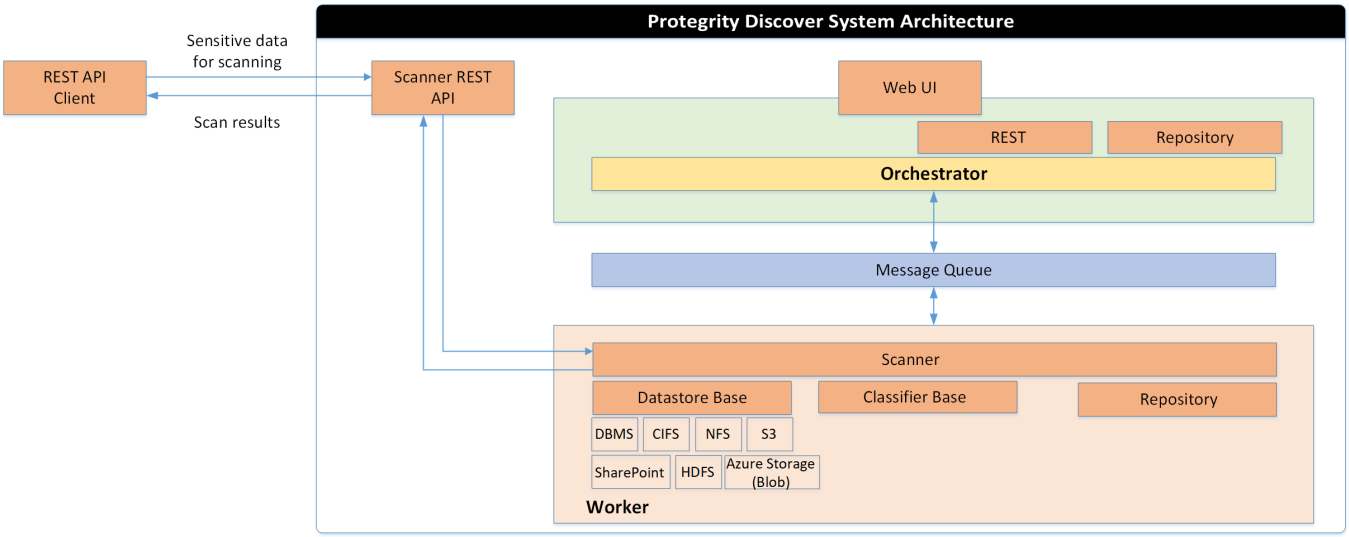


Figure 3-1: Protegrity Discover - System Architecture

Orchestrator and the Worker: The Orchestrator is the main point of entry to the application. It is a service that distributes the work to the Workers. The Workers are the component parts that actually do the work of searching and scanning a system, and using the classifiers. Classifiers validate and classify sensitive data from the sampled records of the scanned system.

The Orchestrator and the Worker communicate with each other via Message Queue, which is a part of the framework. The Orchestrator ensures that there are enough workers available and running for the system to function seamlessly. The minimum number of Workers is one. The Orchestrator dynamically adjusts the number of Workers according to the number of CPU cores, amount of free memory, and average system load.

The Orchestrator starts and stops the Workers using the following logic:

While the Scan is Running	Action taken by Orchestrator
Free memory < 1 GB OR	Shuts down a Worker

While the Scan is Running	Action taken by Orchestrator
1 minute average system load > Number of CPU cores * 2	
Free memory > 1 GB AND 1 minute average system load < Number of CPU cores * 2	Starts a new Worker

Repository: The Repository is utilized by both Orchestrator and *Scanner* processes. It is used to store metadata, results of the scan, and logs.

The information that is stored within the repository, includes the following:

- Classification and analysis details
- Job and scan details
- Datastore details and referential data
- Keytab file, which contains the list of Kerberos principals and their keys.

  For more information about the keytab file, refer to the section *Kerberos*.

Scanner: The Scanner drives the scan process with its essential elements including the Datastore Base, Classifier Base, and Repository.

Datastore Base: The Datastore Base defines the following systems that are supported for the data discovery scan:

- *Database systems:*
  - DB2/UDB
  - DB2/zOS
  - EXAsol
  - Hive
  - Microsoft SQL Server
  - MySQL
  - Oracle
  - PostgreSQL
  - Teradata
- *File-based systems:*
  - Common Internet File System (CIFS)
  - Hadoop Distributed File System (HDFS)
  - Network File System (NFS)
  - SharePoint

    > **Important:** Protegrity Discover supports only the cloud-based SharePoint Online. It does not support the locally hosted SharePoint Server.
- *Cloud storage system*
  - Amazon Web Services (AWS) S3
  - Azure Storage (Blob)

You can add additional systems to the Protegrity Discover Datastore Base. For more information about extending the support to any other system, refer to the section *Extending the Support to Other Systems*.

Classifier Base: The Classifier Base includes the following default classifiers configured and provided with the product:

- User details, such as, first name, last name, date of birth, or Credit Card Number (CCN)
- Location-specific details, such as, city, state, or country code
- Unique identifiers for any user, such as Social Security Number (SSN)
- Uniquely identified user's bank account details, such as International Bank Account Number (IBAN)

For more information about the default classifiers, refer to the *Default Classifiers* appendix.

You can add your own classifier to the Protegrity Discover Classifier Base. For more information about adding the classifier configuration, refer to the section *Managing Classifiers*.

Scanner REST API: Protegrity Discover provides a REST API that is used to scan sensitive data provided in the request body of the REST API. The Scanner REST API component internally uses the Scanner component to scan the data.

For more information about the Scanner REST API, refer to the section *Protegrity Discover REST APIs*.

Chapter 4

Installing Protegrity Discover

Install Protegrity Discover as an On-Premise system or a cloud system, using the information provided in this section.

4.1 Hardware Requirements

This section describes the hardware requirements for the Protegrity Discover.

Before you proceed with installing Protegrity Discover, ensure that the platform requirements are met. The following table lists the minimum and recommended hardware requirements for installing Protegrity Discover:

Component	Minimum Hardware Requirements	Recommended Hardware Requirements
Processor*1	4 Core x86 64-bit processor	8 Core x86 64-bit processor
Hard Disk	200 GB	400 GB
RAM	16 GB	32 GB
Network Interface	1 Network Interface	1 Network Interface

*1 - The processor must support Advanced Vector Extensions (AVX), which are extensions to the x86 instruction set architecture.

> **Note:** The actual hardware configuration depends on the actual usage or amount of data and logs expected.

For any additional system requirements, refer to the section *System Requirements* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

4.2 Installing Protegrity Discover On-Premise

The Protegrity Discover installation on-premise follows a series of steps, which are listed in this section.

To install Protegrity Discover:

1. Insert or mount the Protegrity Discover installation media, DVD or ISO into your system disk drive.
2. Restart the machine and ensure that it boots up from the installation media.

The following splash screen appears.

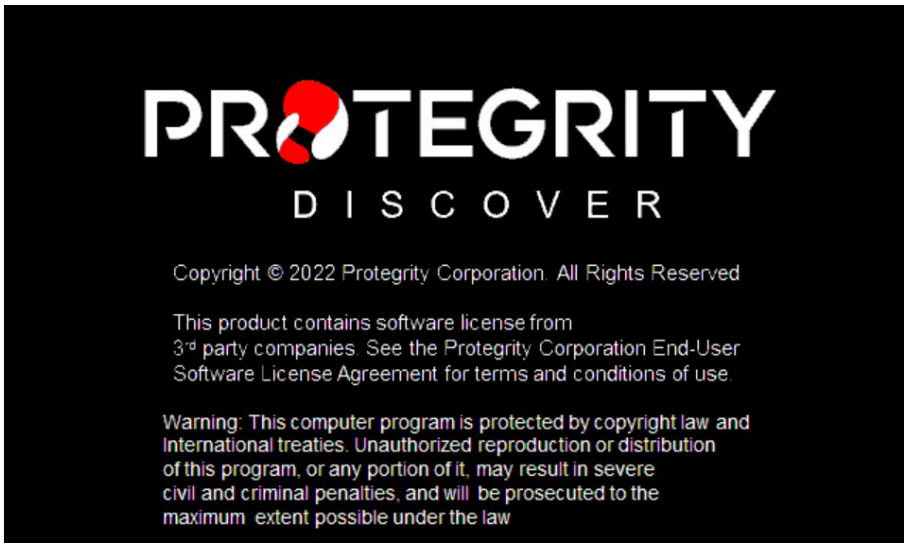


Figure 4-1: Protegrity Discover - Splash Screen

3. Press **ENTER** to start the installation procedure.
   The following screen with installation options appears.

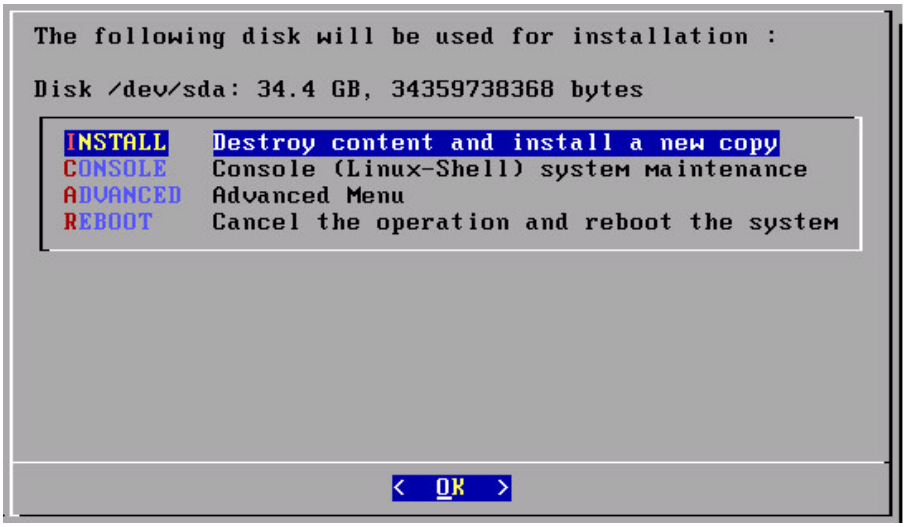


Figure 4-2: Protegrity Discover - Installation Options

4. Using the arrow and tab keys, select **INSTALL** and press **ENTER**.
   The following dialog appears for the DHCP settings.

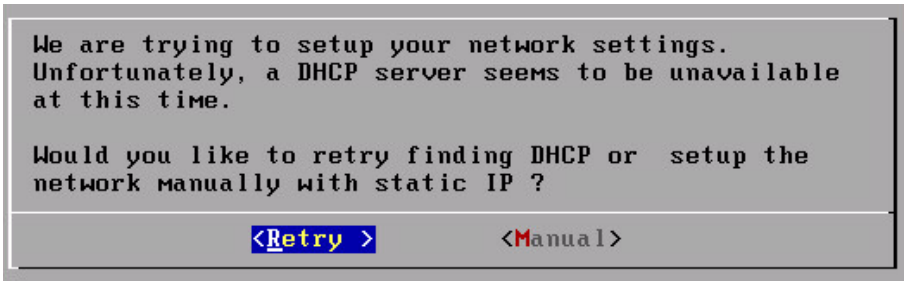


Figure 4-3: Protegrity Discover - Installation DHCP/Static IP Settings

5. Select **Manual** to enter the IP address manually or **Retry** to try finding the DHCP server.
   If you select **Manual**, then the following **Network Configuration Information** dialog appears.

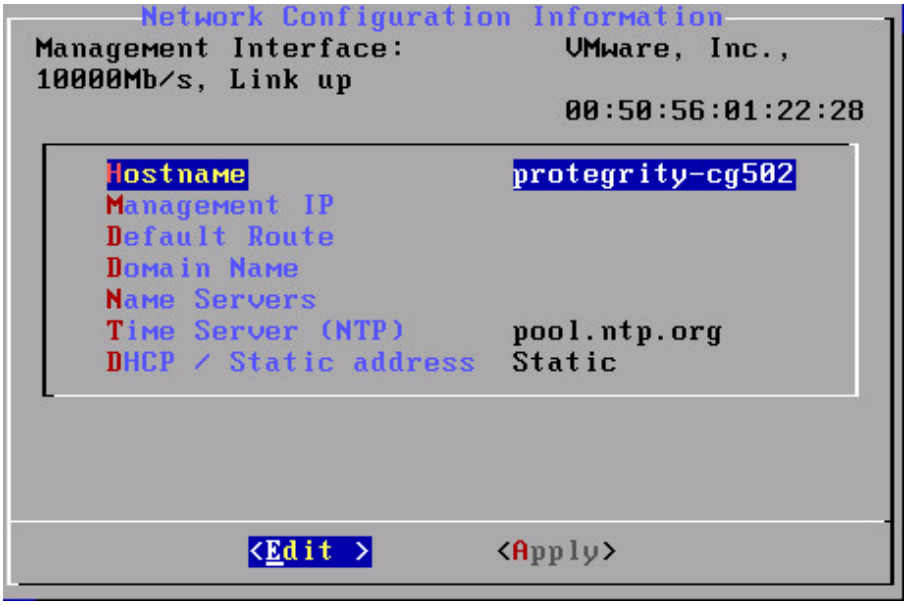


Figure 4-4: Protegrity Discover - Network Configuration Information

6. Enter the network configuration settings and select **Apply** to continue.

> **Note:** Protegrity Discover setup verifies the settings provided for **Management IP**, **Default Route**, **Name Servers** and **Time Server** before the installation proceeds.

7. Select the time zone of the host. Select **Next** and press **ENTER**.

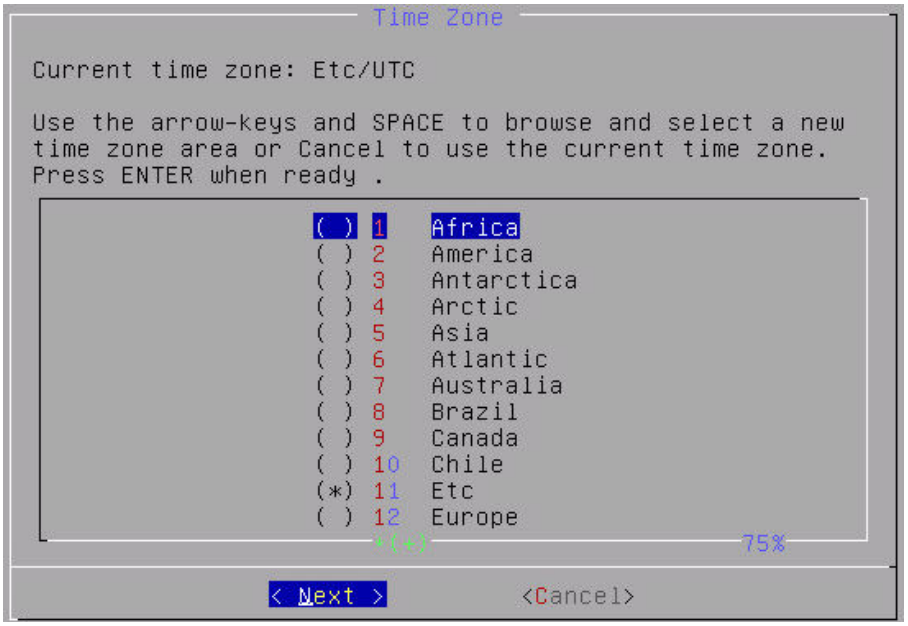


Figure 4-5: Protegrity Discover - Set Time Zone

8. Select the nearest location. Select **OK** and press **ENTER**.

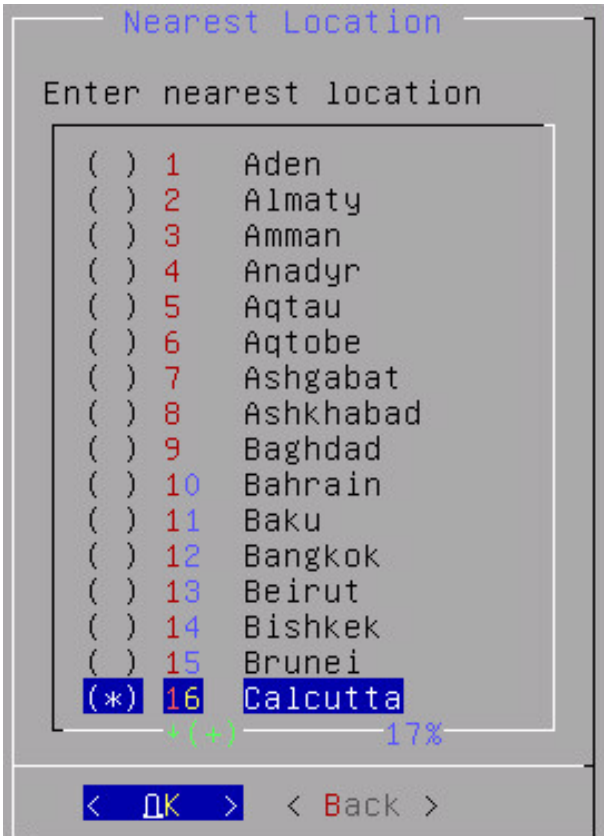


Figure 4-6: Protegrity Discover - Set nearest location

9. Review and update the initial server settings. Select **OK** and press **ENTER** to continue.

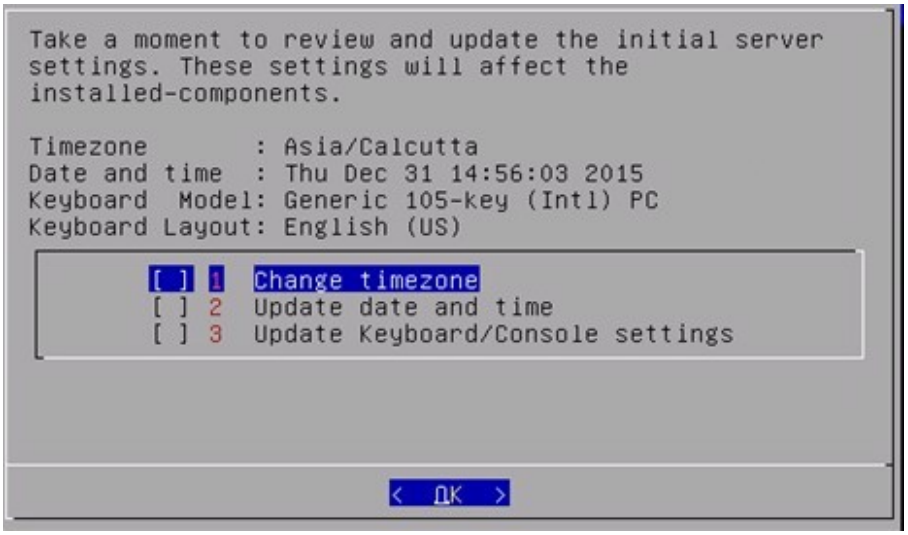


Figure 4-7: Protegrity Discover - Initial Server Settings

10. Configure the GRUB (GRand Unified Bootloader) settings.

    In the Protegrity appliances, GRUB version 2 (GRUB2) is used for loading the kernel. If you want to protect the boot configurations, then you can secure it by enforcing a username and password combination for the GRUB menu.

Figure 4-8: Protegrity Discover - GRUB Settings

Press **spacebar** to select **Enable** to display the **GRUB Credentials** screen. You can then specify the username and password for the GRUB credentials.

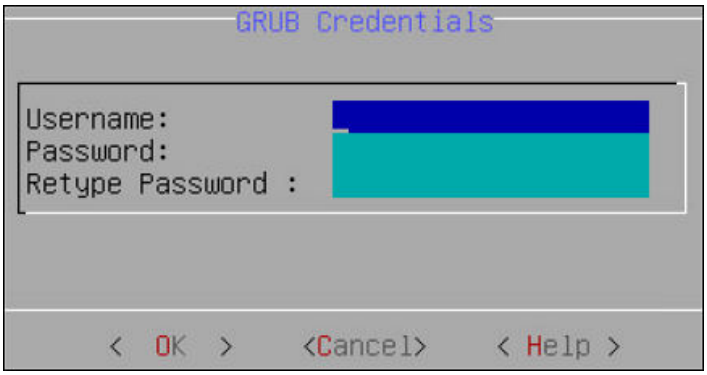
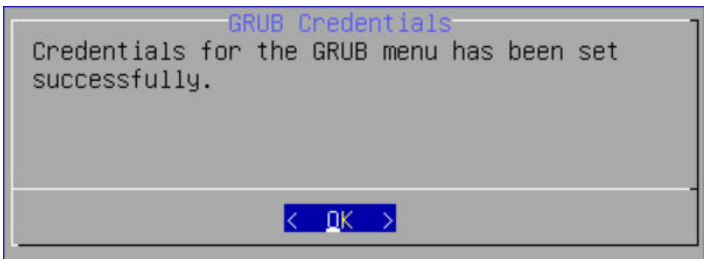


Figure 4-9: Protegrity Discover - GRUB Credentials screen

The following screen appears after you have configured the GRUB credentials successfully. Select **OK** to continue.



For more information about configuring the GRUB settings, refer to the section *Configuring GRUB Settings* in the *Protegrity Installation Guide 9.2.0.0*.

For more information about GRUB, refer to the section *Securing the GRand Unified Bootloader (GRUB)* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

11. Configure the passwords for the default user, and select **Apply** to continue.

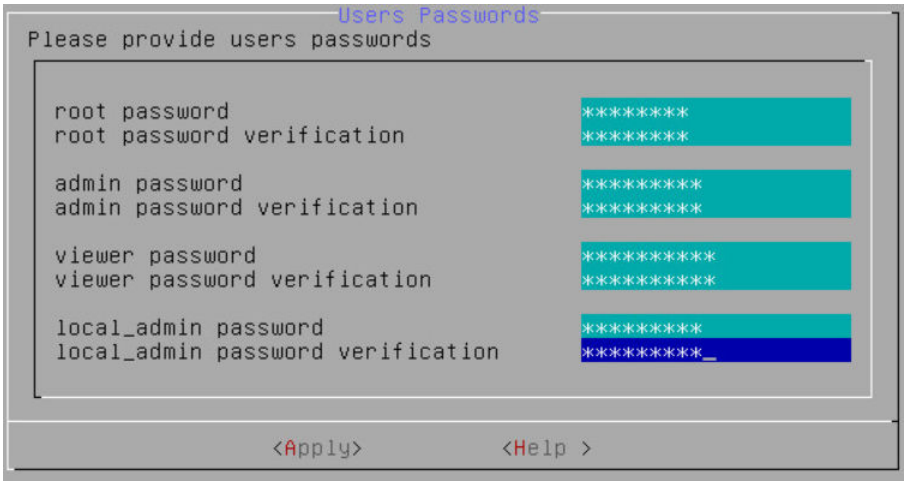
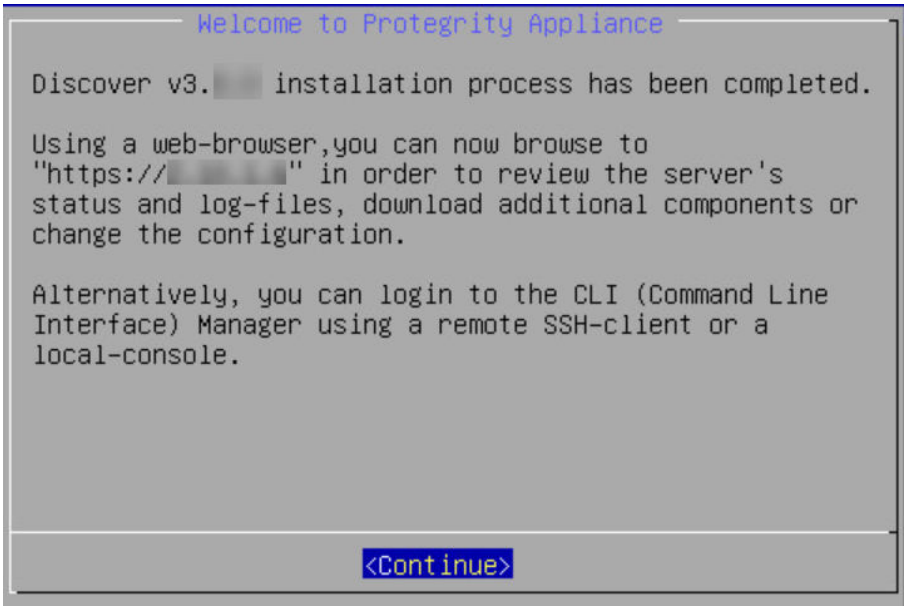


Figure 4-10: Protegrity Discover - Set User Passwords

12. The **Welcome to Protegrity Appliance** dialog appears. Select **Continue** to complete the Protegrity Discover installation.



To access the Protegrity Discover Web UI:

13. Using a web browser, navigate to the Management IP address for Protegrity Discover using the HTTPS protocol.

    This is the same IP address that appears on the **Welcome to Protegrity Appliance** dialog box.

    The following Protegrity Discover Web UI login screen appears.

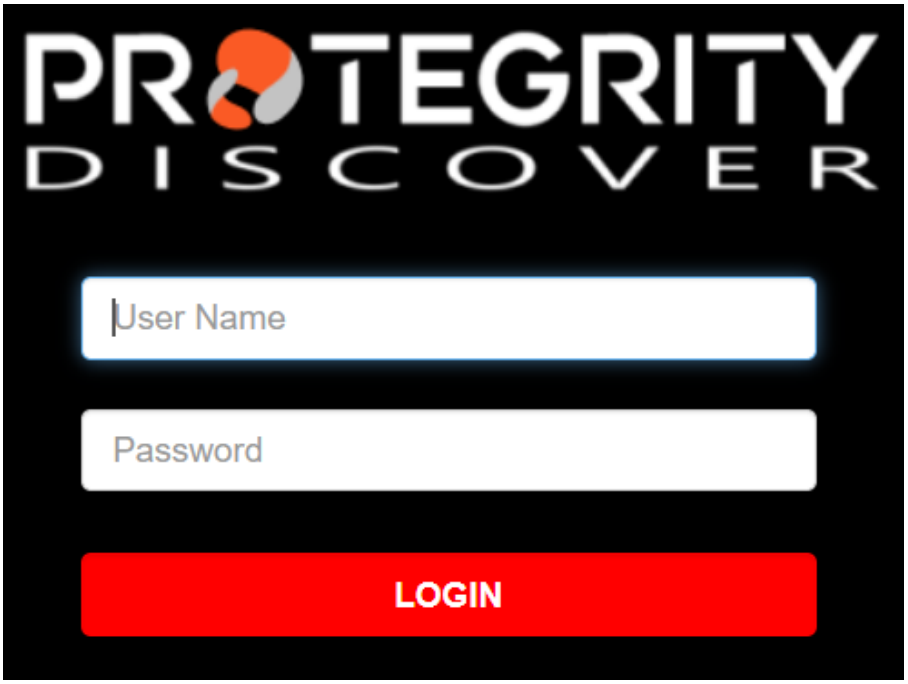


Figure 4-11: Protegrity Discover Web UI Login Screen

> **Note:** If you want to access the Appliance Web UI, then click the icon on the top-right corner of the Login screen.
> For more information about the Appliance Web UI, refer to the section *Managing the Appliance Information*.

14. Enter the user credentials for the *admin* or *viewer* user, and click **LOGIN**.
    The Protegrity Discover Web UI appears.

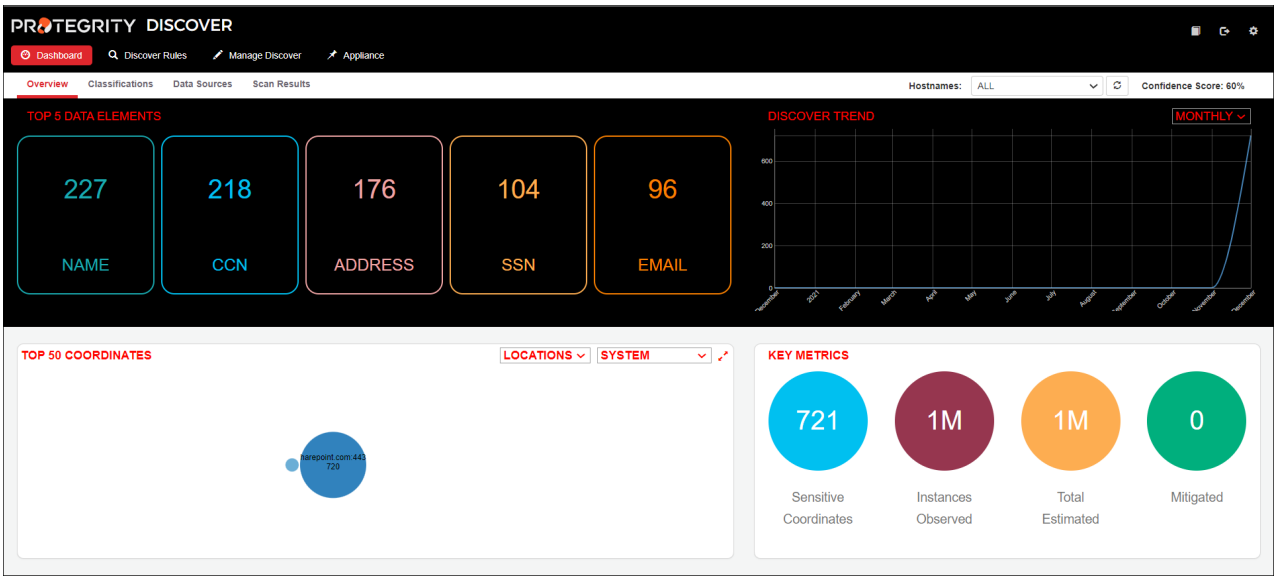


Figure 4-12: Protegrity Discover Web UI

For more information about the Protegrity Discover Web UI, refer to the section *Protegrity Discover Web UI*.

If you want to access the Protegrity Discover online documentation, then click the **Documentation** icon on the top-right corner of the Web UI. The online documentation consists of the following documents:

- Protegrity Discover Guide 3.0.2.0
- Protegrity Appliances Overview Guide 9.2.0.0

If you want to log out of the Protegrity Discover Web UI, then click the **Logout**  icon on the top-right corner of the Web UI.

> **Note:** The login requests are sent to the Service Dispatcher, which is an Apache Multi-Processing Module (MPM) Worker, that is used to route the requests between various Protegrity Discover components. If you want to view any logs related to authentication or authorization, then you need to view the Service Dispatcher logs.
>
> For more information about viewing the Service Dispatcher logs, refer to the section *Viewing Service Dispatcher Logs* in the *Protegrity Appliances Overview Guide 9.2.0.0*.
>
> For more information about Apache MPM Worker, refer to *https://httpd.apache.org/docs/2.4/mod/worker.html*.

4.3 Installing Protegrity Discover on Cloud Platforms

This section describes installing the Protegrity Discover on Cloud platforms, such as, AWS, Azure, or GCP. For installing the Protegrity Discover on cloud platforms, you must mount the image containing the Protegrity Discover on a cloud instance or a virtual machine. After mounting the image, you must run the finalization procedure to install the appliance components.

The following steps must be completed to run Protegrity Discover on a cloud platform:

1. *Configure the cloud instance*
2. *Finalize installation*

4.3.1 Configuring Cloud Instances

Before mounting the Protegrity Discover images on cloud platforms, you must create virtual machines or instances on them. This procedure involves configuring network settings on cloud platforms, creating accounts, containers, and so on.

The process to create instances varies between the different platforms, such as, AWS, Azure, and GCP.

For more information about installing Protegrity Discover on AWS, refer to the section *Appendix: Using Protegrity Appliances on Amazon Web Services (AWS)* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

For more information about installing Protegrity Discover on Azure, refer to the section *Appendix: Using Protegrity Appliances on Azure* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

For more information about installing Protegrity Discover on GCP, refer to the section *Appendix: Using Protegrity Appliances on Google Cloud Platform (GCP)* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

> **Note:** By default, when you install Protegrity Discover, the license is expired. You need to rotate the appliance OS keys to renew the license. This procedure also enables you to change the passwords and create complex passwords for the *root*, *admin*, *local_admin*, and *viewer* accounts.
>
> For more information about rotating the appliance OS keys, refer to the section *Finalizing the Installation of Protegrity Appliance*.

4.3.2 Finalizing the Installation of Protegrity Appliance

When you install the appliance, it generates multiple security identifiers, such as, keys, certificates, secrets, passwords, and so on. These identifiers ensure that sensitive data is unique between two appliances on a network. When you receive a Protegrity

appliance image, the identifiers are generated with certain values. If you use the security identifiers without changing their values, then security is compromised and the system might be vulnerable to attacks. Using the **Rotate Appliance OS Keys** tool, you can randomize the values of these security identifiers for an appliance. During the finalization process, you run the key rotation tool to secure your appliance.

> **Important:** You must rotate the keys to renew the license.

The following sensitive data can be configured:

- Hostname
- Appliance certificates
- Service credentials
- SSH Keys
- Appliance security identifiers

4.3.2.1 Finalizing Protegrity Discover Installation

You can finalize the installation of Protegrity Discover after signing in to the CLI Manager.

> **Caution:** Ensure that the finalization process is initiated from a single session only. If you start finalization simultaneously from a different session, then the *Finalization is already in progress.* message appears. You must wait until the finalization of the instance is successfully completed.
>
> Additionally, ensure that the appliance session is not interrupted. If the session is interrupted, then the instance becomes unstable and the finalization process is not completed on that instance.

> **Important:** If you have installed Protegrity Discover on AWS, then the **Authentication Type** for SSH is set to **Publickey** by default. Ensure that you login to the CLI Manager for the first time using the *local_admin* user and the *Public key*. You can change the authentication type from the Protegrity Discover Web UI, after the installation is finalized.

➤ To finalize Protegrity Discover installation:

1. Sign in to the CLI Manager of the instance created using the administrator credentials.

   > **Important:** If you have installed Protegrity Discover on AWS, then perform the SSH operation on the AWS instance using the key pair utilizing the following command.
   >
   > ```
   > ssh -i <path of the private key pair> local_admin@<IP address of the AWS instance>
   > ```
   >
   > Ensure that you use the *local_admin* user to perform the SSH operation.

   The following screen appears.

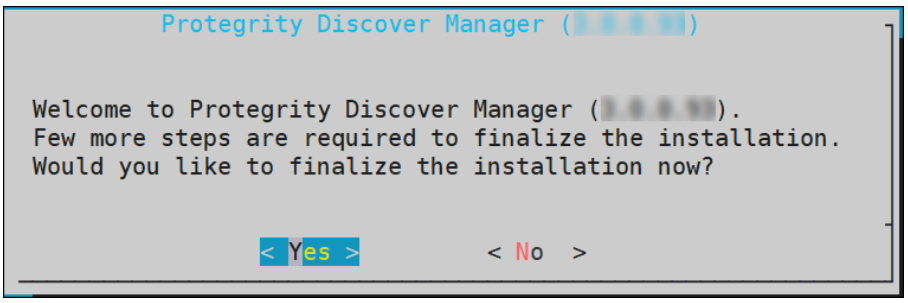


Figure 4-13: Finalizing Installation Confirmation screen

2. Select **Yes** to initiate the finalization process.
   The screen to enter the administrative credentials appears.

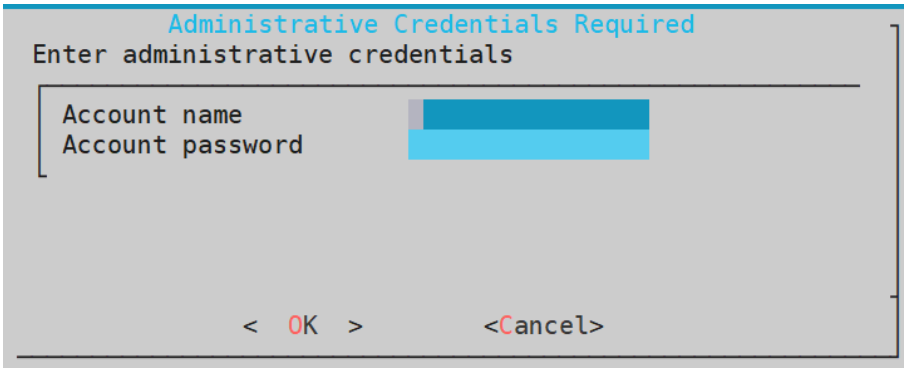


Figure 4-14: Entering Administrative Credentials

> **Important:** If you have logged in to the AWS instance using the *local_admin* user and public key, then the **Administrative Credentials Required** screen does not appear. Instead, the *Appliance OS Key Rotation Screen* screen appears.

> **Note:** If you select **No**, then the finalization process is not initiated.
> To manually initiate the finalization process, you can navigate to **Tools** > **Finalize Installation** and press **ENTER**.

3. Enter the administrative credentials and select **OK**.
   The following screen appears.

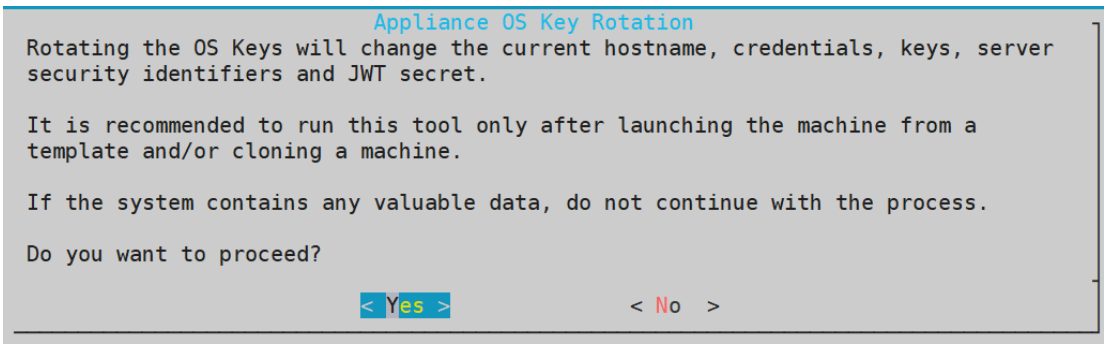


Figure 4-15: Appliance OS Key Rotation Screen

> **Note:** If you select **No**, then the appliance OS Keys will not be rotated.
> For more information about rotating the Appliance OS Keys, refer to *Running the Appliance-Rotation-Tool*.

4. Select **Yes** to rotate the Appliance OS Keys.
   The following screen appears.

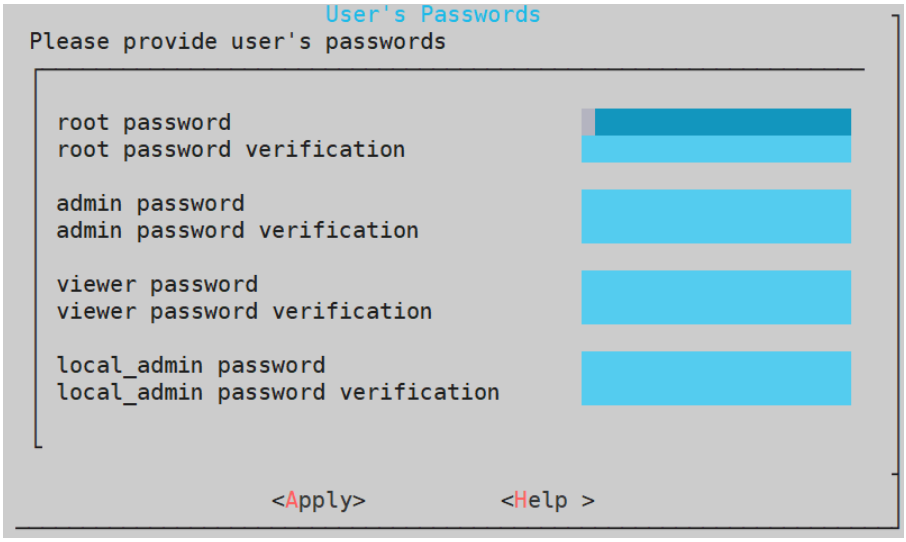


Figure 4-16: User's Passwords Screen

5. Provide the credentials for the following users:
   - root
   - admin
   - viewer
   - local_admin
6. Select **Apply**.
   The user passwords are updated and the finalization process is completed. The Protegrity Discover license is also renewed.

4.3.3 Running the Appliance-Rotation-Tool

This section describes how to run the *Appliance-rotation-tool*.

The *Appliance-rotation-tool* modifies the required keys, certificates, credentials, and passwords for the appliance to differentiate the sensitive data on the appliance from other similar instances.

> **Note:** If you are configuring a Protegrity Discover appliance instance, then you must run the *Appliance-rotation-tool* after creating the instance of the appliance.

> **Note:** Ensure that you do not run the appliance rotation tool when the appliance OS keys are in use.
> For example, you must not run the appliance rotation tool when two-factor authentication is enabled, external users are enabled, and so on.

Perform the following steps to rotate the required keys, certificates, credentials, and passwords for the appliance.

➤ To rotate keys, certificates, credentials, and passwords for the Protegrity appliance:

1. Using SSH, connect to the machine where you have installed Protegrity Discover.
2. On the Protegrity Discover CLI, navigate to **Protegrity Discover Manager** > **Tools** > **Rotate Appliance OS Keys**.

The root password dialog box appears.

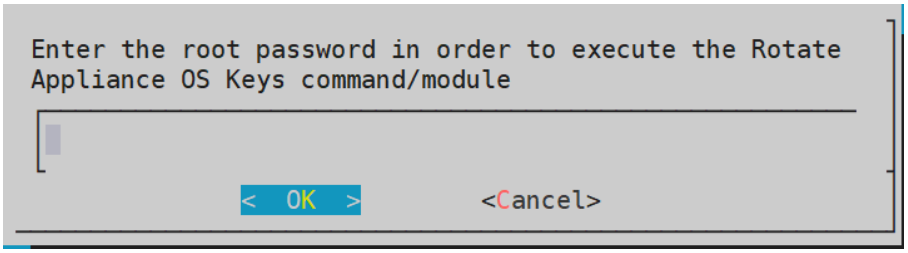


Figure 4-17: Root Password Dialog Box

3. Enter the appliance root password.
4. Select **OK**.
5. Press **ENTER**.

   The **Appliance OS Key Rotation** dialog box appears.

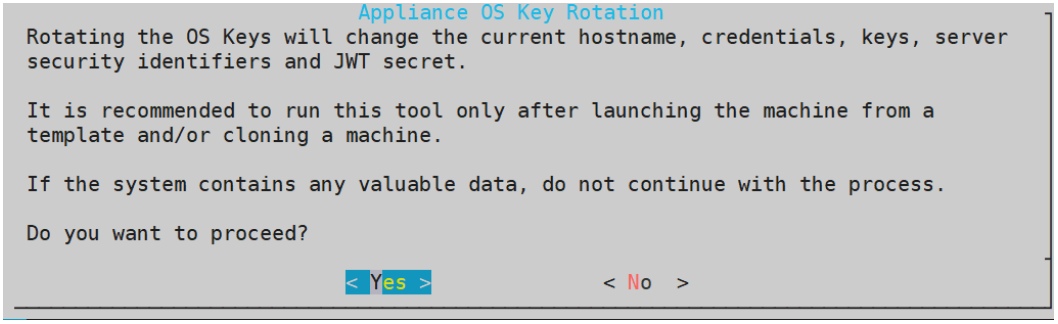


Figure 4-18: Appliance OS Key Rotation Dialog box

6. Select **Yes**.
7. Press **ENTER**.

   The administrative credentials dialog box appears.

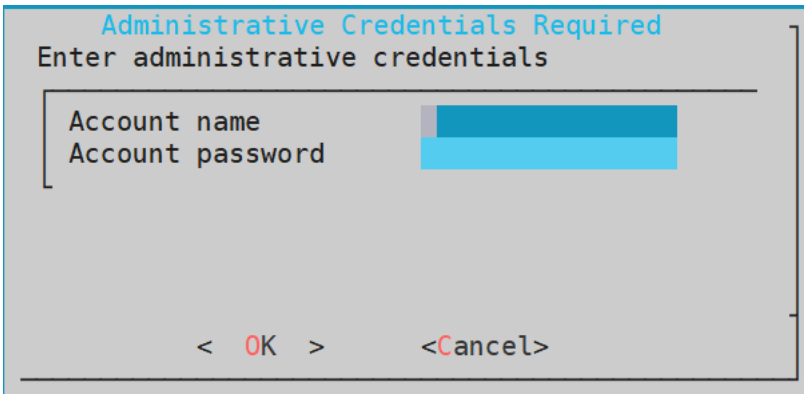


Figure 4-19: Administrative Credentials Dialog box

8. Enter the required Account name on the appliance.
9. Enter the required Account password on the appliance.
10. Select **OK**, and press **ENTER**.

    The following screen appears.

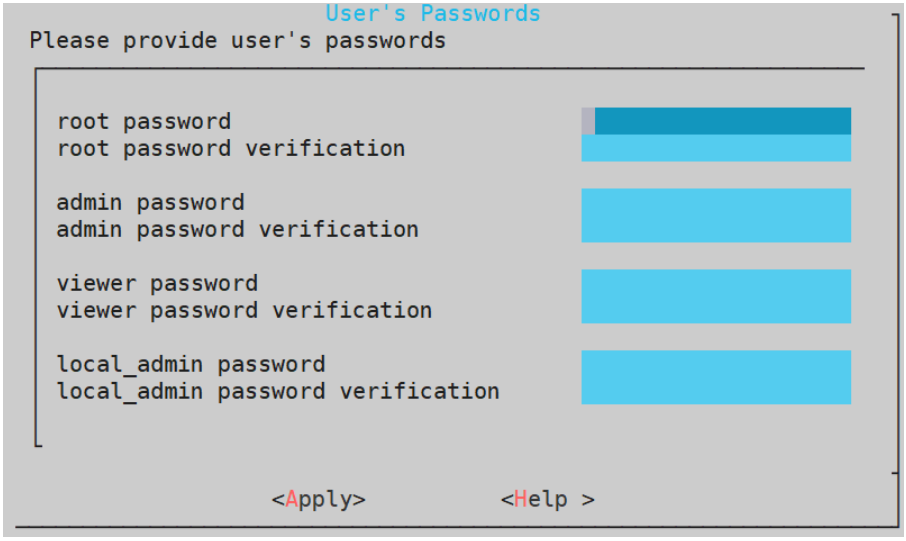


Figure 4-20: User's Passwords Screen

11. Provide the credentials for the following users:
    - root
    - admin
    - viewer
    - local_admin
12. Select **Apply**, and press ENTER.

    The process to rotate the required keys, certificates, credentials, and other identifiers on the appliance starts.

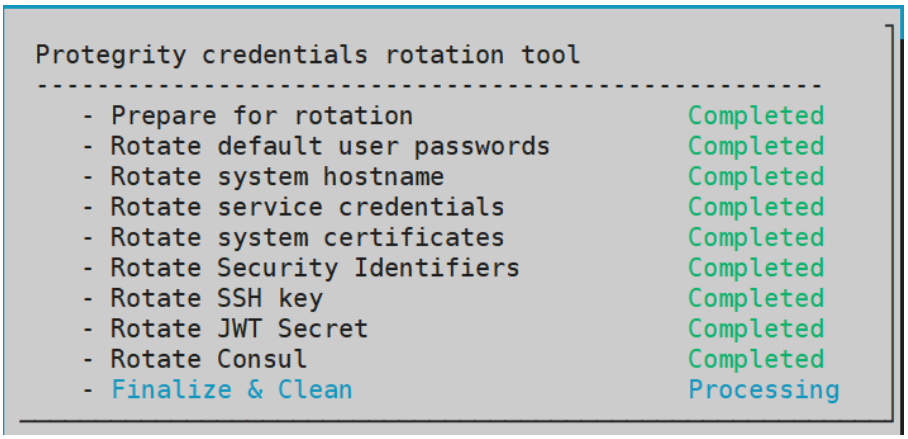


Figure 4-21: Protegrity Credentials Rotation Tool Status screen

After the parameters are rotated, the **Tools** menu appears.

4.4 Migrating Protegrity Discover Data from Version 3.0.0.0 to 3.0.2.0

This section describes how you can migrate the Protegrity Discover data, such as, the repository and classifier data, from version 3.0.0.0 to 3.0.2.0.

➤ To migrate Protegrity Discover data:

1. Login as *root* user to the OS Console on the machine where you have installed Protegrity Discover version 3.0.2.0.

For more information about accessing the OS Console, refer to the section *Accessing the OS Console* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

2. Navigate to the following directory.
   `/opt/protegrity/sureloc/backup_restore`

   The *backup_restore* directory contains the following files:

   - *backup_restore.pyc* - Script for restoring the Protegrity Discover 3.0.0.0 data
   - *readme.txt* - Text file specifying the steps to backup and restore the Protegrity Discover data
   - *compiled3.6/backup_restore.pyc* - Script for backing up the Protegrity Discover 3.0.0.0 data
3. Copy the *compiled3.6/backup_restore.pyc* file to the `/opt/protegrity/sureloc/backup_restore/` directory on the machine where you have installed Protegrity Discover version 3.0.0.0.

   Important: If you have deployed Protegrity Discover 3.0.2.0 on AWS, then before copying the files from a Discover 3.0.2.0 machine to a Discover 3.0.0.0 machine, ensure that the SSH Authentication type is set to **Password** and the SSH mode is set to **Open**.

   After copying the files, you can revert the permissions.

   For more information about setting the authentication, refer to the section *Working with Secure Shell (SSH) Keys* in the Protegrity Appliances Overview Guide 9.2.0.0.

4. Perform the following steps to backup the Protegrity Discover data.
   a. Login as *root* user to the OS Console on the machine where you have installed Protegrity Discover version 3.0.0.0.
   b. Navigate to the `/opt/protegrity/sureloc/backup_restore/` directory.
   c. Run the following command to setup the Python3 virtual environment.
      ```
      source /opt/protegrity/sureloc/python3/bin/activate
      ```
   d. Run the following command to set the *PYTHONPATH* environment variable.
      ```
      export PYTHONPATH=/opt/protegrity/sureloc/bin
      ```
   e. Run the following command to execute the backup script.
      ```
      python backup_restore.pyc --create-backup
      ```
      Alternatively, you can also run the following command to backup the data.
      ```
      python backup_restore.pyc -b
      ```
      You are prompted to enter a password for securing the data.

      Note: Run the `python backup_restore.pyc -h` command to view the help options.

   f. Type the password, and then press *ENTER*.
      A file named *discover.pdb*, which contains the backup data, is created in the `/opt/protegrity/sureloc/backup_restore/` directory.
5. Copy the *discover.pdb* file to the machine where you have installed Protegrity Discover version 3.0.2.0.
6. Perform the following steps to restore the Protegrity Discover backed up data.
   a. Login as *root* user on the machine where you have installed Protegrity Discover version 3.0.2.0.
   b. Navigate to the `/opt/protegrity/sureloc/backup_restore` directory.
   c. Run the following command to set the Python3 virtual environment.
      ```
      source /opt/protegrity/sureloc/python3/bin/activate
      ```
   d. Run the following command to setup the *PYTHONPATH* environment variable.
      ```
      export PYTHONPATH=/opt/protegrity/sureloc/bin
      ```

e. Run the following command to execute the restore script.

```
python backup_restore.pyc --restore-backup --backup-file=<Full path of the discover.pdb file>
```

Alternatively, you can also run the following command to backup the data.

```
python backup_restore.pyc -r -f=<Full path of the discover.pdb file>
```

You are prompted to enter the password that you have specified in *step 4f*.

f. Type the password, and then press *ENTER*.

If the restore process is successful, then the following message appears on the console.

```
INFO Repository restore Successful.. !
   INFO Please restart all discover services.. !
```

7. Restart the Protegrity Discover services.

For more information about restarting Protegrity Discover services, refer to *To manage system services* in the section *Managing the Appliance Information*.

8. Login to the Protegrity Discover 3.0.2.0 Web UI and verify whether the data from Protegrity Discover version 3.0.0.0 is available.

Chapter 5

Capabilities and Support

Protegrity Discover is bundled with out-of-the-box capabilities to search sensitive data present in the cleartext format within a wide range of systems.

Protegrity Discover supports identification of sensitive data from structured and unstructured data.

Data Discovery from Structured Data

The following systems are supported, by default:

- DB2/UDB
- DB2/zOS
- EXAsol
- Hive
- Microsoft SQL Server
- MySQL
- Oracle
- PostgreSQL
- Teradata

The support can also be extended to any Relational Database Management System (RDBMS), which is equipped with its own Open Database Connectivity (ODBC) driver. The ability to support different database systems is possible by uploading the specific ODBC driver and updating the ODBC configuration file settings.

> **Note:** The driver files for the Hive system are not packaged with Protegrity Discover.

For more information about extending the Protegrity Discover support to other systems, refer to the section *Extending the Support to Other Systems*.

Protegrity Discover can retrieve the database schemas, which are a part of the RDBMS, such as database names, tables names, or column names. It can sample the data based on this retrieved schema that considers a sample of records from the database table.

Data Discovery from Unstructured Data

Sensitive data is identifiable in both unstructured and semi-structured file formats.

The unstructured file formats include text files, Microsoft Excel files, Microsoft Word files, PDF documents, and images. The semi-structured file formats include XML, JSON, SAS, CSV, Apache Avro, and Apache Parquet.

For more information on the file formats supported by Protegrity Discover, refer to the section *Supported File Formats*.

Protegrity Discover can scan sensitive data from the following image types:

- Check
- Identification cards
- Flowcharts
- Face
- Credit card numbers
- Documents, which are images containing alphabetic characters. For example, documents include letters and notices.

The following mediums are supported for data discovery from unstructured data:

- AWS S3
- Azure Storage (Blob)
- Common Internet File System (CIFS)
- Hadoop Distributed File System (HDFS)
- Network File System (NFS)
- SharePoint

> **Important:** Protegrity Discover supports only the cloud-based SharePoint Online. It does not support the locally hosted SharePoint Server.

The final confidence on the sensitive data classification can be boosted based on the following keywords:

- User details, such as, first name, last name, date of birth, or Credit Card Number (CCN)
- Location-specific details, such as, city, state, or country code
- Unique identifiers for any user, such as Social Security Number (SSN)
- Uniquely identified user's bank account details, such as International Bank Account Number (IBAN)

5.1 Extending the Support to Other Systems

This section describes how you can extend the Protegrity Discover support to other system types.

If you want to extend the Protegrity Discover support to a specific database type, then ensure that the database has an ODBC driver for 64-bit Linux or 64-bit Debian, if specifically mentioned. You must add the database-specific ODBC driver to the Protegrity Discover ODBC setup. The supported driver package formats are *.deb*, *.tgz*, *.tar.gz*, and *.tar*.

To add the ODBC driver to the Protegrity Discover ODBC setup, you must perform the following steps:

1. Upload the ODBC driver and add the system configuration settings for the ODBC driver.

   For more information about uploading the ODBC driver and configuring the ODBC settings, refer to the section *ODBC Setup and System Configuration Settings*.

2. Specify additional settings, such as, connection, sampling, encoding, and query templates, for the required system. You can configure these settings through **Datastore** screen from the Web UI.

   For more information about adding the settings through the **Datastore** screen, refer to the section *Adding a New Datastore*.

5.1.1 ODBC Setup and System Configuration Settings

You must upload the ODBC driver and configure the system configuration settings through the Web UI.

Uploading the ODBC driver:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **ODBC**.
2. Click **Upload Driver**.
   The **ODBC Driver Upload** dialog box appears.

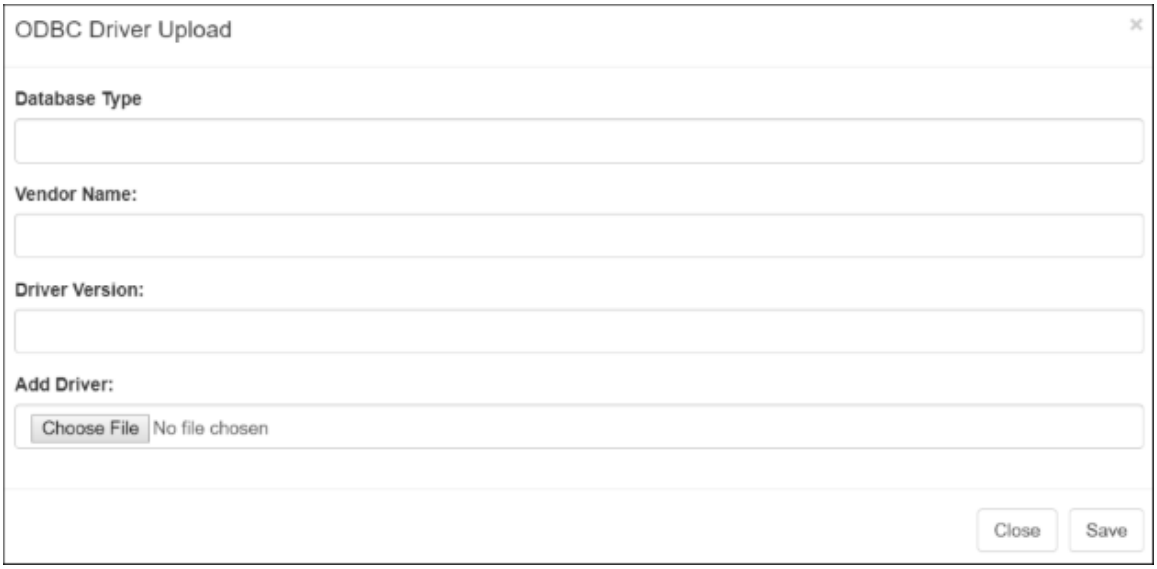


Figure 5-1: ODBC Driver Upload

3. Enter the following details.

Table 5-1: ODBC Driver Upload

Values	Description
Database Type	Type of the database. For example, Hive or MariaDB.
Vendor Name	Name of the database vendor. For example, Cloudera or Hortonworks.
Driver Version	Version number of the driver.

Important: The values that you specify for the database type, vendor name, and driver version are concatenated to create:

- the section heading for the specific driver in the ODBC INI (*odbcinst.ini*) file.
  For more information about the structure of the ODBC INI file, refer to the section *ODBC INI File Structure*.
- the path where the driver is uploaded to the Protegrity Discover machine.
  For example, if you have specified the database type as *MariaDB*, vendor name as *MariaDB_Foundation*, and the version number as *3.0.8*, then the ODBC driver is automatically uploaded to the `/opt/odbc/MariaDB/MariaDB_Foundation/3.0.8/lib/` location on the Protegrity Discover machine.

4. Click **Choose File**, and then select the ODBC driver package from your local machine.
   The supported driver package formats are *.deb*, *.tgz*, *.tar.gz*, and *.tar*. Ensure that the driver package contains at least one *.so* driver file.
5. Click **Save**.
   The **Choose the appropriate driver** dialog box appears, displaying the list of available drivers.
   - If no valid driver file is found in the package, then **`No driver file found!`** message is displayed.

- If the driver already exists, then a message is displayed prompting you to override the driver. Click **Yes**.

6. Select the required driver, and then click **Save**.
   The following message appears on the screen:

   ```
   ODBC ini file has been updated
   ```

7. Click **OK** or **Close** to close the **ODBC Driver Upload** dialog box.

Configuring the ODBC INI (odbcinst.ini) file:

8. Click ✎ next to the **ODBC INI File** text on the **ODBC** screen to add the configuration settings for the particular ODBC driver.

   By default, for any new ODBC driver that you have uploaded, the following lines are automatically added to the ODBC INI file, which include the section heading and the Driver-Path key-value pair.

   ```
   [Section Heading]
   Driver=Path
   ```

   For more information about the structure of an ODBC INI file, refer to the section *ODBC INI File Structure*.

   For example, if you have uploaded an ODBC driver for MariaDB, and you have specified the **Database Type** as *MariaDB*, the **Vendor Name** as *MariaDB_Foundation*, and the **Driver Version** as *3.0.8*, then the ODBC INI file displays the following updated entry:

   ```
   [MariaDB-MariaDB_Foundation-3.0.8-1]
   Driver=/opt/odbc/MariaDB/MariaDB_Foundation/3.0.8/lib/libmaodbc.so
   ```

   The section heading for a new ODBC driver is created by concatenating the values specified in the **ODBC Driver Upload** dialog box as follows:

   <Database Type>-<Vendor Name>-<Driver Version>-<List number of the driver>

   The list number identifies the driver number. For example, if the driver package contains two *.so* driver files, and you select the second driver file, then 2 is concatenated to the section heading. If only one driver file is present, then *1* is concatenated to the section heading.

   Important: You must specify the same section heading in the **Driver** setting of the connection string that you define while adding a new datastore. If the value specified in the **Driver** setting does not match the section heading of the uploaded ODBC driver, then Protegrity Discover cannot recognize the specific ODBC driver for the corresponding datastore, and will not be able to communicate with the datastore.

   For more information about adding a new datastore, refer to the section *Adding a New Datastore*.

   Important: Each ODBC driver can require additional configuration keys, apart from the default *Driver* key that is added to the ODBC INI file. For information on the additional keys and their required values, refer to the documentation from the driver vendor.

   Note: In case you upload a Hortonworks or Cloudera ODBC driver for Apache Hive, then Protegrity Discover automatically creates a softlink or symbolic link at the */usr/lib/hive* or */opt/cloudera* locations respectively. These softlinks always point to the path of the latest ODBC driver that you have uploaded. Therefore, if you want to re-use an ODBC driver that you have previously uploaded, then you must re-upload the driver.

5.1.2 Managing Datastores

This section describes how you can add, modify, and delete datastores.

5.1.2.1 Adding a New Datastore

This section describes how you can add a new datastore, and configure the ODBC and datastore settings for the new datastore through the **Datastore** screen from the **Appliance** menu.

Before you begin

The datastore base defines the different systems that are supported for the data discovery scan.

> **Important:** You can add a datastore only for database systems, and not for file-based or cloud storage systems.

> **Important:** You can also update the datastore settings by modifying the *datastore.json* system file. You can access this system file from the Appliance Web UI.
>
> For more information on modifying the Protegrity Discover system files, refer to the section *Managing the Appliance Information*.

➤ To add a new datastore:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **Datastore**.
2. Click **+**.
   The **Add ODBC Datastore** dialog box appears.

Figure 5-2: Protegrity Discover - Add ODBC Datastore

3. Enter the name of the datastore in the **Datastore Name** text box.

> **Note:** It is recommended to set a datastore name with no spaces. If a space is found, then it is replaced with an underscore.

4. Click **Next**.
   The **Connection String** text box appears.

Figure 5-3: Connection String

5. In the **Connection String** text box, enter the connection string used to connect to the datastore.

   Important: The connection string and its parameters are specific for each ODBC driver that is used to connect to the required datastore.

   For information regarding the connection string parameters and their default values, refer to the documentation provided by the vendor of the specific ODBC driver.

   You can use the existing connection templates for the default ODBC datastores that are provided in the **Connection Template** text box on the **Connection Settings** tab as a reference.

   For example, the following snippet displays the default connection string provided in the **Datastore** > **Connection Settings** > **Connection Template** section for connecting to the Hive database:

   ```
   Driver={Hive};HOST=%(hostname)s;PORT=%
   (port)i;HiveServerType=2;AuthMech=1;ThriftTransport=SASL;Schema=%(schema)s;KrbRealm=%
   (krbrealm)s;KrbHostFQDN=%(krbhostfqdn)s;KrbServiceName=%(krbservicename)s
   ```

   Important: Ensure that the value of the *DRIVER* setting in the connection string is set to the value of the section heading specified for the required ODBC driver in the *odbcinst.ini* file, and the value is included in braces {}.

   For example, if the value of the section heading in the *odbcinst.ini* file for an ODBC driver for MariaDB is *MariaDB-MariaDB_Foundation-3.0.8-1*, then you must specify the value of the *Driver* setting in the connection string as *{MariaDB-MariaDB_Foundation-3.0.8-1}*.

   For more information about the value of the section heading for a specific ODBC driver, refer to the section *ODBC Setup and System Configuration Settings*.

   You can also specify the following variables as values for the connection string parameters:

   - *%(hostname)s* - Specify the host name or IP address of the targeted system, where the datastore is installed
   - *%(port)i* - Specify the port number for accessing the datastore

     Note: In case of the **IBM DB2/zOS** system, you need to specify the TCP port for accessing the DB2 system. You can access the port number from the value of the *TCPPORT* field in the DSNL004I message, which appears on the system console when DB2 is started.

     For more information about the DSNL004I message, refer to the *DB2 for z/OS* documentation.

   - *%(username)s* -Specify the user name for accessing the datastore
   - *%(password)s* - Specify the password for accessing the datastore
   - *%(database)s* - Specify the database name for the **IBM DB2/UDB**, **MySQL**, and **PostgreSQL** systems

     Note: In case of the **IBM DB2/zOS** system, you need to specify the location name for the DB2 system. You can access the location name from the value of the *LOCATION* field in the DSNL004I message, which appears on the system console when DB2 is started.

     For more information about the DSNL004I message, refer to the *DB2 for z/OS* documentation.

   - *%(servicename)s* - Specify the Oracle database service name. By default, the value of this service is set to *orcl*.
   - *%(schema)s* - Specify the Kerberos schema
   - *%(krbrealm)s* - Specify the Kerberos realm for the Hive system
   - *%(krbhostfqdn)s* - Specify the host name of the Hive server as the fully qualified name for Kerberos authentication. The default value *_HOST* is used to indicate the host name of the Hive server.

- *%(krbservicename)s* - Specify the Kerberos service name of the Hive system. By default, the value of this service is set to *hive*.

At runtime, these variables are automatically replaced by the values that you specify for the corresponding parameters while creating a discover job. For example, the values that you specify for the user name and password in the **Credentials** field of the **Add Discover Job** dialog box are automatically populated in the *%(username)s* and *%(password)s* variables at runtime. Similarly, the *%(port)i* and *%(krbhostfqdn)s* variables are populated with the values that you specify for the *port* and *krbhostfqdn* settings respectively, as part of the advanced configuration settings while creating a discover job.

For more information about creating a discover job, refer to the section *Creating a Discover Job*.

For more information about the advanced configuration settings, refer to the section *Advanced Configuration Settings*.

Important: All the variables used in the connection string are case-sensitive.

Ensure that you use the same variables in the connection string.

Note:

If you want to connect to an SSL-enabled Hive server, then refer to the Cloudera ODBC Driver for Apache Hive documentation or the Hortonworks Hive ODBC Driver documentation for configuration details.

In this scenario, you need to add the following attributes to the connection string:

- Set the *SSL* attribute to *1*, if you want to enable the SSL connection.
- Set the *AllowSelfSignedServerCert* attribute to *1*, if you want to allow self-signed certificates from the Hive server.
- Set the *AllowHostNameCNMismatch* attribute to *1*, if you do not want the common name of a CA-issued SSL certificate to match the host name of the Hive server.
- Set the *TrustedCerts* attribute to the full path of the *.pem* file, if you have uploaded a trusted CA certificate on Protegrity Discover.

  For example, set the value of the *TrustedCerts* attribute to **`/etc/ksa/certs/<CA certificate ID>.pem`**.

  You can obtain the ID of the CA certificate from the **Certificate Repository** screen.

  For more information about the Certificate Repository, refer to the section *Certificate Repository* in the *Protegrity Certificate Management Guide 9.2.0.0*.

  For more information about uploading a certificate, refer to the section *To upload certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*.

  If you want to use the trusted CA certificates .pem file that is installed with the driver, then do not specify a value for the *TrustedCerts* attribute.

- Set the *TwoWaySSL* attribute to **`1`**, if you want to enable two-way SSL verification. After setting the attribute to **`1`**, perform the following steps:
  - Set the *ClientCert* attribute to the full path of the *.pem* file containing the client's certificate.

    For example, set the value of the *ClientCert* attribute to **`/etc/ksa/certs/<ID of the client certificate>.pem`**.

    You can obtain the ID of the client certificate from the **Certificate Repository** screen.

    For more information about the Certificate Repository, refer to the section *Certificate Repository* in the *Protegrity Certificate Management Guide 9.2.0.0*.

    For more information about uploading a certificate, refer to the section *To upload certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*.

If your certificate also contains the private key, then you need to specify the path where the certificate file has been uploaded.

If you have separately uploaded the certificate file and the private key, then you must also specify the path where the private key has been uploaded as part of the *ClientPrivateKey* attribute.

For more information about specifying the path of the file containing the private key, refer to the following *step*.

- Set the *ClientPrivateKey* attribute to the full path of the file containing the client's private key.
  For example, set the value of the *ClientPrivateKey* attribute to `/etc/ksa/certs/<ID of the file containing the private key>.key`.

  For more information about the Certificate Repository, refer to the section *Certificate Repository* in the *Protegrity Certificate Management Guide 9.2.0.0*.

  For more information about uploading a certificate, refer to the section *To upload certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*.

- If the private key file is protected with a password, then set the *ClientPrivateKeyPassword* attribute to the password.

6. Click **Test**.
   The **Default Values** section appears below the **Connection String** text box.
   The following sample figure displays the placeholders for entering default values for the connection string variables, if you use the default connection string provided by Protegrity Discover for connecting to a Hive datastore.

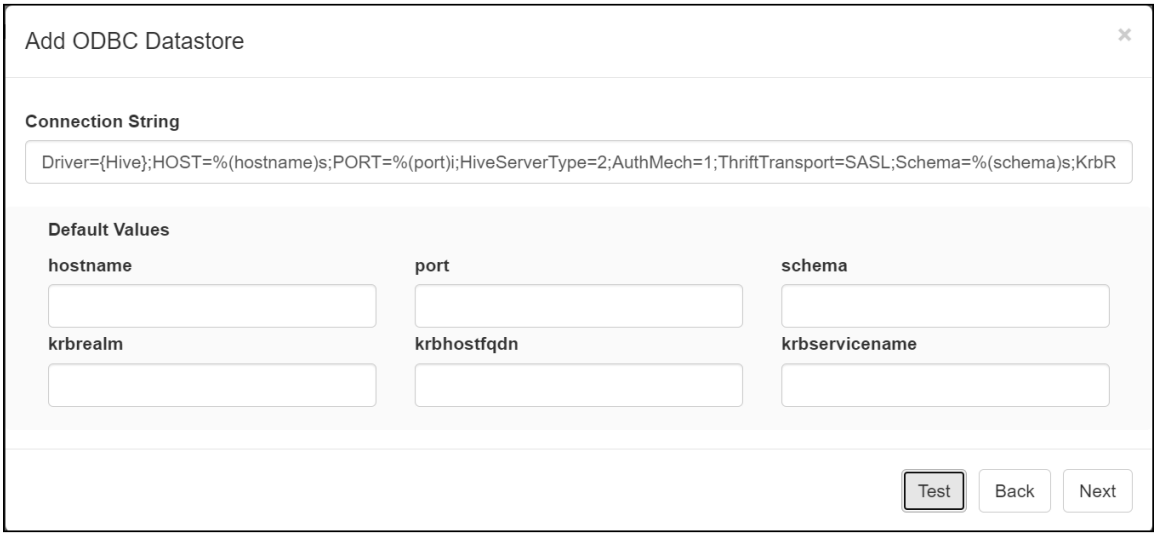


Figure 5-4: Default Values for Connection String Parameters

7. Enter the default values for the connection string variables, if required.

Important: The default values that you specify for the connection string variables are used for testing the connection string. After you create the datastore, these default values also appear in the **Default Values** area on the **Datastore** > **Connection Settings** tab for the corresponding datastore. You can also edit these default values in the **Default Values** area.

As mentioned in *step 5*, during runtime, the default values of the connection string variables are overridden by the values specified for the corresponding parameters when creating a discover job.

8. Click **Test** to test the connection settings with the specified datastore.

   > **Note:** The test uses the basic authentication scheme, which involves validating the username and password details, for authenticating the connection to the specified datastore.

9. Click **Next**.
   The **Get Version**, **Encoding**, and **Decoding** text boxes appear.

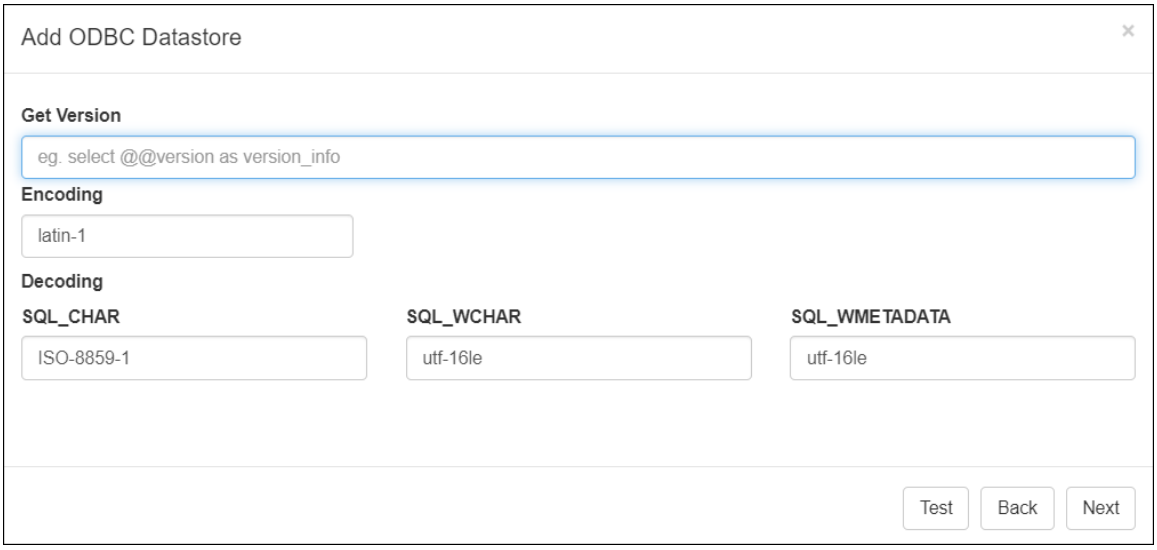


Figure 5-5: Get Version Query, Encoding, and Decoding

10. In the **Get Version** text box, enter the query to retrieve the current version of the datastore.

    > **Important:** This is a mandatory query that is used to validate the connection between Protegrity Discover and the targeted datastore.

    For reference, you can check the *GET_VERSION* query provided for the existing datastores on the **Datastore** > **Queries** tab.

    For example, the following snippet displays the sample *GET_VERSION* query for the Hive database:

```
[
    "SELECT '1.1' AS version_info;"
]
```

    > **Note:** For detailed information about the *GET_VERSION* query and its format, refer to the table *System Queries with Context*.

11. Enter the default encoding setting in the **Encoding** text box.
    The encoding setting specifies the encoding mechanism used for the queries that Protegrity Discover sends to the connected datastore. The encoding mechanism ensures that the queries are correctly translated by the specified ODBC driver for the selected datastore. For example, for a Hive ODBC driver, the default value of the encoding mechanism for encoding query data is *latin-1*.

    By default, the value of the encoding setting is set to *latin-1*.

If you do not specify any value for the encoding setting, then *utf-8* is used as the default encoding mechanism. After you have created a datastore, you can modify the value from the **Encoding** text box on the **Datastore** > **Data Types** tab.

> **Important:** For information about the encoding setting required by your specific ODBC driver, refer to the documentation provided by the vendor of the ODBC driver.

12. Enter the default decoding settings in the **Decoding** text fields.

    The decoding settings specify the encoding mechanism used to decode the data received from the ODBC driver. You must specify the following decoding settings:

    - **SQL_CHAR** - Specify the encoding mechanism for decoding the single byte character string sent by the ODBC driver. For example, for a Hive ODBC driver, the default value of the encoding mechanism for decoding single byte character strings is *ISO-8859-1*.

      By default, the value of the **SQL_CHAR** setting is set to *ISO-8859-1*.

      If you do not specify any value, then *utf-8* is used as the default encoding mechanism to decode the single byte character strings. After you have created a datastore, you can modify the value from the **SQL_CHAR** text box in the **Decoding** section of the **Datastore** > **Data Types** tab.

    - **SQL_WCHAR** - Specify the encoding mechanism for decoding the multibyte character strings sent by the ODBC driver. For example, for a Hive ODBC driver, the default value of the encoding mechanism for decoding multibyte character strings is *utf-16le*.

      By default, the value of the **SQL_WCHAR** setting is set to *utf-16le*.

      If you do not specify any value, then *utf-8* is used as the default encoding mechanism to decode multibyte character strings. After you have created a datastore, you can modify the value from the **SQL_WCHAR** text box in the **Decoding** section of the **Datastore** > **Data Types** tab.

    - **SQL_WMETADATA** - Specify the encoding mechanism for decoding the metadata, such as, database table and column names, that are sent by the ODBC driver. For example, for a Hive ODBC driver, the default value of the encoding mechanism for decoding metadata is *utf-16le*.

      By default, the value of the **SQL_WMETADATA** setting is set to *utf-16le*.

      If you do not specify any value, then *utf-32le* is used as the default encoding mechanism for decoding metadata. After you have created a datastore, you can modify the value from the **SQL_WMETADATA** text box in the **Decoding** section of the **Datastore** > **Data Types** tab.

> **Important:** For information regarding the decoding settings required by your specific ODBC driver, refer to the documentation provided by the vendor of the ODBC driver.

13. Click **Next**.

    A list of queries that can be used by Protegrity Discover to fetch details from the particular system appears.

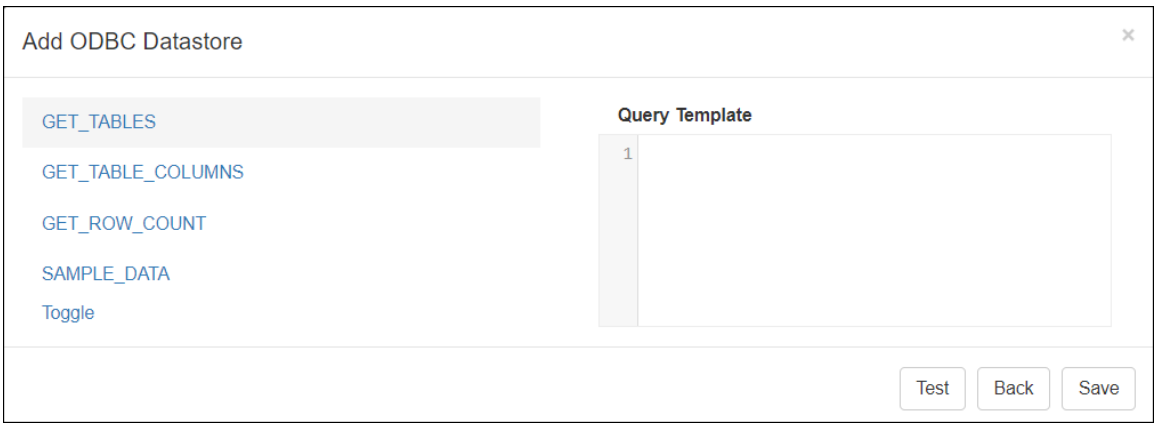


Figure 5-6: List of Queries

14. Select a query from the left pane and then enter the corresponding query details in the **Query Template** text area.

    You can also click **Toggle** to display additional set of queries.

> **Important:** For detailed information about the queries and their format, refer to the table *System Queries with Context*.

The following queries are listed.

Query	Description
GET_DATABASE	Retrieve a list of databases or schemas available on the targeted datastore.
GET_DATABASE_TABLES	Retrieve a list of tables within a database or a schema on the targeted datastore. This query is used together with the *GET_DATABASE* query.
GET_TABLES	Retrieves a list of all the tables across all databases or schemas on the targeted datastore. This query is used as an alternative to the *GET_DATABASE* and *GET_DATABASE_TABLES* queries. This query is used for datastores that support a single query to list all the tables, instead of using two queries.
GET_TABLE_COLUMNS	Retrieves the information about names, data types, and lengths of table columns.
SAMPLE_DATA	Retrieves the sample data in a table.
GET_ROW_COUNT	Retrieves the size of the target table that contains the sample data.
GET_PARITIONED_ROW_COUNT	Retrieves the size of the target partitioned table that contains the sample data. This query is used as an alternative to the *GET_ROW_COUNT* query and is used if the targeted database table is partitioned.
TEMPLATE_PARAMETERS_TRANSFORMATION	Transforms the template parameters.

For reference, you can use any existing query specified in the **Queries** tab.

For example, the following snippet displays the sample *SAMPLE_DATA* query for the Hive database:

```
[
    "SELECT %(columns)s FROM %(database_name)s.%(table_name)s ;"
]
```

15. Click **Test**.

    The **Query Dependents** section appears below the **Query Template** text area. It displays a list of parameters that are part of the specific query.

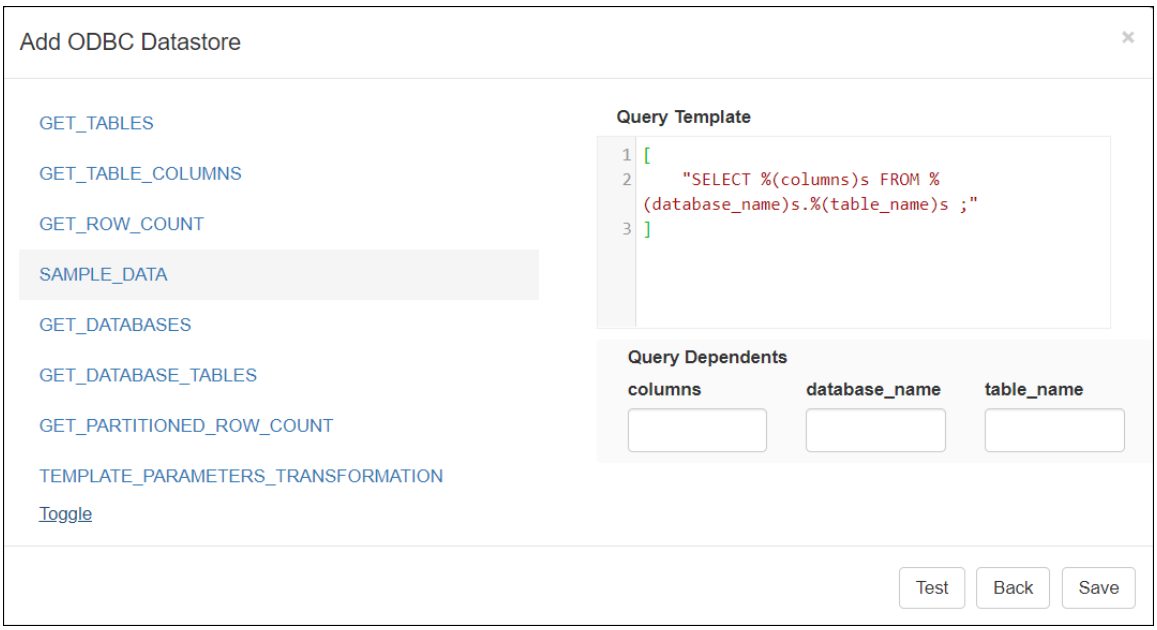


Figure 5-7: Query Dependents

16. In the **Query Dependents** section, enter the default values for the query parameters.

    You can click **Test** to test the connection settings with the updated queries.

    > **Important:** The default values that you specify for the query parameters are used only for testing the connection settings. During the scanning process, the parameters are replaced by the output of other queries.
    >
    > For more information about replacing the query parameters during the scanning process, refer to the table *System Queries with Context*.

17. Click **Save**.

    The datastore appears in the **DATASTORE** list.

    After the datastore is created, the following log message appears on the **Scanner Log** screen.

    ```
    Datastore <Name_of_the_Datastore> created successfully
    ```

5.1.2.2 Modifying the Datastore Settings

This section describes how you can modify the datastore settings through the **Datastore** screen from the **Appliance** menu.

➤ To modify the datastore settings:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **Datastore**.

   By default, the **Connection Settings** tab is selected for the Teradata database.

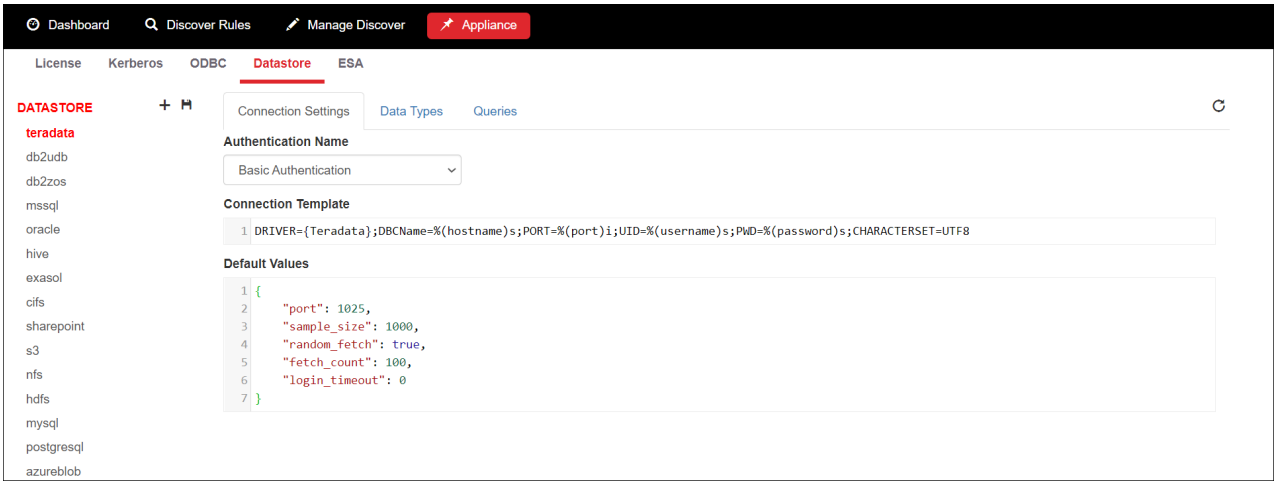


Figure 5-8: Connection Settings

Important: You can also update the datastore settings by modifying the *datastore.json* system file. You can access this system file from the Appliance Web UI.

For more information about modifying the Protegrity Discover system files, refer to the section *Managing the Appliance Information*.

2. Select an existing datastore from the **DATASTORE** list.

   The **Connection Settings** tab appears for the selected datastore.

3. Select an authentication type from the **Authentication Name** list.

   Note: By default, if you are modifying the settings of a datastore that you have created, then only the **Basic Authentication** option is displayed in the **Authentication Name** list. Similarly, if you are modifying the settings of a default datastore, then only the authentication types that have been pre-defined in Protegrity Discover are displayed.

   If you want to specify a different authentication mechanism for your database, then you need to modify the *datastore.json* file.

4. Modify the connection settings in the **Connection Template** text box, if required.

   Note: The **Connection Template** text box is displayed only for database systems.

5. Modify the default values for the configuration parameters in the **Default Values** section, if required.

   The values that you specify for the configuration parameters listed in the **Default Values** section are used as the default values when you are running a discover job. However, you can override these values by specifying custom values for these configuration parameters in the **Config** text area when you are creating a discover job.

   For more information on overriding the default values of the configuration parameters, refer to *step 10* of the section *Creating a Discover Job*.

6. Click **Data Types**.

   The **Data Types** tab appears for the selected datastore.

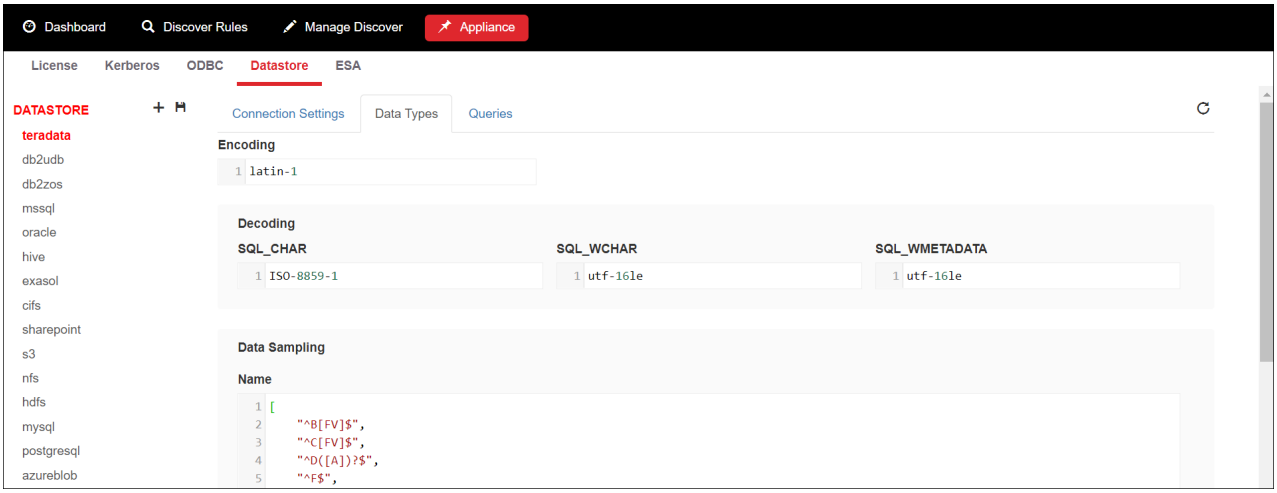


Figure 5-9: Data Types

> **Note:** The **Data Types** tab is displayed only for database systems.

7. Modify the encoding setting in the **Encoding** box, if required.
8. Modify the following decoding settings in the **Decoding** text area, if required:
   - SQL_CHAR
   - SQL_WHCHAR
   - SQL_WMETADATA
9. Modify the following settings for sampling the data in the **Data Sampling** text area, if required:
   - Name
   - Replace Pattern

   For more information on the sampling pattern and replace pattern, refer to the *GET_TABLE_COLUMNS* query in the table *System Queries with Context*.
10. Click **Queries**.

    The **Queries** tab displays the queries that you have selected while creating the selected datastore.

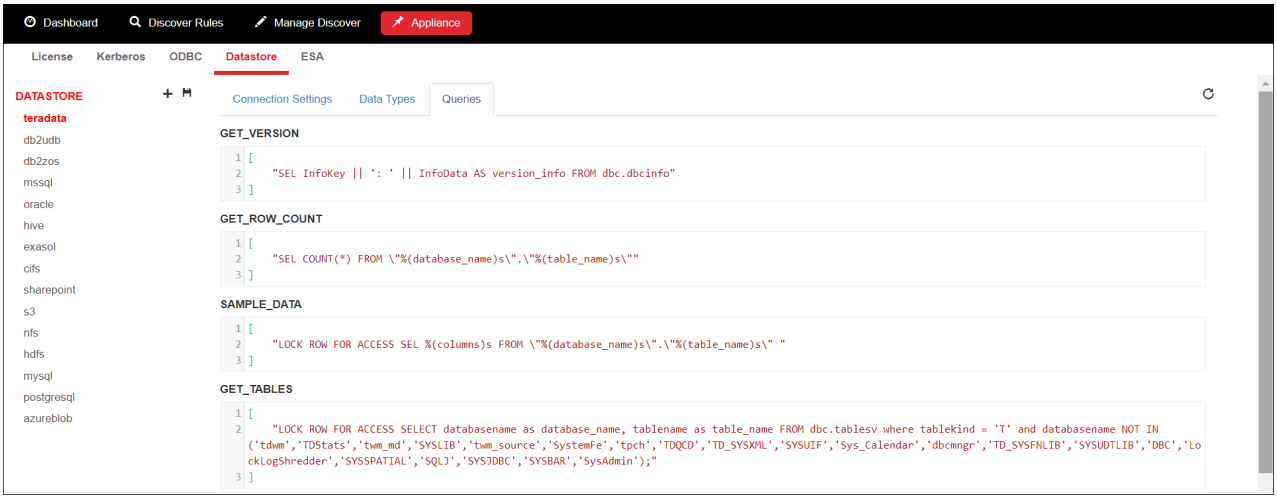


Figure 5-10: Queries

> **Note:** The **Queries** tab is displayed only for database systems.

11. Modify the queries that Protegrity Discover can use to fetch details from the selected datastore, if required.

> **Note:** For more information about queries and their format, refer to the table *System Queries with Context*.

12. Click to save the modified datastore settings.

> **Important:** If you have modified the settings for any default datastore provided by Protegrity Discover and you want to rollback to the default settings, then you must first navigate to the respective tab that you have modified and then click the reset icon at top-right corner. Click to save the changes after you have restored the default settings.

5.1.2.3 Deleting a Datastore

This section describes how you can delete a datastore that you have added.

To delete a new datastore:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **Datastore**.
2. Select a datastore that you have added from the **DATASTORE** list.

> **Important:** You cannot delete the default datastores that have been provided by Protegrity Discover.

3. Click .
   The **Delete Datastore** dialog box appears.

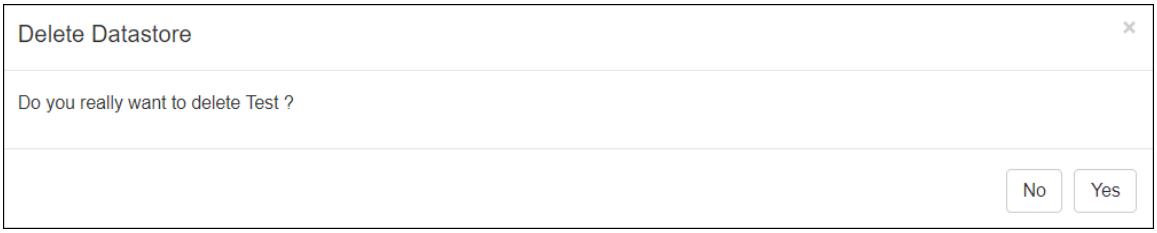


Figure 5-11: Protegrity Discover - Delete Datastore

4. Click **Yes** to delete the selected datastore.

   After the datastore is deleted, the following log message appears on the **Scanner Log** screen.

```
Datastore <Name_of_the_Datastore> removed successfully
```

Chapter 6

Protegrity Discover Web UI

The Web UI is interactive and enables you to perform all Protegrity Discover-related tasks and activities. The following table describes the Web UI elements and its capabilities:

Table 6-1: Protegrity Discover Web UI - Tasks and Capabilities

Menu	Tab	Description
Dashboard	Overview	The Protegrity Discover **Overview** screen provides summary of the following details: • **Top 5 data elements** - Lists the top five data elements classified as sensitive data with a count of location instances where the sensitive data is found. • **Top 50 Coordinates** - Provides a graphical view of the top 50 coordinates that contain sensitive data, categorized by location or estimated sensitive data values. The coordinates can be classified as systems, sub-systems, or compartments. • **Discover Trend** - Depicts a trend chart representing the total number of sensitive data element instances from all the systems reported over a period of time. • **Key Metrics** - Provides a summary of scan results that includes the total count of sensitive coordinates, observed sensitive data, estimated sensitive data, and the mitigated data (sensitive data that was later protected). For more information about the **Overview** screen, refer to the section *Data Discover Overview*.
	Classifications	From the **Classification** screen, you can check the classification results from this screen, with the history records of all previously classified data. For more information about the data classification results, refer to the section *Classifications*.
	Data Sources	The **Data Sources** screen provides an aggregated view of the following details: • Graphical view of the top 50 coordinates that contain sensitive data, categorized by estimated sensitive data values. The coordinates can be classified as systems, sub-systems, or compartments.

Menu	Tab	Description
		• Tabular view of classification results per system, sub-system, or compartment For more information about the **Data Sources** screen, refer to the section *Data Sources*.
	Scan Results	The **Scan Results** screen enables you to view a list of all the scans that you have performed for any datastore or job. You can also view and delete each scan details from the *Repository*. For more information about viewing the list of scans and deleting individual scans, refer to the section *Scan Results*.
Discover Rules	Jobs	From the **Jobs** screen, you can create the discover job to scan a particular coordinate. You can also edit, rerun, stop, refresh, clone, or remove the scan record. For more information about working with discover jobs, refer to the section *Working with Jobs*.
	Classifiers	From the **Classifiers** screen, you can create custom classifiers to identify sensitive data. You can also modify the default classifiers that are used by Protegrity Discover to identify sensitive data. For more information about managing classifiers, refer to the section *Managing Classifiers*.
Manage Discover	Scanner Log	From the **Scanner Log** screen, you can check the logs, which include the system related warning and error messages, for the different Protegrity Discover modules. The log messages help you to troubleshoot any issue. For more information about how to check the Protegrity Discover logs, refer to the section *Viewing Scanner Logs*.
	REST API Log	From the **REST API Log** screen, you can check the logs, which include the system related warning and error messages, for the different Protegrity Discover REST APIs. The log messages help you to troubleshoot any issue. For more information about how to check the Protegrity Discover REST API logs, refer to the section *Viewing REST API Logs*.
	REST API Analytics	From the **REST API Analytics** screen, you can view the statistics for the Protegrity Discover REST API requests. For more information about how to view the statistics for the Protegrity Discover REST API requests, refer to the section *Viewing REST API Analytics*.
Appliance	License	From the **License** screen, you can manage the Protegrity Discover license when it expires or is no longer valid. For more information about managing the Protegrity Discover license, refer to the section *License Manager*.
	Kerberos	From the **Kerberos** screen, you can manage the Kerberos related configuration settings. For more information about Kerberos, refer to the section *Kerberos*. For more information about configuring Kerberos with Protegrity Discover, refer to the section *Kerberos Configuration Manager*.
	ODBC	From the **ODBC** screen, you can upload the ODBC driver that your vendor has provided and add settings to the Protegrity Discover setup to extend the support to other systems.

Menu	Tab	Description
		For more information about extending the Protegrity Discover support to other systems, refer to the section *Extending the Support to Other Systems*.
	Datastore	The **Datastore** screen enables you to add a new datastore, delete a datastore, and configure additional settings, such as, connection, sampling, encoding, and query templates directly through the Web UI. For more information about adding or deleting a datastore, and configuring additional datastore settings, refer to the section *Managing Datastores*.
Managing the Appliance Information		You can check information related to services, disk usage, system files, and appliance logs from the Appliance Web UI. The system user management capabilities are also available from this screen. For more information about managing the appliance information, refer to the section *Managing the Appliance Information*.

6.1 Discover Overview

The **Overview** screen provides a dashboard summary of the sensitive data discovered across all systems through the discovery scans. This section explains the different parts of the dashboard, such as top 5 data elements, discover trend, top 50 coordinates, and key metrics.

> **Note:** The dashboard summary, by default, provides an overview of all scanned element instances with a respective confidence score of 60% or above. You can modify this filter setting using the slider, which is available at the top-right corner of the dashboard.

By default, the **Overview** screen provides a dashboard summary for all the scanned systems. However, you can filter the summary based on a specific system using the **Hostnames** drop-down list, which is available at the top-right corner of the dashboard. You can refresh the list of host names by clicking the icon.

The **Overview** screen is the home page for the Protegrity Discover Web UI and is divided into four parts:

- **Top 5 Data Elements** - Displays the top five sensitive data elements that were found collectively from the discovery scans.

Figure 6-1: Protegrity Discover - Top 5 Data Elements

The number associated with each sensitive data element indicates the number of discovered locations that contain the particular data element. If you click any number, you are redirected to the *Classifications* screen for further analysis.

- **Discover Trend** - Displays a trend chart representing the total number of sensitive data element instances from all systems, reported over a period of time.

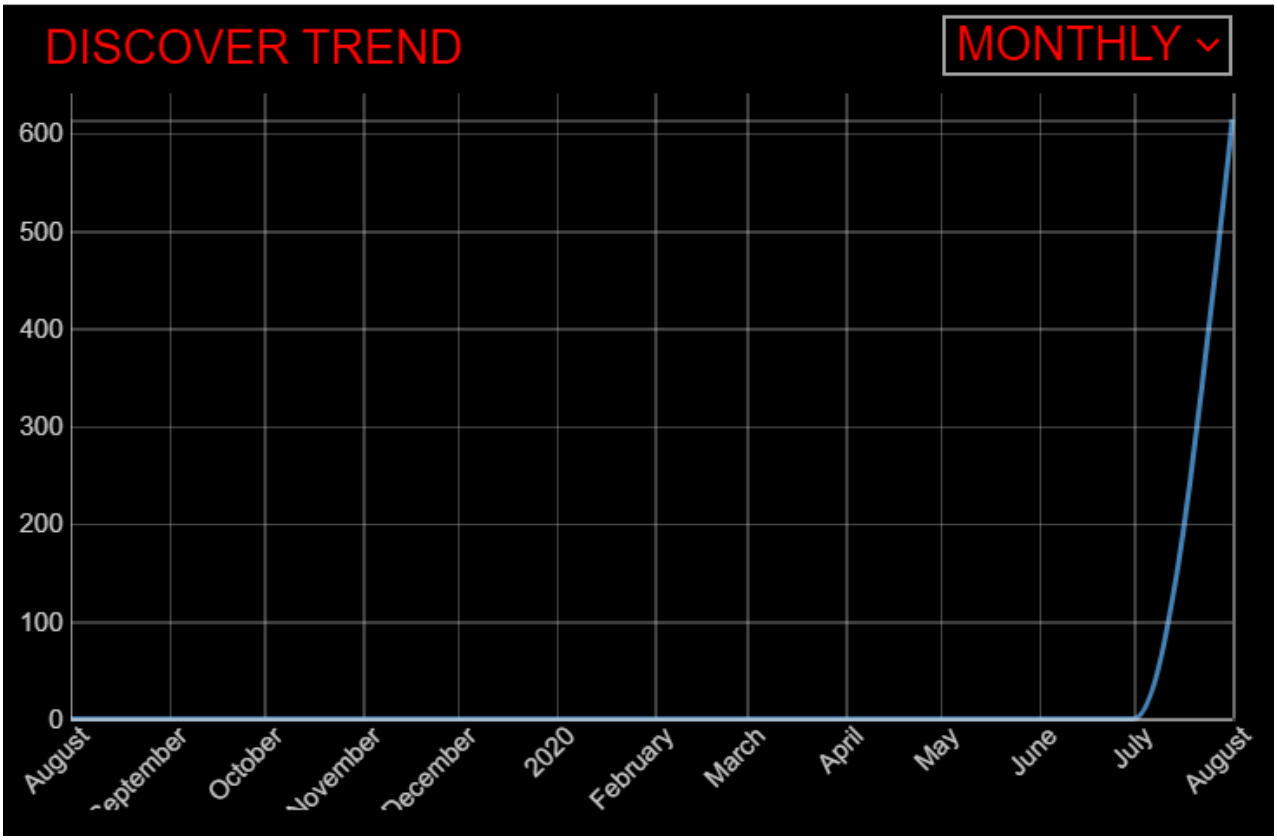


Figure 6-2: Protegrity Discover - Discover Trend

The X-axis denotes the period of time and Y-axis denotes the total number of sensitive data element instances.

The trend chart will reflect and display any changes in real-time. For example, changes can include the user configuring additional systems for the discovery scan or the current systems accumulating the latest discovery results or both. In each case, the latest trend is updated and displayed.

If a discovery scan no longer finds or classifies an instance, which was previously cataloged as receiving data, the trend result will go down. This can happen if sensitive data is now protected using data protection methods, such as encryption and tokenization, or if the sensitive data is removed.

The trend result statistically measures security risks that have been observed for the sensitive data in the clear across all the systems. It also indicates the progress made in addressing the identified security risk.

You can set the trend chart to generate the results for a daily, weekly, or monthly time period by selecting the required option from the provided dropdown.

- **Top 50 Coordinates**: Displays a bubble chart of the top 50 coordinates that contain sensitive data, with the following details:
  a. *Location instances*: It indicates the total number of locations having sensitive data elements, categorized by systems. Each bubble represents a system, which denotes the count of locations. You can filter the bubble chart further to represent data for sub-systems or compartments of a system. Protegrity Discover displays a bubble chart for the top 50 systems, the top 50 sub-systems, and the top 50 compartments.

     The locations containing the sensitive data can be categorized based on a hierarchical level. For example, in the case of structured data, the machine containing the sensitive data can be considered as the topmost level, while the database

column containing the sensitive data can be considered as the lowermost level. The system, sub-system, and compartment filter options represent three separate levels in the hierarchy of locations.

The following table defines the hierarchical level of locations represented by each filter option, for structured, unstructured, and semi-structured data.

Table 6-2: Hierarchical Level of Locations Represented by Filter Options

Filter Option	Structured Data	Semi-Structured Data	Unstructured Data
System	Hostname or IP address of the scanned machine	Hostname or IP address of the scanned machine	Hostname or IP address of the scanned machine
Sub-system	Database name or schema name, depending on the targeted datastore. For example, in case of Oracle, the sub-system represents the schema name. In case of MS SQL Server, the sub-system represents the database containing the sensitive data.	Topmost directory containing the sensitive data. For example, in case of semi-structured files on SharePoint, the sub-system represents the *sites* directory.	Topmost directory containing the sensitive data. For example, in case of unstructured files on SharePoint, the sub-system represents the *sites* directory.
Compartment	Name of the *database table* that contains the sensitive data	Name of the *directory* that contains the files with sensitive data.	Name of the *directory* that contains the files with sensitive data.

In the following figure, each bubble represents a system compartment with the count of locations containing sensitive data elements. In this particular example, the *CCN* directory has 120 location instances with sensitive data.

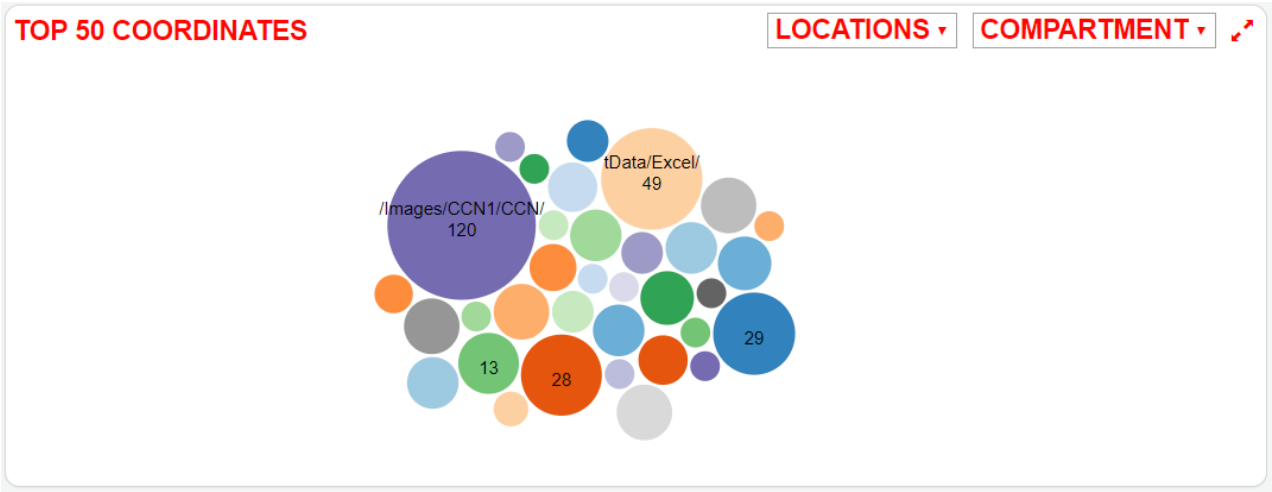


Figure 6-3: Protegrity Discover - Top 50 Coordinates by Location Instances across System Compartments

If you click any bubble, you are redirected to the *Classifications* screen for further analysis.

b. *Estimated values*: The bubble chart representation can be changed to reflect the estimated number of actual sensitive data values.

> **Note:** For more information about how Protegrity Discover derives an estimated sensitive data value, refer to the section *Sample Data Validation*.

Each bubble represents a system, which denotes an estimate of actual sensitive data values. You can filter the bubble chart further to represent data for sub-systems or compartments of a system. Protegrity Discover displays a bubble chart for the top 50 systems, the top 50 sub-systems, and the top 50 compartments.

In the following figure, each bubble represents a system compartment with estimated sensitive data values. In this particular example, the *Address* directory contains 591 thousand sensitive data values.

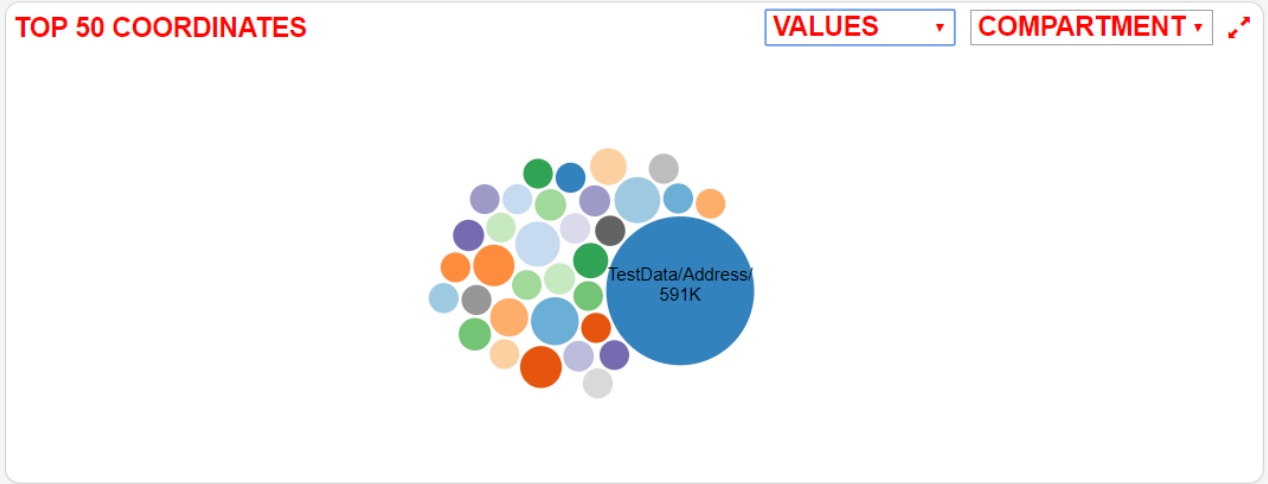


Figure 6-4: Protegrity Discover - Top 50 Coordinates by Estimated Sensitive Data Values across System Compartments

If you click any bubble, you are redirected to the *Classifications* screen for further analysis.

- **Key Metrics**: Displays a summary view of the entire scan results, which includes the total count of sensitive coordinates, observed sensitive data, estimated sensitive data, and the mitigated data (sensitive data that was later protected).

  In the following figure, it is found that 22 coordinates contain sensitive data values with 12K unique values of data instances that were classified as sensitive data. The total estimated sensitive data is 18 million data values at these 22 coordinates.

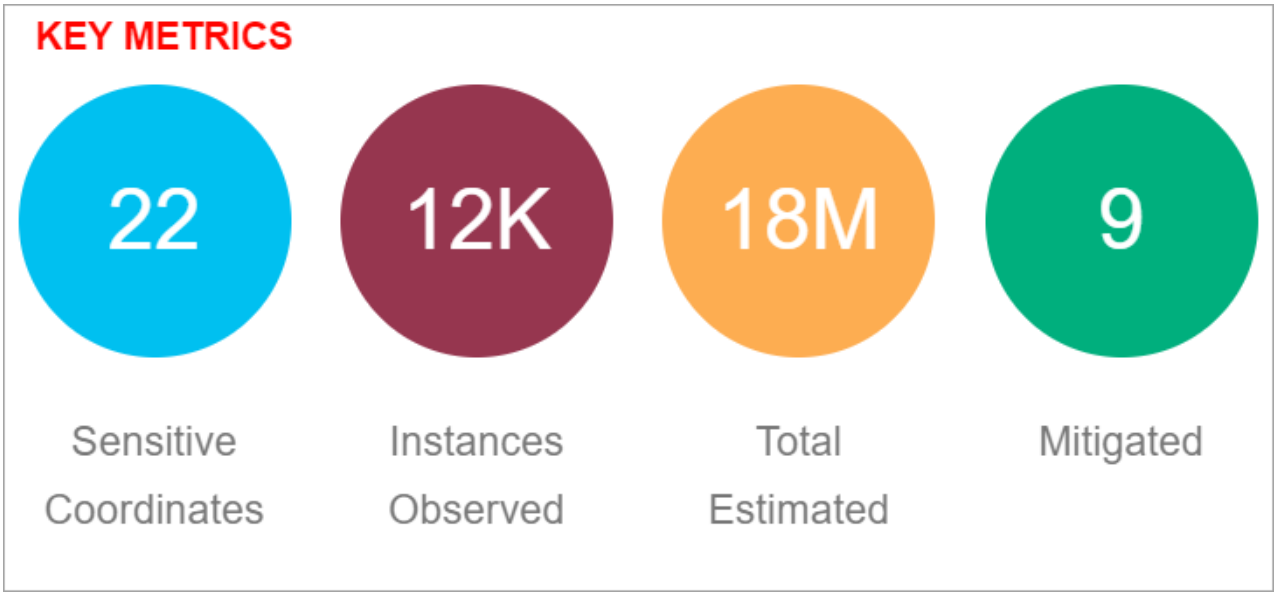


Figure 6-5: Protegrity Discover - Key Metrics

> **Note:** For more information about how Protegrity Discover derives at an estimated sensitive data value, refer to the section *Sample Data Validation*.

The mitigated count, which in this example is nine, indicates the instances that were previously categorized as sensitive data, but as per the latest scan history are now mitigated. This mitigated count is most likely a result of using data protection methods, or from being data that has either been reduced or removed.

The mitigated count varies based on the confidence score that you have specified. If you want to view all the mitigated values on the dashboard, then you must move the confidence score slider to 0%.

If you click any bubble, you are redirected to the *Classifications* screen for further analysis.

6.2 Data Sources

The **Data Sources** screen provides an aggregated summary of the sensitive data discovered across all systems through the discovery scans. This section explains the different parts of the screen, such as, the chart view and the table view.

> **Note:** The **Data Sources** screen, by default, provides an aggregated display considering a confidence score of 60% or above. You can modify this filter setting using the slider, which is available at the top-right corner of the screen.

You can also filter the summary based on a specific system using the **Hostnames** drop-down list, which is available at the top-right corner of the screen. You can refresh the list of host names by clicking the  icon.

The **Data Sources** screen is divided into two parts:

- **Chart view**: Displays a bubble chart of the top 50 coordinates that contain sensitive data, with the estimated number of sensitive data values in each coordinate.

  > **Note:** For more information about how Protegrity Discover derives an estimated sensitive data value, refer to the section *Sample Data Validation*.

  Each bubble represents a system, which denotes an estimate of actual sensitive data values. You can filter the bubble chart further to represent data for sub-systems or compartments of a system. Protegrity Discover displays a bubble chart for the top 50 systems, the top 50 sub-systems, and the top 50 compartments.

  For more information about the hierarchical locations represented by each filter option, refer to the *Hierarchical Level of Locations Represented by Filter Options* table.

  In the following figure, each bubble represents a system compartment with estimated sensitive data values. In this particular example, the *Address* directory contains 591 thousand sensitive data values.

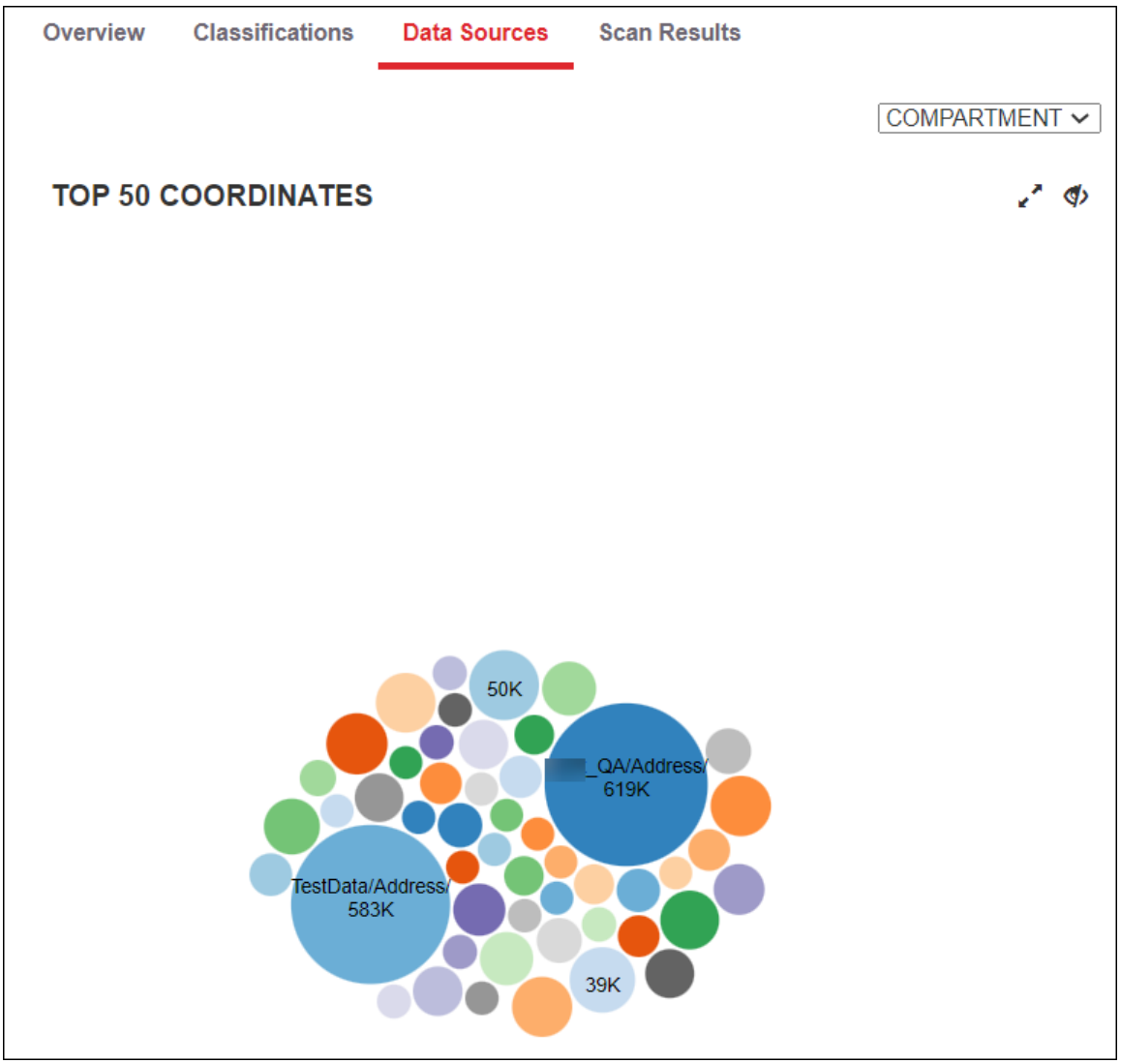


Figure 6-6: Protegrity Discover - Estimated Sensitive Data Values across System Compartments

- **Table view**: Enables you to view the classification records at an aggregated level.

  Each record in the classifications list indicates an estimate of actual sensitive data values, categorized by systems. You can filter the classifications list further to represent data for sub-systems or compartments of a system.

  For more information about the hierarchical locations represented by each filter option, refer to the *Hierarchical Level of Locations Represented by Filter Options* table.

  By default, the classifications list displays up to ten records on the screen. You can choose to display 25, 50, or 100 records on the screen using the **Show entries** drop-down list. Alternatively, you can use the page navigation buttons on the top and bottom of the classifications list to navigate the classifications records across pages.

  The following figure displays the classification records for a SharePoint scan.

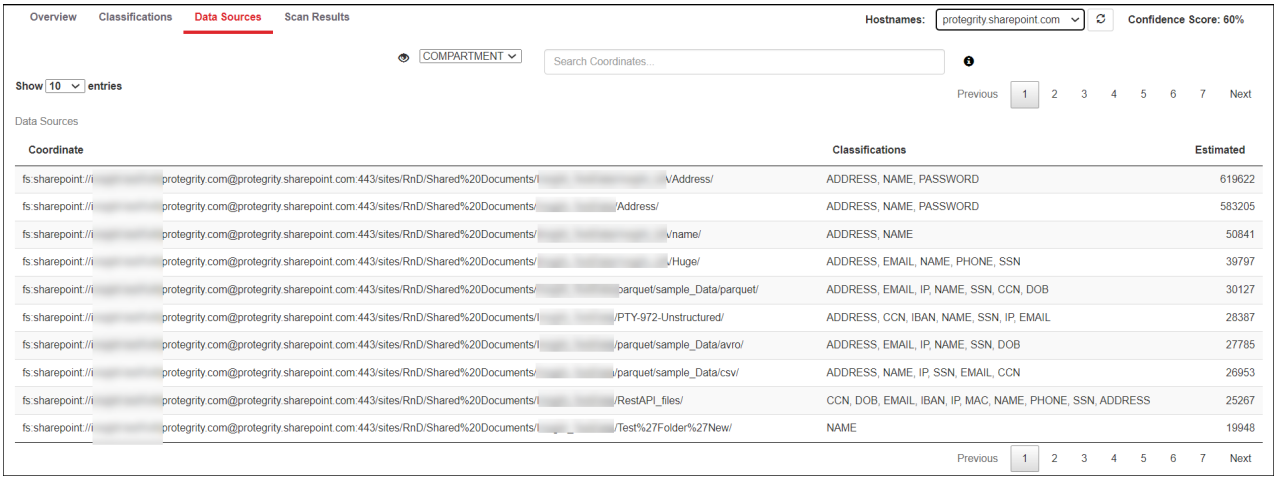


Figure 6-7: Protegrity Discover: Classification Record Findings

The following table explains the different attributes of a classification record.

Table 6-3: Protegrity Discover: Classification Record Findings

Attribute	Description
Coordinate	System for which the scan is performed
Classification	The classification type that reported the sensitive data values
Estimated	Estimated count of unique sensitive data values examined from the sampled data

The following table describes additional actions that you can perform from the **Data Sources** screen.

Table 6-4: Protegrity Discover - Additional Actions through the Data Sources Screen

Action	Description
Search the coordinates list	Search the **Coordinate** column using a text or a string search in the **Search Coordinates** text box. By default, the search is not case-sensitive. However, you can specify the search text or string within double quotes to perform a case-sensitive search. For example, type "Key Word" to perform a case-sensitive search for this exact phrase. You can narrow down the search results by including additional keywords in the search text.
Display only the chart view on the screen	Hide the table view and show only the chart view on the **Data Sources** screen using the button
Hide the chart view from the screen	Hide the chart view and display only the table view on the **Data Sources** screen using button. To display the chart view with the table view, click the button.

6.3 Scan Results

The **Scan Results** screen enables you to view a list of all the scans that you have performed.

> **Note:** The **Scan Results** screen, by default, displays the scan results considering a confidence score of 60% or above. You can modify this filter setting using the slider, which is available at the top-right corner of the screen.

You can also filter the scan results based on a specific system using the **Hostnames** drop-down list, which is available at the top-right corner of the screen. You can refresh the list of host names by clicking the icon.

The **Scan Results** screen is divided into two parts:

- **Scan Table View**: Displays a list of all the scans that have been performed. The latest scan appears at the top of the table.

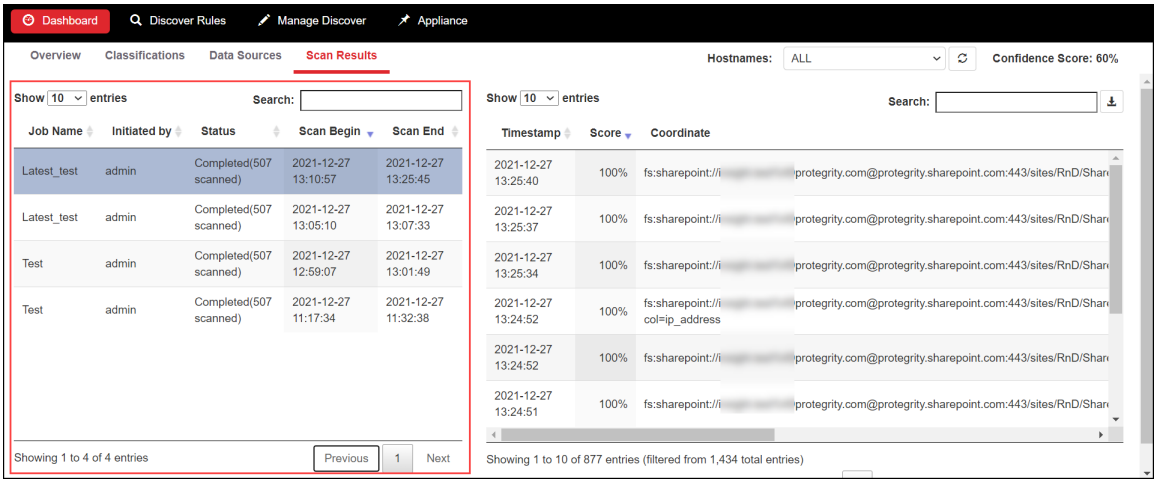


Figure 6-8: Protegrity Discover - Scan Table View

By default, the scan table view displays up to ten scan results on the screen. You can choose to display 25, 50, or 100 results on the screen using the **Show entries** drop-down list. Alternatively, you can use the page navigation buttons on the top and bottom of the classifications list to navigate the scan results across pages.

The following table explains the different attributes of the scan table.

Table 6-5: Protegrity Discover: Scan Table Attributes

Attributes	Description
Job Name	Name of the Discover job associated with the scan. If you have scanned the data using the Protegrity Discover REST APIs, then the job name appears as **RESTAPI** by default. However, you can specify a custom job name as the value of the *Job-Name* header in the REST API request. For more information about scanning data using Protegrity Discover REST APIs, refer to the section *Using the Protegrity Discover REST APIs*.
Initiated by	User who has initiated the scan.
Status	Specifies the present status of the Discover job. For example, *Running* or *Completed*. If you have scanned the data by creating a job in the **Discover Rules** > **Jobs** screen of the Protegrity Discover Web UI, then the status also indicates the number of records that have been scanned. However, if you have scanned the data using the Protegrity Discover REST APIs, then the status does not show how many records were scanned because the results of the scan are not stored in the Repository. If you want to find the number of records that were scanned, then you need to view the *individual scan results*.
Scan Begin	Time instance at which the scan was initiated
Scan End	Time instance at which the scan was completed

The following table describes additional actions that you can perform from the **Scan Table View**.

Table 6-6: Protegrity Discover - Additional Actions through the Scan Table View

Action	Description
Search the scan table	Filter the data in the scan table based on keywords

Action	Description
Show individual scan details	Select a scan from the scan table to display the *Individual Scan Results view* for viewing the scan details.

- **Individual Scan Results View**: Enables you to view the classification records for each scan result.

  The following figure displays the individual scan results.

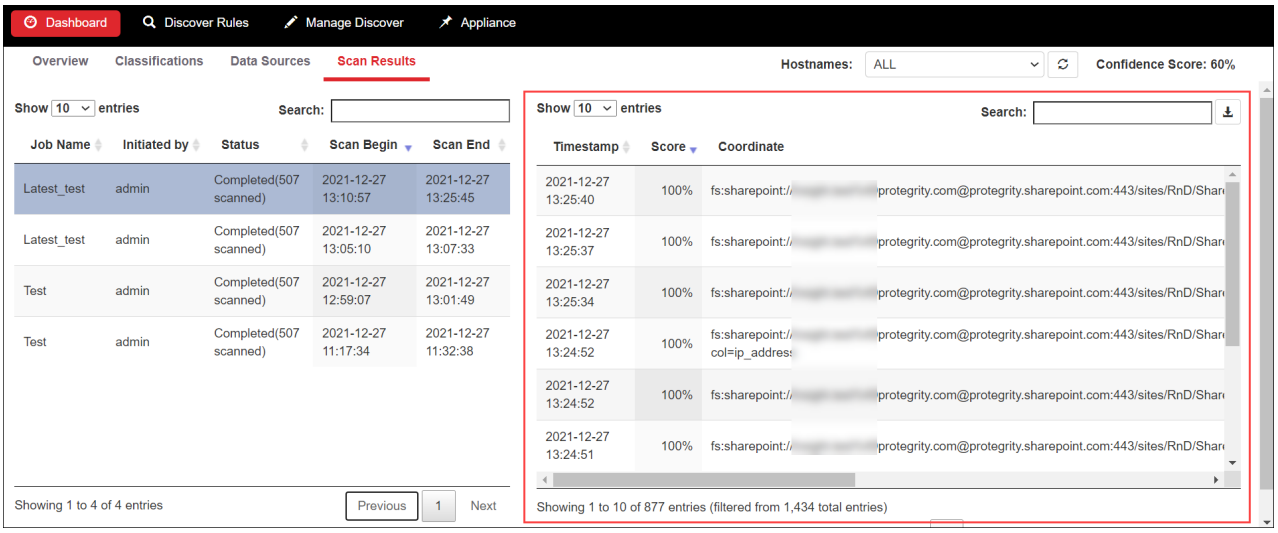


Figure 6-9: Protegrity Discover: Individual Scan Results

By default, the scan result view displays up to ten scan results on the screen. You can choose to display 25, 50, or 100 results on the screen using the **Show entries** drop-down list. Alternatively, you can use the page navigation buttons on the top and bottom of the classifications list to navigate the scan results across pages.

The following table explains the different attributes of a scan record.

Table 6-7: Protegrity Discover: Scan Record Attributes

Attributes	Description
Timestamp	Time instance at which the scan was completed
Score	The final confidence score for the scan derived as per analysis. For each classifier, individual scores are recorded. The scores against each classifier are aggregated to form the final confidence score for the scan. For more information about analysis and score calculation, refer to the section *Analysis*.
Coordinate	Location of the sensitive data
Classifier	Specific classifier name, and its rules, defined for the particular classification.
Classification	The classification type that reported the sensitive data values

> **Note:** If the scan is unable to find any sensitive data in the datastore or if the data has not been modified since the last scan, then the scan record displays the following message.
>
> ```
> No data available in table
> ```

The following table describes additional actions that you can perform from the **Individual Scan Results View**.

Table 6-8: Protegrity Discover - Additional Actions through the Individual Scan Results View

Action	Description
Search the scan table	Filter the data in the individual scan records based on keywords
Download the scan records	Click the  to download the scan records for an individual scan in JSON format. The records are saved in a JSON file named *discover_scan_view.json*.

6.4 Working with Discover Rules

This section describes how you can create discover jobs and customize classifiers.

6.4.1 Working with Jobs

The **Jobs** screen from the Protegrity Discover Web UI enables you to create and run a discover job. It also allows you to edit, refresh, stop, clone, or remove a discover job.

You can create a discover job to search for sensitive data at a specified coordinate. Protegrity Discover can automatically detect the system (for example, database type, etc.) for the specified coordinate. You can optionally create scheduled jobs to be run at specified time intervals, such as daily, weekly, or monthly, as well as set custom cron jobs. You must define the login credentials for the targeted system and can additionally define supporting parameters, such as size of sample data, and so on.

The following table defines the fields that you must set to create a discover job.

Table 6-9: Discover Job Configuration Settings

Settings	Description	Mandatory / Optional
Name	Set this field to define a name for the discover job	Mandatory
Description	Set this field to describe the discover job	Optional
Scheduling	Define a scheduling job pattern on a daily, weekly, or monthly basis. You can also set a cron expression to create a custom job.	Optional
System	Auto detect or assign the type of system using the list of supported systems from the dropdown	Mandatory
Authentication type	Set the mode of authenticating the user credentials for the targeted system.	Mandatory
Hostname	Define the hostname of the targeted system. The port number is optional and can be specified here, if the value is different than the default value. The field label changes depending on the targeted system.	Mandatory
Credentials*1	Specify the login credentials for the targeted system. The system user must have read only permissions.	Mandatory
Config*1	Define the advanced configuration settings for a discover job. For more information about the advanced configuration settings, refer to the section *Advanced Configuration Settings*.	Optional

Note: *1 - The MS SQL Server uses two different types of authentication, which are Basic Authentication and Windows Integrated Authentication. The Basic Authentication is for database users. The Windows Integrated Authentication is for specific Windows users or domain users belonging to group accounts, which are the trusted entities to login to the SQL Server. The Windows Integrated Authentication uses the Kerberos authentication protocol.

To login to a MS SQL Server that uses the Kerberos authentication protocol, the domain users must create a Kerberos ticket before creating the discover job.

6.4.1.1 Creating a Discover Job

This section explains the steps to create a discover job from the Protegrity Discover Web UI and specifies how to manage tasks pertaining to a discover job.

Before you begin

If you are connecting to a CIFS-based file system, then ensure that the following prerequisites are met before running a discover job:

- On a Windows machine: Ensure that the following network browsing features and service are running:
  - Server Message Block (SMB) client and server
  - SMB Direct
  - NetLogon service
- On a Linux machine: Ensure that the following services are accurately installed and configured:
  - SAMBA server
  - SMB service

If you are connecting to an NFS-based file system, then ensure that the following prerequisites are met before running a discover job:

- On a Windows machine: Ensure that the NFS server has been installed on the Windows machine.
- On a Linux machine: Ensure that the following services are installed:
  - nfs-common
  - nfs-kernel-server

To create a discover job:

1. On the **Protegrity Discover Web UI**, navigate to **Discover Rules** > **Jobs**.
2. Click the **Add** icon .
   The following **Add Discover Job** popup appears.

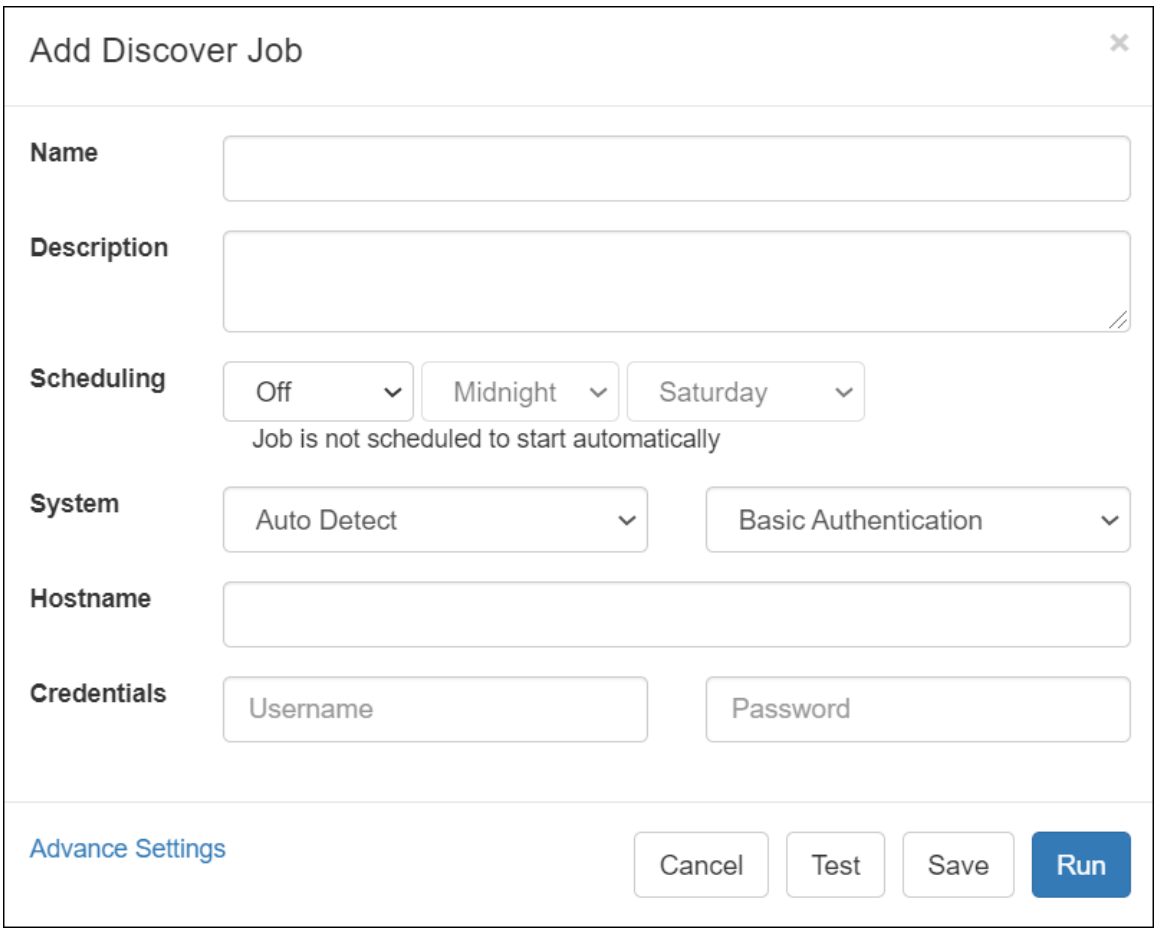


Figure 6-10: Protegrity Discover - Add Discover Job

3. Enter name and description for your job in the **Name** and **Description** text boxes respectively.
4. In the **Scheduling** list, select the required scheduling options to run the discover job at a specified time automatically.

   The scheduling of job option is optional and disabled by default. You can set up a scheduled job on a daily, weekly, or monthly basis. You can also set up a custom job using the *Custom* option from the scheduling list and by setting your own cron job scheduling pattern.

   For example, a cron expression '0 15 10 ? * *' triggers a custom job to be run at 3 pm on the 10th of every month.

5. In the **System** list, click **Auto Detect** to enable auto detection of the system or select the required system from the list.

   If you select **Auto Detect**, then you first need to enter the host name or IP address of the targeted system and the credentials required to access the targeted system, and then click **Test**. Protegrity Discover detects the required system, and the **Add Discover Job** dialog box automatically displays those fields that are applicable to the detected system.

   For example, if Protegrity Discover detects that the connected system is Oracle using the auto detect functionality, then the **Add Discover Job** dialog box automatically displays only those fields that are applicable to an Oracle system.

   If the auto detect functionality fails to detect the targeted system, then you must manually select the required system from the **System** list.

   > **Note:** If you have added a new datastore or system and are not able to view it in this system list, ensure that you refresh the browser window.

6. Choose the required authentication type for the targeted system.

   The available authentication types depend on the targeted system.

System	Authentication Type
AWS S3	• Access KeyID and Secret • IAM Role (Instance profile)
Azure Storage (Blob)	Account Access Keys
EXAsol	Basic Authentication
Hadoop File System (HDFS)	• No User Authentication • Kerberos Authentication (password) • Kerberos Authentication (keytab) • Token Authentication (password) • Token Authentication (keytab)
Hive	• Kerberos Authentication (password) • Kerberos Authentication (keytab) • Basic Authentication
IBM DB2/UDB	Basic Authentication
IBM DB2/zOS	Basic Authentication
Microsoft SQL Server	• Windows Authentication (password) • Windows Authentication (keytab) • Basic Authentication
MySQL	Basic Authentication
Network File System (NFS)	• No User Authentication • Kerberos Authentication (password) • Kerberos Authentication (keytab)
Oracle	Basic Authentication
PostgreSQL	Basic Authentication
Sharepoint	Basic Authentication **Important:** Protegrity Discover does not support multi-factor authentication for SharePoint. If your administrator has enforced multi-factor authentication for SharePoint, then you need to create an app password connect to the SharePoint without multi-factor authentication. For more information about creating app passwords, refer to the section *Create new app passwords* in the Microsoft Azure documentation.
Teradata	Basic Authentication
Windows Share (CIFS)	Basic Authentication

7. In the **Hostname** field, enter the host name or the IP address of the targeted system.

   The port number is optional and must be specified if the value is different from the default value.

   - AWS S3 - In case of AWS S3, the **Hostname** field specifies the AWS region that contains the S3 buckets that you want to scan. By default, the value of this field is set to *s3.amazonaws.com*, which indicates that S3 buckets are in the US East (North Virginia) region. However, you can specify a different region by modifying the host name to a valid endpoint name for a particular region.

     For example, enter *s3-us-east-2.amazonaws.com* to specify the US East (Ohio) region.

- Azure Storage (Blob) - In case of Azure Storage (Blob), the **Hostname** field specifies the storage account that is used to store the Azure data objects that you want to scan. By default, the value of this field is set to *<storage account name>.blob.core.windows.net*.
- HDFS - In case of HDFS, you need to specify the host name or IP address of the Name Node or Data Node of the HDFS server, based on the service that you want to use to access the data from HDFS. In HDFS, a Name Node manages the metadata of the file system, while a Data Node stores the data.

  For more information about HDFS, refer to the *Apache Hadoop documentation*.

  In addition, you need to specify the port number, depending on the service that you use to access data.

  For example, you can use either the WebHDFS service or the HttpFS service to access the data from HDFS.

  - WebHDFS - If you want to use the WebHDFS service, then you must specify the host name in the following format:
    ```
    <Host name or IP address of the Name Node>:<port>
    ```

    If you do not specify any port number, then *50070* is used as the default port number. If the WebHDFS service in your Hadoop cluster uses a different port number, then you can specify a different value.

    For more information about the WebHDFS service, refer to the *WebHDFS REST API* documentation.

  - HttpFS - If you want to use the HttpFS service, then you must specify the host name in the following format:
    ```
    <Host name or IP address of the Name Node or Data Node>:<port>
    ```

    If you are using the HttpFS service, then it is mandatory to specify a port number. By default, the HttpFS service uses *14000* as the port number. If the HttpFS service in your Hadoop cluster uses a different port number, then you can specify a different value.

    For more information about the HttpFS service, refer to the *HttpFS* documentation.

  For more information about the default values of the port number that needs to be specified, refer to the section *Scan Job Advanced Configuration Settings*.

The field label changes depending on the targeted system. For example, instead of the **Hostname** field, the following fields are displayed based on your targeted system:

- **Server** - Specify the server name if the targeted system is Microsoft SQL Server.
- **DBCName** - Specify the IP address of the Teradata server or its alias, if the targeted system is Teradata.

> **Note:** You can enter a maximum number of 256 characters in the **Hostname**, **Server**, or **DBCName** fields.

8. In the **Credentials** field, enter the following login credentials for accessing the targeted system, based on the selected authentication type.

Authentication Type	Credentials
Basic Authentication	Enter the user name and password for accessing the system. **Note:** You can enter a maximum number of 256 characters in the Username and Password fields. **Note:** If you have created an app password for connecting to SharePoint, then you must specify your user name and the app password for accessing the SharePoint system.

Authentication Type	Credentials
	For more information about creating app passwords, refer to the section *Create new app passwords* in the Microsoft Azure documentation.
No User Authentication	Enter the name of the user who has permissions to access HDFS. If you select *No User Authentication*, then you are not required to specify the password.
Windows Authentication	Enter the Windows user account name and the password for accessing the Microsoft SQL Server system. You must use the full domain name with the user name. For example, you must specify *username@domain*. You can choose one of the following methods for specifying the credentials: • *Password* - Specify the service principal name (SPN) and password. • *Keytab* *1 - Specify the SPN and the keytab. You can select an existing keytab from the list, or you can click **<upload keytab>** to browse for and select a keytab from your local machine. For more information about the principal and keytab, refer to the table *Protegrity Discover - Required Kerberos Configuration Settings*. **Note:** You can enter a maximum number of 256 characters in the Username and Password fields. **Important:** When you specify the credentials for the Windows Authentication and test or run the job, then a Kerberos ticket is automatically generated. *2
Access KeyID and Secret	Enter the access key ID and secret access key for accessing the AWS S3 system. If the specified access key id and secret access key are correct, then the name of the AWS user is automatically displayed in the **Username** field after saving the job. You can also scan AWS S3 buckets in another account or the same account, by using the IAM role that has permissions to access the S3 buckets for that account or instance. This is known as assuming a role. For more information regarding assuming a role, refer to the section *Assuming another role*. **Note:** If you want to use the IAM role for scanning AWS S3 buckets, then the AWS S3 user must have the *assume-role* permissions. For more information about the *assume-role* permissions, refer to the *AWS CLI Command Reference* documentation.

Authentication Type	Credentials
IAM Role (Instance profile)	If an IAM role is attached to the AWS EC2 instance where you have deployed Protegrity Discover, then the role name is automatically displayed in the **Username** field after saving the job. If you have created an IAM role that allows you to access the AWS S3 buckets, then you can attach the same role to the AWS EC2 instance. As a result, you can scan the AWS S3 buckets using the **IAM Role (Instance profile)** authentication. **Assuming another role**: By default, you can use the IAM role to scan the AWS S3 buckets belonging to the same account as that of the AWS EC2 instance. However, if you want to scan the AWS S3 buckets using the IAM role in a different account, then you need to specify the ARN of the IAM role associated with that account in the *role_arn* parameter in the **Advanced Settings** section. When you click **Test** or **Save**, the IAM role name associated with that instance automatically appears in the **Username** field. This scenario is known as assuming the role of another account or instance. For more information about the *role_arn* parameter, refer to the section *Scan Job Advanced Configuration Settings*.
Kerberos Authentication	Enter the Kerberos login credentials for accessing the system. You can choose one of the following methods for specifying the credentials: • *Password* - Specify the Kerberos principal and the password. • *Keytab* *1- Specify the Kerberos principal and the keytab. You can select an existing keytab from the list, or you can click **<upload keytab>** to browse for and select a keytab from your local machine. For more information about the principal and keytab, refer to the table *Protegrity Discover - Required Kerberos Configuration Settings*. **Note:** You can enter a maximum number of 256 characters in the Username and Password fields. You can also configure Kerberos from the **Kerberos** screen. For more information about setting the Kerberos-specific configurations, refer to the section *Kerberos Configuration Manager*. You can also select the *Kerberos Authentication* option for HDFS and NFS. In these cases, you do not have to provide any Kerberos-specific credentials on the **Add Discover Job** dialog box. For example, in case of HDFS, you need to specify the name of the user who has permissions to access HDFS. **Important:** When you specify the credentials for the Kerberos Authentication and test or run the job, then a Kerberos ticket is automatically generated. *2

Authentication Type	Credentials
Token Authentication	Enter the credentials of the user who has permissions to access HDFS. You can choose one of the following methods for specifying the credentials: • *Password* - Specify the SPN and the password. • *Keytab* [*1]- Specify the SPN and the keytab. You can select an existing keytab from the list, or you can click **<upload keytab>** to browse for and select a keytab from your local machine. **Important:** The Token Authentication uses a token provided by the HDFS Name Node to access the HDFS data. This token is known as the delegation token. When you specify the credentials for the Token Authentication and test or run the job, then a Kerberos ticket is automatically generated. [*2]

Note: [*1] - When you create a new job, the keytab is automatically associated with the job.

If an existing keytab is compromised, then you can create a new keytab with the same name as that of the existing keytab, and then upload the new keytab. The existing keytab is overwritten and the new keytab is then associated with the job. The *update_keytab* parameter in the **Advanced Settings** section determines whether an existing keytab can be overwritten.

For more information about the *update_keytab* parameter, refer to the section *Advanced Configuration Settings*.

Note: [*2] - If your existing Kerberos ticket is valid, then Protegrity Discover renews your ticket. If your existing Kerberos ticket has expired or is invalid, then Protegrity Discover generates a new ticket.

For more information about requesting a Kerberos ticket, refer to the section *Kerberos Configuration Manager*.

Caution: Ensure that the user has read-only permissions to the system.

Important: If you are using a third-party application, such as, CyberArk, for storing the password of the datastore that you want to scan, then you can configure Protegrity Discover to communicate with CyberArk for retrieving the password during runtime.

In this case, leave the **Password** field blank, as Protegrity Discover does not store the datastore password. Instead, Protegrity Discover uses the password retrieved from CyberArk during runtime to connect to the datastore.

For more information about integrating Protegrity Discover with CyberArk, refer to the section *Appendix H: Integrating Protegrity Discover with CyberArk*.

Note: If you have selected the authentication type as *No User Authentication* for the HDFS or NFS in *step 6*, then the **Credentials** field is not available.

9. Enter the additional details in the following fields based on the targeted system, if required:
   - **Access Key** - Specify the storage account keys that allow you to access the Azure Storage (Blob) data.
   - **Realm** - Specify the Kerberos realm for the Hive system.
   - **Default Database** - Enter the name of the default database for the IBM DB2/UDB, MySQL, and PostgreSQL systems.

     Note: In case of the IBM DB2/zOS system, you need to specify the name of the DB2 location that you want to access in the **Default Database** field. You can access the location name from the value of the LOCATION field in the DSNL004I message, which appears on the system console when DB2 is started.

> For more information about DSNL004I message, refer to the *DB2 for z/OS* documentation.

- **Service** - Enter the service name depending on the targeted system:
  - *Hive* - Specify the Kerberos service name of the Hive system. By default, the value of this service is set to *hive*.
  - *Oracle Database* - Specify the Oracle database service name. By default, the value of this service is set to *orcl*.
- **Path** - Specify the path details depending on the targeted system:
  - *AWS S3* - Enter the name of the bucket, directory, or file that you want to scan, in the following format:

    ```
    /[bucket]/[directory]/[filename]
    ```

    If you do not specify any file or directory name, then Protegrity Discover scans all the files and directories within the specified bucket. Similarly, if you do not provide any of the three values, then Protegrity Discover scans all the buckets specified in the selected AWS S3 region.

  - *Azure Storage (Blob)* - Enter the name of the container and blob that you want to scan, in the following format:

    ```
    /<container_name>/[blob_name]
    ```

    If you do not specify the blob name, then Protegrity Discover scans all the blobs within the specified container. Similarly, if you do not provide both the values, then Protegrity Discover scans all the containers specified in the selected storage account.

  - *Hadoop File System (HDFS)* - Enter the path of the directory or the file that needs to be scanned.
  - *Network File System (NFS)* - Enter the path of the folder that you want to scan, in the following format:

    ```
    /nfs_share_directory[/directory]
    ```

    Protegrity Discover cannot scan the folders present in the root directory of the NFS server on Linux. Therefore, you must first create a directory, *nfs_share_directory*, which is shared on the NFS server.

    > **Important:** Protegrity Discover scans only those folders that have been shared on the NFS server.

  - *Sharepoint* - Enter a valid SharePoint path for scanning the SharePoint location. Ensure that the path includes the term *sites* followed by the exact path.

    For example, you can specify the SharePoint path as `/sites/<directory>`.

  - *Windows Share (CIFS)* - Enter the path of the directory that needs to be scanned.
  - *IBM DB2/zOS* - Enter the specific database table that you want to scan, in the following format:

    ```
    /<CREATOR>/<Table Name>
    ```

    For example, you can specify the path as */PTY/CUSTOMER*, where *PTY* is the CREATOR name, while *CUSTOMER* is the name of the database table. In this case, you want to scan the *PTY.CUSTOMER* table in the DB2 subsystem at the location that is specified in the **Default Database** field.

  - Remaining datastores - Enter the specific database table that you want to scan, in the following format:

    ```
    /database/table
    ```

    If you want to scan a specific database column, then you must specify the path in one of the following formats:

    Format 1

    ```
    /<database>/<table>/<column name>
    ```

Format 2

```
/<database>/<table>/?col=<column name>
```

> **Note:** If you want to use Format 2 for specifying the column name, and if the column name contains any special characters, then you must percent-encode the special characters using UTF-8 encoding before running the discover job.
>
> For more information about percent-encoding, refer to the section *2.1 Percent-Encoding* in the *W3 Uniform Resource Identifier* specification.
>
> For more information about how you can percent-encode commonly used special characters, refer to the section *Appendix I: Percent-Encoding Special Characters*.

> **Note:** If you are either using Format 1 or Format 2, and if the database, table, or column name contains a *%* character followed by two hexadecimal digits, then you must perform percent-encoding by replacing the *%* character with *%25* before running the discover job.
>
> For example, if the datastore path is */foo/%20*, then you must change the path to */foo/%2520*.

10. Click **Advanced Settings** to display the **Config** text area, where you can override default configuration parameters for each datastore.

    By default, the **Config** text area displays the configuration parameters specific for each datastore as a placeholder.
11. Specify the advanced configuration settings for the targeted system.

    For more information about the advanced configuration settings, refer to the section *Advanced Configuration Settings*.
12. If you want to send the scan information in a POST response from the Protegrity Discover machine to an external URL each time the discover job identifies sensitive data in the selected datastore, then you need to use a webhook.

    For more information about using webhooks, refer to the section *Using Webhooks*.
13. Click **Test** to test the system connectivity.
14. Click **Save** to save the created job.
15. Click **Run** to run the created discover job manually.

The created discover job is listed in the **Jobs** screen with its status reflected in the *Status* column. The results of the discover job in *running* status start appearing in the overview dashboard screen instantly.

6.4.1.2 Manage Discover Jobs

After the discover job is created, you can manage the existing jobs. This section lists the additional actions that you can perform pertaining to a job.

After you create a discover job, it is listed in the **Jobs** screen as illustrated in the following figure.

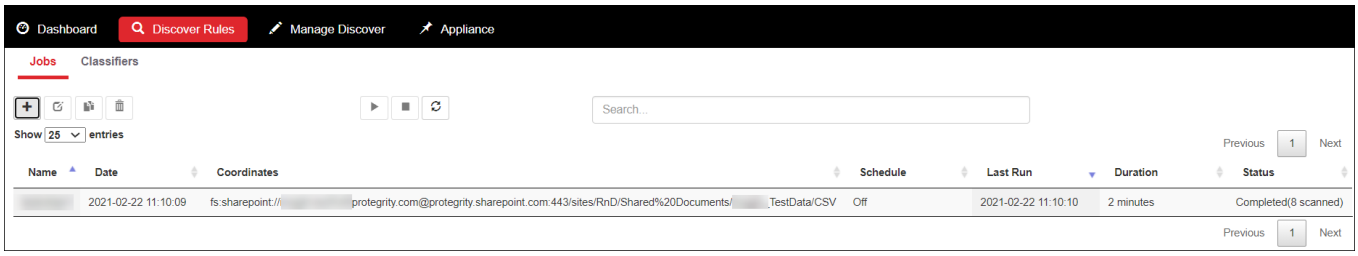


Figure 6-11: Protegrity Discover - List of Discover Jobs

You can manage the created discover jobs from the Web UI using additional actions, as mentioned in the following table.

Table 6-10: Protegrity Discover - Manage Discover Jobs

Action	Icon	Description
Edit a discover job		Enables you to edit a discover job from the list. You can edit the scan parameters mentioned in the *Discover Job Configuration Settings* table.
Clone a discover job		Enables you to create a copy of an existing discover job. If you select an existing job and click the Clone icon, then the **Clone Discover Job** dialog box appears. The text *clone* is appended to the name of the copied job. All the existing scan parameters of the copied job, except the authentication password or secret access key, are retained. You can choose to edit the scan parameters. For more information on editing the scan parameters, refer to the *Discover Job Configuration Settings* table.
Remove a discover job		Enables you to remove a discover job from the list.
Refresh a discover job		Enables you to refresh a discover job from the list.
Run a discover job		Enables you to run a discover job.
Stop a discover job		Enables you to stop an ongoing discover job from the list.

You can also search for a discover job from the list.

6.4.2 Managing Classifiers

Classifiers validate and classify sensitive data from the sampled records on the targeted system. This section describes how you can add and modify classifiers.

For more information about how the classifiers validate the data, refer to the section *Classifier Configuration*.

6.4.2.1 Creating Classifiers

This section describes how you can create custom classifiers for identifying sensitive data.

To create classifiers:

1. On the Protegrity Discover Web UI, navigate to **Discover Rules** > **Classifiers**.
   The **Classifiers** screen displays the list of default classifiers.

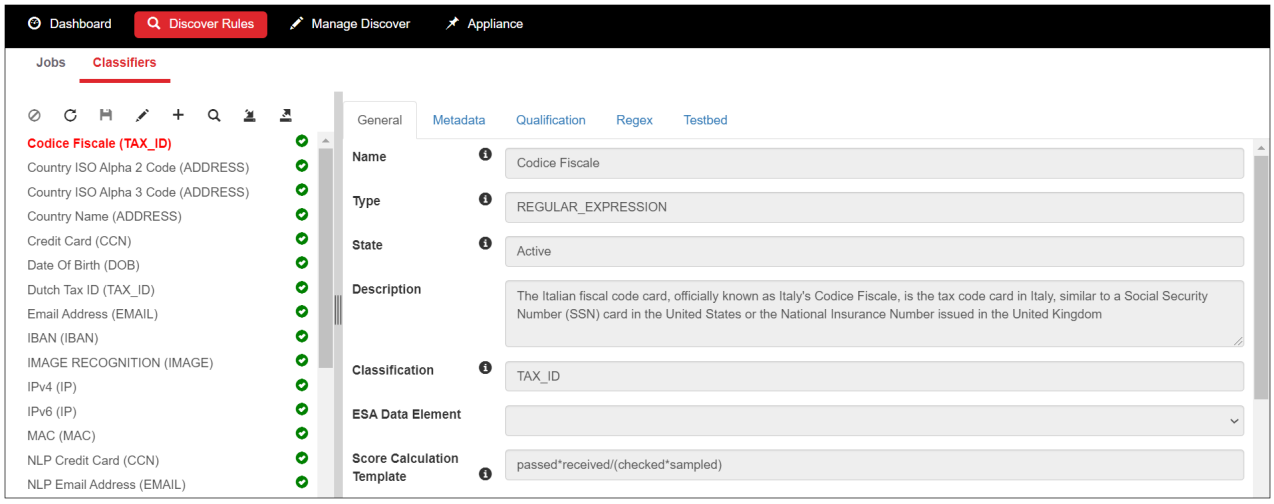


Figure 6-12: Classifiers Screen

2. Click ➕.
   The **Add New Classifier** dialog box appears.
3. In the **Name** field, enter a unique name for the classifier.

   Note: Do not include the following special characters in the **Name** field:
   - /
   - \
   - :
   - *
   - ?
   - <
   - >
   - |

4. In the **Type** list, select one of the following classifier types:
   - *User Defined* - Write custom code in Python to identify sensitive structured data.

     Warning: Protegrity Discover enables you to add custom Python code so that you can create your own classifiers. Any incorrect, improper, or malicious use of the custom Python code can cause serious, system-wide problems or result in security vulnerabilities. Protegrity is not responsible for any damages caused due to the use of custom Python code. Use this option at your own risk.

   - *Regular Expression* - Use regular expressions to identify sensitive structured data.
   - *Dictionary* - Upload a reference table that contains a list of words or phrases that you want to identify. For example, cities, countries, postal codes, and country codes. This classifier is used to identify sensitive structured data.
   - *NLP* - Create spaCy patterns to identify sensitive unstructured data using natural language processing. spaCy is a free and open-source library for advanced natural language processing in Python.

     For more information about spaCy, refer to the *spaCy* website.

   The classifier name appears in the **Classifiers** list. The classifier opens in edit mode, and the **General** tab appears by default. The details required for creating a classifier depend on the type of classifier.

5. Enter the details in the applicable tabs based on the classifier type.

Type of Classifier	Applicable Tabs
User Defined	• *General* • *Metadata* • *Qualification* • *Source Code*
Regular Expression	• *General* • *Metadata* • *Qualification* • *Regex*
Dictionary	• *General* • *Metadata* • *Qualification* • *Reference*
NLP	• *General* • *Metadata* • *Spacy Patterns*

6. Test the classifier.
   For more information about testing the classifier, refer to the section *Testing the Classifier*.
7. Click to save the classifier.
   Alternatively, if you do not want to save the changes done to the classifier, then perform *step 8*.
8. Perform the following steps if you do not want to save the changes done to the classifier.
   a. Click .
   The **Save Changes** dialog box appears that prompts you to save the changes, discard the changes, or close the dialog box.
   b. Click **Discard** to discard the changes.
   The classifier is deleted and is removed from the **Classifiers** list.

   Alternatively, if you want to delete the classifier, then click . The classifier is deleted and removed from the **Classifiers** list.

6.4.2.1.1 Updating the General Tab

This section describes how to update the details in the **General** tab.

To update the **General** tab:

1. On the **Classifiers** screen, click **General**.
   The following **General** tab appears.

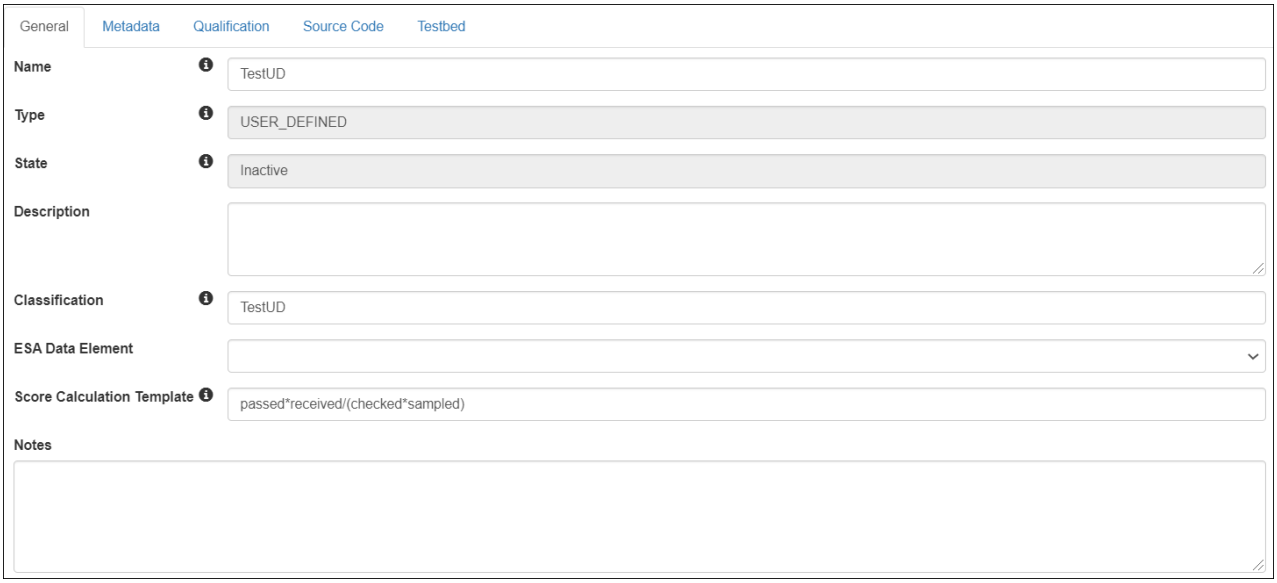


Figure 6-13: General Tab

By default, the **General** tab appears when you create a new classifier.

2. Specify the following details.

Field	Description
Name	Specify a unique name to identify the classifier.
Type	Indicates the type of classifier. This field is populated by default and is uneditable. This field can have one of the following values for a new classifier: • USER_DEFINED • REGULAR_EXPRESSION • DICTIONARY • NLP This field can have one of the following values for the default classifiers: • REGULAR_EXPRESSION • DICTIONARY • CREDIT_CARD • DATE • DUTCH_TAX_ID • EMAIL_ADDRESS • USER_DEFINED • IMAGE_RECOGNITION • NLP_CREDIT_CARD • NLP_EMAIL_ADDRESS • NLP_NAME • NLP_PHONE_NUMBER • NLP_US_ADDRESS • NLP_US_SOCIAL_SECURITY_NUMBER • PHONE_NUMBER • US_SOCIAL_SECURITY_NUMBER
State	Identifies the state of the classifier. It can have two values:

Field	Description
	• *Inactive* - By default, when you create a new classifier, its state is set to *Inactive*. Protegrity Discover does not use an inactive classifier to identify sensitive data while performing a scan. If you want to activate an inactive classifier, then you must ensure that the classifier passes the testbed test. An inactive classifier is also indicated by the  icon appearing next to the classifier in the **Classifiers** list. For more information about the testbed test, refer to the *Testing the Classifier* section. • *Active* - Protegrity Discover uses an active classifier to identify sensitive data while performing a scan. An active classifier is also indicated by the  icon appearing next to the classifier in the **Classifiers** list. **Note:** The default classifiers are always in active state.
Description	Type a description for the classifier.
Classification	Specify a tag name or data element name to categorize the classifier. You can specify the same classification name for multiple classifiers. Protegrity Discover groups together the scan results of all the classifiers that have the same classification name. Protegrity Discover uses the classification name to categorize the scan results on the **Overview** screen, where it displays the top five data elements. For more information on the Overview screen, refer to the *Data Discover Overview* section. By default, when you create a new classifier, the value of the **Classification** field is set to the value of the **Name** field.
ESA Data Element	Select the ESA data element that you want to associate with the classifier. By default, this list is empty. After you retrieve the data elements from the ESA, this list is auto-populated with the retrieved data elements. For more information about retrieving data elements from the ESA, refer to the section *Retrieving ESA Data Elements*.
Score Calculation Template	Specify the template for calculating the confidence score. By default, the value of this field is set to `passed*received/(checked*sampled)`. You can choose to modify the score calculation template. However, only the following variables are supported in the score calculation template: • sampled • received • qualified • checked • passed • duplicate • empty For more information on the values supported in the score calculation template, refer to the *Details* attribute in the *Protegrity Discover: Classification Record Findings* table. **Note:** The score calculation template is not applicable for NLP classifiers. Therefore, this field displays *N/A* value, which is grayed out.
Notes	Add any notes for the given classifier.

6.4.2.1.2 Updating the Metadata Tab

This section describes how to use the **Metadata** tab to specify regex patterns, a tool used to identify schema keywords within close proximity to data that is being searched by the classifier. Each keyword is associated with a boost or score that adjusts the final confidence score when a match is found.

➤ To update the **Metadata** tab:

1. On the **Classifiers** screen, click **Metadata**.

   The following **Metadata** tab appears.

Figure 6-14: Metadata Tab

2. Click + to add a new Regex pattern in the **Keyword Patterns** area.

   The following **Add New Metadata Keyword** dialog box appears.

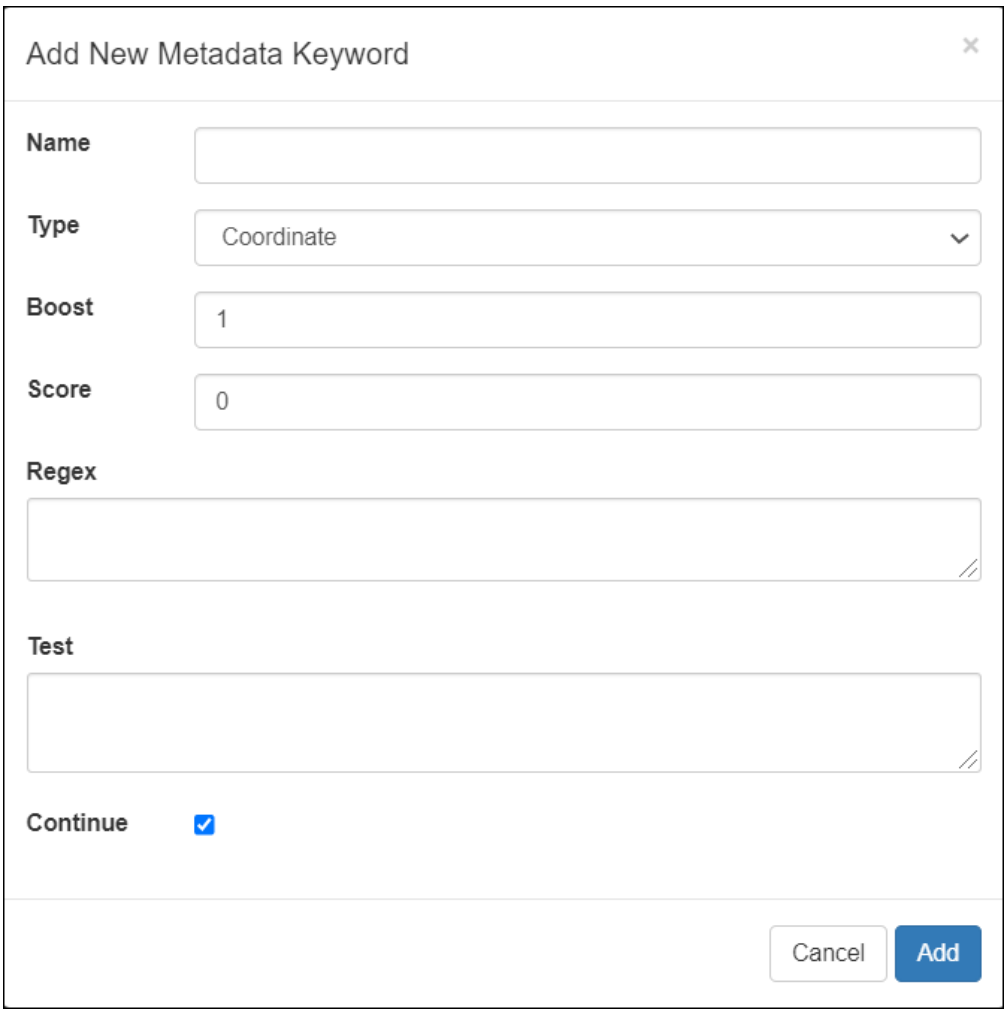


Figure 6-15: Add New Metadata Keyword - Coordinate Metadata

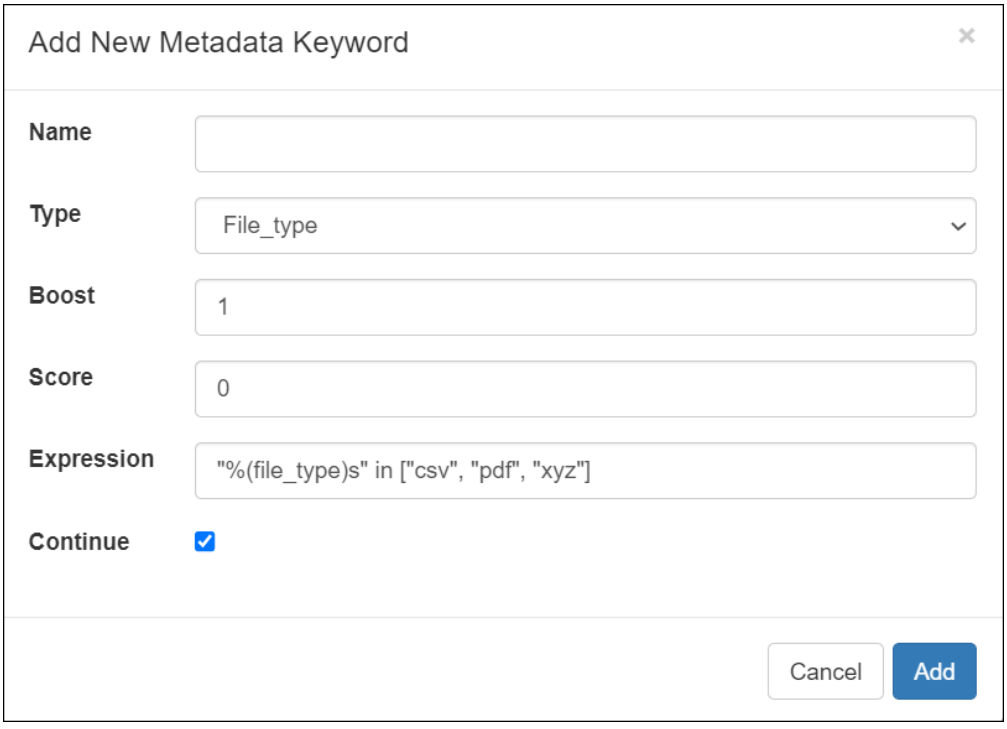


Figure 6-16: Add New Metadata Keyword - File Metadata

3. Specify the following details in the **Add New Metadata Keyword** dialog box.

Field	Description
Name	Specify a unique name for the Regex pattern.
Type	Specify one of the following metadata types: • Coordinate - The regular expression is applicable to the data store coordinates. If you select **Coordinate**, then you need to specify the regular expression in the *Regex* field. • File metadata: The Python boolean expression is applicable to the following file metadata: • File_type • Modified_time • File_size • Owner • Group • Created_time • Permission • Storage_class • Aws_metadata • Aws_tags • Azure_metadata • Azure_tags If you select a file metadata, then you need to specify the Python boolean expression in the *Expression* field. For more information about the list of file metadata and the applicable filestores, refer to the section *Appendix J: File Metadata Collected in Filestores*.
Boost	Specify a value to boost the confidence score. If the Regex pattern matches with any metadata, then the boost value is applied to the confidence score. You can also set the value of this field to *0*, if you want to filter out certain keywords. For example, in the Date of Birth classifier, the boost value for identifying the *start* or *end* keyword is set to *0*. Therefore, if Protegrity Discover detects a column containing dates that has a column name as *start* or *end*, then it will not classify the dates as date of birth. For more information on how the boost value is used to calculate the final confidence score, refer to the section *Analysis*.
Score	Specify the confidence score value. If the Regex pattern matches with any metadata, then the score value is added to the final confidence score. By default, the score value is set to *0*. The maximum score value can be set to *1*, where 1 represents 100 percent. Typically, you can set the value of this field to greater than 0 in cases where you want to detect the location of the data by keywords only, even if no data is detected. For example, consider that in a database you have a column named *birthdate*. However, this column does not have any data. If you have created a regex pattern for identifying the *birthdate* keyword and set its confidence score value to more than *0*, then Protegrity Discover can detect and classify this column even if it does not contain any data. **Important:** If the classifier returns the confidence score as 1, then the classifier accurately identifies the file using the metadata. In this case, the classifier does not scan the contents of the file.

Field	Description
	For more information on how the confidence score is used to calculate the final confidence score, refer to the section *Analysis*.
Regex	Specify the Regex pattern for identifying the Coordinate metadata keyword. For example, the following Regex pattern is used to identify the *birthdate* keyword: `(?i)b(irth)?[\s_-]?d(ate)?` **Note:** This field is applicable only to the Coordinate metadata.
Expression	Specify the Python boolean expression for identifying the file metadata keyword. The classifier provides the following default placeholders for creating custom expressions: • *File_type* - `"%(file_type)s" in ["csv", "pdf", "xyz"]` Specify the extension of the file type that you want to classify. **Important:** Ensure that you specify the file extension in lowercase. For example, if you want to search for *.mp3* or *.mp4* files, then you can specify the expression as `"%(file_type)s" in ["mp3", "mp4"]`. • *Modified_time* - `%(modified_time)s < datetime.datetime(2021, 1, 31, 23, 50, 59).timestamp()` You can classify the files based on the time that they were modified. You can specify the input date in the *YYYY M DD HH MM SS* format. For example, if you want to find out all the files that were modified before 11:59:59 PM on January 10, 2020, then you can specify the value of the expression as `%(created_time)s < datetime.datetime(2020, 1, 10, 23, 59, 59).timestamp()`. Protegrity Discover uses the Python *datetime.timestamp()* method in the *datetime* module to convert the input data into an Epoch time, which is also known as the Unix time or POSIX time. **Note:** Epoch time specifies the number of seconds that have elapsed after the Unix Epoch, which is 00:00:00 UTC on 1 January 1970. For more information about the *datetime.timestamp()* method, refer to the *Python documentation*. • *File_size* - `%(size)s > 1024**3` You need to specify the value of the file size in bytes. By default, the value is set to `1024**3`, which is equal to 1 GB. • *Owner* - `"%(owner)s" == "<Add your expected value>"` Specify the name of the user who has created the file. You can specify the value based on the format used for the owner name in your file system.

Field	Description
	For example, if your file system specifies the owner name as *First_name Last_name*, then specify the value of the expression as `"%(owner)s" == "First_name Last_name"`. However, if your file system specifies the owner name as first_name.last_name, then specify the value of the expression as `"%(owner)s" == "first_name.last_name"`. You can also specify multiple values using an array. For example, if you want to search files whose owners are User 1 and User 2, then you can specify the expression as `"%(owner)s" in ["User1_first_name.last_name", "User2_first_name.last_name"]`. • *Group* - `"%(group)s" == "<Add your expected value>"` Specify the name of the group whose members have permissions to access the file. You can specify the value based on the format of the group name used in your file system. • *Created_time* - `%(created_time)s < datetime.datetime(2021, 1, 31, 23, 50, 59).timestamp()` You can classify the files based on the time that they were created. The *created_time* metadata uses the same format as that of the *modified_time* metadata. For more information about the formatting the *created_time* metadata, refer to the *modified_time*. • *Permission* - `"%(permission)s" == "<Add your expected value>"` Specify the file permission values in case of HDFS and NFS. For example, if you want to search for files where the owner of the file has read and write access, then you can specify the expression as `"%(permission)s" == "644"`. • *Storage_class* - `"%(storage_class)s" == "<Add your expected value>"` Specify the name of the AWS S3 storage class where the file has been stored. For example, if you are using **Standard** storage class to store the files in AWS S3, then you can specify the value of the expression as `"%(storage_class)s" == "Standard"`. For example, if you are using both **Standard** and **Standard-IA** storage classes to store the files in AWS S3, then you can specify the value of the expression as `"%(storage_class)s" in ["Standard", "Standard-IA"]`. • *Aws_metadata* - `"%(aws_metadata)s" == "<Add your expected value>"` Specify the metadata created by users in AWS. **Important:** Only user-defined metadata is supported. • *Aws_tags* - `"%(aws_tags)s" == "<Add your expected value>"` Specify the tags created by end users in AWS. • *Azure_metadata* - `"%(azure_metadata)s" == "<Add your expected value>"` Specify the metadata created by users in Azure. • *Azure_tags* - `"%(azure_tags)s" == "<Add your expected value>"`

Field	Description
	Specify the tags created by end users in AWS.
Test	Specify test values for testing the Regex pattern. While you are typing the keywords, Protegrity Discover automatically checks whether they match the Regex pattern. **Note:** This field is applicable only to the Coordinate metadata.
Continue	Select this check box to include this pattern and all subsequent patterns in the **Keyword Patterns** area to identify the metadata. If you clear this check box, then the corresponding Regex pattern is included for identifying the metadata. However, all subsequent Regex patterns are excluded from identifying the metadata. For more information about how the **Continue** check box is used to include or exclude the metadata keywords during the scanning process, refer to *step 7*.

Important: The metadata keyword appears on the **Classifications** screen in the Protegrity Discover Web UI. If the metadata keyword contains any sensitive data, then the sensitive data will appear on the **Classifications** screen.

4. Click **Add** to add the Regex pattern.
   The Regex pattern is added to the **Keyword Patterns** area.
5. Repeat steps *2* to *4* to add multiple Regex patterns for identifying the metadata.
6. If you want to edit any metadata keyword, then perform the following steps:
   a. Select the metadata keyword.
   b. Click  .
      The following **Edit Metadata Keyword** dialog box appears.

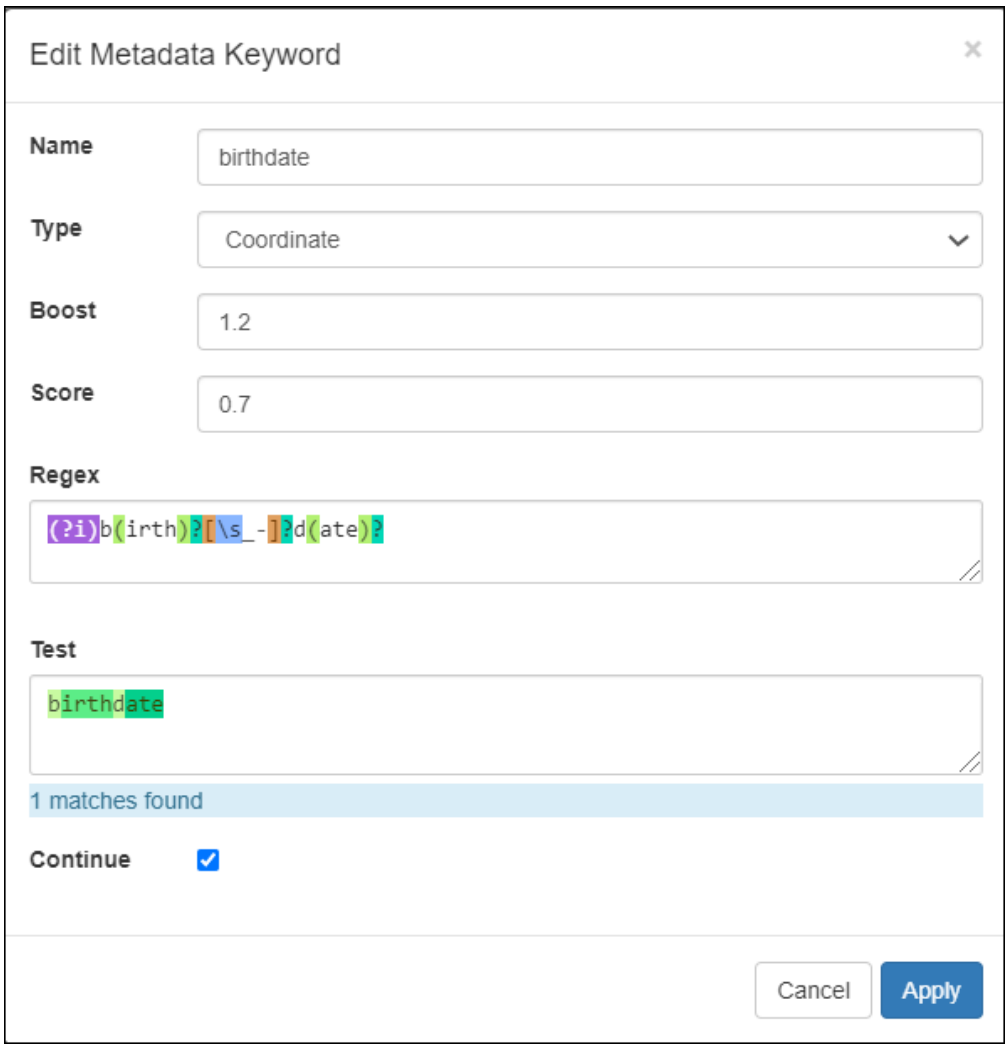


Figure 6-17: Edit Metadata Keyword

c. Modify the details, if required.

d. Click **Apply** to apply the changes.

7. If you want to delete any metadata keyword, then perform the following steps:

   a. Select the metadata keyword.

   b. Click ✖.

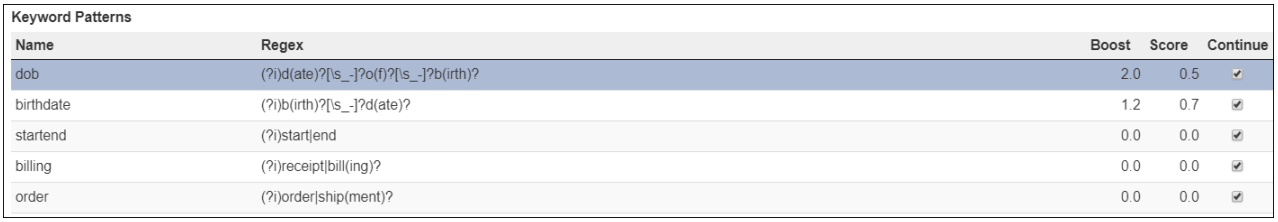
   The **Remove Metadata Keyword** dialog box appears.

   c. Click **Remove** to delete the metadata keyword.

8. Select a keyword from the **Keyword Patterns** area, and then click ⬇ or ⬆ to change the order of the keyword, if required.

   If you have cleared the **Continue** check box for a particular keyword, then the order of the remaining keywords in the **Keyword Patterns** area determines whether they are included in the scanning process.

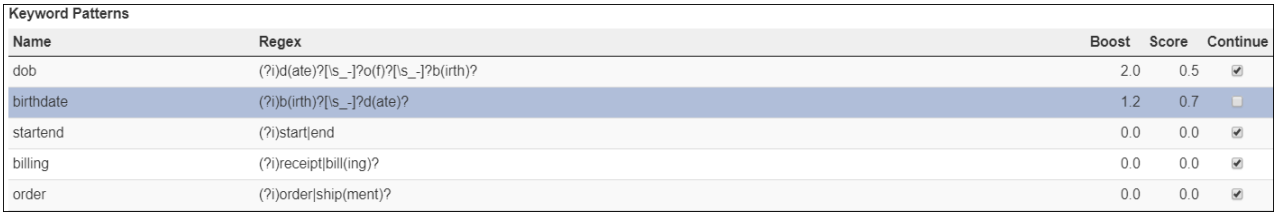
   For example, the following figure shows the sample keyword patterns that are created to identify the metadata keywords for the *Date Of Birth* classifier. In this case, Protegrity Discover will use all five keyword patterns, namely, *dob*, *birthdate*, *startend*, *billing*, and *order*, to identify the corresponding metadata keywords during the scanning process.



Keyword Patterns

Name	Regex	Boost	Score	Continue
dob	(?i)d(ate)?[\s_-]?o(f)?[\s_-]?b(irth)?	2.0	0.5	☑
birthdate	(?i)b(irth)?[\s_-]?d(ate)?	1.2	0.7	☑
startend	(?i)start|end	0.0	0.0	☑
billing	(?i)receipt|bill(ing)?	0.0	0.0	☑
order	(?i)order|ship(ment)?	0.0	0.0	☑

If you clear the **Continue** check box corresponding to the *birthdate* keyword pattern, as shown in the following figure, then all the keyword patterns that are below the *birthdate* keyword pattern will be skipped from the scanning process. For example, Protegrity Discover will not use the *startend*, *billing*, and *order* keyword patterns for identifying the corresponding metadata keywords.



Keyword Patterns

Name	Regex	Boost	Score	Continue
dob	(?i)d(ate)?[\s_-]?o(f)?[\s_-]?b(irth)?	2.0	0.5	☑
birthdate	(?i)b(irth)?[\s_-]?d(ate)?	1.2	0.7	☐
startend	(?i)start|end	0.0	0.0	☑
billing	(?i)receipt|bill(ing)?	0.0	0.0	☑
order	(?i)order|ship(ment)?	0.0	0.0	☑

9. Perform the following steps to add the test data values for testing the Regex patterns:
   a. Click **Add Test Data**.
      The **Add Keyword Test Data** dialog box appears.
   b. In the **Keyword Test Data** area, specify the test values for testing the Regex patterns.

      > **Note:** It is recommended to use a comma or a new line to separate the test values.

   c. Click **Add**.
      The test values are added to the **Keywords Test** area.
10. Click **Test** to match the test data values that are listed in the **Keywords Test** area against all the Regex patterns specified in the **Keywords Pattern** area.

6.4.2.1.3 Updating the Qualifications Tab

This section describes how to specify a regex pattern in the **Qualifications** tab for qualifying a value before evaluating it. The regex can be used to search for or rule out values that should not be scanned by the classifier.

To update the Qualification tab

1. On the **Classifiers** screen, click **Qualification**.
   The following **Qualification** tab appears.

Figure 6-18: Qualification Tab

2. Specify the required regex pattern in the **Qualification Pattern** area.

   For example, the Password classifier has the following qualification pattern:

   ```
   ^.{6,16}$
   ```

   In this case, the qualifying pattern specifies that the value should not be less than 6 characters and more than 16 characters. Therefore, the Password classifier will restrict the data set used to identify whether a particular value is a password to only those values that match the criteria specified by the qualification pattern.

3. Specify test data values in the **Qualifying Values** area to test whether the regex pattern can identify the data.

   Ensure that any new data value should be on a new line.

   As you type the data in the **Qualification Values** area, Protegrity Discover tries to match the test data values with the regex pattern and highlights the data if there is a match.

6.4.2.1.4 Updating the Source Code Tab

This section describes how you can add a custom Python code in the **Source Code** tab for identifying sensitive data.

To update the **Source Code** tab

1. On the **Classifiers** screen, click **Source Code**.
   The following **Source Code** tab appears.

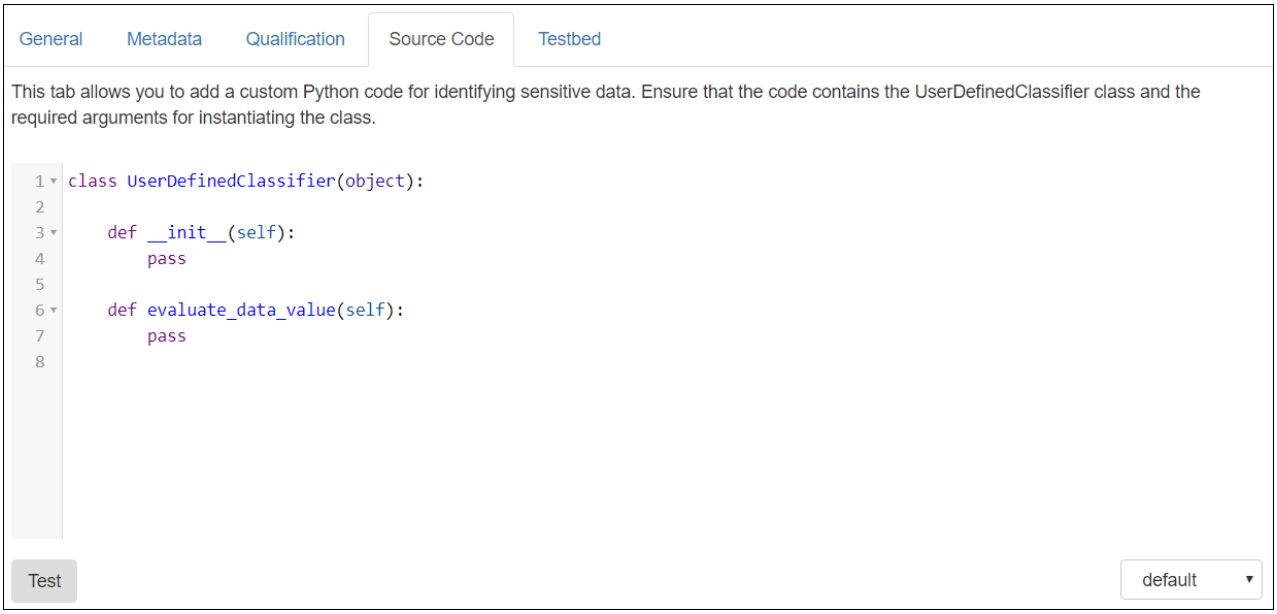


Figure 6-19: Source Code Tab

By default, the following code block appears:

```
class UserDefinedClassifier(object):

    def __init__(self):
        pass

    def evaluate_data_value(self):
        pass
```

Important: Ensure that your code contains the ***`UserDefinedClassifier`*** class.

2. Add your custom Python code to identify the sensitive data.

   For example, the following snippet shows the sample Python code written for the default *IBAN* classifier used to identify the international bank account number:

```
class UserDefinedClassifier(object):

    LETTER_DICT = {'A': 10, 'B': 11, 'C': 12, 'D': 13, 'E': 14, 'F': 15, 'G': 16, 'H':
17, 'I': 18, 'J': 19, 'K': 20,
                   'L': 21, 'M': 22, 'N': 23, 'O': 24, 'P': 25, 'Q': 26, 'R': 27, 'S':
28, 'T': 29, 'U': 30, 'V': 31,
                   'W': 32, 'X': 33, 'Y': 34, 'Z': 35, 'a': 10, 'b': 11, 'c': 12, 'd':
13, 'e': 14, 'f': 15, 'g': 16,
                   'h': 17, 'i': 18, 'j': 19, 'k': 20, 'l': 21, 'm': 22, 'n': 23, 'o':
24, 'p': 25, 'q': 26, 'r': 27,
                   's': 28, 't': 29, 'u': 30, 'v': 31, 'w': 32, 'x': 33, 'y': 34, 'z':
35, '0': 0, '1': 1, '2': 2,
                   '3': 3, '4': 4, '5': 5, '6': 6, '7': 7, '8': 8, '9': 9}

    IBAN_SETTING_DICT = {15: ['NO'], 16: ['BE', 'BI'], 18: ['DK', 'FI', 'FO', 'GL',
'NL'], 19: ['MK', 'SI'],
                         20: ['AT', 'BA', 'EE', 'KZ', 'LT', 'LU', 'XK'], 21: ['CH', 'HR',
'LI', 'LV'],
                         22: ['BG', 'BH', 'CR', 'DE', 'GB', 'GE', 'GG', 'IE', 'IM', 'JE',
'ME', 'RS'],
                         23: ['AE', 'GI', 'IL', 'IQ', 'TL'],
                         24: ['AD', 'CZ', 'DZ', 'ES', 'MD', 'PK', 'RO', 'SA', 'SE', 'SK',
'TN', 'VG'],
                         25: ['AO', 'CV', 'GW', 'MZ', 'PT', 'ST'], 26: ['IR', 'IS', 'TR'],
                         27: ['CF', 'CG', 'CM', 'DJ', 'EG', 'FR', 'GA', 'GQ', 'GR', 'IT',
```

```
'KM', 'MC', 'MG', 'MR', 'SM',
                                'TD'],
                           28: ['AL', 'AZ', 'BF', 'BJ', 'BY', 'CI', 'CY', 'DO', 'GT', 'HN',
'HU', 'LB', 'MA',
                                'ML', 'NE', 'PL', 'SN', 'SV', 'TG'],
                           29: ['BR', 'PS', 'QA', 'UA'], 30: ['JO', 'KW', 'MU'], 31: ['MT',
'SC'], 32: ['LC', 'NI']}

    def __init__(self):
        import re
        self.exp = re.compile('[^\\w]*')
        # here setting will be read from config rather than hard code when we create a
IBAN classifier
        self.letters = {ord(k): str(v) for k, v in self.LETTER_DICT.items()}

    def evaluate_data_value(self, coordinate, data_value):
        if not isinstance(data_value, str):
            data_value = str(data_value)

        characters = self.exp.sub('', data_value)

        length_characters = len(characters)
        isIBAN = False

        if length_characters in self.IBAN_SETTING_DICT.keys() and characters[:2].upper()
in self.IBAN_SETTING_DICT[length_characters]:
            # Move first 4 chars to end of string
            iban_string = characters[4:] + characters[:4]
            # Assemble digit string, translate LETTERS to digits
            digit_string = iban_string.translate(self.letters)
            # MOD 97 checksum
            isIBAN = (1 == int(digit_string) % 97)

        return isIBAN and self.check_digit(characters)

    def check_digit(self, data_value):
        check_digit = int(data_value[2:4])
        characters = data_value[:2] + '00' + data_value[4:]
        iban_string = characters[4:] + characters[:4]
        digit_string = iban_string.translate(self.letters)
        return check_digit == (98 - int(digit_string) % 97)
```

Warning: Protegrity Discover enables you to add custom Python code so that you can create your own classifiers. Any incorrect, improper, or malicious use of the custom Python code can cause serious, system-wide problems or result in security vulnerabilities. Protegrity is not responsible for any damages caused due to the use of the custom Python code. Use this option at your own risk.

Note: If you modify the default code, then a message is displayed below the text editor informing you that the source code has changed, and you must test the classifier to validate the source code.

3. If required, then change the default theme of the native code editor by choosing a value from the drop-down list located at the bottom-right of the screen.
4. Click **Test**.

   Protegrity Discover compiles the source code and validates the Python syntax. It also validates whether the code complies with the requirements of the user-defined class, such as, ensuring that the code contains the **UserDefinedClassifier** class and the required arguments for instantiating the class. Any success or error message appears below the source code.

6.4.2.1.5 Updating Regex Tab

This section describes how to specify the regex patterns for identifying the sensitive data.

To update the **Regex** tab:

1. On the **Classifiers** screen, click **Regex**.
   The following **Regex** tab appears.

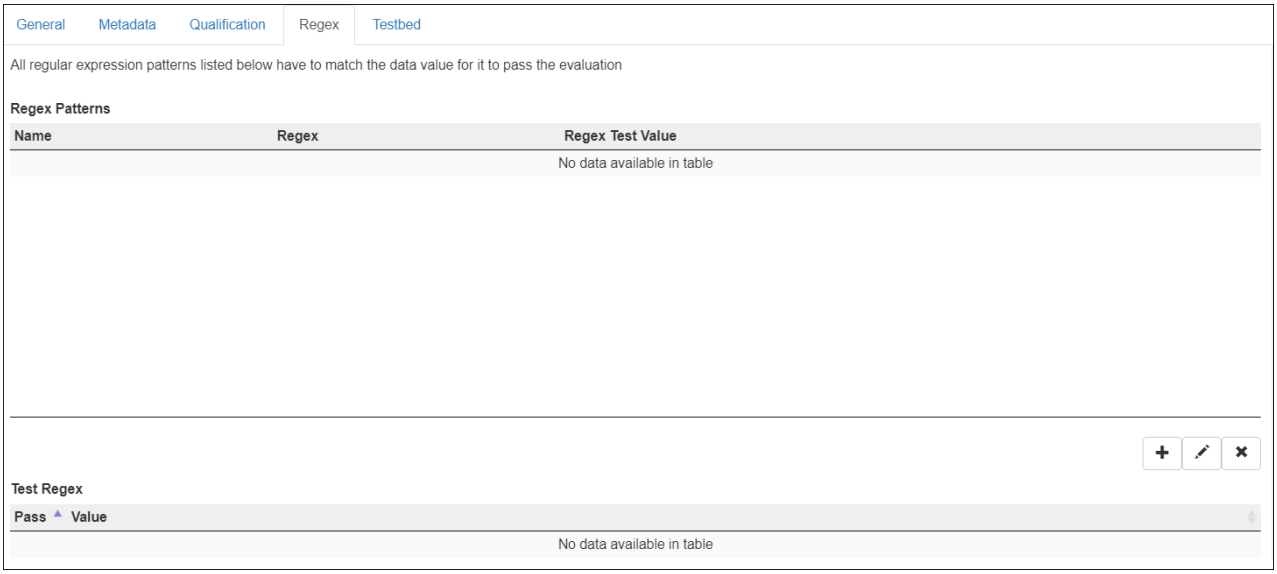


Figure 6-20: Regex Tab

2. Click **New** to add a new regex pattern in the **Regex Patterns** area.
   The following **Add New Regex Pattern** dialog box appears.

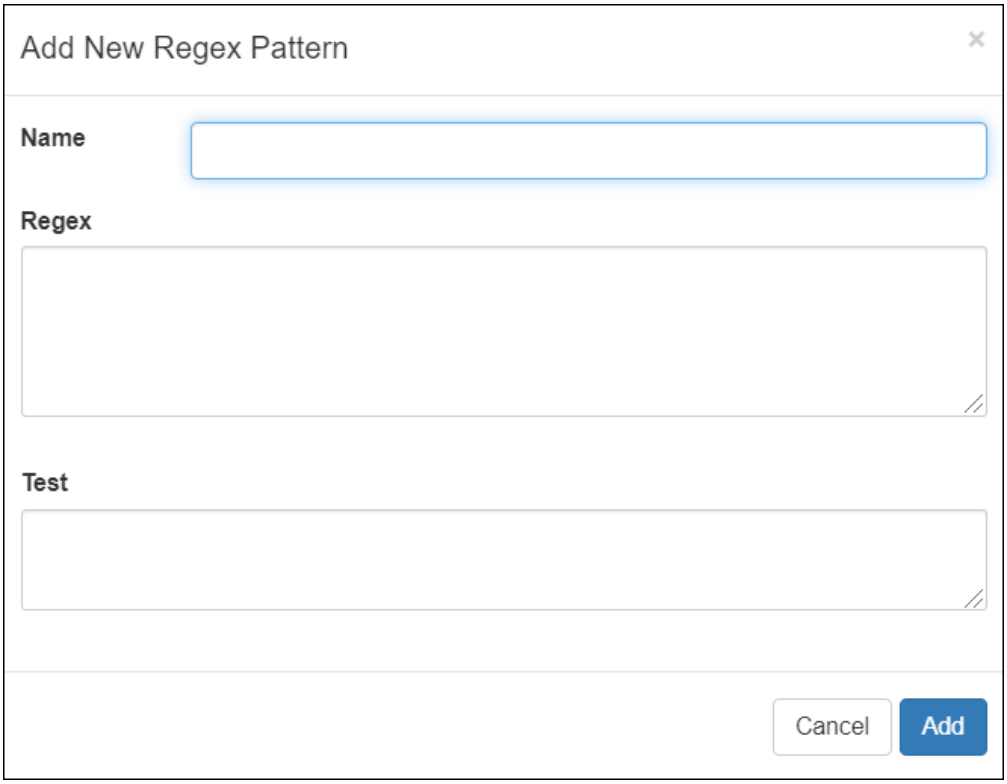


Figure 6-21: Add New Regex Pattern

3. Specify the following details in the **Add New Regex Pattern** dialog box.

Field	Description
Name	Specify a unique name for the regex pattern.

Field	Description
Regex	Specify the regex pattern for identifying the sensitive data. For example, the following snippet shows the regex pattern specified in the Password classifier for identifying uppercase letters. `[A-Z]`
Test	Specify test values for testing the regex pattern. Ensure that you specify each data value on a new line.

4. Click **Add** to add the regex pattern to the **Regex Patterns** area.
5. Repeat *step 2* to *step 4* to add multiple regex patterns.
6. If you want to edit any regex patterns, then perform the following steps:
   a. Select the regex pattern.
   b. Click .
      The following **Edit Regex Pattern** dialog box appears.

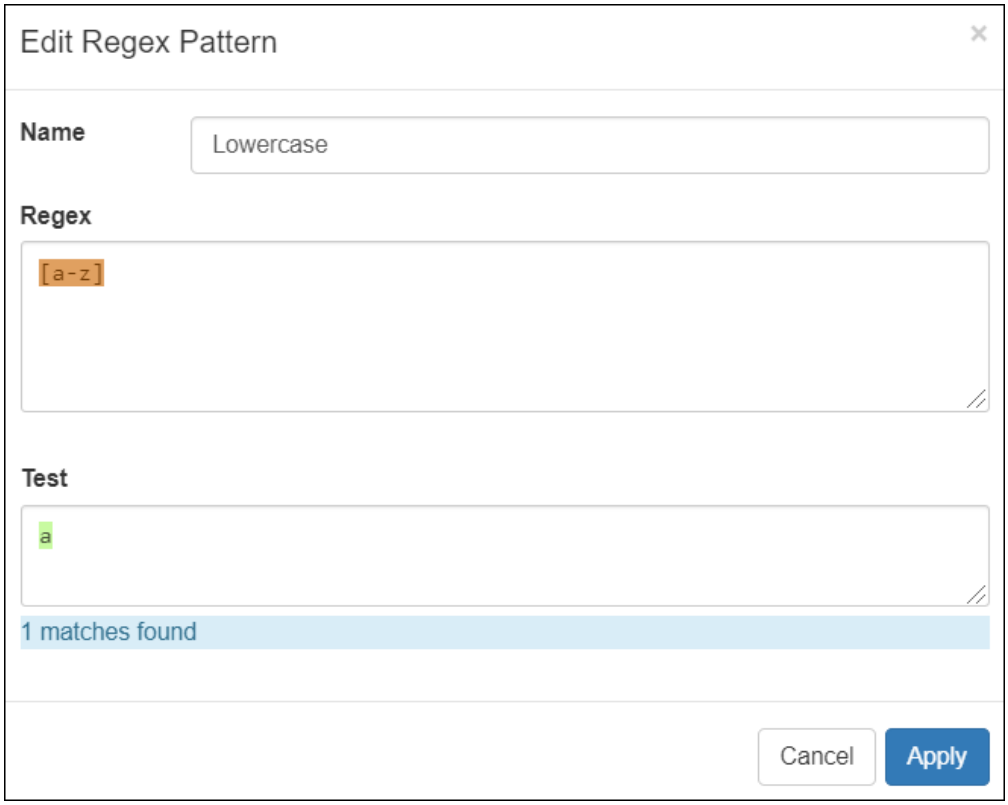


Figure 6-22: Edit Regex Pattern

   c. Modify the regex pattern and test values, if required.
   d. Click **Apply** to apply the changes.
7. If you want to delete any regex pattern, then perform the following steps:
   a. Select the regex pattern.
   b. Click .
      The **Remove Regex Pattern** dialog box appears.
   c. Click **Remove** to delete the regex pattern.

8. Click **Add Test Data** to add test data to verify the regex patterns.
   The following **Add Pattern Test Data** dialog box appears.

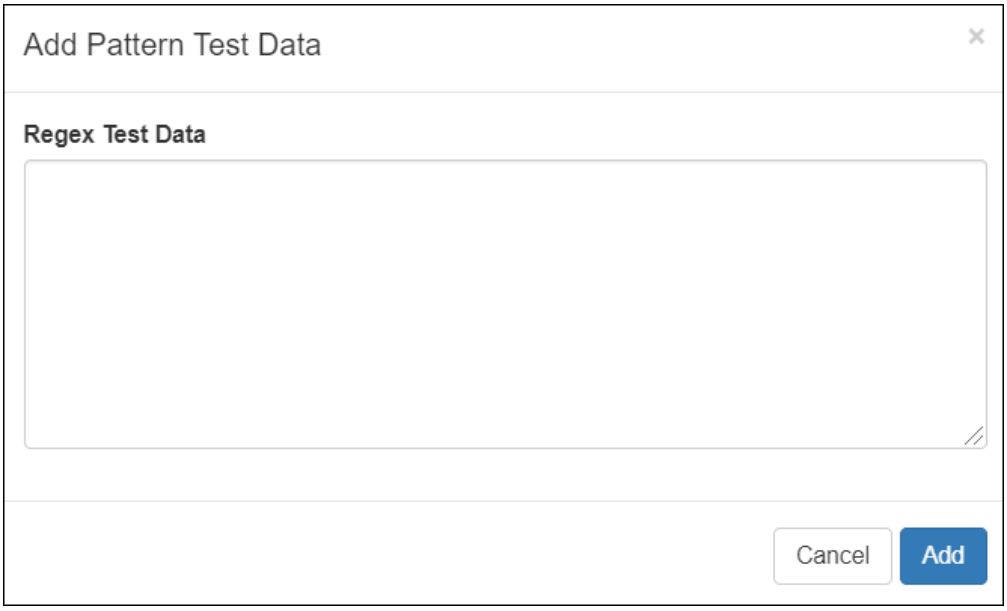


Figure 6-23: Add Pattern Test Data Dialog Box

9. Specify the test data in the **Regex Test Data** area.
   The test data that you specify here is applicable for all the regex patterns that you specified in the **Regex Pattern** area.
   Protegrity Discover considers any data value that is on the same line as part of a single data item. Ensure that you add multiple test data values on separate lines.

10. Click **Add** to add the test data.
    The test data appears in the **Test Regex** area.
11. Click **Test** to test the data.
    The status of the test appears in the **Pass** column of the **Test Regex** table. If the test data matches the regex pattern, then a tick symbol appears in the **Pass** column. If the test data does not match the regex pattern, then a cross symbol appears in the **Pass** column next to the test data.

6.4.2.1.6 Updating the Reference Tab

This section describes how you can create a pre-defined list of reference data values, such as names of cities, countries, and states. Protegrity Discover uses this pre-defined list to identify sensitive data. The file that contains the reference data to be uploaded must either be in *CSV* format or in *TXT* format with pre-defined data separators for separating the individual data items.

➤ To update the **Reference** tab:

1. On the **Classifiers** screen, click **Reference**.
   The **Reference** tab appears. You can choose to create a new table or append the data to an existing table.

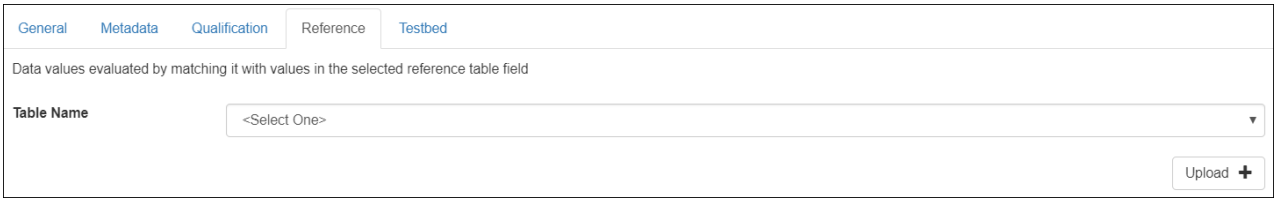


Figure 6-24: Reference Tab

2. Click **Upload**.

   The following **Upload Reference Table** dialog box appears.

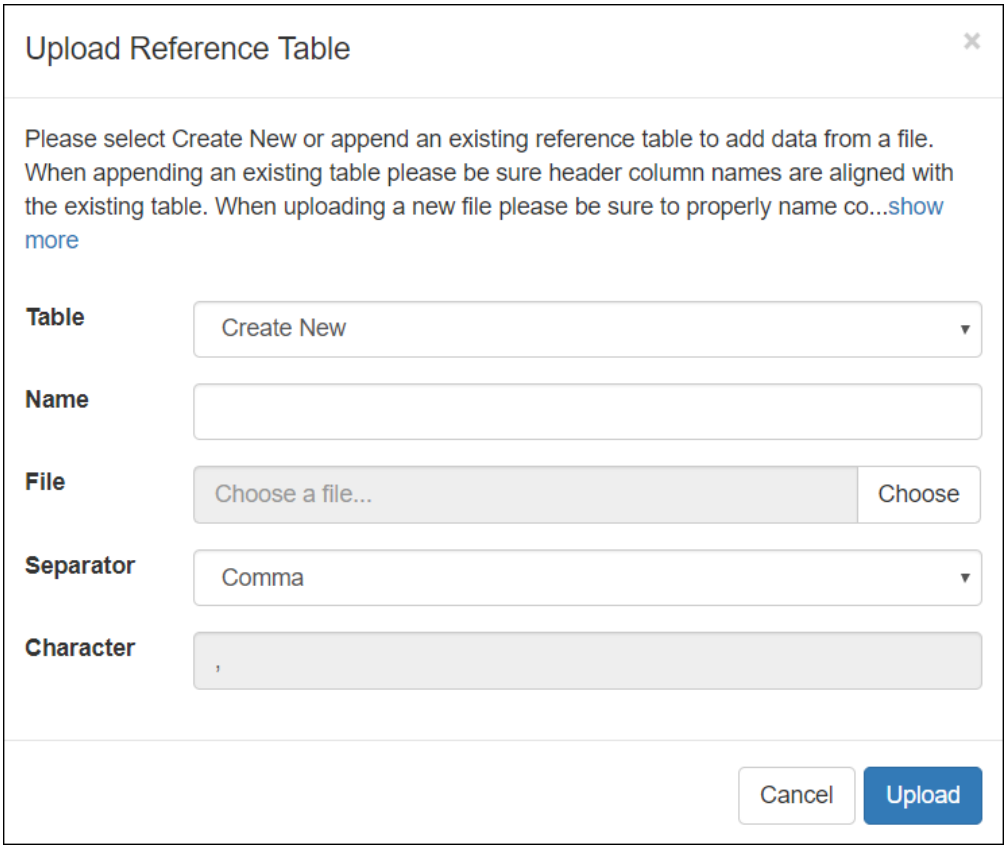


Figure 6-25: Upload Reference Table

3. Enter the following details in the **Upload Reference Table** dialog box.

Field	Description
Table	Choose one of the following values: • **Create New** - Create a new table. • **Append Existing** - Append the reference data to an existing table. **Important:** If you append data to an existing table, then you cannot revert the changes. Therefore, if you want to add reference data, then it is recommended to create a custom classifier and then create a new reference table. For example, if you want to add address data for your country that is currently not supported in Protegrity Discover, then you can create a custom classifier of type *Dictionary*, and then create a new reference table, instead of appending the data to the existing *Address* table.
Name	If you are creating a new table, then type the name of the table.

Field	Description
	If you are appending the reference data to an existing table, then choose a table that you have previously created or choose one of the following default tables: • Phone Area Code • Street • Address • Country • Name For more information about the countries and languages supported out-of-the-box for the default reference tables, refer to the classifier description for the Reference data method in the table *Protegrity Discover - Classifier methods and types*.
File	Click **Choose** to browse for a text or CSV file from your local machine. **Important:** Column names are mandatory. **Note:** If you want to upload reference data that contains address information, such as city or street names, and you want to associate this information with a particular country, then you must also add a *Country* column with the ISO Alpha 2 or ISO Alpha 3 code of the corresponding country.
Separator	Choose one of the following separators to separate the individual data items: • Comma • Semi-colon • Tab • Other - Specify a custom data separator in the **Character** field.
Character	Displays the character of the data separator selected in the **Separator** field. If you have chosen **Other** in the **Separator** field, then you must type the character or characters that you want to use as data separators.

4. Click **Upload**.
   The reference table is uploaded to Protegrity Discover and is displayed in the **Sneak Peek** table.
   After you have uploaded the reference table, you can check whether a particular value is included in the reference table.
5. Select the table pertaining to the data you want to check, from the **Table Name** field.
6. In the **Field Name** field, choose a column name from the reference table.
   Protegrity Discover tries to search the test value from this chosen column.
7. In the **Matching Type** field, choose one of the following values:

Table 6-11: Matching Type Fields

Field	Description
Match	Perform a search for the given keywords within the specified column of the reference table. If you have provided a single keyword, then Protegrity Discover checks whether any column entry exactly matches this keyword. For example, if you specify a keyword, and the column entry contains the same keyword with a leading or trailing space, then Protegrity Discover does not identify the keyword. If you have provided multiple keywords, then Protegrity Discover performs the following tasks:

Field	Description
	• Checks whether the individual keywords exactly match any entry in the specified column • Checks whether the multiple keyword phrase exactly matches any entry in the specified column **Note:** The search operation is not case-sensitive.
Match Phrase	Perform a search for the given phrase within the specified column of the reference table. In this scenario, Protegrity Discover does not perform a one-to-one match between the text specified in the **Test Value** field and the column entries. Instead, it returns a positive match even if the value specified in the **Test Value** field is part of a larger phrase in the column entries. **Note:** The search operation is not case-sensitive.
Exact Match	Perform a case-sensitive search for the given keyword or phrase within the specified column of the reference table. In this scenario, Protegrity Discover performs a one-to-one match of the text specified in the **Test Value** field with the column entries. For example, if you specify a phrase, and the column entry contains the same phrase with a leading or trailing space, then Protegrity Discover does not identify the phrase.

Example of Matching Type Fields

Consider that you have uploaded the following reference table to Protegrity Discover on the **Reference** tab:

Table 6-12: Uploaded Reference Table

Country Code	Place Name	Postal Code
US	Twentynine Palms	92278
US	Palms	48465
US	Isle of Palms	29451
US	Thousand Palms	92276
US	Twentynine Palms	92277

The following section describes the test results if you choose the *Place Name* column in the **Field Name** field and then choose one of the following options in the **Matching Type** field:

- *Match*
  - *Sample 1* - Type `Palms` in the **Test Value** field, and click **Test**.

    The test result shows 1 hit. In this case, the text `Palms` exactly matches the value in the second row of the **Place Name** column.

  - *Sample 2* - Type `Twentynine Palms` in the **Test Value** field, and click **Test**.

    The test result shows 3 hits.

    In this case, Protegrity Discover first checks whether the individual keywords `Palms` and `Twentynine` match exactly with any column entry. In this case, the text `Palms` exactly matches with the value in the second row of the **Place Name** column.

Protegrity Discover then checks whether the complete phrase ***Twentynine Palms*** exactly matches with any column entry. In this case, the phrase ***Twentynine Palms*** exactly matches with the values in the first and fifth row of the **Place Name** column.

Therefore, Protegrity Discover displays 3 hits in the **Test Results** area.

- *Match Phrase*
  - *Sample 1* - Type ***Palms*** in the **Test Value** field, and click **Test**.

    The test result shows 5 hits. Protegrity Discover checks whether the text ***Palms*** is part of any row in the **Place Name** column. In this case, the text ***Palms*** is a part of all the five rows of the **Place Name** column.

  - *Sample 2* - Type ***Twentynine Palms*** in the **Test Value** field, and click **Test**.

    The test result shows 2 hits. Protegrity Discover checks whether the complete phrase ***Twentynine Palms*** is part of any row in the **Place Name** column. In this case, the complete phrase is part of the first and fifth row of the **Place Name** column.

- *Exact Match*
  - *Sample 1* - Type ***Palms*** in the **Test Value** field, and click **Test**.

    The test result shows 1 hit. Protegrity Discover performs a case sensitive search on the **Place Name** column for the text ***Palms***, and checks whether the text exactly matches any column entry. In this case, the text ***Palms*** exactly matches the value in the second row of the **Place Name** column.

  - *Sample 2* - Type ***Twentynine Palms*** in the **Test Value** field, and click **Test**.

    The test result shows 2 hits. Protegrity Discover performs a case sensitive search on the **Place Name** column for the complete phrase ***Twentynine Palms***, and checks whether the phrase exactly matches any column entry. In this case, the complete phrase ***Twentynine Palms*** exactly matches the first and fifth row of the **Place Name** column.

  - *Sample 3* - Type ***Twentynine palms*** in the **Test Value** field, and click **Test**.

    The test result shows 0 hits. Protegrity Discover performs a case sensitive search on the **Place Name** column for the complete phrase ***Twentynine palms***, and checks whether the phrase exactly matches any column entry. In this case, the complete phrase ***Twentynine palms*** does not match any row of the **Place Name** column.

8. In the **Test Value** field, type the test data that you want to search within the reference table.
9. Click **Test**.

   The **Test Results** table displays the results of the test. It lists the number of occurrences of the particular test value, and categorizes the data based on the country name.
10. You can also filter the reference data in the **Sneak Peak** table and the test results in the **Test Results** table using the **Search** field that is present above each table.

6.4.2.1.7 Updating the Spacy Patterns Tab

This section describes how to specify spaCy patterns to identify data using Natural Language Processing (NLP).

➤ To update the **Spacy Patterns** tab:

1. On the **Classifiers** screen, click **Spacy Patterns**.
   The following **Spacy Patterns** tab appears.

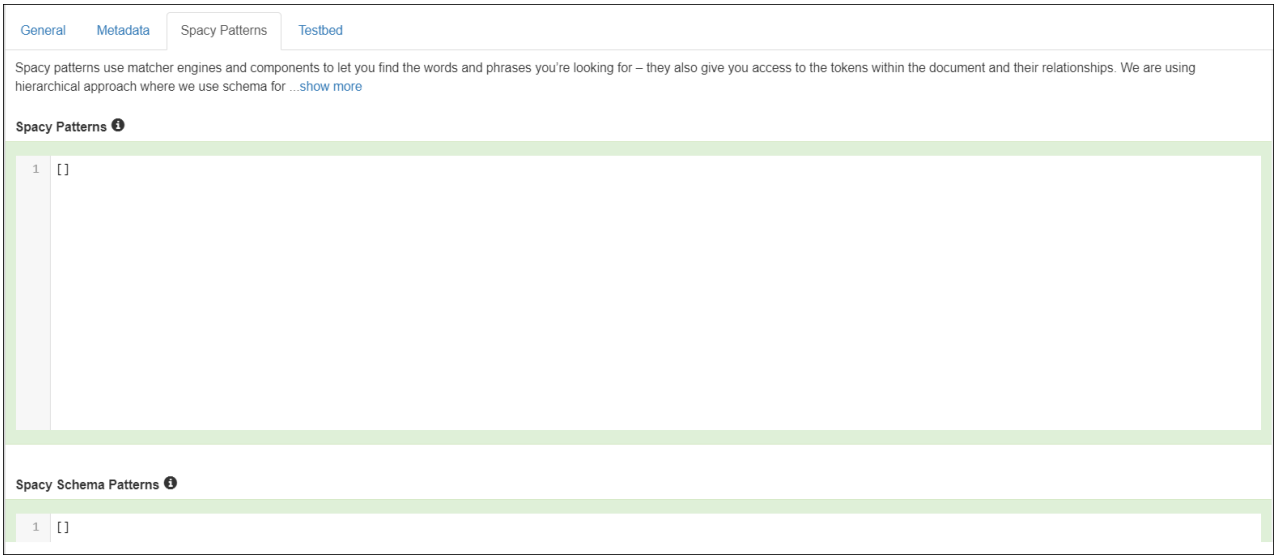


Figure 6-26: Spacy Patterns Tab

2. Enter the qualifying pattern in the **Spacy Schema Patterns** area.

   The schema patterns are used to identify the metadata that is adjacent to the data.

   You must specify the spaCy schema patterns as a list of lists, as shown in the following code snippet. The outer square brackets are mandatory.

```
[
 [
  {
   "Attribute 1": "Value"
  }
 ],
 [
  {
   "Attribute 2": "Value"
  }
 ]
]
```

   For example, the following snippet shows the spaCy schema pattern for the default NLP Phone Number classifier:

```
[
 [
  {
   "LOWER": "phone"
  }
 ],
 [
  {
   "LOWER": "number"
  }
 ]
]
```

   Using this spaCy schema pattern, the classifier checks whether the words *phone* and *number*, irrespective of the case, are included in the name of the file that is being scanned for sensitive data. If either of the words are found, then Protegrity Discover boosts the confidence score of the data.

   For detailed information about the attributes used in the spaCy schema patterns, refer to the *spaCy* website.

3. Enter the spaCy pattern in the **Spacy Patterns** area to find the exact text.

You must also specify the spaCy patterns as a list of lists, similar to the spaCy schema patterns in *step 2*.

For example, the following snippet shows the spaCy pattern for the default NLP Phone Number classifier:

```
[
 [
  {
   "ORTH": "(",
   "OP": "+"
  },
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": ")",
   "OP": "+"
  },
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": "-",
   "OP": "?"
  },
  {
   "SHAPE": "dddd"
  }
 ],
 [
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": ".",
   "OP": "+"
  },
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": ".",
   "OP": "+"
  },
  {
   "SHAPE": "dddd"
  }
 ],
 [
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": "-",
   "OP": "+"
  },
  {
   "SHAPE": "ddd"
  },
  {
   "ORTH": "-",
   "OP": "+"
  },
  {
   "SHAPE": "dddd"
  }
 ],
 [
  {
   "SHAPE": "ddd"
  },
  {
```

```
    "SHAPE": "ddd"
   },
   {
    "SHAPE": "dddd"
   }
  ],
  [
   {
    "IS_DIGIT": true,
    "LENGTH": 10
   }
  ],
  [
   {
    "IS_DIGIT": true,
    "LENGTH": 7
   }
  ]
 ]
```

For detailed information about the attributes used in the spaCy patterns, refer to the *spaCy* website.

You can modify the *nlp_max_length* parameter, which is found in the *classifiers.json* file, to configure the size of the input data that can be processed by the spaCy module at one time.

You can update the *tokenizer_config* parameter, which is found in the *classifiers.json* file, to specify the delimiters that can be used to split the input data into individual tokens that can be processed by the spaCy module.

The *tokenizer_config* parameter includes the following properties, which are regular expressions, that can be used to specify the delimiters for splitting the input data into individual tokens.

Property	Description	Default Values
prefix	Identify the prefix delimiter for splitting the input data.	["[0-9]+", "-", "\\+", "/", ",", "\\."]
infix	Identify the delimiter for splitting the input data internally. **Important:** Do not add a period to the *infix* property. If you add a period, then you will be unable to identify an email address using the spaCy module.	["-", "\\+", "/", ","]
suffix	Identify the suffix delimiter for splitting the input data.	["[0-9]+", "-", "\\+", "/", ",", "\\."]

For example, if the input data is a phone number **`+49-332-1234567`**, then the spaCy module uses the default values of the *prefix*, *infix*, and *suffix* properties to split the phone number into the following individual tokens:

```
["+", "49", "-", "332", "-", "1234567"
```

You can choose to add custom delimiters to the *tokenizer_config* parameter if the spaCy module is unable to split the input data correctly using the default delimiters.

Important: By default, the state of the NLP object is stored in the `/opt/temp/nlp.pkl` file. This file is used to cache the NLP configuration.

If you modify the *tokenizer_config* parameter, then you must delete the *nlp.pkl* file. When the Protegrity Discover services are restarted, a new *nlp.pkl* file is automatically created that contains the updated NLP configuration.

> If you do not delete the *nlp.pkl* file, then the spaCy module uses the old NLP configuration even if you have modified the *tokenizer_config* parameter.

For more information about modifying the *classifiers.json* file, refer to the section *To manage system files*.

6.4.2.1.8 Testing the Classifier

This section describes how to test the classifier configuration using test data.

To update the **Testbed** tab:

1. On the **Classifiers** screen, click **Testbed**.
   The following **Testbed** tab appears.

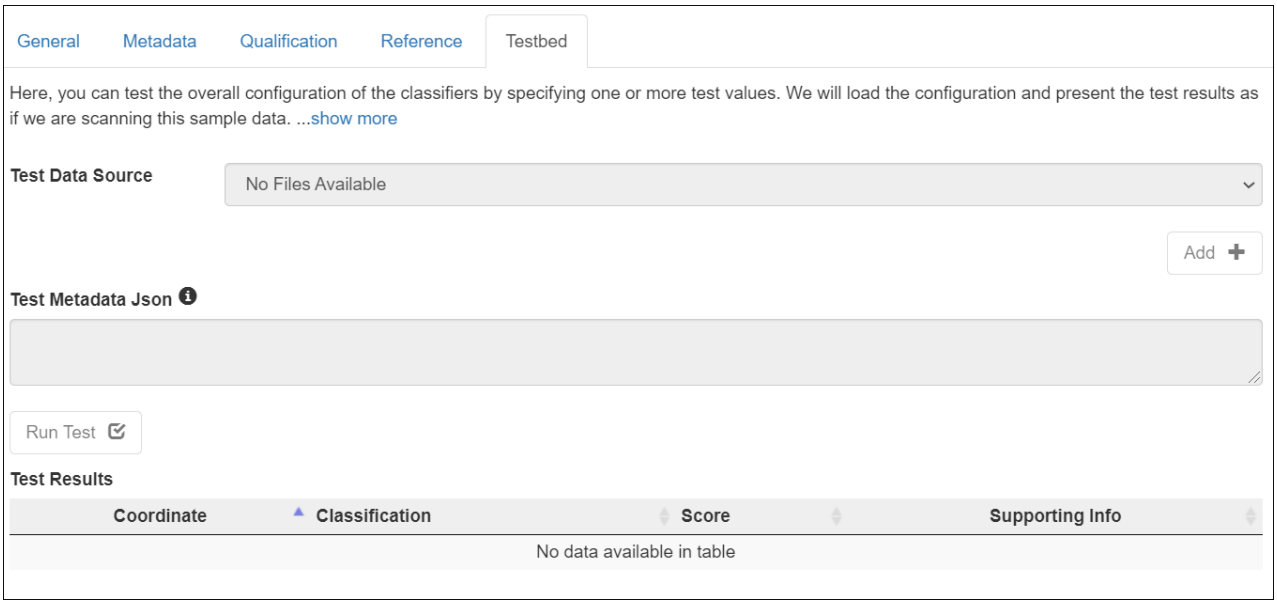


Figure 6-27: Testbed Tab

2. Click **Add** to add the test data.
   The following **Add Test Data** dialog box appears.

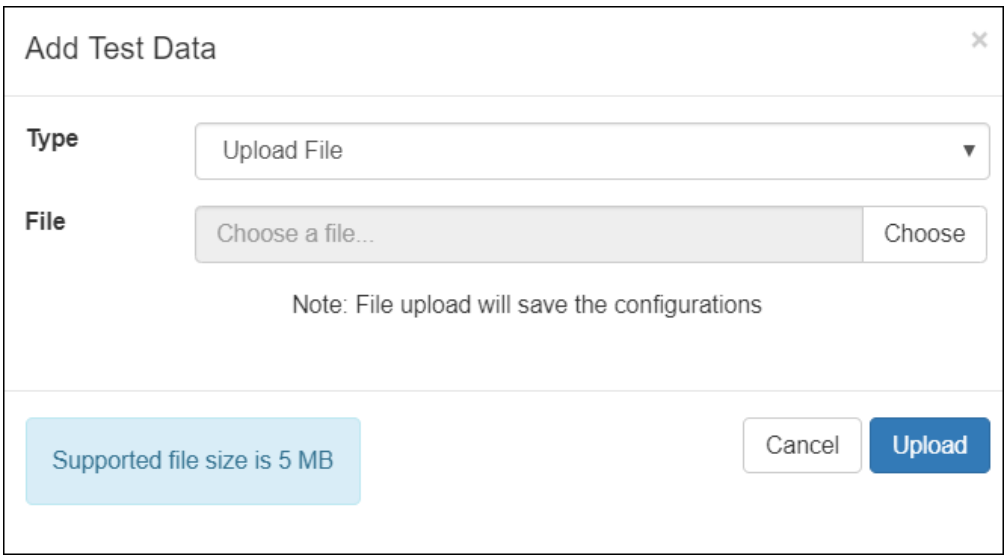


Figure 6-28: Add Test Data Dialog Box

3. Select one of the following values from the **Type** list:
   - **Upload File** - Upload a file containing test data. Click **Choose** to browse for and select the file, and then click **Upload**. If you try to upload a file whose name matches with a file that has already been uploaded for testing, then you are prompted to override the existing file by clicking **Confirm**.

     The supported data types are listed in the *Supported File Formats* appendix. After the file is uploaded, a sample of the uploaded file appears in the **Test Data File Text** area.

     However, only *CSV* and *txt* data are displayed on-screen. To view all other data types, you need to download the file by clicking **Download**.

     The maximum file size that you can upload is 5 MB.

     > **Note:** If you have created a new classifier that you have not yet saved, then it will automatically get saved after you upload the test file.

   - **Free Text** - Click **Add** to test the classifier using free text. Type text data in the **Free Text** area that appears after you select **Free Text**. You can enter a maximum of 500 characters.

   > **Important:** Do not use production data as test data.

   > **Note:** It is recommended to use a comma or a new line for separating individual test values.

4. If you have entered free text data in the **Free Text** area, and you want to save this data, then click **Save**.
5. If you want to test the metadata expressions listed in the **Metadata** tab, then specify the required key-value pairs in the **Test Metadata JSON** area.

   > **Important:** You need to specify the file metadata type in *lowercase* as the key in the key-value pair.
   > For more information about the supported key values, refer to the list of *File Metadata*.

   For example, if you have created a file metadata entry named *owner* in the **Metadata** tab, then you can specify the following text in the **Test Metadata JSON** area:

   ```
   {"owner": "<owner_name>"}
   ```

You can also specify multiple key-value pairs.

For example, if you have created two file metadata entries named *file_size* and *modified_time* in the **Metadata** tab, then you can specify the following text in the **Test Metadata JSON** area:

```
{"file_size": "<file_size_in_bytes>", "modified_time": "<Timestamp_in_Epoch_time_format>"}
```

Note: In case of *modified_time* and *created_time* metadata, you need to specify the value of the timestamp as Epoch time, which specifies the number of seconds that have elapsed after the Unix Epoch, which is 00:00:00 UTC on 1 January 1970.

To convert a timestamp value to an Epoch time, refer to an online Epoch time converter or use the *datetime.timestamp()* in the Python *datetime* module.

For more information about the *datetime.timestamp()* method, refer to the *Python documentation*.

In addition to file metadata, you can also test the coordinate metadata by specifying the following text in the **Test Metadata JSON** area:

```
{"coordinate_path": "<Value>"}
```

Important: The coordinate_path parameter is applicable only if you are testing the data specified in the **Free Text** area or the **Test Metadata JSON** area.

The classifier evaluates the Python boolean expression specified in the **Metadata** tab using the values specified in the **Test Metadata JSON** area. If the evaluation is successful, then the results appear in the **Test Results** area.

For more information about the **Metadata** tab, refer to the section *Updating the Metadata Tab*.

6. Click **Run Test** to check whether the classifier can classify using metadata or data.

   The results are displayed in the **Test Results** table. The test results depend on whether you are testing for data or metadata.

   If you have specified both the data and metadata, then the classifier first tests the metadata specified in the **Test Metadata JSON**.

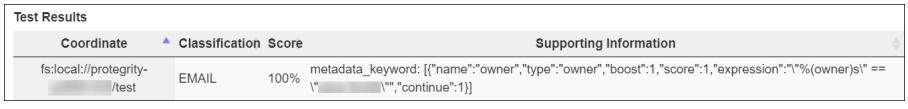
   If the confidence score of the metadata test results is 100%, then the classifier does not test the data and displays the results based on the metadata test. If the confidence score of the metadata test results is less than 100%, then the classifier tests the data and displays the results based on whether the classifier can identify the data.

   The test results displays the following information:

   - Coordinate
   - Classification
   - Score
   - Supporting information

   The supporting information differs depending on whether the classifier tests the data or the metadata.

   - Results of testing the metadata: The metadata test results only appear if the confidence score is 100%.

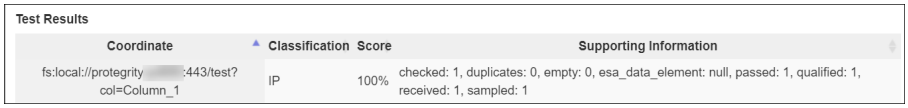


   The following supporting information appears for the metadata.

Field	Description
name	Name of the file metadata
type	Type of the file metadata
Boost	Value to boost the confidence score. If the Regex pattern matches with any metadata, then the boost value is applied to the confidence score.
Score	Confidence score value
expression	The regex pattern or conditional expression for identifying the metadata keyword.
Continue	Indicates that the Continue check box has been selected for the metadata in the **Metadata** tab.

For more information about the supporting information, refer to the *step 3* in the section *Updating the Metadata Tab*.

- Results of testing the data:



The following supporting information appears.

Key	Description
checked	Total number of data entries that have been tested.
duplicates	Total number of data entries that have been identified as *duplicated*.
empty	Total number of entries that do not have any data or are blank.
passed	Total number of data entries that have been successfully identified by the classifier.
qualified	Total number of data entries that have passed the qualification regex check.
received	Total number of data values received from the given set of sampled values.
sampled	Total number of data values (records) configured for the scan from the overall data set.

Important: To activate the classifier, at least one data or metadata value must be passed in the test.

7. If you want to remove the test files that you have uploaded or the free text that you have entered, then perform the following steps:
   a. From the **Test Data Source** list, choose either **Free Text** or the test file that you have uploaded.
   b. Click **Remove**.
      The **Remove** dialog box appears, prompting you to confirm whether you want to delete the selected free text data or uploaded test file.
   c. Click **Remove**.

6.4.2.2 Modifying Existing Classifiers

This section describes how you can modify the existing classifiers or the custom classifiers that you have created or both.

To modify existing classifiers:

1. On the Protegrity Discover Web UI, navigate to **Discover Rules** > **Classifiers**.

The following **Classifiers** screen displays the list of existing classifiers.

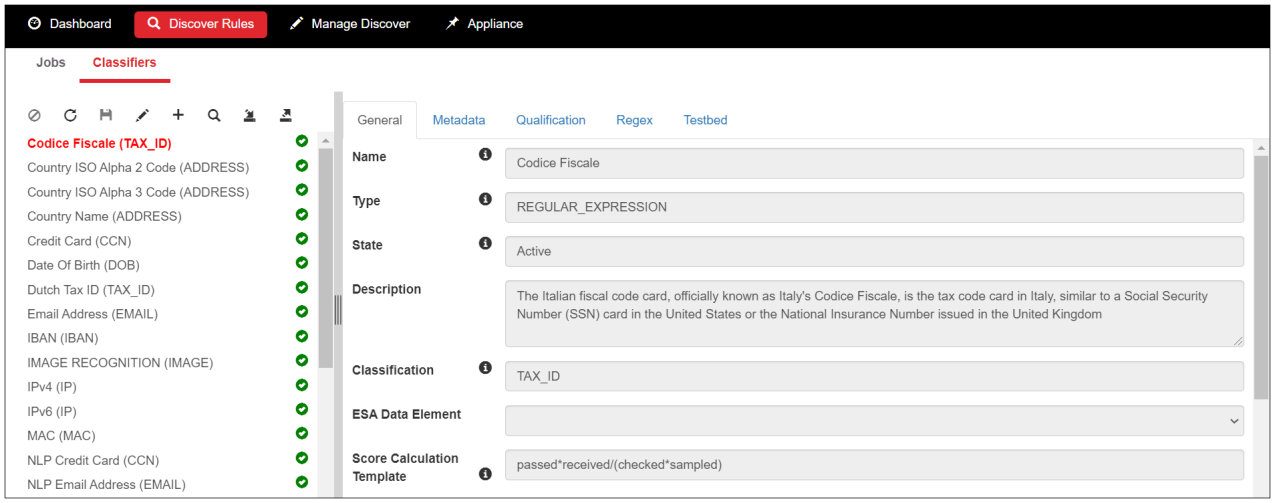


Figure 6-29: Classifiers Screen

2. If you want to search for a classifier, then perform the following steps.
   a. Click 🔍.
      The following **Search** text box appears.

      Figure 6-30: Search Text Box
   b. Type the first few letters of the classifier that you want to search.
      The classifiers list is filtered to display only those classifiers that match your search criteria.
3. Select the classifier.
4. Click ✎.
   The classifier opens in edit mode.
5. Modify the details for the specific classifier.
   The tabs that are displayed depend on the classifier type.

To modify the details in the...	Refer to the following section...
General tab	*Updating the General Tab*
Metadata tab	*Updating the Metadata Tab*
Qualification tab	*Updating the Qualification Tab*
Source Code tab	*Updating the Source Code tab*
Regex tab	*Updating the Regex Tab*
Reference tab	*Updating the Reference Tab*
Spacy Patterns tab	*Updating the Spacy Patterns Tab*
Settings tab	*Updating the Settings Tab*

Important: If you modify any field other than the **Name**, **Description**, or the **Notes** field in the **General** tab, then the classifier is deactivated. If you want to re-activate the classifier, you must test the classifier and ensure that at least one data value passes the test.

Important: If you modify a classifier and save it while a scan process is running, then Protegrity Discover reloads the ongoing scan process with the modified configuration. For example, if you deactivate a classifier and save it while a scan process is running, then Protegrity Discover excludes the classifier from the remainder of the scan process.

6. If you want to deactivate the classifier, then click .
7. Test the classifier, if required.

   At least one data value must pass the test to activate the classifier.

   For more information on testing the classifier, refer to the section *Testbed*.

   Important: If you activate a classifier and save it during a scanning process, then Protegrity Discover includes the classifier in the remainder of the scanning process.

8. Click to save the classifier.

   If you do not want to save the changes made to the classifier, then click . The **Save Changes** dialog box appears that prompts you to either save the changes, discard the changes, or close the dialog box. Click **Discard** to discard the changes.

   If you want to delete the classifier, then click . The classifier is deleted and removed from the **Classifiers** list.

9. If you want to reset the settings of a default classifier, then select the classifier and click .

   The out-of-the-box settings of the default classifier are restored. Click to save the changes.

   For a list of the default classifiers, refer to the *Default Classifiers* appendix.

   Note: The reset functionality is not applicable to the custom classifiers that you have created.

6.4.2.2.1 Updating the Settings Tab

This section describes how you can modify the **Settings** tab, which is only applicable for the Date of Birth, Email Address, NLP Email Address, and Phone Number classifiers.

To update the **Settings** tab:

1. On the Protegrity Discover Web UI, navigate to **Discover Rules** > **Classifiers**.
   The **Classifiers** screen displays the list of existing classifiers.
2. Select one of the following classifiers:
   - Date Of Birth
   - Email Address
   - NLP Email Address
   - Phone Number
3. Click .
   The classifier opens in edit mode.
4. Click **Settings**.
   The **Settings** tab appears.
5. Enter the following details based on the classifier you have selected.

Classifier	Field	Description
Date Of Birth	Range Minimum	Specify the minimum value of the date range that is used to qualify a date as a valid date of birth. This value is expressed as the number of years in the past from the current date. Any date that falls within

Classifier	Field	Description
		the lower and upper limits specified by this date qualification range is considered a valid date of birth. For example, if today's date is July 29, 2019, and you set the value of the **Range Minimum** field to *10*, then Protegrity Discover will not identify an inputted date as a valid date of birth if it falls between July 29, 2019 and July 29, 2009. For example, Protegrity Discover will not identify September 15, 2014 as a valid date of birth. By default, the value of this field is set to *10*. If you do not specify any value in this field, then a default value of *0* is used.
	Range Maximum	Specify the maximum value of the date range that is used to qualify a date as a valid date of birth. This value is expressed as the number of years in the past from the current date. Any date that falls within the lower and upper limits specified by this date qualification range is considered a valid date of birth. For example, if today's date is July 29, 2019, and you set the value of the **Range Maximum** field to 92, then Protegrity Discover will not identify a date as a valid date of birth if it falls beyond July 29, 1927. For example, Protegrity Discover will not identify September 15, 1921 as a valid date of birth. By default, the value of this field is set to *92*. If you do not specify any value in this field, then a default value of *100* is used. **Note:** Protegrity Discover does not identify a date as a valid date of birth if it falls beyond January 1, 1900.
• Email Address • NLP Email Address	Verify MX record	Validate the mail exchanger (MX) record of the hostname for an email address. The MX record identifies the mail server that is responsible for receiving the email messages. In case of the Email Address classifier, if this field is selected, then an email is not considered valid if a DNS query determines that the hostname has no MX record. **Important:** In case of the NLP Email Address classifier, if this field is selected and a DNS query determines that the hostname has no MX record, then the email address is still considered as valid. However, in this case, the confidence score is decreased. Conversely, if this field is selected and a DNS query determines that the hostname has a MX record, then the confidence score is increased.
	DNS Servers	Specify a comma separated list of DNS server hostnames or IP addresses. If this field is empty, then Protegrity Discover uses a system configured DNS server.
Phone Number	Default Country Code	Specify the default country code. By default, this value is set to *1* for United States.

6.4.2.3 Exporting Classifiers

This section describes how you can export the existing classifiers or custom classifiers. This allows you to create a copy of a classifier that can then be imported to another instance of Protegrity Discover.

To export existing classifiers:

1. On the Protegrity Discover Web UI, navigate to **Discover Rules** > **Classifiers**.
   The following **Classifiers** screen displays the list of existing classifiers.

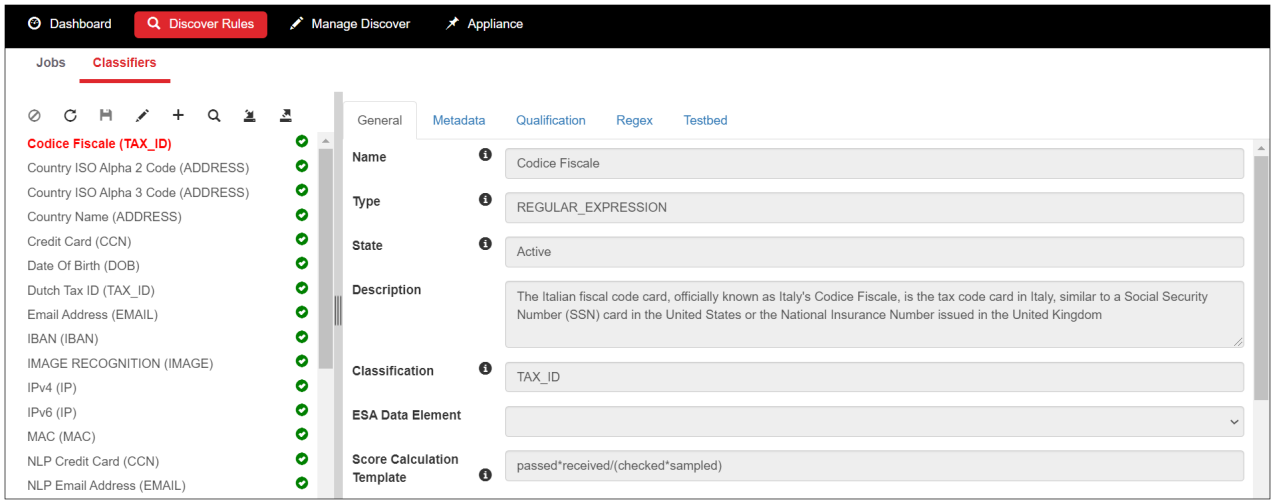


Figure 6-31: Classifiers Screen

2. Select the classifier.
3. Click .
   The following **Export Classifier** dialog box appears.

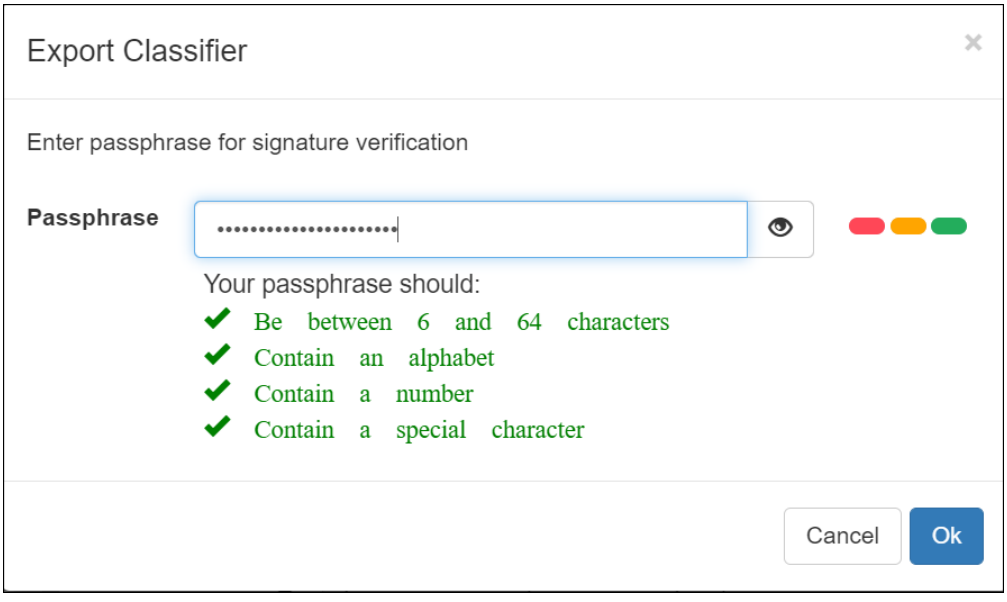


Figure 6-32: Export Classifier Dialog Box

4. In the **Passphrase** text box, type a passphrase that will be used for verifying the signature when you import the classifier.

Important:

- Ensure that the length of the passphrase is between 6 and 64 characters.
- It is recommended to include at least one alphabet, one digit, and one special character to improve the passphrase strength.

Click the icon to show the passphrase. Click the to hide the passphrase.

5. Click **OK**.

   A *tar.gz* file is created on your local machine. For example, if you have exported the Codice Fiscale classifer, then the *codice_fiscale.tar.gz* file is created.

   The *tar.gz* file contains the following files:

   - Checksum information
   - Version information about Protegrity Discover
   - A *.pdc* file that contains the configuration for the exported classifier
   - Reference table, if you have a uploaded a reference table to a classifier of type Dictionary

     For more information about uploading a reference table, refer to the section *Updating the Reference Tab*.

   - Test data file, if you have added a test data file to the testbed of the classifier

     For more information about adding a test data file, refer to the section *Testing the Classifier*.

   You can then upload the *tar.gz* file to another instance of Protegrity Discover.

   In addition, if you have successfully exported the classifier, then a Warning log is generated. You can access this log from the **Scanner Log** screen.

   For more information about viewing scanner logs, refer to the section *Viewing Scanner Logs*.

6.4.2.4 Importing Classifiers

This section describes how you can import the classifiers that have been exported.

To import classifiers:

1. On the Protegrity Discover Web UI, navigate to **Discover Rules** > **Classifiers**.
   The following **Classifiers** screen displays the list of existing classifiers.

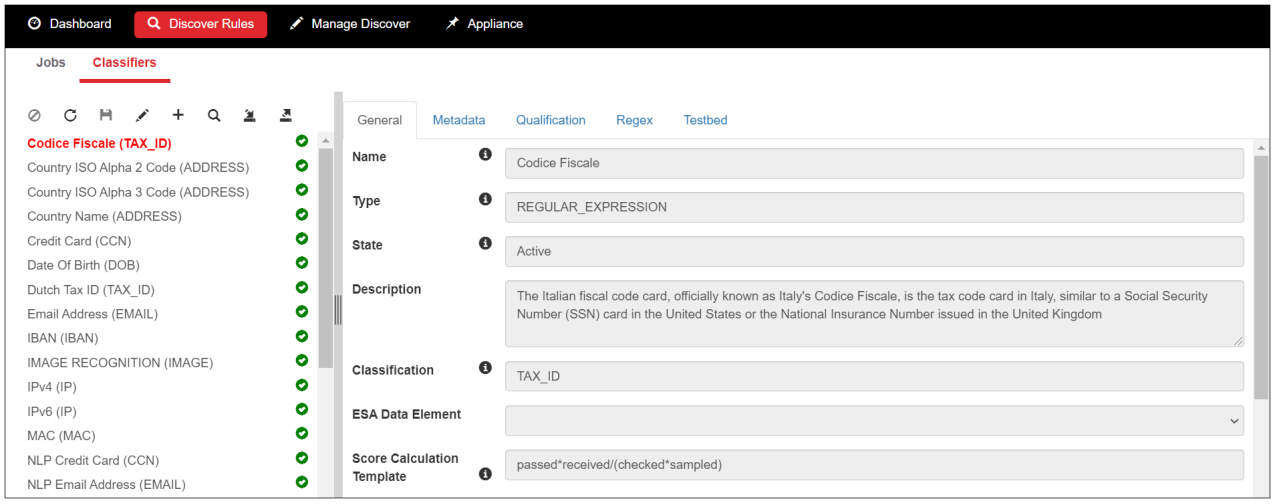


Figure 6-33: Classifiers Screen

2. Click .

   The **Upload Classifier** dialog box appears.

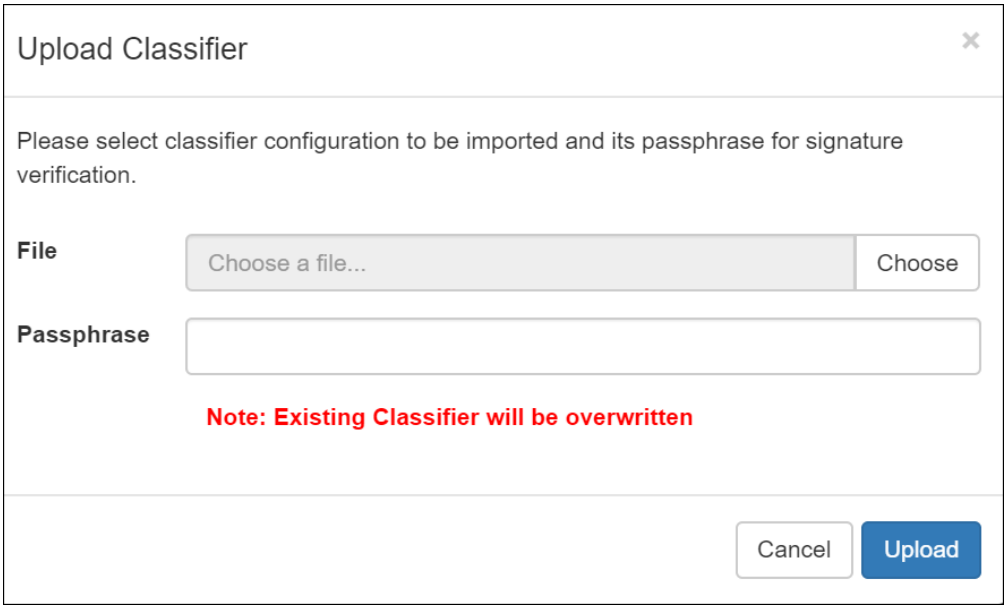


Figure 6-34: Upload Classifier

3. In the **File** text box, click **Choose** to browse for the *tar.gz* classifier file that you want to import.
4. In the **Passphrase** text box, type the passphrase that you had previously specified when exporting the classifier.
5. Click **Upload**.

   The classifier is imported into Protegrity Discover, and the following message appears.

   Classifier imported successfully

   In addition, if you have successfully imported the classifier, then a Warning log is generated. You can access this log from the **Scanner Log** screen.

   For more information about viewing scanner logs, refer to the section *Viewing Scanner Logs*.

By default, if you import a classifier, then it is set to an *Inactive* state. This ensures that if you import a classifier to a Protegrity Discover machine where an existing scan is running, then the imported classifier is not included automatically in the running scan. As a result, you need to manually activate the classifier.

If you want to activate an inactive classifier, then you must ensure that the classifier passes the testbed test.

For more information about testing the classifier, refer to the section *Testing the Classifier*.

> **Important:** If you import an existing classifier, then it will be overwritten.

6.5 Viewing Logs and Statistics

This section describes how you can view the scanner logs, REST API logs, and the REST API request statistics.

It includes the following sub-sections:

- *Scanner Logs* - Displays the logs related to the scanning jobs performed from the Protegrity Discover Web UI.
- *REST API Logs* - Displays the logs related to the scanning jobs performed from the Protegrity Discover REST API.
- *REST API Analytics* - Displays the statistics related to the REST API requests sent to the Protegrity Discover REST API.

6.5.1 Viewing Scanner Logs

Protegrity Discover enables you to refer to the log messages for success or failure events. It also reports information or warning messages that might require attention.

The Protegrity Discover log messages are divided into three levels, which are *warning*, *information*, and *debug* levels. The following figure illustrates the **Manage Discover** > **Scanner Log** screen from the Protegrity Discover Web UI.

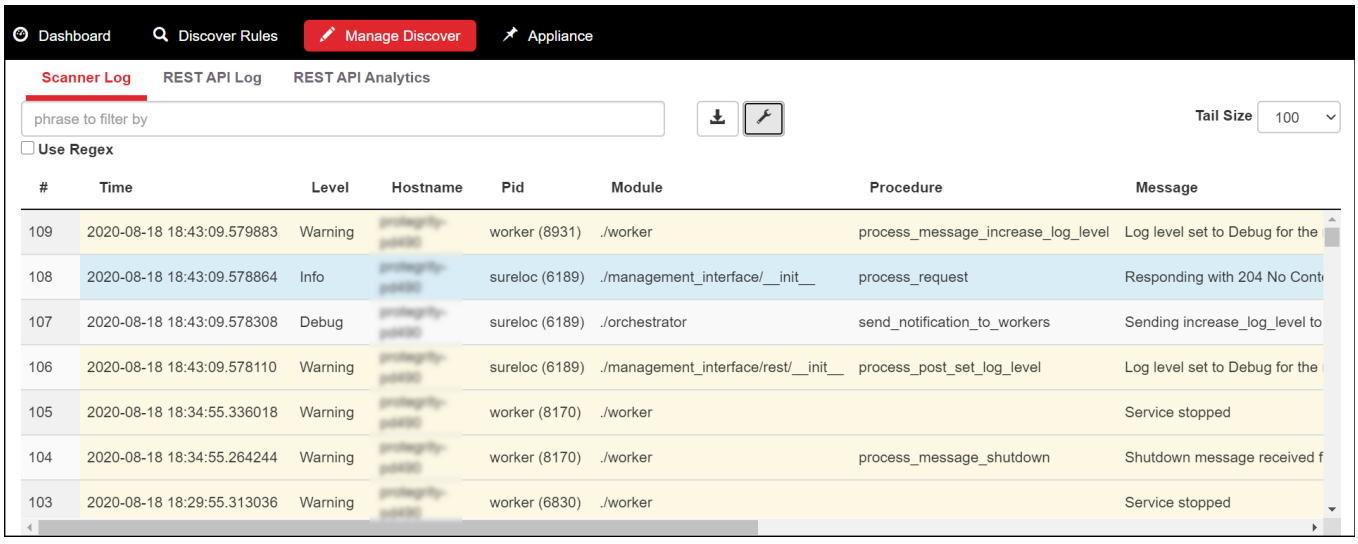


Figure 6-35: Protegrity Discover Scanner Logs

The following table describes the fields that are listed with Protegrity Discover scanner logs.

Table 6-13: Protegrity Discover Scanner Logs

Field name	Description
Time	Timestamp details for the log entry **Note:** The log timestamp follows the *Coordinated Universal Time (UTC)* time zone, which is a standard by which the world regulates time.
Level	Type of log message: warning, information, or debug
Hostname	Host name for the Protegrity Discover system
Pid	Process id for the process that is reporting the log message
Module	File name or script from which a function is generating the log message
Procedure	Function that is generating the log message
Message	Actual log message for the log entry

The logs can be filtered using *string* or *phrase* search. It can also be set at a system log level of warning, info, or debug level. You can also download the log results for further review, as mentioned in the table *Protegrity Discover - Manage Logs*.

You can manage the log messages from the Web UI, as mentioned in the following table.

Table 6-14: Protegrity Discover - Manage Logs

Action	Icon/Field	Description
Filter logs by *string* or *phrase* search		Filter the logs using a string or phrase in the search field. It automatically refines the results as per the provided input.
Set system log levels		Modify the system log level to *warning*, *info* or *debug* level. **Note:** The system log level is automatically reset to *warning* level after 60 minutes.
Download log results		Download the Protegrity Discover logs to your machine. The log file is a text file named *<Host_name_of_the_Protegrity Discover_Machine>-Protegrity-Discover.log*. The content of the log file *<Host_name_of_the_Protegrity Discover_Machine>-Protegrity-Discover.log* matches the content of the non-archived log file *sureloc.log* that you can download from the **Support** screen of the Appliance Web UI. For more information about the **Support** screen, refer to the section *Support* in the *Protegrity Appliances Overview Guide 9.2.0.0*. Protegrity Discover checks the logs every hour and archives the log file if it exceeds 10 MB size. The following entry is added to the log

Action	Icon/Field	Description
		file, after the existing log file is archived or rotated: `Application log file rotation completed successfully` The archived log content is saved in a new file *sureloc.log.1*. A new archive file, with a sequentially incremented number, is created every hour, if the log size exceeds 10 MB. For more information about downloading the archived and non-archived log files, refer to the section *To download troubleshooting information*.

6.5.2 Viewing REST API Logs

This section describes how you can refer to the log messages for success or failure events related to the data processed by the Protegrity Discover REST APIs. It also reports information or warning messages that might require attention.

The Protegrity Discover REST API log messages are divided into five levels, which are *critical*, *error*, *warning*, *information*, and *debug* levels. The following figure illustrates the **Manage Discover** > **REST API Log** screen from the Protegrity Discover Web UI.

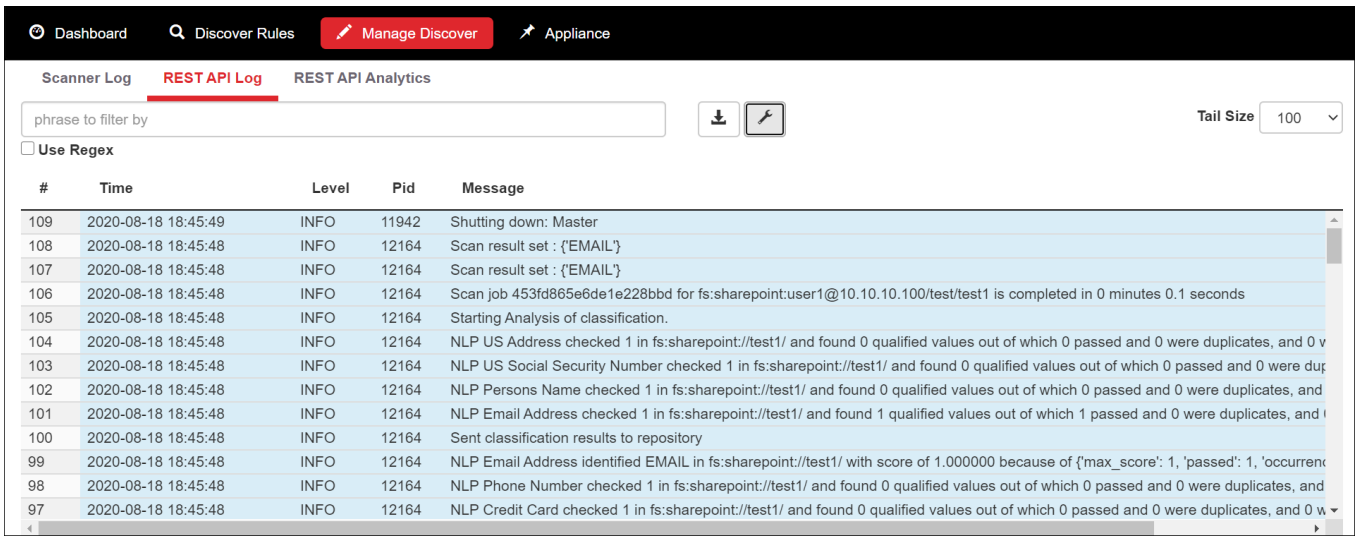


Figure 6-36: Protegrity Discover REST API Logs

The following table describes the fields that are listed in the Protegrity Discover REST API logs.

Table 6-15: Protegrity Discover REST API Logs

Field name	Description
Time	Timestamp details for the log entry **Note:** The log timestamp follows the *Coordinated Universal Time (UTC)* time zone, which is a standard by which the world regulates time.
Level	Type of log message: *critical*, *error*, *warning*, *information*, or *debug*

Field name	Description
Pid	Process id for the process that is reporting the log message
Message	Actual log message for the log entry

The logs can be filtered using *string* or *phrase* search. It can also be set at a system log level of *warning*, *info*, or *debug* level. You can also download the log results for further review, as mentioned in the table *Protegrity Discover - Manage Logs*.

You can manage the log messages from the Protegrity Discover Web UI, as mentioned in the following table.

Table 6-16: Protegrity Discover - Manage Logs

Action	Icon/Field	Description
Filter logs by *string* or *phrase* search		Filter the logs using a string or phrase in the search field. It automatically refines the results as per the provided input.
Set system log levels		Modify the system log level to *critical*, *error*, *warning*, *info* or *debug* level. **Note:** The system log level is automatically reset to *warning* level after 60 minutes. **Caution:** Every time you set the system log level, the Protegrity Discover REST service restarts.
Download log results		Download the Protegrity Discover REST API logs to your machine. The log file is a text file named *<Host_name_of_the_Protegrity Discover_Machine>-Protegrity-Discover-RESTful.log*. The content of the log file *<Host_name_of_the_Protegrity Discover_Machine>-Protegrity-Discover-RESTful.log* matches the content of the non-archived log file *sureloc-restful.log* that you can download from the **Support** screen of the Appliance Web UI. For more information about the **Support** screen, refer to the section *Support* in the *Protegrity Appliances Overview Guide 9.2.0.0*. Protegrity Discover checks the logs every hour and archives the log file if it exceeds 10 MB size. The following entry is added to the log file, after the existing log file is archived or rotated: `Restful log file rotation completed successfully` The archived log content is saved in a new file *sureloc-restful.log.1*. A new archive file, with

Action	Icon/Field	Description
		a sequentially incremented number, is created every hour, if the log size exceeds 10 MB. For more information about downloading the archived and non-archived log files, refer to the section *To download troubleshooting information*.

6.5.3 Viewing REST API Analytics

This section describes how you can view the statistics for the REST API requests.

To view the REST API statistics, from the Protegrity Discover Web UI, navigate to **Manage Discover** > **REST API Log** screen.

> **Important:**
> The REST API requests data is updated every 10 minutes.

The REST API Analytics screen consists of the following graphs:

- **Graph 1**: View number of requests and number of bytes sent every 10 minutes, over a specific time period. It also lists the number of requests that contain sensitive and non-sensitive data.
- **Graph 2**: View the number of possible requests and bytes that have been sent per second every 10 minutes, over a specific time period.

> **Important:** Graph 2 does not show the actual number of requests and bytes sent per second, but derives these values based on the actual number of requests and bytes sent and the actual time taken to send these requests. These values can be used to evaluate the performance of the REST API.
>
> For example, if a user sends *x* number of requests and *y* number of bytes, and the time taken to send these requests is *z* milliseconds, then Protegrity Discover calculates the possible number of requests, that is, `(x/z) * 1000`, and the possible number of bytes, that is, `(y/z) * 1000`, that could have been sent within a second, and displays these calculated values in Graph 2.

The following section visually explains Graph 1.

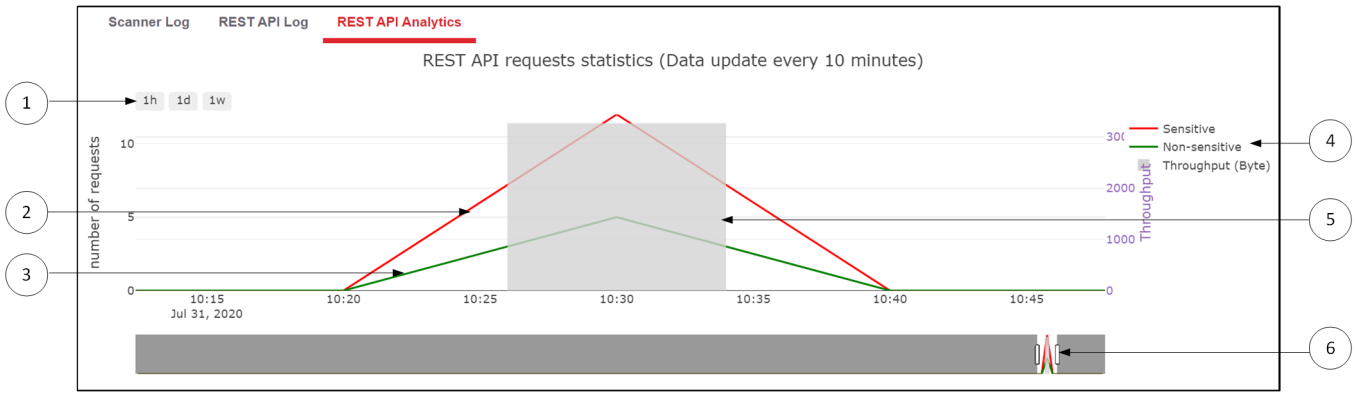


Figure 6-37: Graph 1

Item	Description
1	Specify the time period for which you want to view the REST API request graph. You can select one of the following options:

Item	Description
	• *1h* - Specifies the REST API statistics for the last hour • *1d* - Specifies the REST API statistics for the last 24 hours • *1w* - Specifies the REST API statistics for the last week By default, the graph displays the REST API statistics for the time period spanning from the last week till the current date.
2	Specifies the number of REST API requests that were identified to contain sensitive data
3	Specifies the number of REST API requests that were identified to contain non-sensitive data
4	Specifies the legends for the REST API request graph. Click the legend to hide the corresponding graph. For example, if you click the *Sensitive* legend, then the graph related to the number of requests containing sensitive data disappears. Double-click the legend to focus on the specific graph and hide the other graphs. For example, if you double-click the *Sensitive* legend, then only the graph related to the number of requests containing the sensitive data is visible. Both the graphs containing the throughput data and the number of requests containing the non-sensitive data disappear.
5	Specifies the number of bytes sent in the REST API requests in the specified time. This is known as the *throughput*.
6	Specifies the range slider that enables you to display the data for a specific time period within the graph. For example, you can click and move the right and the left handles of the range slider to zoom in to the data within a specific time period. By default, the range slider displays the data for the last 18 hours. However, if you select one of the pre-defined time periods, such as, *1h*, *1d*, or *1w*, then the time period specified by the range slider changes accordingly.

The following section visually explains Graph 2.

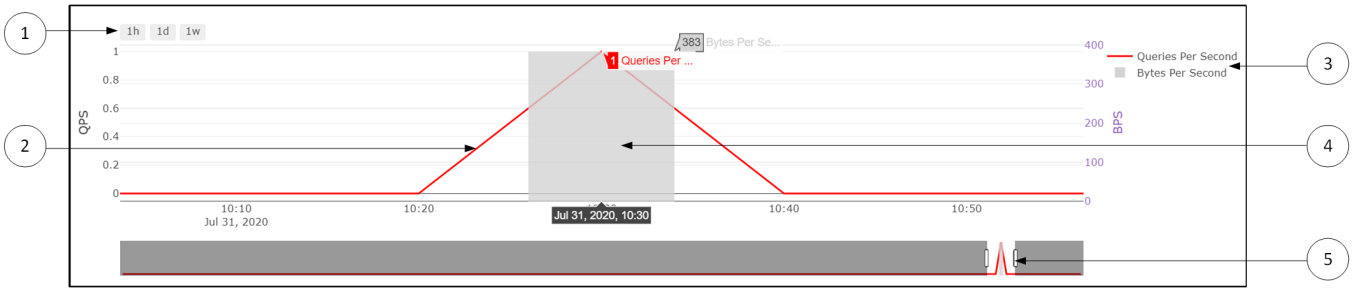


Figure 6-38: Graph 2

Item	Description
1	Specify the time period for which you want to view the REST API request graph. You can select one of the following options: • *1h* - Specifies the REST API statistics for the last hour • *1d* - Specifies the REST API statistics for the last 24 hours • *1w* - Specifies the REST API statistics for the last week By default, the graph displays the REST API statistics for the time period spanning from the last week till the current date.
2	Specifies the number of REST API requests sent per second, within the specified time period **Note:** This value does not represent the actual number of requests sent per second, but is calculated based on the actual number of requests sent and the actual time taken to send these requests.

Item	Description
3	Specifies the legends for the REST API request graph. Click the legend to hide the corresponding graph. For example, if you click the *Queries Per Second* legend, then the graph related to the queries per second data disappears. Double-click the legend to focus on the specific graph and hide the other graphs. For example, if you double-click the *Queries Per Second* legend, then only the graph related to the queries per second is visible. The graph containing the bytes per second data disappears.
4	Specifies the number of bytes sent in the REST API requests per second, within the specified time. **Note:** This value does not represent the actual number of bytes sent per second, but is calculated based on the actual number of bytes sent and the actual time taken to send these bytes.
5	Specifies the range slider that enables you to display the data for a specific time period within the graph. For example, you can click and move the right and the left handles of the range slider to zoom in to the data within a specific time period. By default, the range slider displays the data for the last 18 hours. However, if you select one of the pre-defined time periods, such as, *1h*, *1d*, or *1w*, then the time period specified by the range slider changes accordingly.

You can also download the REST API requests data to a CSV file by clicking the button.

The following figure displays a snapshot of a sample CSV file.

	positive	negative	throughput	milliseconds	qps	bps
2020-07-31T10:00:00.000Z	0	0	0	0	0	0
2020-07-31T10:10:00.000Z	0	0	0	0	0	0
2020-07-31T10:20:00.000Z	0	0	0	0	0	0
2020-07-31T10:30:00.000Z	12	5	3268	8519	1	383
2020-07-31T10:40:00.000Z	0	0	0	0	0	0
2020-07-31T10:50:00.000Z	0	0	0	0	0	0

Figure 6-39: Sample Downloaded CSV File

The following table describes the columns of the CSV file that contain the statistics regarding the REST API requests.

Name	Description
Timestamp	Specifies the timestamp at which the REST API request data has been collected
positive	Specifies the number of REST API requests that were identified to contain sensitive data. For example, the sample CSV file shows that the Protegrity Discover REST API received 12 requests in the 10 minutes between 10:20:00 and 10.30:00 that were identified to have contained sensitive data.
negative	Specifies the number of REST API requests that were identified to contain non-sensitive data. For example, the sample CSV file shows that the Protegrity Discover REST API received 5 requests in the 10 minutes between 10:20:00 and 10.30:00 that were identified to have contained non-sensitive data.

Name	Description
throughput	Specifies the number of bytes sent in the REST API requests in the specified time. For example, the sample CSV file shows that the Protegrity Discover REST API received 3268 bytes of data in the 10 minutes between 10:20:00 and 10.30:00.
milliseconds	Specifies the time in milliseconds during which the REST API requests were received by the Protegrity Discover REST API
qps	Specifies the number of REST API requests sent per second, within the specified time period. The unit of the bps parameter is integer. **Note:** This value does not represent the actual number of requests sent per second, but is calculated based on the actual number of requests sent and the actual time taken to send these requests. `qps = integer value of [ {([positive] + [negative]) / milliseconds} * 1000 ]` For example, the sample CSV file shows that the Protegrity Discover REST API received 17 requests within 8519 milliseconds, in the 10 minutes between 10:20:00 and 10.30:00. In the sample CSV file: `qps = integer value of [ {(12 + 5) / 8519} * 1000 ]` `= integer value of [ {17 / 8519} * 1000 ]` `= integer value of [ 0.00199 * 1000 ]` `= integer value of [ 1.99 ]` `= 1`
bps	Specifies the number of bytes sent in the REST API requests per second, within the specified time. The unit of the bps parameter is integer. **Note:** This value does not represent the actual number of bytes sent per second, but is calculated based on the actual number of bytes sent and the actual time taken to send these bytes. `bps = integer value of [ {(throughput) / milliseconds} * 1000 ]` For example, the sample CSV file shows that the Protegrity Discover REST API received 3268 bytes of data within 8519 milliseconds, in the 10 minutes between 10:20:00 and 10.30:00. In the sample CSV file: `bps = integer value of [ {3268 / 8519} * 1000 ]` `= integer value of [ 0.38361 * 1000 ]` `= integer value of [ 383.61 ]` `= 383`

6.6 License Manager

If your Protegrity Discover license expires, then you must request a new license to continue to use the product. This section explains how you can manage your Protegrity Discover license from the **License Manager** screen.

To request a new license:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **License**.
2. Click **Download License Request**.
3. Send the downloaded license request to Protegrity.
   After you receive the new license from Protegrity, you must upload it to the Protegrity Discover system.
4. On the Protegrity Discover Web UI, navigate to **Appliance** > **License**.
5. Click **Activate License**.
   The following **License Activation** dialog box appears.

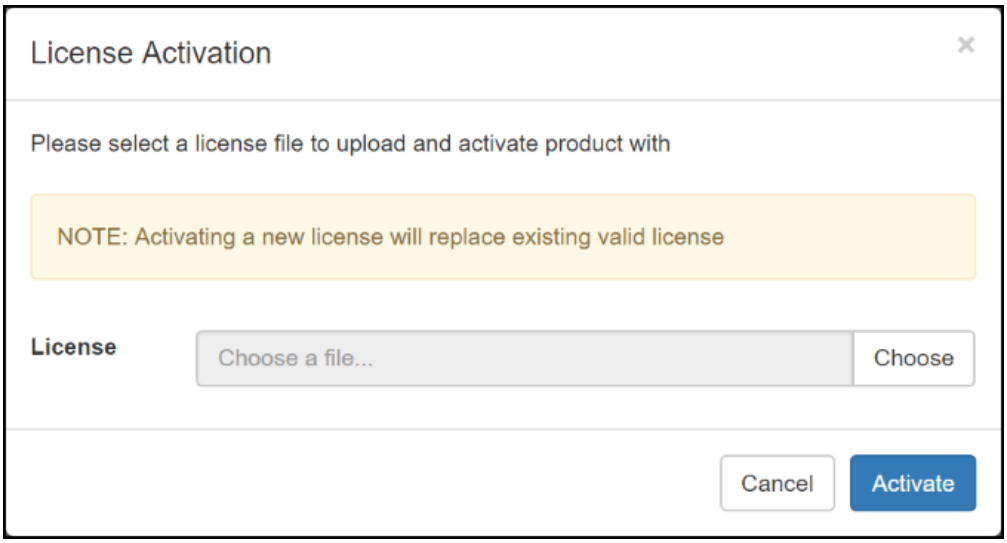


Figure 6-40: Protegrity Discover - License Activation

> **Important:** Successful activation of the Protegrity Discover license will replace the existing one. A new validity period will be set as per the new activated license.

6. In the **License** field, click **Choose** to browse and select the license file from your system.
7. Click **Activate**.
   The **License Manager** screen is updated with the new license details, as shown in the following screenshot.

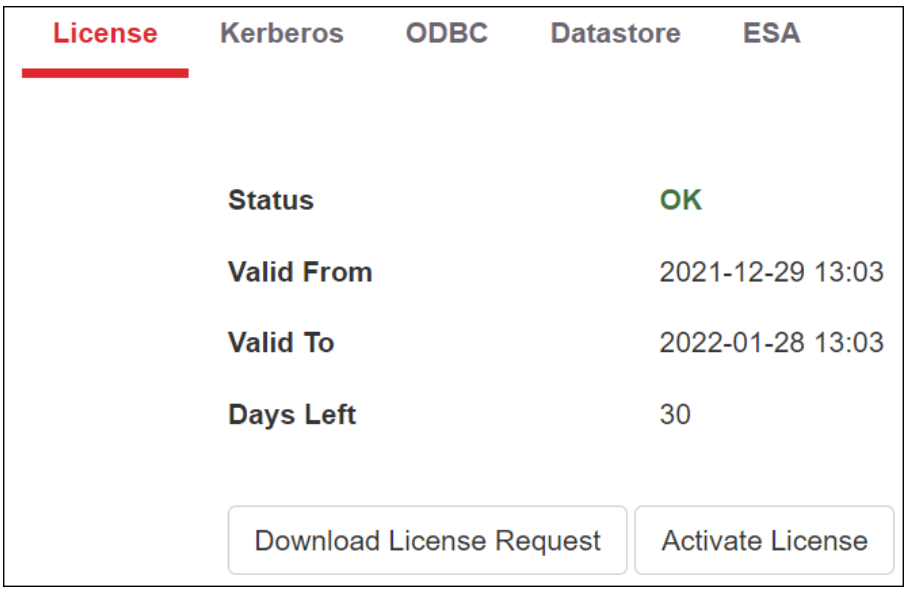


Figure 6-41: Protegrity Discover - License details

> **Note:** Ensure that the license status label **Status** on the **License Manager** screen appears as OK, after you upload the license activation file.

6.7 Kerberos

Protegrity Discover provides you the option to configure Kerberos, a widely adopted network authentication protocol that is widely adopted with most systems today. This section provides an overview about Kerberos and explains how its capabilities are used and configured with Protegrity Discover.

Kerberos © manages service requests through tickets. It enables the authentication of nodes that are trying to communicate with each other a non-reliable network, by helping the nodes identify one another in a secure manner. It was originally developed by Massachusetts Institute of Technology (MIT) and comprises of three main components, which are a client, a server, and a Key Distribution Center (KDC). The systems using Kerberos rely on the KDC, which is a trusted third-party authentication mechanism. In addition to providing the authentication, the KDC is also a ticket granting service that issues tickets to the nodes to identify each other securely. The communication channel between the nodes is encrypted using a shared secret key, disallowing any information packets traveling through the network from being interpreted or modified.

For more information about Kerberos authentication protocol, refer to the *MIT's Kerberos official website*.

The systems using Kerberos as an authentication protocol require other systems communicating with it to pass its authentication rules. Protegrity Discover communicates with its datastores for a data discovery scan. Therefore, if the Kerberos authentication protocol is enabled for a datastore, then you need to configure the Kerberos settings for Protegrity Discover so that it can communicate with the datastore.

The following table lists the Kerberos-specific configuration settings that are required with Protegrity Discover.

Table 6-17: Protegrity Discover - Required Kerberos Configuration Settings

Configuration Setting	Description
Principal	It is a unique identifier for the system user or host to which the Kerberos' KDC can assign tickets for authentication. A principal identifier value contains a number of segments, each of which is separated by a '/' separator. The following sample provides the principal identifier definition: `hive/master.localdomain@EXAMPLE.COM`

Configuration Setting	Description
	The principal identifier comprises of three segments: • The first part of the identifier is the user name, in case of a user. If it is a host, then the first part is the word *host*. • The second part of the identifier is an instance that qualifies the first part and is separated by a '/' separator. The instance is *null* for a user. If it is a host, the second part is the fully qualified host name. • The last part of the identifier is the Kerberos realm in uppercase letters. It is generally the domain name of the Kerberos system. **Important:** The domain name is case sensitive.
Keytabs	The keytab (short form for the term *key table*) is a file that stores the keys for a list of principals. The keytab file is used by server applications to both retrieve the client principal credentials, and to identify the different client principals that are intending to communicate.
krb5.conf	The *krb5.conf* file contains the Kerberos configuration settings, which includes the following: • Default realm • KDC and admin server locations for every Kerberos realm • Kerberos realms linked to every hostname

Note: For in-depth information about Kerberos configuration settings, refer to the *MIT's Kerberos official website*.

6.7.1 Kerberos Configuration Manager

The Kerberos settings are configurable through the Web UI from the **Kerberos** screen. This section explains the series of steps that you must follow to successfully configure Kerberos with Protegrity Discover.

Editing the krb5.conf file:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **Kerberos**.
2. Click edit to edit the *krb5.conf* file.

   The **Krb5 configuration** popup appears. Modify the configuration as per your requirements.

   The following code block shows a sample *krb5.conf* file.

```
[libdefaults]
default_realm = EXAMPLE.COM
dns_lookup_realm = false
dns_lookup_kdc = true
ticket_lifetime = 24h
forwardable = yes
default_ccache_name = DIR:/tmp/

[realms]
EXAMPLE.COM = {
kdc = kdcserver.example.com
admin_server = kdcserver.example.com
default_domain = EXAMPLE.COM
}

[domain_realm]
```

```
.example.com = EXAMPLE.COM
example.com = EXAMPLE.COM
```

Note: If Kerberos has been installed on the targeted machine that is being scanned by Protegrity Discover, then you can re-use the same configuration specified in the *krb5.conf* file of the targeted machine. Typically, the *krb5.conf* file is installed in the */etc* directory.

Important: Ensure that the case of the domain name defined in the *krb5.conf* file matches the case of the domain name of the Kerberos system.

Requesting a new ticket:

3. Click **Request A New Ticket**.

   The following **Request Ticket** dialog box appears.

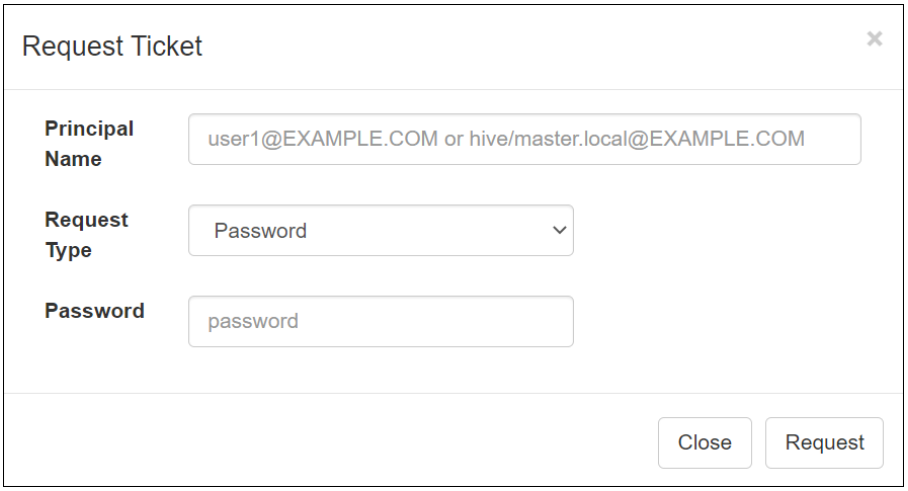


Figure 6-42: Protegrity Discover - Request a new Kerberos ticket

4. Enter the system specific principal in the **Principal** text box.

   The principal value is optional and is required here only if it is not defined in the *krb5.conf* file.

5. In the **Request Type** list, select whether you want to specify a password or upload a keytab.

6. If you have specified **Password** in the **Request Type** list, then specify the password in the **Password** text box required to access the Kerberos system. Else navigate to *step 7*.

7. If you have specified **Keytab** in the **Request Type** list, then click **Choose File** in the **Upload Keytab** field to select the Kerberos keytab file from your system.

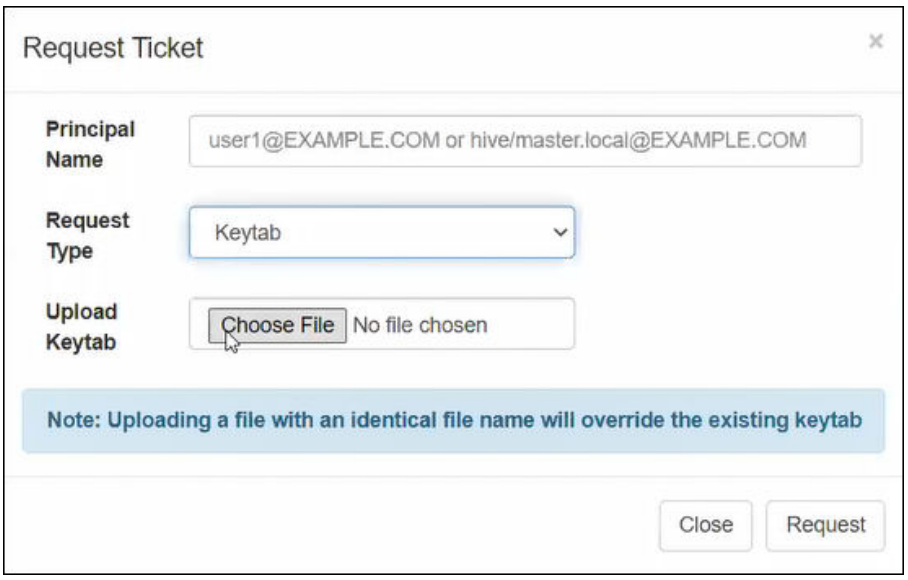


Figure 6-43: Protegrity Discover - Request a new Kerberos ticket with request type as Keytab

Important: Uploading a file with an identical file name will override the existing keytab.

The keytab file is optional and is required here only if its configuration is not defined in the *krb5.conf* file.

Note: Before adding the keytab, ensure that you have created a keytab in your system.

In a Windows system, you can create a keytab using the *ktpass* command, as shown in the following snippet.

```
ktpass -princ <Principal> -mapuser <Host name>\<User> -pass <Password> -out  <Keytab
name>
```

For example:

```
ktpass -princ sqluser1@TEST-PROTEGRITY.COM -mapuser TEST-PROTEGRITY\sqluser1 -pass
protegrity -out  sqluser1.keytab
```

For more information about the *ktpass* command, refer to the *Microsoft Documentation*.

In a Linux system, you can create a keytab using the *ktutil* command, as shown in the following snippet.

```
ktutil
addent -password -p <Principal> -k <Key version number> -e <Encryption type>
wkt /<Path to the Keytab file>/<Keytab name>
```

For example:

```
ktutil
addent -password -p sqluser1@TEST-PROTEGRITY.COM -k 1 -e <Encryption type>
wkt /<Path to the Keytab file>/sqluser1.keytab
```

For more information about the *ktutil* command, refer to the *MIT Kerberos Documentation*.

Ensure that the principal has been defined in the Kerberos database.

8. Click **Request**.

The requested ticket with keytab details is listed on the following Web UI.

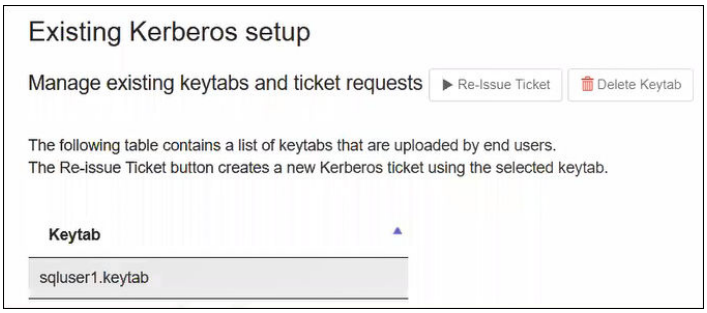


Figure 6-44: Protegrity Discover - List of Keytabs

The tickets remain active for a specific period of time, as specified by the *ticket_lifetime* parameter in the *krb5.conf* file.

> **Note:** When a ticket is created or renewed, a log of type *INFO* is generated.

To re-issue a ticket:

9. Select a keytab, and then click **Re-Issue Ticket**.
   A dialog box appears that prompts you to specify the name of the principal.

10. In the **Principal Name** text box, specify the principal name.
11. Click **Request** to re-issue the Kerberos ticket for the specified principal using the selected keytab.

To delete a keytab:

12. Select a keytab, and then click **Delete Keytab**.

> **Caution:** You cannot delete a keytab that is associated with an existing job.

> **Note:** When a keytab is deleted, a log of type *warning* is generated.

6.8 Retrieving ESA Data Elements

You can integrate Discover with the ESA, which allows you to associate the data elements configured in the ESA with the classifiers. You need to specify the ESA hostname and the login credentials for connecting to the ESA.

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **ESA**.
   The **ESA** screen appears.

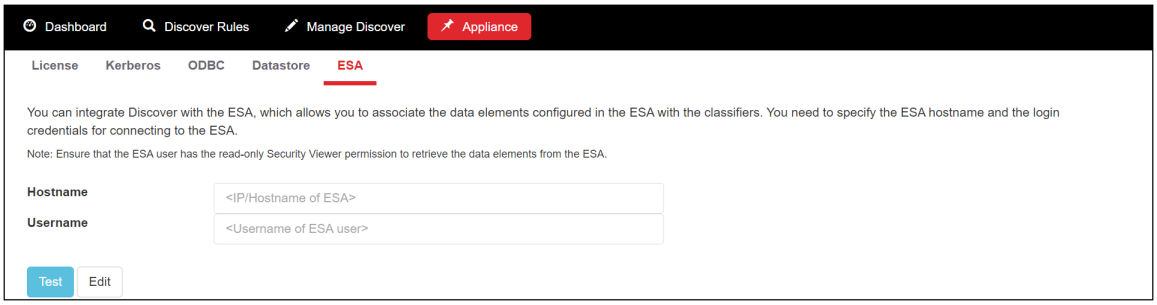


Figure 6-45: ESA Screen

2. Click **Edit**.

   The **ESA Config** dialog box appears.

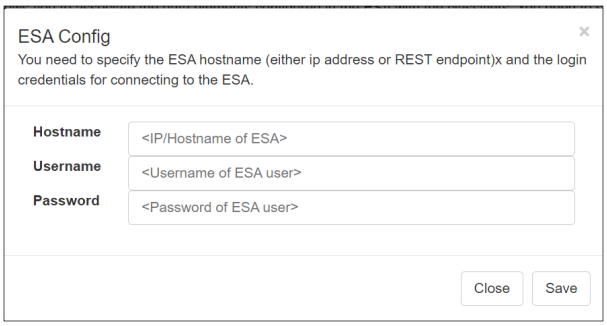


Figure 6-46: ESA Config Dialog Box

3. Specify the following details in the **ESA Config** dialog box.

Table 6-18: ESA Config Fields

Field	Description
Hostname	Specify the IP address or the hostname of the ESA from where you want to retrieve the data elements. **Important:** You can retrieve the data elements only from the ESA version 7.2.1 or later.
Username	Specify the name of the user for logging into the ESA. **Important:** Ensure that the ESA user has the read-only Security Viewer permission to retrieve the data elements from the ESA.
Password	Specify the password of the ESA user. **Note:** Ensure that you specify the correct password. If you type an incorrect password multiple number of times and then click **Save**, then the user will get locked out from the ESA.

4. Click **Save**.

   The user credentials are sent to the ESA. If the credentials are accurate, then Protegrity Discover connects with the ESA and downloads the list of data elements from the ESA.

   After the data elements are downloaded from the ESA, they are populated in the **ESA Data Element** list in the **Discover Rules** > **Classifiers** > **General** tab of each classifier. You can then associate a data element from the list to the corresponding classifier.

For more information about associating an ESA data element to a classifier, refer to *step 2* of the section *Updating the General Tab*.

5. If you want to verify whether the list of data elements in the ESA have been updated, then click **Test** to connect to the ESA and retrieve the current list of data elements.
6. If you want to connect to another ESA, then click **Edit** and modify the value of the fields in the **ESA Config** dialog box in *step 3*.

   After you successfully retrieve the data elements from another ESA, the data elements retrieved from any earlier ESA are removed from the **ESA Data Element** in the **General** tab of each classifier.

6.9 Managing the Appliance Information

The Protegrity Discover appliance information is available from the Appliance Web UI. To access this information, click on the **Settings** icon on the top-right corner of the Protegrity Discover Web UI. Alternatively, if the Protegrity Discover service is down, then you can access the Appliance Web UI by navigating to the *https://{Management IP}/index.html* URL. This section explains the information that you can refer to and manage from the Appliance Web UI.

To check disk usage information:

1. On the **Dashboard** menu, check the **Disk Usage** section, as shown in the following screenshot.

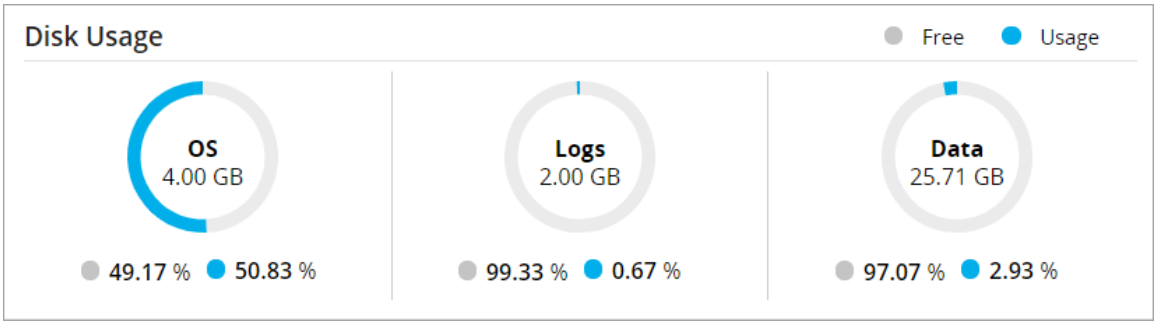


Figure 6-47: Protegrity Discover - Disk Usage

To manage system services:

2. Navigate to **System** > **Services**.

Figure 6-48: Protegrity Discover - System Services

3. Click stop or restart to stop or restart the specific service, respectively.

To manage system files:

4. Navigate to **Settings** > **System** > **Files**.

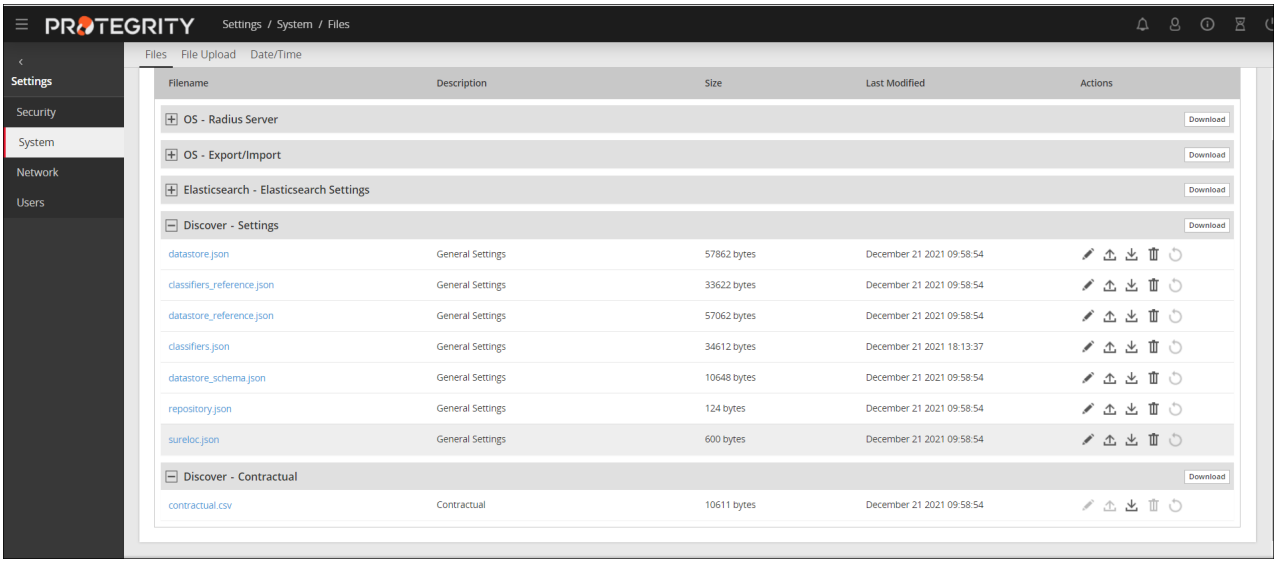


Figure 6-49: Protegrity Discover - System Files

The following system files are listed:

- *datastore.json* - Contains ODBC and datastore settings.

  You can also configure the datastore settings from the **Appliance** > **Datastore** menu of the Protegrity Discover Web UI.

  For more information about configuring the datastore settings from the Protegrity Discover Web UI, refer to the section *Managing Datastores*.

- *classifiers_reference.json* - Contains the default value for the classifier settings. If you have modified the classifier settings from the *classifiers.json* file or from the Protegrity Discover Web UI, and you want to refer to the default values, then you can access the *classifiers_reference.json* file.
- *datastore_reference.json* - Contains the default values for the datastore settings. If you have modified the datastore settings from the *datastore.json* file or from the Protegrity Discover Web UI, and you want to refer to the default values, then you can access the *datastore_reference.json* file.
- *classifiers.json* - Contains the classifier settings.

  You can also configure the classifier settings from the **Discover Rules** > **Classifiers** menu of the Protegrity Discover Web UI.

  For more information about configuring the classifier settings from the Protegrity Discover Web UI, refer to the section *Managing Classifiers*.

  You can also configure the size of the input data that can be processed by the spaCy module at one time, using the *nlp_max_length* parameter in the settings for the NLP GENERIC classifier. By default, the value of this parameter is set to *200,000*.

  For example, consider an input document that contains 2,000,000 contiguous characters, without blank lines. In this case, Protegrity Discover will break the input document into 10 chunks of 200,000 characters each, based on the value that you have set in the *nlp_max_length* parameter. Each chunk is then sent to the NLP classifiers for processing.

  In the *nlp_max_length* parameter, you can specify an integer value between *1* and *1,000,000*, both values inclusive.

However, if you specify the value of this parameter as *0*, then Protegrity Discover uses the default chunk size of *1,000,000* characters, which is internally defined in the application.

> **Note:** The *nlp_max_length* parameter is applicable to all the discover jobs in Protegrity Discover.

For more information about how spaCy processes data, refer to the section *Language Processing Pipelines* on the spaCy website.

- *datastore_schema.json* - Contains the schema of the *datastore.json* file.
- *repository.json* - Contains the following attributes related to the repository:
  - *verify_certs* - Ensures whether Protegrity Discover and the repository are communicating securely. By default, this value is set to *true*.
  - *timeout* - Specifies the time for which Protegrity Discover tries to communicate with the repository, after which the connection times out. By default, this value is set to *45*. The unit of the *timeout* parameter is seconds.
  - *hosts* - Specifies the IP address or hostname of the machine on which the repository has been installed. By default, this value is set to *localhost*, which indicates that both the repository and Protegrity Discover have been installed on the same machine. You can specify a remote IP address if you have installed the repository on a different machine.
  - *max_retries* - Indicates the maximum number of attempts by Protegrity Discover to recommunicate with the repository after the connection times out. By default, this value is set to *10*.
  - *retry on timeout* - Specifies whether Protegrity Discover attempts to recommunicate with the repository after the connection times out. By default, this value is set to *true*.
- *sureloc.json* - Contains the following attributes:
  - *balance_workers_intervals* - Specifies the time interval, in seconds, after which Protegrity Discover checks whether additional workers are required to scan a job. By default, this value is set to *5*.
  - *default_score_calculation_template* - Specifies the default formula for calculating the confidence score for all classifiers. By default, this value is set to:

    ```
    passed*received/(checked*sampled)
    ```

    This same value is populated in the **Score Calculation Template** field on the **General** tab for all classifiers.

    For more information about updating the **General** tab, refer to the section *Updating the General Tab*.
- *contractual.csv* - Specifies the license information for all the third-party components used in Protegrity Discover. You can access this file from the **Discover - Contractual** section on the **Files** tab.

5. Click to make changes to a file.

To upload a file:

6. Navigate to **Settings** > **System** > **File Upload**.

Figure 6-50: Protegrity Discover - File Upload

7. Click **Choose File**.
8. Select the file from your system, and then click **Upload**.

To check appliance logs:

9. Navigate to **Logs** > **Appliance**.
   The current event logs are displayed as follows:

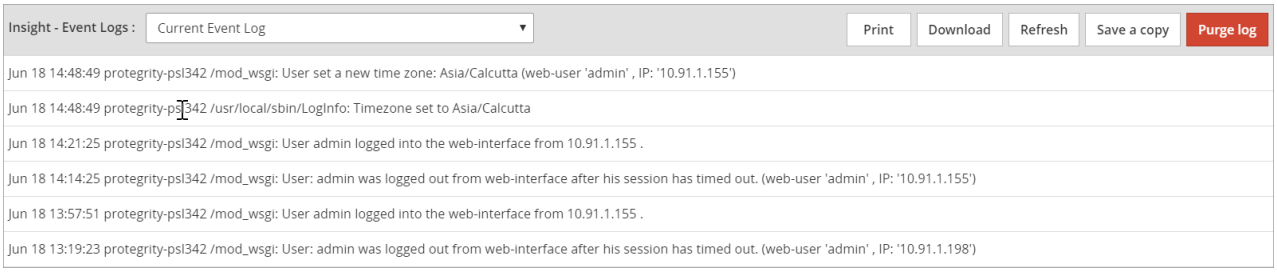


Figure 6-51: Protegrity Discover - Appliance Logs

10. Select an event from the list box to refer to the specific event log.
11. Click **Download** or **Save a copy** to save log results to a local file.

To manage system date/time:

12. Navigate to **Settings** > **System** > **Date/Time**.

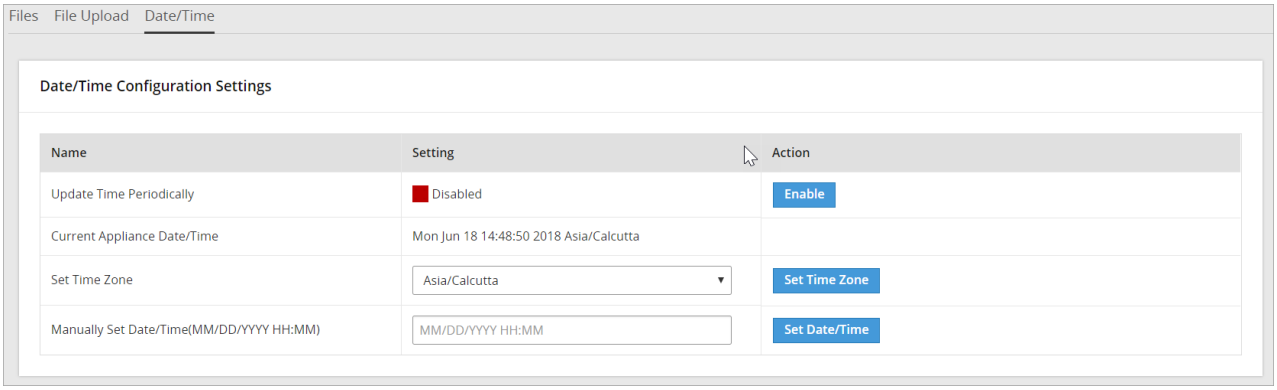


Figure 6-52: Protegrity Discover - System Date/Time

13. Select the time zone from the list box, and then click **Set Time Zone**.
14. Enter the date/time manually in the text box, and then click **Set Date/Time**.

To manage network settings:

15. Navigate to **Settings** > **Network** > **Network Settings**.

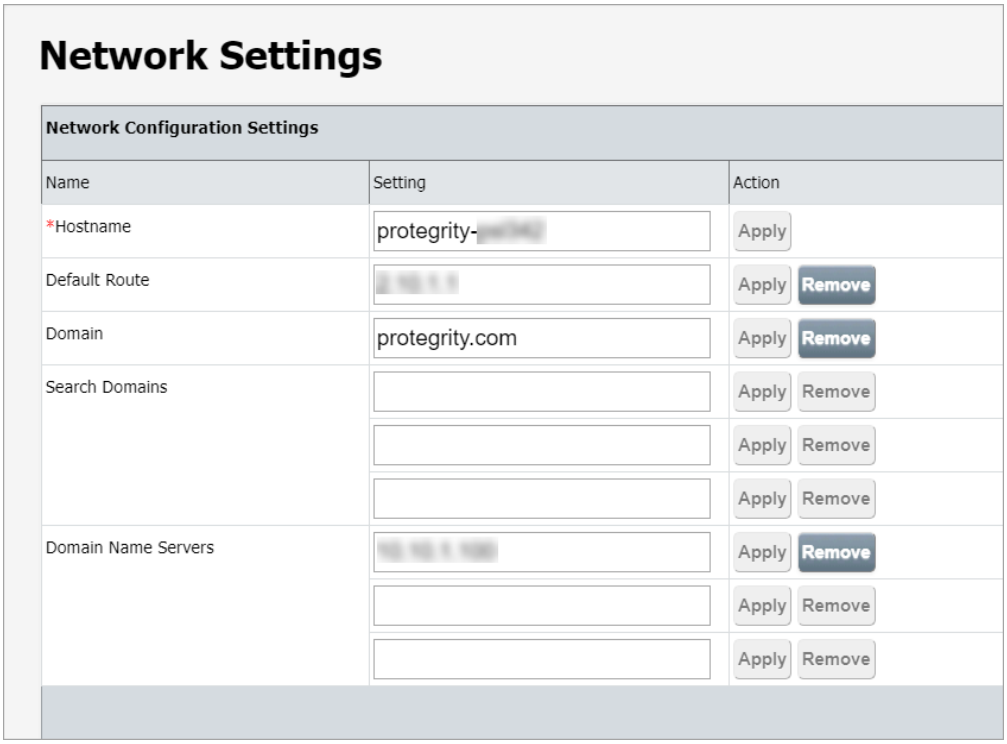


Figure 6-53: Protegrity Discover - Network Settings

16. After making any changes, click **Apply**.

To manage system users or define password policy:

17. Navigate to **Settings** > **Users** > **User Management**.

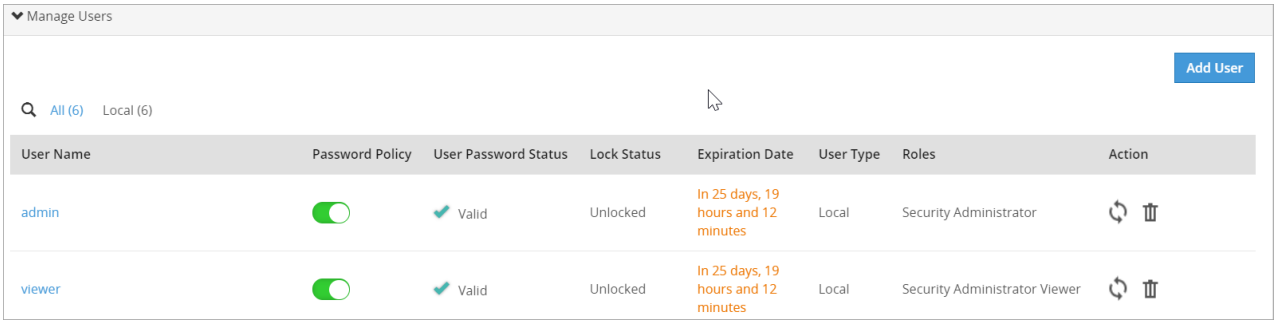


Figure 6-54: Protegrity Discover - Manage Users

18. Click Reset to reset *admin* or *viewer* password.
    The panel to set a new password appears.
19. Enter a new password and retype the new password in the text boxes, and then click **Ok**.
20. Click Edit to define password policy.
21. Make changes to your password policy, and then click **Apply Changes**.

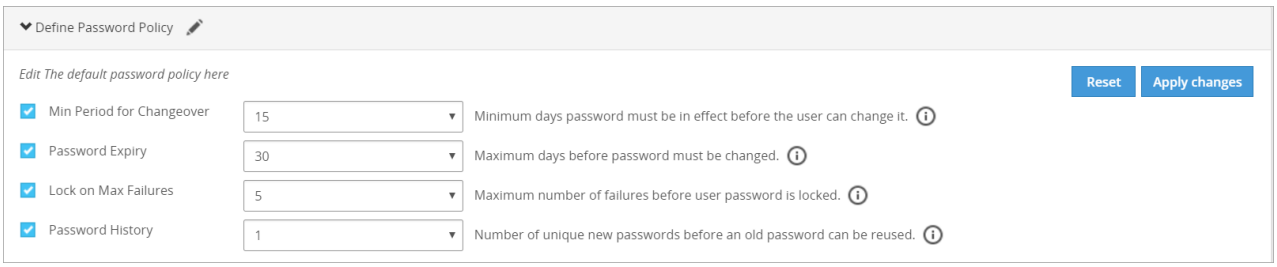


Figure 6-55: Protegrity Discover - Define Password Policy

For more information about defining a password policy, refer to the section *Password Policy Configuration* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

22. Click **Reset** to reset your password policy to default settings.
23. If you want to add users, then click **Add User**.

    The users who want to use Protegrity Discover must be assigned the required Protegrity Discover-specific permissions through roles. Roles are templates that include permissions, and users can be assigned one or more roles.

    By default, the *Security Administrator* role includes the *Discover Admin* and *Discover RESTAPI* permissions, while the *Security Administrator Viewer* role includes the *Discover Viewer* permission.

    You can also create custom roles, and assign them Protegrity Discover-specific permissions, as shown in the following figure.

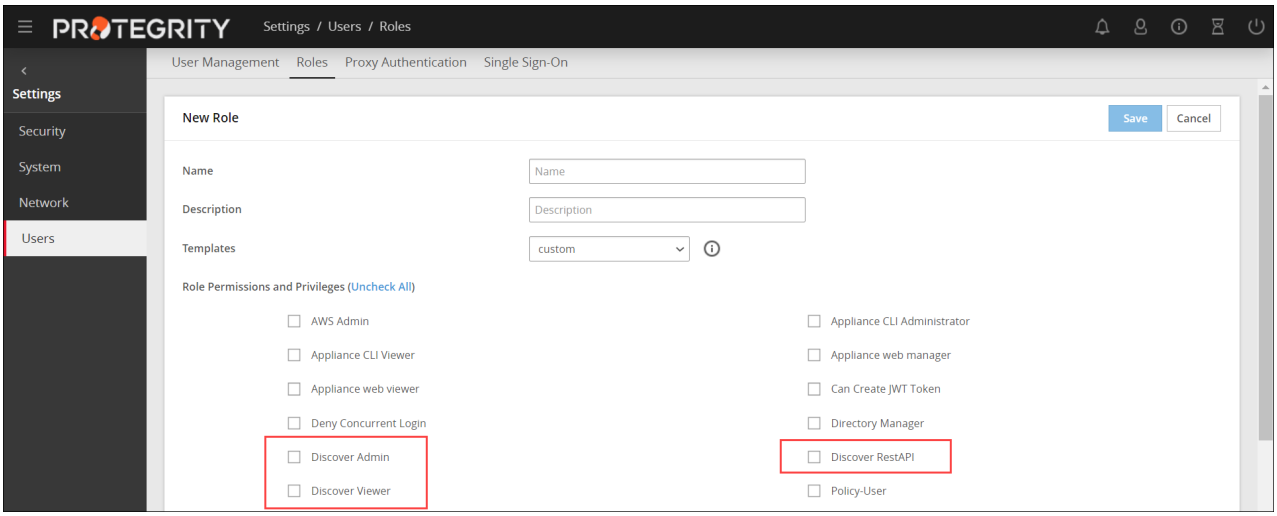


Figure 6-56: Create Roles

For more information about Protegrity Discover-specific permissions, refer to the section *Protegrity Discover-specific Permissions*.

For more information about managing roles, refer to the section *Managing Roles* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

For more information about managing users, refer to the section *Managing Users* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

To download troubleshooting information:

24. On the top-right corner of the Appliance Web UI, click > **Support**.

    The following **Support** screen appears.

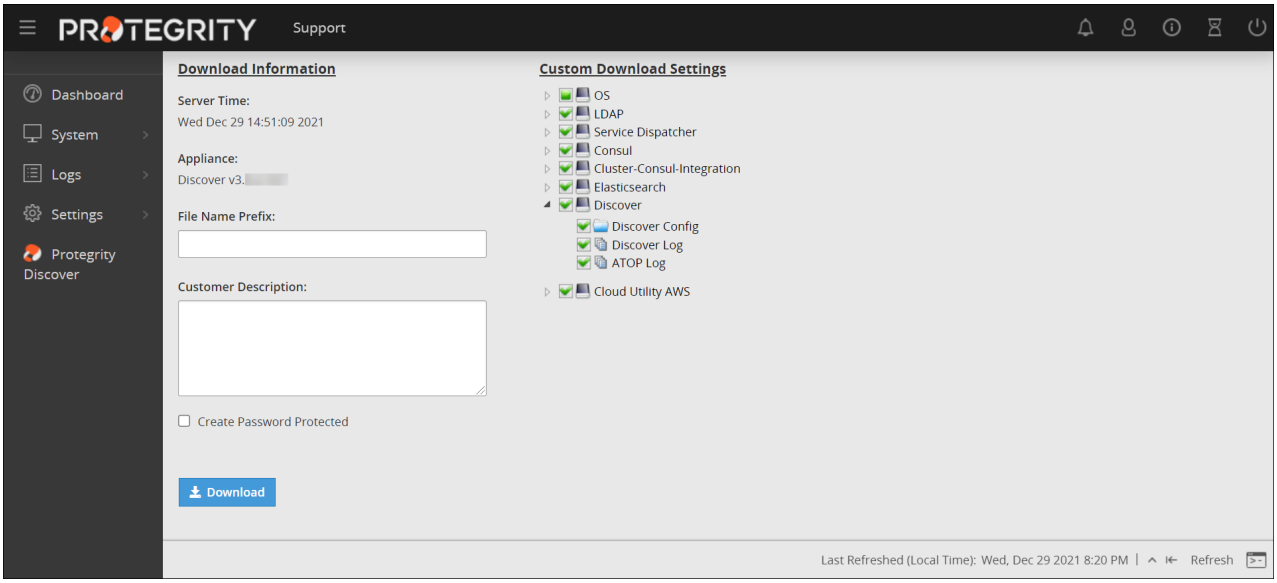


Figure 6-57: Support Screen

To assist in troubleshooting errors, you can choose to download information, such as status of the appliance and other services, repository configuration and logs, and Protegrity Discover configuration and logs. This information can help you in troubleshooting any issue.

The Protegrity Discover section enables you to download the following information:

- *Discover Config* - Consists of configuration files for Protegrity Discover
- *Discover Log* - Consists of archived and non-archived Protegrity Discover Web UI and REST API log files.

  For more information about Protegrity Discover log files, refer to the section *Log Manager*.

- *ATOP Log* - Consists of log files that are required for monitoring the status history of the system where Protegrity Discover has been installed.

> **Tip:** If you want to download the troubleshooting information regarding Protegrity Discover, then ensure that you clear the **ATOP Log** check box. You can select the **ATOP Log** check box only if you want to troubleshoot the system performance.

For more information about the **Support** screen, refer to the section *Support* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

25. In **Custom Download Settings**, select one or more check boxes that correspond to the settings that you want to download.
26. Click **Download** to download the settings to a zip file.

    After you extract the zip file, you can access the repository, Protegrity Discover Web UI, and Protegrity Discover REST API logs from the *var\log* directory.

To navigate to the Protegrity Discover dashboard:

27. On the left navigation pane, click **Protegrity Discover** to navigate to the Protegrity Discover dashboard.

Chapter 7

Calculating the Confidence Score

Protegrity Discover data discovery engine, which is a part of the platform itself, uses a confidence scoring method for the data classification. This section provides an overview of the confidence scoring method, including how the confidence score is calculated.

Confidence scores help to evaluate our confidence in the classification findings, and aim for *zero false positives or false negatives*.

The following section describes the methods - *confidence scoring* and *analysis* - used by Protegrity Discover to classify the data.

7.1 Confidence Scoring

A confidence score determines the confidence or severity level in the data classification findings. To arrive at a confidence score, Protegrity Discover investigates a coordinate. A coordinate represents the location of sensitive data, which can be any system, database, schema, table, column, or file path. The following figure illustrates the steps involved in the calculation of a confidence score.

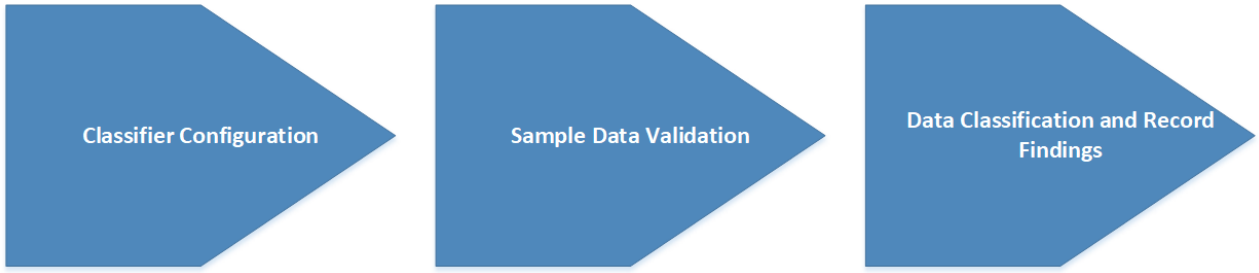


Figure 7-1: Protegrity Discover - Confidence Score Calculation Steps

7.1.1 Classifier Configuration

Classifiers validate and classify sensitive data from the sampled records of the targeted system. The following table lists the methods that are used by the classifiers to validate the data.

Table 7-1: Protegrity Discover - Classifier methods and types

Method	Classifier description
Regular Expressions	Protegrity Discover provides two types of regular expressions: • *Data Qualifier Regex* - Used to identify and qualify the data before evaluation. • *Regex Patterns* - Used to evaluate and classify the data.

Method	Classifier description
	For more information about classifiers with regular expressions, refer to the section *Regular Expression*. For more information about updating the data qualifier regex for qualifying the data, refer to the section *Updating the Qualifications Tab*. For more information about updating the regex patterns for evaluating the data, refer to the section *Updating the Regex Tab*.
Reference Data	The reference data classifies common names and list of cities, states, or postal codes effectively with a dictionary containing the valid geolocation data of all the required entries. Protegrity Discover provides inbuilt support for the referential data, with preloaded lists of names and addresses being embedded with the system. In case of structured classifiers, Protegrity Discover supports the following language-specific and country-specific reference data: • *Address* - Protegrity Discover currently supports address data for USA, Sweden, Italy, Germany, Netherlands, and Turkey. • *Phone numbers* - Protegrity Discover currently supports area codes for USA and Germany. • *Name* - Protegrity Discover currently supports names in English, Dutch, German, and Italian languages. In case of unstructured NLP classifiers, Protegrity Discover only supports address data for USA. For more information about updating the reference data for classifiers, refer to the section *Updating the Metadata Tab*.
Logical Tests	The variable categories, such as Credit Card Number (CCN), phone number, and email address, are classified using logical tests using a predefined function containing the classification logic. For more information about classifiers that use logical tests, refer to the section *Logical Tests*.
Schema Keyword Identification	The schema specific keywords are provided that include a regex and a boost. The regex defines a pattern to identify or classify the data (column name) by investigating whether the data is in close proximity to the pattern. The boost is the weight given to the classifier, which acts as a catalyst to measure the confidence score. For more information about classifiers with schema keywords, refer to the section *Schema Keyword Identification*. For more information about adding schema keywords for classifiers, refer to the section *Updating the Metadata Tab*.
Natural Language Processing	Protegrity Discover uses spaCy patterns to identify sensitive, unstructured data using natural language processing. For more information about spaCy, refer to the *spaCy* website.
User-Defined	Protegrity Discover enables you to write custom code in Python for identifying sensitive data. For example, Protegrity Discover uses custom code for identifying IBAN (International Bank Account Number). For more information about writing custom code for user-defined classifiers, refer to the section *Updating the Source Code Tab*.

Method	Classifier description
Image	Protegrity Discover uses an in-built model that implements TensorFlow Lite for performing image recognition. For more information about TensorFlow Lite, refer to the *TensorFlow Lite* website.

As a first step to meet your classification requirements, you can add your own custom settings to the classifier configuration to include a new type of classifier.

You can modify existing classifiers or create new classifiers from the **Discover Rules** > **Classifiers** screen.

For more information about managing classifiers from the Protegrity Discover Web UI, refer to the section *Managing Classifiers*.

Alternatively, you can also customize the classifier configuration using the *classifiers.json* configuration file.

For more information about creating or modifying the classifiers in the *classifiers.json* configuration file, refer to the section *Customizing the Classifier Configuration*.

The following sections explain the different classifier methods and its settings, as listed in *Table 1. Protegrity Discover - Classifier methods and types*.

7.1.1.1 Regular Expressions

The *Data Qualifier Regex* contains the logic or pattern to identify and evaluate the sensitive data values residing in the system. *Data qualifier Regex* is used to qualify a value before evaluating it and to filter out any values that you are not searching.

For example, a *Password* classifier can be set to search for password values that are between 6 and 16 characters in length. In this case, the qualifying pattern can be **^.{6,16}$** as shown in the following figure.

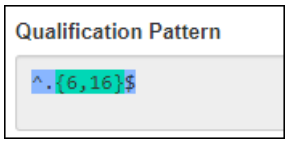


Figure 7-2: Data Qualifier Regex for Password Classifier

The *Regex Patterns* enable you to define multiple Regex patterns for identifying data. For example, a *Password* classifier can define Regex settings with patterns, such as uppercase, lowercase, allowed digits, and symbols. The following figure is a sample of the regex settings that are set for the *Password* classifier.

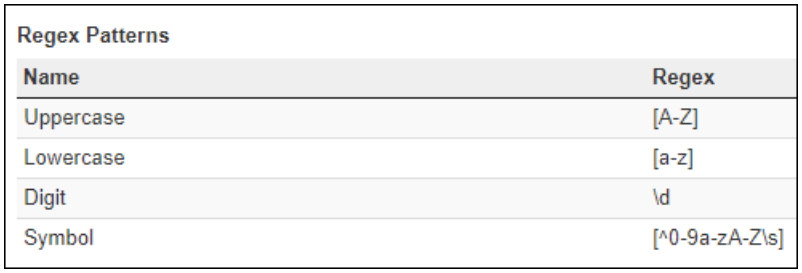


Figure 7-3: Regex Patterns for Password Classifier

In the given figure, the *Password* classifier is set to accept combinations of uppercase letters, lowercase letters, and digits, but any special characters are not allowed.

For more information about updating the data qualifier Regex for qualifying data, refer to the section *Updating the Qualifications Tab*.

For more information about updating the regular expressions class for evaluating data, refer to the section *Updating the Regex Tab*.

Alternatively, you can modify the data qualifier Regex and Regex patterns within the *classifier.json* configuration file.

For more information about modifying the *classifiers.json* configuration file, refer to the section *Customizing the Classifier Configuration*.

7.1.1.2 Logical Tests

Logical tests use a pre-defined function that contains the classification logic. Protegrity Discover uses logical tests with specific types of classifiers to classify the data. The variable categories, such as Credit Card Number (CCN), phone number, and email address, are classified using logical tests.

For example, a credit card number must be of a specific length. It must begin with a valid Bank Identification Numbers (BIN) range and must pass the *Luhn* checksum. The email address must have a specific format, valid domain name, and the user account must exist in the domain's mail system.

7.1.1.3 Schema Keyword Identification

This section provides information about the schema keyword and metadata keyword identification.

In the case of structured data, the schema or metadata keywords identify the column name of the database table. In the case of unstructured and semi-structured data, the schema or metadata keywords are used to identify the keywords present in the file name or directory name.

You can assign specific keywords to a classifier configuration so that the classifier can accurately identify the column using pattern-matching. Keywords can have different weights (boost) based on the database metadata definition, data dictionary, or as per the standard naming convention rules of the organization. A greater keyword accuracy, as per organization rules, results in a higher classifier weight (boost), and consequently a higher confidence score.

For example, the social security number defined as *SSN* or *social_security_number* might have a higher weight as compared to *social* or *soc*.

The following figure is a sample of the schema or metadata keywords for the *Password* classifier.

Keyword Patterns

Name	Regex	Boost	Score	Continue
password	(?i)pass[\W_\s]*?(word|code)	4.0	0.0	☑
pwd	(?i)pwd	4.0	0.0	☑

Figure 7-4: Metadata Keywords for Password Classifier

In the sample, the metadata keywords identify any column that matches *password* or *pwd* string based on the pattern matching expressions defined in the regex. Additionally, both strings are given equal boosts. If an additional string, for example *passwd* is required, you can specify a lower boost in comparison to the *password* and *pwd* string, as per the accuracy and standard naming convention rules.

For more information about adding schema keywords for classifiers from the **Classifiers** screen, refer to the section *Updating the Metadata Tab*.

Alternatively, you can also set the schema keywords by modifying the *classifiers.json* file.

For more information about setting the schema keywords in the *classifiers.json* configuration file, refer to the section *Customizing the Classifier Configuration*.

7.1.1.4 Customizing the Classifier Configuration

This section explains how you can set or customize classifier configurations by modifying the *classifier.json* file.

To customize the classifier configuration:

1. On the Protegrity Discover Web UI, click the top-right icon for settings.
2. Navigate to **Settings** > **System** > **Files**.

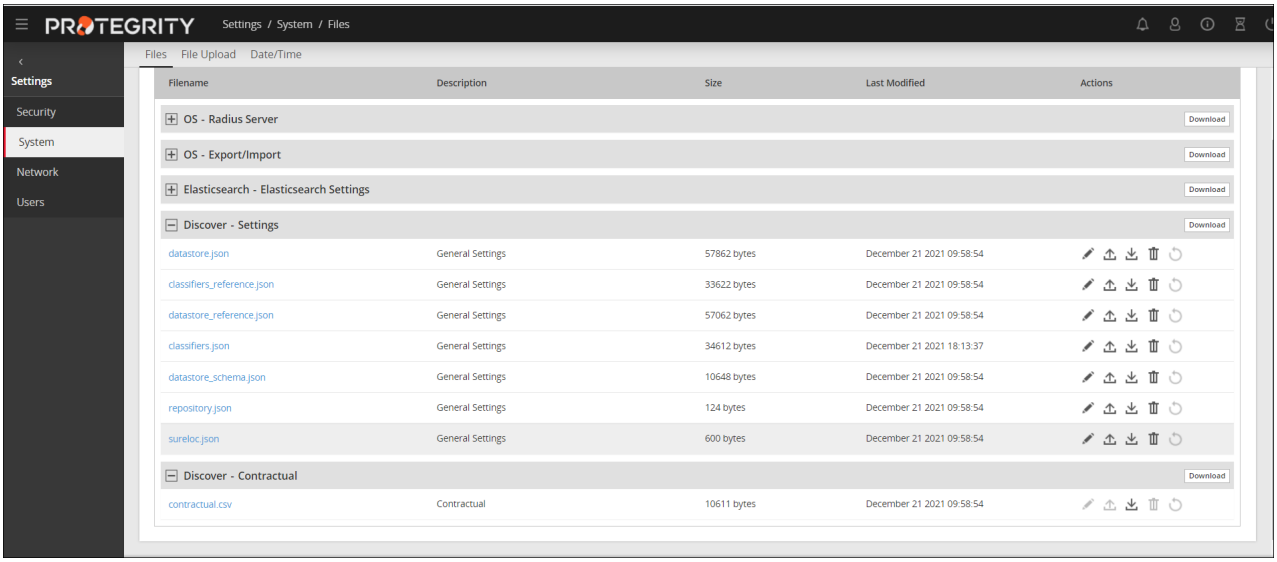


Figure 7-5: Protegrity Discover Appliance Web UI - Classifier Configuration

3. In the **Discover - Settings** section, click **Edit**, adjacent to the *classifiers.json* configuration file, to edit the file.

   The following snippet is a sample of the IBAN classifier configuration:

```
{
    "name": "IBAN",
    "description": "International Bank Account Number (aka IBAN) is a number attached to
all accounts in the EU countries identify the country the account belongs to, the account
holder's bank and the account number itself",
    "data_element_name": "IBAN",
    "class": {
      "type": "USER_DEFINED",
      "programing_language": "Python",
      "source_code": "class UserDefinedClassifier(object):<IBAN classifier logic>"
    "data_qualifier_regex": "(?i)^\\w{2}[\\s\\-]?\\d{2}[\\s\\-]?[\\w\\s]{11,37}$",
    "schema_keywords": [
      {"name": "iban", "regex": "(?i)iban", "boost": 1.5},
      {"name": "international", "regex": "(?i)inter(national)?", "boost": 1.1},
      {"name": "bank", "regex": "(?i)bank", "boost": 1.5},
      {"name": "account", "regex": "(?i)acc(ount)?", "boost": 1.5},
      {"name": "account", "regex": "(?i)num(ber)?", "boost": 1.1}
    ]
  },
```

Important: Alternatively, you can also update the classifier configuration from the Protegrity Discover Web UI from the **Discover Rules** > **Classifiers** screen.

> For more information about managing classifiers from the Protegrity Discover Web UI, refer to the section *Managing Classifiers*.

4. Click **Save** after you have modified the file.

7.1.2 Sample Data Validation

Sample data consists of the configured data size, or the number of records that you set, when you are creating a discovery scan.

The sampled records exist within the Protegrity Discover system while the classification is in progress, and are cleared when the classification is completed. The enabled classifiers inspect the sample data collected by the system to ensure that the data has the right format and valid type. For example, the *Credit Card* classifier can skip any alpha (a-z) values, and the *Name* classifier can skip any alpha-numeric (a-z, 0-9) values. The Protegrity Discover system also reports null and repeating data values.

The sample data that passes this validation is the observed sensitive data and it forms the basis for the *classification* stage. The classification stage reports the estimated number of sensitive data values at the coordinate.

The estimate is determined using either the total size of the data container or the volume of the data at the coordinate. For example, if you sample 1000 data values (records) from a table containing 10K records, and scan results indicate that 700 data values are sensitive, then it implies that 70 percent of the sampled data classifies as *sensitive* data.

However, the sample data is a fraction of the data container with 10K records. The estimated sensitive data that is present at the coordinate or a location, is derived from the following formula:

```
Estimated sensitive data = Volume of data at the coordinate * (observed sensitive data /
sample data)
```

The estimated sensitive data is calculated by using the ratio of observed sensitive data to sampled data, multiplied by the total size of the data container. Therefore, as per this example, the estimated sensitive data count is 7000 data values.

7.1.3 Data Classification and Record Findings

The **Classifications** screen, which is accessible at the Web UI path **Dashboard** > **Classifications**, enables you to check classification records and download them for further analysis.

After validating the sample data, classifiers validate the observed sensitive (validated) data. For more information about how the classifiers validate the data, refer to the section *Classifier Configuration*.

The associated classifiers record every positive result for the classification. The record can't be modified and is permanent. You can also check the classification history for a coordinate, which provides the reasoning behind any previously determined conclusion.

> **Note:** The classification records in the classifications list, by default, are filtered to show records with a confidence score of 60% or above. You can modify this filter setting using the slider, which is available at the top-right corner of the classifications list.

By default, the classifications list displays up to 25 records on the screen. You can choose to display 10, 50, or 100 records on the screen using the **Show entries** drop-down list. Alternatively, you can use the page navigation buttons on the top and bottom of the classifications list to navigate the classifications results across pages.

The following figure displays the classification results for a SharePoint scan.

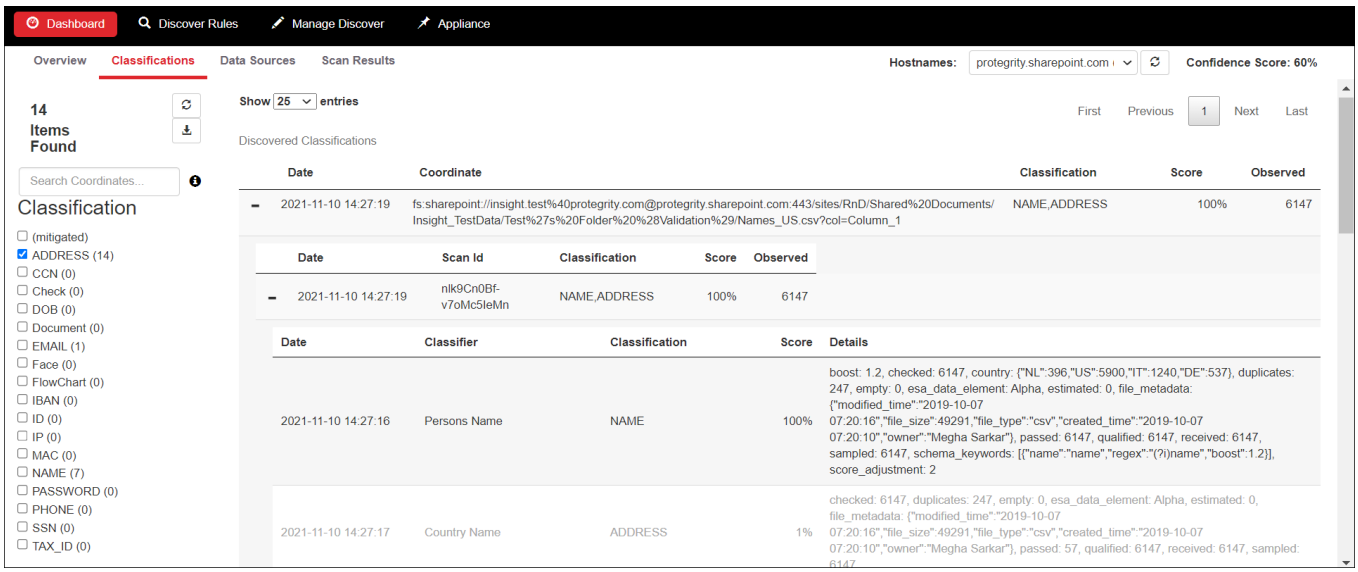


Figure 7-6: Protegrity Discover: Classification Record Findings

The following table explains the different attributes of a classification record.

Table 7-2: Protegrity Discover: Classification Record Findings

Attribute	Description
Date	Date when the scan is performed.
Coordinate	System for which the scan is performed. **Important:** In the case of structured data, such as, CSV, SAS files, Apache Avro, and Apache Parquet, Protegrity Discover automatically detects whether the first row of the data table is a column header, and includes this value as the column name in the coordinate attribute. Therefore, you must ensure that the first row of any structured data does not contain sensitive data. Otherwise, the sensitive data will appear on the Protegrity Discover Web UI.
Classification	The classification type that reported the sensitive data values. The classification result with the highest confidence score appears at the top of the classifications list.
Score	The final confidence score for the scan derived as per analysis. For each classifier, individual scores are recorded. The scores against each classifier are aggregated to form the final confidence score for the scan. For more information about analysis and score calculation, refer to the section *Analysis*.
Observed	Observed count of unique sensitive data values examined from the sampled data.
Expanding any classification record results into classification details for each classifier.	
Date	Date when the scan is performed.
Scan Id	Generated scan id for the scan job.
Classification	Classification type for the scan record.
Classifier	Specific classifier name, and its rules, defined for the particular classification.
Details	Additional details derived using the specific classifier, which include: • Classifier boost value, which is a weight assigned to the classifier • *Schema keywords or regex settings* that led to this result • Metadata for files - The *file_metadata* attribute displays metadata related to the files, such as, file type, date created, file size, and date modified. **Important:** The metadata depends on the filestore type.

Attribute	Description
	For more information about the metadata collected for each filestore, refer to the section *Appendix J: File Metadata Collected in Filestores*. • Sampled values - Total number of data values (records) configured for the scan from the overall data set • Received values - Number of data values received from the given set of sampled values • Empty values - Count of data values that are either all spaces, null, NaN, none, zero length string, or spaces/tabs only, and so on. • ESA Data Element - The *esa_data_element* attribute displays the ESA data element associated with the classification type. This parameter is applicable only if you have associated an ESA data element with the classifier, in the **General** tab of the classifier. For more information about associating an ESA data element to a classifier, refer to *step 2* of the section *Updating the General Tab*. For more information about retrieving data elements from the ESA, refer to the section *Retrieving ESA Data Elements*. • Estimated values - The estimated count of sensitive data values is derived using the following formula: `Estimated sensitive data = Volume of data at the coordinate * (observed sensitive data / sample data)` • Duplicate values - Count of data values that appear more than once in the data set • Checked values - Count of data values examined (received values excluding the empty values) • Qualified values - Count of data values that passed the data qualification pattern or regex check • Passed values - Count of data values that passed all data checks or evaluation rules, and is likely the data the classifier is looking for.
Score	Confidence score result for the specific classifier. The default confidence score is calculated as per the following formula: `Default confidence score = passed*received/(checked*sampled)`

The following table describes additional actions that you can perform from the **Classifications** screen.

Table 7-3: Protegrity Discover - Additional Actions through the Classifications Screen

Action	Description
Search the classifications list	Search the **Coordinate** column using a text or a string search in the **Search Coordinates** text box. By default, the search is not case-sensitive. However, you can specify the search text or string within double quotes to perform a case-sensitive search. For example, type "Key Word" to perform a case-sensitive search for this exact phrase. You can narrow down the search results by including additional key words in the search text.
Refresh the classifications list	Refresh the classifications list using the button.
Download the classifications list	Download the classifications list using the button.
Filter the classifications list	Filter the classifications list on the basis of either Classification type or scan date. You can also filter the classifications list on the basis of the hostname using the **Hostnames** drop-down list, which is available at the top-right corner of the classifications list.

7.2 Analysis

After the initial classifications are completed, context analysis is performed in the analysis stage. The context analysis, which becomes the basis of initial findings, fits a particular pattern of context.

For example, details, such as, city, postal code, state are part of one context, which is the address. If classifiers identify any of the elements in a context, the confidence scores are impacted for all of these elements.

The following additional factors are considered that have the potential to impact the confidence score.

- Overall weight adjustment for the classifier - You can adjust the overall score for the classifier to avoid any false positive result. You can specify the score adjustment value using the *data_evaluation_score_adjustment* variable in the *classifiers.json* configuration file.

```
{
    "name": "Date Of Birth",
    "data_element_name": "DOB",
    "class": {
      "type": "DATE",
      "range_max": 92,
      "range_min": 10
    },
    "data_evaluation_score_adjustment": 0.5,
    "schema_keywords": [
      {"name": "dob", "regex": "(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?", "boost": 2}
    ]
  },
```

> **Note:** The score adjustment value must be less than one.

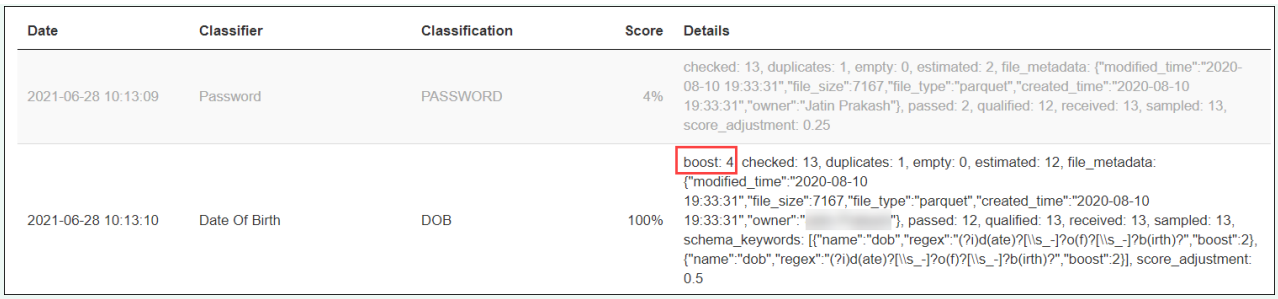
  For more information about how to modify the *classifiers.json* configuration file, refer to the section *Customizing the Classifier Configuration*.

- Schema or metadata keyword settings and proposed classification settings - If the schema or metadata keyword settings specify a *boost* value and the classifier settings specify *score adjustment* value, then the boost value is multiplied with the score adjustment value.

The confidence score is derived using the following formula:

```
Final Confidence Score = [(passed * received)/(checked * sampled)] * boost * score adjustment
```

> **Note:** If schema keywords contain multiple boost values for different keywords, then it is aggregated to display the final boost value on the **Classifications** screen.
>
> The following figure displays the final boost value of *4*, which is calculated by multiplying the two individual boost values *2* and *2*.



Date	Classifier	Classification	Score	Details
2021-06-28 10:13:09	Password	PASSWORD	4%	checked: 13, duplicates: 1, empty: 0, estimated: 2, file_metadata: {"modified_time":"2020-08-10 19:33:31","file_size":7167,"file_type":"parquet","created_time":"2020-08-10 19:33:31","owner":"Jatin Prakash"}, passed: 2, qualified: 12, received: 13, sampled: 13, score_adjustment: 0.25
2021-06-28 10:13:10	Date Of Birth	DOB	100%	boost: 4 checked: 13, duplicates: 1, empty: 0, estimated: 12, file_metadata: {"modified_time":"2020-08-10 19:33:31","file_size":7167,"file_type":"parquet","created_time":"2020-08-10 19:33:31","owner":"[redacted]"}, passed: 12, qualified: 13, received: 13, sampled: 13, schema_keywords: [{"name":"dob","regex":"(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?","boost":2}, {"name":"dob","regex":"(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?","boost":2}], score_adjustment: 0.5

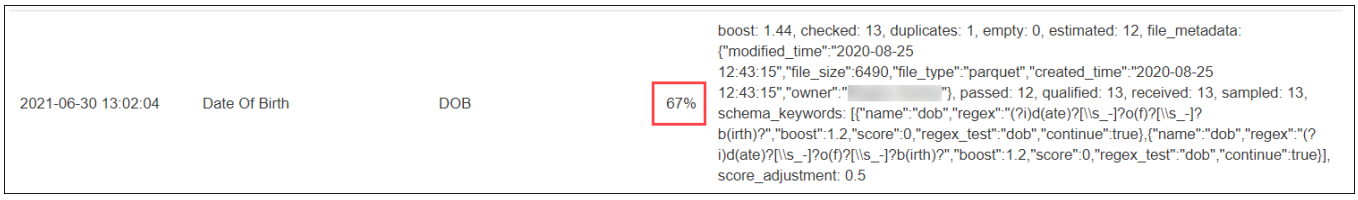
Figure 7-7: Protegrity Discover - Final Boost Value

If you have specified a confidence score for a schema keyword, and Protegrity Discover identifies the keyword as part of the schema or metadata, then the confidence score is added to the calculated confidence score before applying the required boost and score adjustment values. Therefore, the final confidence score for a single schema keyword is derived using the following formula:

```
Final Confidence Score = ([(passed * received)/(checked * sampled)] + Confidence score of
schema keyword) * boost * score adjustment
```

If you have specified confidence score values for multiple schema keywords that are detected by Protegrity Discover as part of the same schema or metadata, then all of the confidence score values are added to the calculated confidence score before applying the boost and score adjustment values.

The following figure shows the classification records for the *Date of Birth* classifier.



2021-06-30 13:02:04	Date Of Birth	DOB	67%	boost: 1.44, checked: 13, duplicates: 1, empty: 0, estimated: 12, file_metadata: {"modified_time":"2020-08-25 12:43:15","file_size":6490,"file_type":"parquet","created_time":"2020-08-25 12:43:15","owner":"[redacted]"}, passed: 12, qualified: 13, received: 13, sampled: 13, schema_keywords: [{"name":"dob","regex":"(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?","boost":1.2,"score":0,"regex_test":"dob","continue":true},{"name":"dob","regex":"(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?","boost":1.2,"score":0,"regex_test":"dob","continue":true}], score_adjustment: 0.5

Figure 7-8: Confidence Score for Multiple Schema Keywords

In this example, the confidence score of the classifier is calculated using the following formula:

```
Calculated Confidence score = [(passed * received)/(checked * sampled)]
                            = [(12 * 13) / (13 * 13)]
                            = 92.30769
```

The confidence score for the *dob* schema keyword is 0, which represents 0 percent. The confidence score for the *birthdate* schema keyword is 0, which represents 0 percent.

Therefore, the final confidence score is calculated using the following formula:

```
Confidence score = (Calculated Confidence Score + Confidence Score of Schema Keyword 1 +
Confidence Score of Schema Keyword 2)
                   * (Boost value of Schema Keyword 1 * Boost value of Schema Keyword 2) *
Score Adjustment
                 = (92.3 + 0 + 0) * (1.2 * 1.2) * 0.5
                 = 92.30769 * 1.44 * 0.5
                 = 66.46154
```

The final confidence score is rounded to 67 as shown in the figure.

For more information about adding schema keywords for classifiers from the **Classifiers** screen, refer to the section *Updating the Metadata Tab*.

The highest final confidence score is published at the top of the *data classifications* list with an estimate of the sensitive data found at the coordinate.

Chapter 8

Analyzing False Positive and False Negative Results

Data discovery solutions often involve predictive analytics, which can lead to false positive and false negative results. This section describes how Protegrity Discover deals with false positive and false negative results.

The prediction or classification results provide data that are classified with the highest accuracy. However, when dealing with data in the real world, the logical decision of *Is sensitive data* or *Is Not sensitive data* is not always possible. A false positive or false negative result can occur when data is classified with insufficient information, leading to an incorrect conclusion.

For example, consider data that consists of a collection of 16-digit values, in which a subset of these values might match a pattern that can be identified as a CCN. In such a scenario, the data belongs to the *false positive* category. Additionally, a different type of card, such as, a Brand Loyalty card might be skipped from the classification, resulting in a *false negative* situation.

Protegrity Discover solves this problem by applying an approach in which different methods are used to classify the data. The methods include sampling the data, using logical classifiers, or pattern matching. Each method of classification contributes its own weight towards establishing confidence with the final results.

Instead of categorizing data as *false positive* or *false negative*, Protegrity Discover relies on this confidence in the final results, using a confidence scoring method based on the score calculation. As a result, data is neither missed as *false negative* nor incorrectly identified as *false positive*. Rather, the data is now simply constituted as *low confidence* finding. If even a single test returns a positive result, it is recorded and the confidence score is calculated.

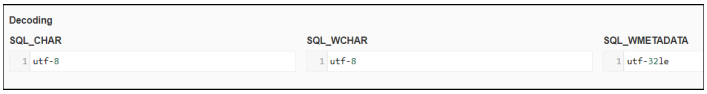
The recorded results, collectively, include all classification results performed on the data, as well any low confidence scores. The low confidence findings can be evaluated to check whether it requires further investigation.

Chapter 9

Troubleshooting Issues

This section includes the issues that you might encounter and the recovery actions to troubleshoot these issues.

Table 9-1: Troubleshooting Issues

Issue	Recovery Action
If the Protegrity Discover license is expired or invalid, you cannot view any records on the classification page and the page is blurred. Additionally, you cannot download the *analysis_data.csv* file.	To request a new license, you must perform the following steps: 1. On the Protegrity Discover Web UI, navigate to **Appliance** > **License**. 2. Click **Download License Request**. 3. Send the downloaded license request to Protegrity. 4. After you receive the new license from Protegrity, you must upload it to the Protegrity Discover system using the Web UI. 5. On the Protegrity Discover Web UI, navigate to **Appliance** > **License**. 6. Click **Upload License Activation File**. 7. Select the license file from your machine.
When you scan a DB2/zOS system, you receive the following error message: `An attempt to connect to the database server failed because of a licensing problem. (SQLSTATE=42968; SQL1598N).`	The DB2/zOS system must be configured accurately and activated for remote connectivity. The system also requires an active license. For more information to resolve this issue, refer to the IBM documentation.
When you scan a Hive cluster, the scan gets aborted with the following error message: `Failed to detect %s SQL_WMETADATA decoding charset', 'hive'`	You must modify the decoding for the Hive SQL_WMETADATA configuration to *utf-32le* from the **Appliance** > **Datastore** screen. The following figure shows this configuration:  Additionally, for the Cloudera Distribution Including Apache Hadoop (CDH) cluster, you must modify the *minimum user id (min.user.id)* configuration in the Yet Another Resource Negotiator (YARN) configuration file to a value of *100*. Restart the stale services.
You might encounter low disk space error with SQL Server database, causing the query execution and scan to fail.	This issue is caused due to the minimum disk space available on the SQL database server. The sampling query tries to fetch a number of records that exceeds this minimum disk space. You must edit the datastore configuration for Microsoft SQL Server to remove the 'order by' clause in the data sampling query. You first need to navigate to **Appliance** > **Datastore** and select **mssql** from the **DATASTORE** list. You then need to click the **Queries** tab to edit the SAMPLE_DATA query for Microsoft SQL Server.

Issue	Recovery Action
The data discovery scan for multiple systems using multiple Kerberos tickets, fail within the same Protegrity Discover machine.	Edit the *krb5.conf* file, and set the *default_ccache_name* profile variable under the *libdefaults* section to ***DIR:/<directory to store multiple credential caches>*** For example: `default_ccache_name = DIR:/tmp/user1/`
When you scan any system, you receive the following error message: ***[unixODBC][Driver Manager]Can't open lib '<Driver Name>' : file not found (0)***	The following possible issues can be the root cause: • Driver file is not found - You must download the driver file from the vendor's website and upload it to the Protegrity Discover Web UI using the Web UI path **Appliance** > **ODBC**. For more information about uploading the ODBC driver and setting the configuration settings, refer to the section *Extending the Support to Other Systems*. • Driver file is incompatible or corrupted - Ensure that the database has the ODBC driver for Linux to be compatible with Protegrity Discover. • Mismatch with driver names - Driver name in the *ODBC INI* file does not match with the driver name in the **Connection Template** field on the **Appliance** > **Datastore** > **Connection Settings** tab.
You encounter no results after a scan is completed.	This might happen due to the following possible reasons: • In cases where you might have a confidence score lower than 60%, the dashboard or classifications list does not display associated records due to the slider being set at 60% or higher, by default. • You have sampled a lesser number of records and thus, observe less numbers of derived sensitive data values. • The target coordinate did not contain any sensitive data, as per the defined scope of the associated data discovery job. • The target location is an incorrect coordinate. For additional assistance, contact Protegrity Support along with the log file details.
You encounter a Connection Refused error when you try and test the connectivity to a CIFS-based file system.	This might happen due to the following reasons: • The firewall is on • On Windows - The following network browsing features and service are not running: • Server Message Block (SMB) client and server • SMB Direct • NetLogon service • On Linux - The following services have not been installed and correctly configured: • SAMBA server • SMB service In case of Linux, you can check whether these services are installed by running the ***service smb status*** command.
You encounter a ***Failure to connect to repository*** error when you restart the machine on which Protegrity Discover is running.	This issue occurs because the repository takes a few milliseconds to start. As a result, it is possible that the repository has not yet started while Protegrity Discover has already started.

Issue	Recovery Action
When you scan a SQL Server database, you receive the following error message: ***VIEW DATABASE STATE permission denied in database*** .	This error occurs when you query a dynamic management view in Microsoft SQL Server, and you do not have the VIEW DATABASE STATE permission required to execute this query. For example, if you want to determine the total number of records available in a database table across partitions, then you must query the *sys.dm_db_partition_stats* dynamic management view using the GET_PARTITIONED_ROW_COUNT query from the **Appliance** > **Datastore** > **Queries** tab. However, if you do not have the required VIEW DATABASE STATE permission to execute this query, then the ***VIEW DATABASE STATE permission denied in database*** error is displayed. If the database table does not have any partitions, then you can delete the GET_PARTITIONED_ROW_COUNT query, and then rescan the database to avoid this error. You can use the following query in the GET_ROW_COUNT query to retrieve the total number of records in the table: ***"SELECT COUNT(*) as row_count FROM %(table_name)s;"*** For more information about the GET_PARTITIONED_ROW_COUNT and GET_ROW_COUNT queries, refer to the section *Configuring Datastore Queries*.
When you try and test the connectivity to a Hive datastore, you encounter the following error: ***SASL(-1): generic failure: GSSAPI Error: Unspecified GSS failure. Minor code may provide more information (Server hive/<server name>@<Kerberos realm> not found in Kerberos database). (34) (SQLDriverConnect)***	This issue occurs if the Kerberos principal is not registered with the KDC. To resolve this issue, you must add the principal to the KDC. For more information about the Kerberos settings required for Protegrity Discover, refer to the section *Kerberos*.
When you try to request a new Kerberos ticket, you encounter the following error: ***kinit command failed. kinit: keytab contains no suitable keys for <Principal> while getting initial credentials***	This issue occurs if the Kerberos principal has not been defined in the keytab file. To resolve this issue, you must include the principal name in the keytab file. To list all the principals that have been included in a keytab file, use the ***klist -kte <Keytab_name>*** command. For more information about the *klist* command, refer to the *MIT Kerberos Documentation*. For more information about the Kerberos settings required for Protegrity Discover, refer to the section *Kerberos*.
You are unable to connect to a DB2 server on a z/OS machine because of licensing issues.	To connect Protegrity Discover to the DB2 server on the z/OS machine, activate the DB2 Connect Unlimited Edition for zSeries license certificate with the *db2connectactivate* utility. For more information about using the *db2connectactivate* utility, follow the steps provided in the section *Activating the DB2 Connect Unlimited Edition* in the *DB2 for z/OS* documentation.

Issue	Recovery Action
When you try to connect to a DB2 server on a z/OS machine, you get the following error: `SQL0805N package <Package name> was not found`	This issue occurs if you are connecting to the z/OS system using ODBC for the first time. To troubleshoot this issue, ensure that you bind the SQL0805N package to the DB2 database on the z/OS machine. For more information about binding a package to the DB2 on the z/OS machine, refer to the *DB2 for z/OS* documentation.
When you try to request a new Kerberos ticket, you encounter the following error: `Command '['kinit', 'user1@EXAMPLE.COM']' timed out after 100 seconds` This timeout issue can occur if the KDC takes more than 100 seconds to issue a ticket.	Use one of the following commands in the CLI to manually create a Kerberos ticket. • `kinit [-principal]` • `kinit [-principal] -kt [keytab_file_location]` For more information about the *kinit* command, refer to the *MIT Kerberos Documentation*. For more information about the Kerberos settings required for Protegrity Discover, refer to the section *Kerberos*.

Chapter 10

Configuring Datastore Queries

This chapter describes how to configure the datastore queries when creating a datastore, or for modifying the settings of an existing datastore.

> **Note:** The queries are applicable only for ODBC-based datastores, and are not applicable for datastores that are based on a file system or cloud storage.

Protegrity Discover provides default query templates for each query through the Web UI for the out-of-the-box datastores. The default query templates contain pre-defined parameters that can be used to locate a schema instance, fetch records, and retrieve sample data. The query templates can be used as is, or can be modified as per your requirements. If you add a new datastore, then you must define the query templates for each query. You can refer to the default query templates for the out-of-the-box datastores.

The following flowchart specifies the sequence in which the datastore queries are executed.

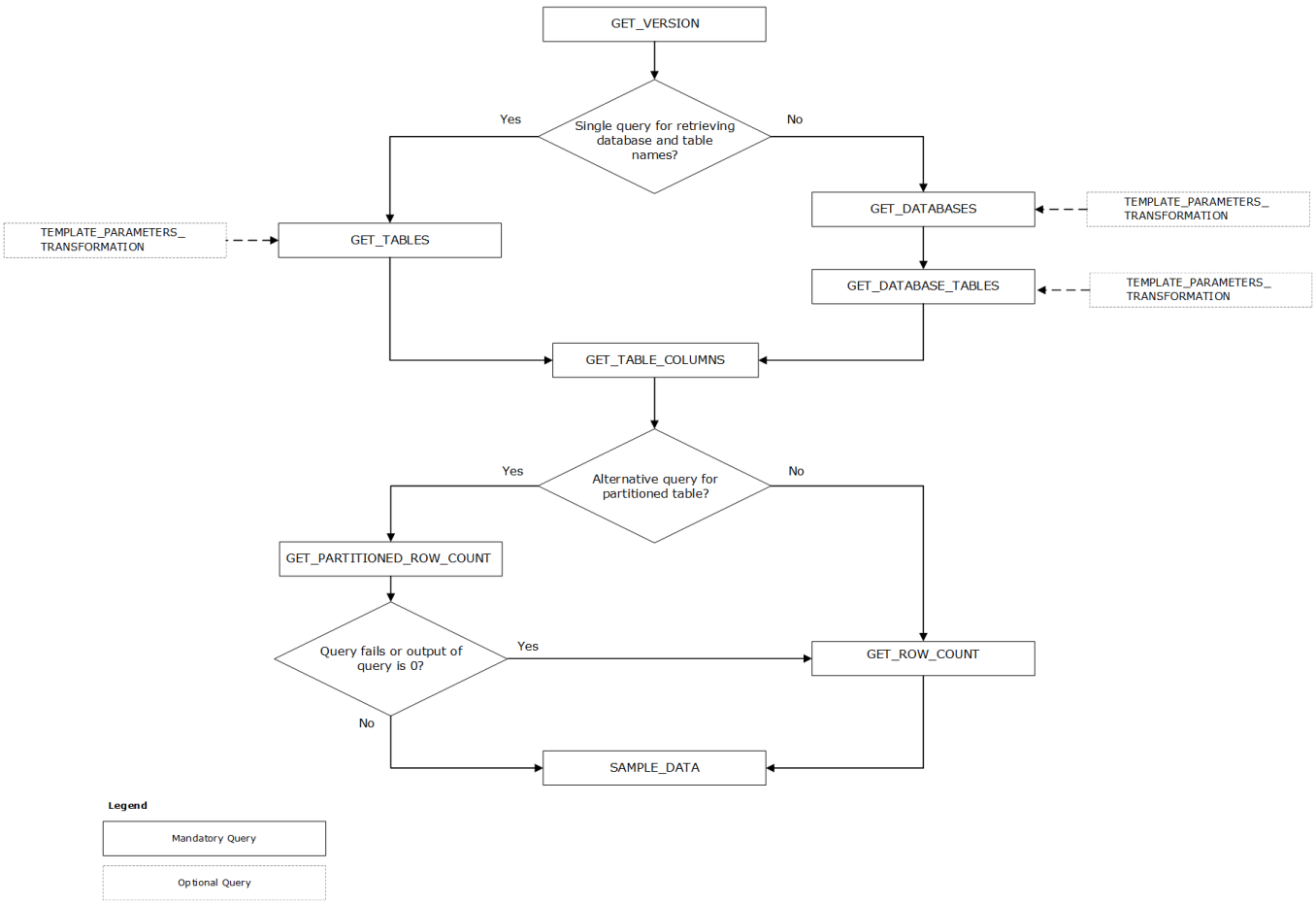


Figure 10-1: Sequence of Query Flow Execution

The context in which these queries are used is listed in the following table:

Table 10-1: Protegrity Discover - System Queries with Context

Query	Contextual meaning with Protegrity Discover
GET_VERSION	The GET_VERSION query is executed to check for a working database connection with version information as the output. The output of this query is logged into the application logs. The version information is checked against the specified encoding and decoding configuration settings before proceeding with the execution of other queries. **Important:** The execution of this query is a mandated prerequisite for every system.
GET_DATABASES GET_DATABASE_TABLES	The GET_DATABASES query provides a list of database names as an output, which is then used to query for the list of tables in each of the databases using the GET_DATABASE_TABLES query. For databases that do not support the use of a single query, such as the GET_DATABASESThese two queries are used as an alternative to the GET_TABLES query, for those databases that do not support the use of a single query to list all the tables across all database schemas. The GET_DATABASES query should not include any parameters. It returns the database name as its output. The GET_DATABASE_TABLES query uses the *%(database_name)s* parameter to fetch tables for each database instance. During the scanning process, the *%(database_name)s* parameter is replaced by the output of the GET_DATABASES query.
GET_TABLES	The GET_TABLES query is executed to retrieve a list of databases and tables for the sensitive data scan. This query can be used for databases that support the use of a single query to list all the tables across all database schemas. This is an alternative to using the GET_DATABASES and GET_DATABASE_TABLES queries together. **Important:** Do not include any parameters in this query. This query returns the database name and the table names in the output. **Note:** The query excludes any system internal tables, views, and other schema instances, which are not required to be scanned for the particular data source. You can modify the query to exclude additional instances or remove any default excluded instance for the scan.
GET_TABLE_COLUMNS	The GET_TABLE_COLUMNS query is executed to retrieve information about the structure of a specific database table, such as column names, data types, and data length. The output of the query is a list of columns from the specified database tables that need to be scanned. This query uses the *$(database_name)s* and *$(table_name)s* parameters. During the scanning process, the *$(database_name)s* and *$(table_name)s* parameters are replaced by the output of the GET_TABLES query, or by the output of the GET_DATABASES and GET_DATABASE_TABLES queries, respectively, depending on which queries you have used to retrieve the database and table names. **Sampling Pattern** The data type of the database columns is checked against the list of data types that are set using the **Data Sampling** > **Name** section on the **DATA TYPES** tab for the selected datastore. The GET_TABLE_COLUMNS query excludes any non-matching data types. *Example - Hive database*

Query	Contextual meaning with Protegrity Discover
	The following snippet displays the default data sampling pattern for the Hive database. The sampling pattern lists the regex pattern for the supported data types. `[` `"^((tiny|small|big)?int|decimal|numeric)(.*)?$",` `"^(((var)?char)\|string)(.*)?$",` `"^(double|float)?$",` `"^timestamp$",` `"^date$",` `"^interval$",` `"^boolean$"` `]` The column names that are retrieved using the GET_TABLE_COLUMNS query are filtered based on whether the data types of those columns match the regex pattern of the data types listed in the default sampling pattern for Hive. For example, if a database column contains data of type *tinyint* or *boolean*, then it is included for retrieving the sample data because it matches with the `"^((tiny\|small\|big)?int\|decimal\|numeric)(.*)?$"` and `"^boolean$"` regex patterns, respectively. *Example - Teradata database* The following snippet displays the default data sampling pattern for the Teradata database. `[` `"^B[FV]$",` `"^C[FV]$",` `"^D([A])?$",` `"^F$",` `"^I([128])?$",` `"^AT$",` `"^T[SZ]$",` `"^S[CZ]$",` `"^Y[RM]$",` `"^MO$",` `"^D[YHMS]$",` `"^H[RMS]$",` `"^M[IS]$",` `"^P[DMSTZ]$",` `"^N$",` `"^UT$"` `]` However, in Teradata, the sampling pattern lists the regex pattern for the supported data type codes. For example, the BF data type code represents the BYTE data type, while CV represents the VARCHAR data type. When you execute the GET_TABLE_COLUMNS query on the target system, the query lists the data types used. For more information about the data types associated with each data type code, refer to the Teradata documentation. The column names that are retrieved using the GET_TABLE_COLUMNS query are filtered based on whether the data types of those columns match the regex pattern of the data type codes listed in the default sampling pattern for Hive. For example, if a database column contains data of type *BYTE* or *VARCHAR*, then it is included for retrieving the sample data, because it matches with the `"^B[FV]$"` and `"^C[FV]$"` regex patterns respectively. **Replace Pattern** The replace pattern is a regex pattern that is specified in the **Data Sampling** > **Replace Pattern** section on the **Data Types** tab for a selected datastore. The replace pattern is applied on the column names that are retrieved using the GET_TABLE_COLUMNS query. The replace pattern enables you to specify delimiters for the column name identifier, based on the specific datastore. This is useful if the column names include spaces or special characters. For example, you can choose to add square brackets, single quotes, or double quotes as delimiters for the column names.

Query	Contextual meaning with Protegrity Discover
	For example, in Teradata, the default replace pattern for column names is specified as *"\1"*. If the name of the column that is retrieved using the GET_TABLE_COLUMNS query is *Test*, then the default replace pattern is applied on the column name. Therefore, in the SAMPLE_DATA query that is used to retrieve the sample data, the column name used is *"Test"*.
SAMPLE_DATA	The SAMPLE_DATA query is executed to retrieve the sample data from a targeted database table. The parameters used in this query as described as follows: • *%(columns)s* - to fetch the columns. During the scanning process, this parameter is replaced by the output of the GET_TABLE_COLUMNS query, after the list of columns have been filtered for matching data types. If the replace pattern has been provided, then it is applied on the filtered output. • *%(database_name)s.%(table_name)s* - to represent the targeted table. During the scanning process, these two parameters are replaced by either the output of the GET_TABLES query or outputs of the GET_DATABASES and GET_DATABASE_TABLES queries, depending on which queries you have chosen to retrieve the database and table names, respectively. • *%(sample_count)i* - to specify the sample number of records to fetch from the targeted table. During the scanning process, this parameter is replaced by the value of the *sample_size* configuration parameter that you have specified while creating a discover job. For more information about the *sample_size* configuration parameter, refer to the section *Scan Job Advanced Configuration Settings*. By default, the SAMPLE_DATA query retrieves the top records. For example, if you have specified the number of sample records to be retrieved as *1000*, then the SAMPLE_DATA query retrieves the top 1000 records from the selected database column. Therefore, it is recommended to use a randomization function in the query to ensure that random records are retrieved as part of the sample data. For example, the default SAMPLE_DATA query template for Oracle uses the **DBMS_RANDOM.VALUE** function to retrieve the sample data randomly, as shown in the following snippet: `[ "SELECT %(columns)s FROM (SELECT %(columns)s FROM \"%(database_name)s\".\"%(table_name)s\" T1 WHERE ROWNUM < %(sample_count)i * 10 ORDER BY DBMS_RANDOM.VALUE ) T1 WHERE ROWNUM < (%(sample_count)i + 1);" ]`
GET_ROW_COUNT	The GET_ROW_COUNT query is used to retrieve the total number of records from a table. The estimated number of sensitive data values at a coordinate is calculated using the total number of records. For example, the sensitive data value obtained from a sampled set is a direct indicator of the total sensitive data values at the coordinate. The GET_ROW_COUNT query utilizes the *%(database_name)s* and *%(table_name)s* parameters to obtain the row count for a database table. The following snippet shows the default query template used in the GET_ROW_COUNT query for Teradata: `[ "SEL COUNT(*) FROM \"%(database_name)s\".\"%(table_name)s\"" ]` When the sample data is expressed as a percentage, the GET_ROW_COUNT is used to calculate the *%(sample_count)i* parameter.
GET_PARTITIONED_ROW_COUNT	The GET_PARTITIONED_ROW_COUNT query is used as an alternative to the GET_ROW_COUNT query to retrieve the total number of records from a partitioned table. This query is used in cases where the database vendor supports an alternative method of querying a partitioned table, as compared to querying a non-partitioned table.

<table>
<tr><th>Query</th><th>Contextual meaning with Protegrity Discover</th></tr>
<tr><td></td><td>For example, in the case of using a Microsoft SQL Server, the sys.dm_db_partition_ stats dynamic management view can determine the total number of records available in a database table across partitions. The following snippet shows the default query template used in the GET_PARTITIONED_ROW_COUNT query for a Microsoft SQL Server:

<pre>[
    "USE [%(database_name)s];",
    "SELECT SUM(row_count) as row_count FROM sys.dm_db_partition_stats WHERE
object_id=OBJECT_ID ('%(table_name)s') AND (index_id=0 or index_id=1);"
]</pre>
The GET_PARTITIONED_ROW_COUNT query utilizes the %(database_name)s and %(table_name)s parameters to obtain the row count for a partitioned table.

Important: If the GET_PARTITIONED_ROW_COUNT query fails to execute, or the output of the GET_PARTITIONED_ROW_COUNT query is 0, then the GET_ROW_COUNT query is executed.</td></tr>
<tr><td>TEMPLATE_PARAMETERS_TRANSFORMATION</td><td>The TEMPLATE_PARAMETERS_TRANSFORMATION query specifies a regex pattern that replaces the parameter values with query format before execution. For example, some database systems represent table names in the format schemaname.tablename. However, when using the parameters with the query format, the schemaname is required to be separated from the tablename as [schemaname].[tablename].</td></tr>
</table>

Chapter 11

Protegrity Discover REST APIs

The Protegrity Discover REST APIs include the Scanner REST API, which is used to scan the input data for sensitive data.

11.1 Accessing Protegrity Discover using the REST APIs

The following authentication mechanisms can be used to access Protegrity Discover:

- Basic authentication with user name and password
- Client Certificate-based authentication
- Token-based authentication

For more information about accessing Protegrity Discover using these authentication mechanisms, refer to the section *Accessing REST API Resources* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

11.2 Accessing the Protegrity Discover REST API Documentation

This section describes how to access the Protegrity Discover REST API documentation.

➤ To access the Protegrity Discover REST API documentation:

1. On the Protegrity Discover Web UI, navigate to **Appliance** > **REST Api**.
   The **REST Api** screen appears, which displays the REST API Swagger documentation.

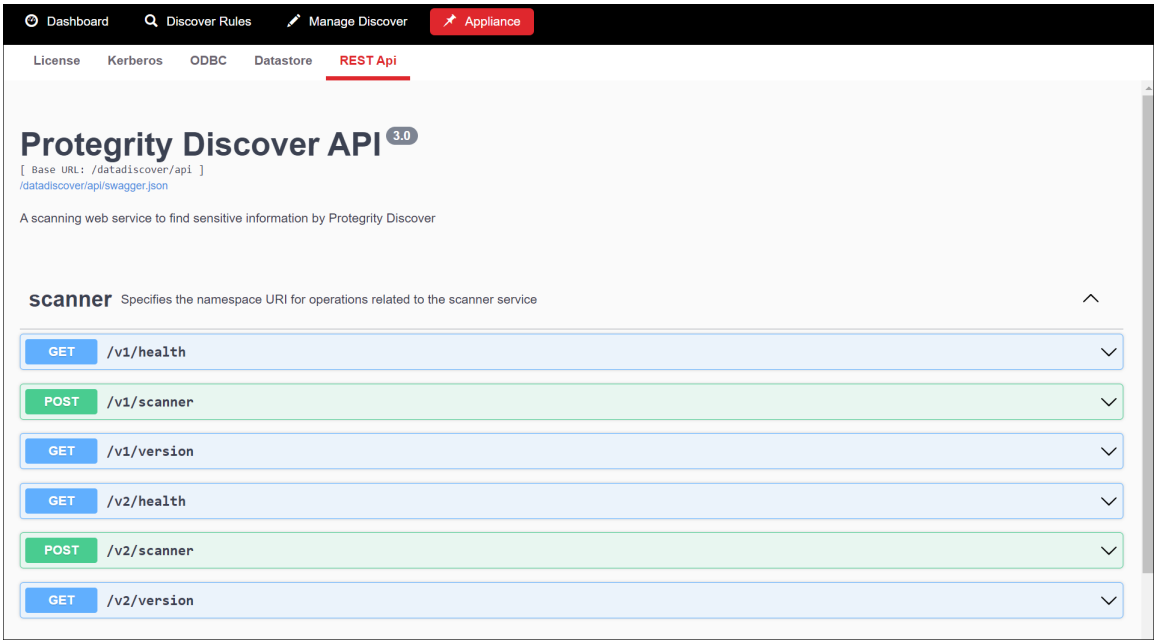


Figure 11-1: REST Api Screen

Important: Users with the Discover Admin and Discover Viewer permissions can access the Protegrity Discover REST API documentation.

For more information about the Protegrity Discover-specific permissions, refer to the section *Appendix G: Understanding Protegrity Discover-specific Permissions*.

Two versions of the Protegrity Discover appear. Both the versions perform identical functions. However, the two versions differ in the response body.

Table 11-1: Protegrity Discover REST API Versions

Version	Response Body	Output Sample
v1	The response body shows a list of classification records. The classification records are not grouped under the respective classifier.	<pre>[{ "classifier": "EMAIL", "classifier_name": "NLP Email Address", "end": 18, "score": 0.72, "start": 0 }, { "classifier": "EMAIL", "classifier_name": "NLP Email Address", "end": 30, "score": 1, "start": 19 }, { "classifier": "NAME", "classifier_name": "NLP Persons Name", "end": 41, "score": 0.92, "start": 34 }]</pre>

Version	Response Body	Output Sample
v2	The response body shows a list of classifiers. The classification records are grouped under the respective classifier.	<pre>{ "EMAIL": [{ "classifier_name": "NLP Email Address", "end": 18, "score": 1, "start": 0 }, { "classifier_name": "NLP Email Address", "end": 30, "score": 1, "start": 19 }], "NAME": [{ "classifier_name": "NLP Persons Name", "end": 43, "score": 0.92, "start": 36 }] }</pre>

Important: Protegrity recommends you to use v2 version of the Protegrity Discover REST APIs.

The v1 APIs have been included for backward compatibility with previous versions of Protegrity Discover.

You can also use the online Swagger documentation to generate cURL samples for the Protegrity Discover REST API.

For more information about generating cURL samples, refer to the section *Generating the Protegrity Discover REST API Samples*.

2. Alternatively, you can also perform the following steps to access the REST API Swagger documentation.
   a. Navigate to the following URL:

      https://{Protegrity Discover IP address}/datadiscover/api/doc

      Protegrity Discover IP address is the IP address of the machine where you have installed Protegrity Discover.

      You are prompted to enter the Protegrity Discover username and password.

   b. Enter the user name and password, and then click **OK** or **Sign in**, depending on the browser you are using to sign in to Protegrity Discover.

      The online Swagger documentation for the Protegrity Discover API appears.

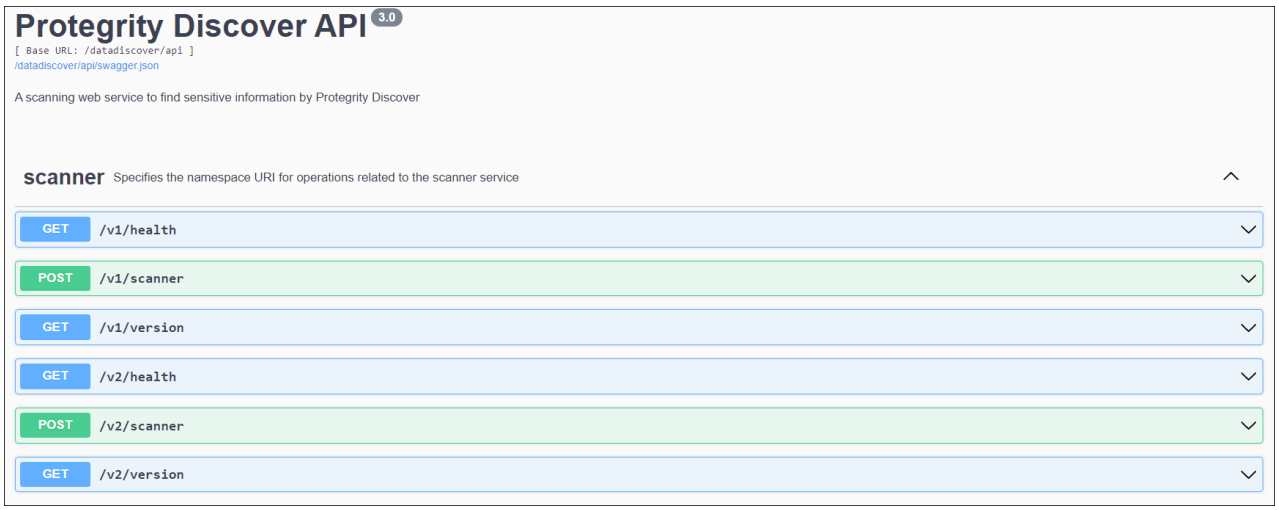


Figure 11-2: Online Swagger Documentation for the Protegrity Discover REST API

3. If you want to download the *swagger.json* file, which describes the API in JSON format, then click the */datadiscover/api/swagger.json* link below the **Base URL** content.

   If you want to directly download the *swagger.json* file, then you can navigate to the following URL:

   https://{Protegrity Discover IP address}/datadiscover/api/swagger.json

   In this case, you need to specify the Protegrity Discover user name and password. After you enter the Protegrity Discover user name and password, the *swagger.json* file is downloaded on your local machine.

11.3 Using the Protegrity Discover REST APIs

This section explains the usage of the Protegrity Discover APIs with some generic samples.

Table 11-2: REST API Samples

REST API	Section Reference
Scan Sensitive Data	*Scanning Sensitive Data*
Scan Sensitive Data within a File	*Scanning Sensitive Data in a File*

11.3.1 Scanning Sensitive Data

This section explains how you can scan sensitive data using the Protegrity Discover REST API.

Base URL

https://{Protegrity Discover IP address}/datadiscover/api/

In the base URL, the Protegrity Discover IP address specifies the IP address of the machine where you have installed the Protegrity Discover application.

Path

/v2/scanner

Method

POST

Sample Request 1

```
curl -k --user <username>:<password> -X POST "https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner" -H "Content-Type: text/plain" -d "data=hello@protegrity.com"
```

Sample Response 1

```
[{"classifier":"EMAIL","classifier_name":"NLP
Email Address","end":25,"esa_data_element":null,"score":1,"start":5},
{"classifier":"EMAIL","classifier_name":"Email
Address","col":"0","col_idx":0,"esa_data_element":null,"score":1.0}]
```

The response indicates that the Protegrity Discover has classified the sensitive data as an email, with a confidence score of 1. The *start_index* specifies the starting point of the location where the sensitive data is present. The *end_index* specifies the end point of the location where the sensitive data is present.

Note: The classifier used to identify the data depends on the content-type of the raw data. If the content-type of the data is text, then the data is considered as unstructured and the NLP classifiers are used to identify the sensitive data.

If you specify the content-type of the raw data as JSON or XML, then the data is considered as semi-structured, and accordingly non-NLP classifiers are used to identify the sensitive data.

Important: If you want the results to appear on the **Classifications**, **Data Sources**, and **Scan Results** screens on the Protegrity Discover Web UI, then you must specify the scan-coordinate in the header of the request sent to the REST API. If you do not specify the scan coordinate, then Protegrity Discover scans the requested data and displays the result only on the console or the UI of the REST API client that was used to send the request.

If you want to specify the scan-coordinate, then you must specify one of the following protocol values as valid prefixes to the scan-coordinate:

- *dbms* - Indicates that the datastore is a database.
- *fs* - Indicates that the datastore is a filestore.
- *restful* - Indicates that the data is part of a restful application. For example, the sensitive data that needs to be scanned is part of a chat application.

If you are using the *dbms* or *fs* protocols, then you also need to specify the sub-protocols as valid prefixes to the scan-coordinate. For example:

- *dbms:teradata*
- *dbms:oracle*
- *fs:sharepoint*
- *fs:nfs*

Important: If you do not specify any one of the valid prefixes to the scan-coordinate, then the REST API returns an `Issue with co-ordinate` error in the response.

Sample Request 2

```
curl -k --user <username>:<password> -X POST "https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner"
-H "Scan-Coordinate:fs:sharepoint://user1@<IP address of the datastore>/test/test1" -H "Content-Type: text/plain" -d
"data=hello@protegrity.com"
```

Sample Response 2

```
[{"classifier":"EMAIL","classifier_name":"NLP
Email Address","end":25,"esa_data_element":null,"score":1,"start":5},
{"classifier":"EMAIL","classifier_name":"Email
Address","col":"0","col_idx":0,"esa_data_element":null,"score":1.0}]
```

As the scan-coordinate has been specified in the request header, the results also appear on the **Classifications**, **Data Sources**, and **Scan Results** screens on the Protegrity Discover Web UI.

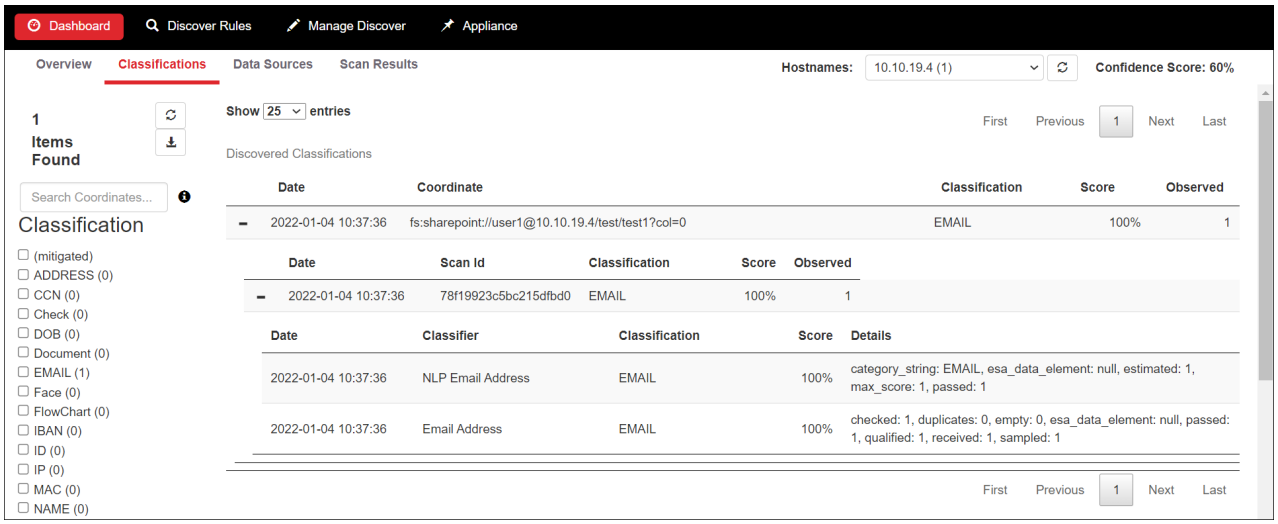


Figure 11-3: Classifications Screen

For more information about the **Classifications** screen, refer to the section *Data Classification and Record Findings*.

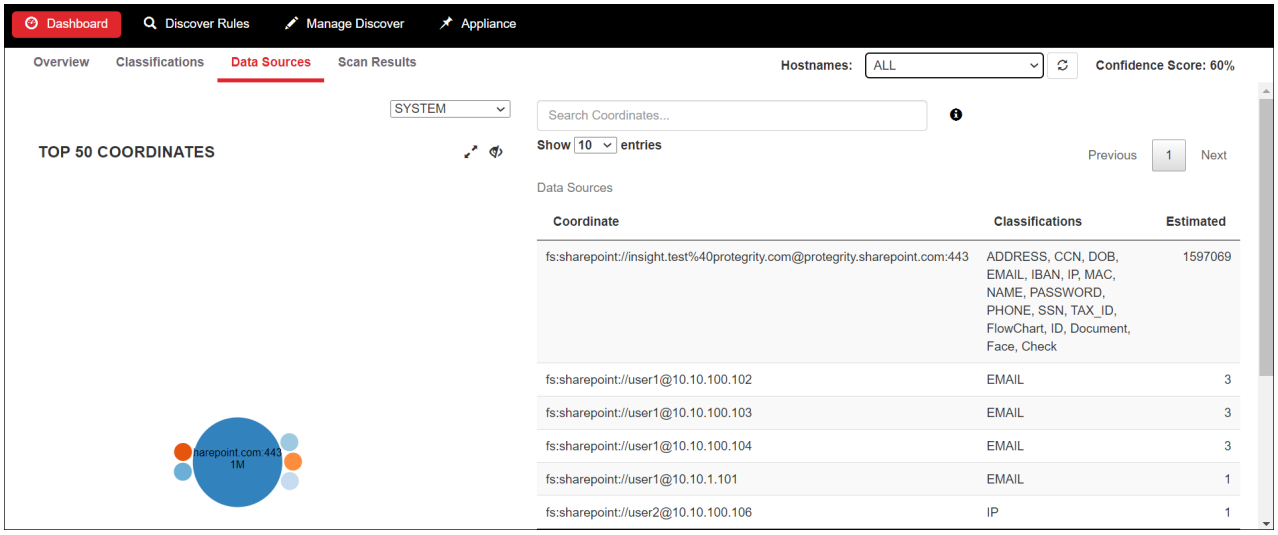


Figure 11-4: Data Sources Screen

For more information about the **Data Sources** screen, refer to the section *Data Sources*.

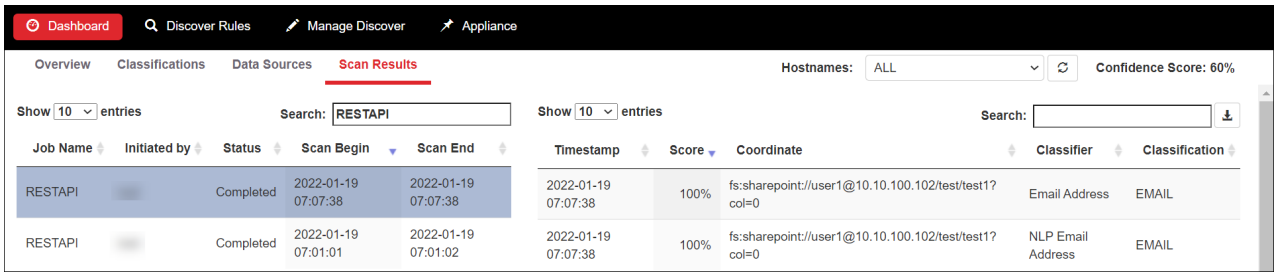


Figure 11-5: Scan Results Screen

For more information about the **Scan Results** screen, refer to the section *Scan Results*.

By default, the name of the job appears as **RESTAPI** in the **Scan Results** screen. If you want to display a custom job name in the **Scan Results** screen, then you can specify the job name as the value of the *Job-Name* header in the REST API request, as shown in the following code snippet.

```
curl -k --user <username>:<password> -X POST "https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner" -H "Scan-Coordinate:fs:sharepoint://user1@<IP address of the datastore>/test/test1" -H "Content-Type: text/plain" -H "Job-Name: REST_scan" -d "data=hello@protegrity.com"
```

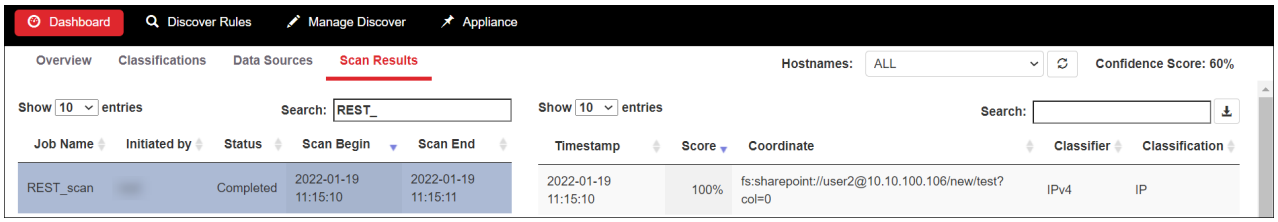


Figure 11-6: Scan Results Screen

11.3.2 Scanning Sensitive Data in a File

This section explains how you can scan sensitive data in a file using the Protegrity Discover REST API.

Base URL

https://{Protegrity Discover IP address}/datadiscover/api

In the base URL, the Protegrity Discover IP address specifies the IP address of the machine where you have installed the Protegrity Discover application.

Path

/v2/scanner

Method

POST

Sample Request

```
curl -k -u <username>:<password> -X POST https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner -H "Scan-Coordinate:restful:RESTAPI//<IP address of the Chat server>/support_conv" -F "file=@Email.csv"
```

> **Important:** Ensure that you specify the *@* symbol before the file name.

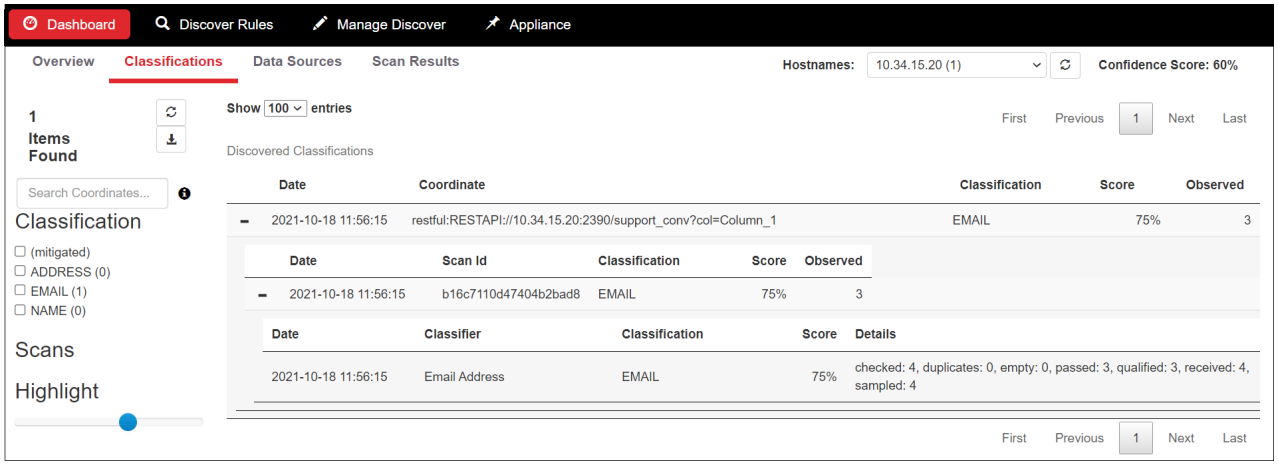


Figure 11-7: Classifications Screen

For more information about the file types supported by the Protegrity Discover REST API, refer to the section *Appendix D: Supported File Formats*.

You can also choose to specify the *Content-Type* header in the request to explicitly identify the file type. However, it is optional to specify the *Content-Type* header, except in case of the Apache Parquet file type. If you do not specify the *Content-Type* header, then the Protegrity Discover REST API automatically tries to detect the file type.

For example:

```
curl -k -u <username>:<password> -X POST https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner -H "Content-Type: text/csv" -H "Scan-Coordinate:restful:RESTAPI//<IP address of the Chat server>/support_conv" -F "file=@Email.csv"
```

The following table lists the values that you can specify for the *Content-Type* header for each supported file type.

Table 11-3: Values for the Content-Type Header

File Type	Content-Type Header
Text	• text/plain • application/text
CSV	• text/csv • application/csv • application/vnd.ms-excel • text/x-csv
HTML	• text/html • application/html
XML	• text/xml • application/xml
Image files	• image/jpeg • image/jpg • image/bmp • image/x-ms-bmp • image/gif • image/jfif • image/png
JSON	• application/json • application/ld+json • text/json
PDF	• application/pdf • text/pdf
Word files	• application/msword • application/vnd.openxmlformats-officedocument.wordprocessingml.document
Excel files	• application/vnd.ms-excel • application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
Apache Parquet	• application/parquet

File Type	Content-Type Header
	• text/parquet
SAS files	• application/sas7bdat • application/sd7
Apache Avro	• application/avro • avro/binary

Important: If you want to scan an Apache Parquet file, then you must include the *Content-Type* header in the request, and specify the value of the header as *application/parquet* or *text/parquet*.

For example, you can send the following sample request:

```
curl --insecure -u <username>:<password> -XPOST https://{Protegrity Discover
IP address}/datadiscover/api/v2/scanner -F "file=@test_file.parquet;type=application/
parquet"
```

Important: If you want to scan an Apache Avro file, then you must include the *Content-Type* header in the request, and specify the value of the header as *application/avro* or *avro/binary*.

For example, you can send the following sample request:

```
curl --insecure -u <username>:<password> -XPOST https://{Protegrity Discover IP
address}/datadiscover/api/v2/scanner -F "file=@test_file.avro;type=application/avro"
```

The scan-coordinate parameter is optional. However, if you want the scan results to appear on the **Classifications**, **Data Sources**, and **Scan Results** screens on the Protegrity Discover Web UI, then you must specify the scan-coordinate.

If you are specifying a scan-coordinate, then you must specify a valid prefix.

For more information about specifying the valid prefix, refer to the section *Scanning Sensitive Data*.

Sample Response

```
[{"classifier": "EMAIL", "score": 0.17, "start_index": "col=Column_1",
"end_index": ""}]
```

The response indicates that the Protegrity Discover has identified that the *Email.csv* file contains email addresses in the first column, with a confidence score of 0.17.

If you have specified the scan-coordinate in the request header, then the results also appear on the **Classifications**, **Data Sources**, and **Scan Results** screens on the Protegrity Discover Web UI.

11.4 Debugging the Protegrity Discover REST APIs

This section describes how to debug the Protegrity Discover REST APIs.

To debug the Protegrity Discover REST APIs:

1. On the Protegrity Discover Web UI, navigate to **Manage Discover** > **REST API Log**.

The **REST API Log** screen appears.

For more information about the **REST API Log** screen, refer to the section *Viewing REST API Logs*.

2. Click the icon to set the system log levels.
   The **Set RestAPI Log Level** screen appears.

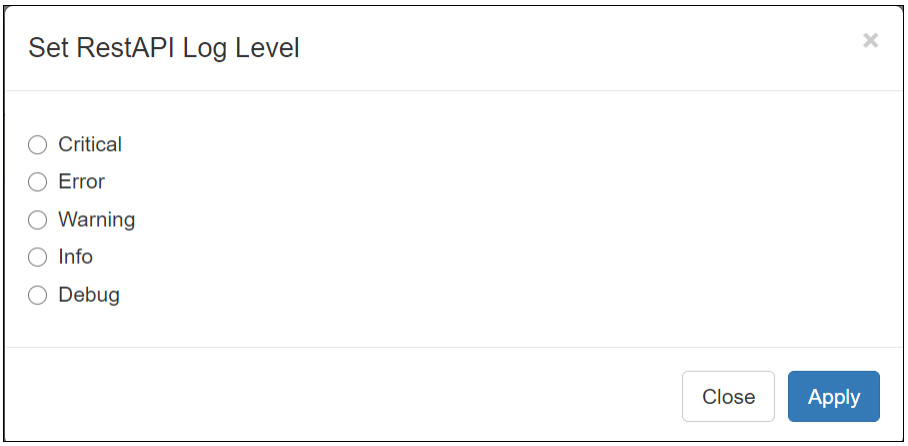


Figure 11-8: Set REST API Log Level

3. Set the log level to **Debug**, and then click **Apply**.
4. Send a REST API request for scanning data or a file, and ensure that you include the *-i* parameter in the cURL command.
   The *-i* parameter includes the HTTP response headers in the output, helping you to debug the request.

   For example:

```
curl -k -i --user <username>:<password> -X POST "https://{Protegrity Discover IP address}/datadiscover/api/v2/scanner" -H "Scan-Coordinate:fs:sharepoint:user1@<IP address of the datastore>/test/test1" -d "data=hello@protegrity.com"
```

   You can view the debug logs on the **REST API Log** screen.

11.5 Generating the Protegrity Discover REST API Samples

You can generate cURL samples using the online Swagger documentation.

Perform the following steps to generate samples using the online Swagger documentation.

1. Access the online Swagger documentation for the Protegrity Discover REST API.
   For more information about accessing the online Swagger documentation, refer to the section *Accessing the Protegrity Discover REST API Documentation*.

2. Click **Try it out**.
3. Enter the parameters for the API request.
4. Select the content type from the **Parameter content type list**.

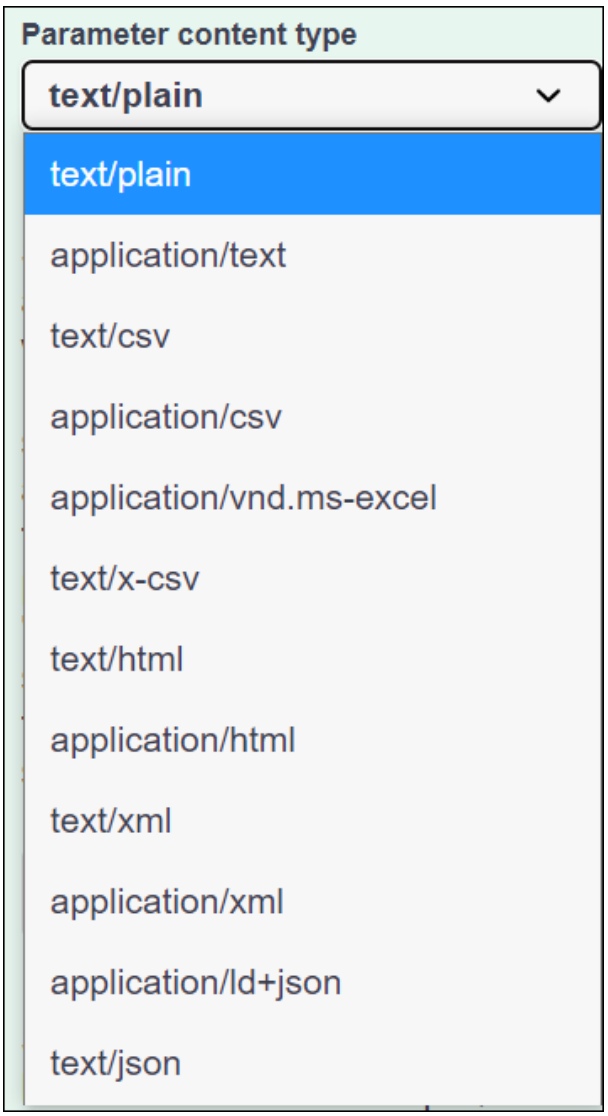


Figure 11-9: Parameter Content Type List

5. Click **Execute**.

   The generated cURL command and the URL for the request appears in the *Responses* section.

   The following snippet shows a sample cURL command.

```
curl -X 'POST' \
  'https://10.37.3.93/datadiscover/api/v2/scanner' \
  -H 'accept: application/json' \
  -H 'Content-Type: text/plain' \
  -d 'a.b@protegrity.com'
```

> **Important:** By default, a \ (backslash) character is added to the command. If you want to copy the cURL command and run it on a console, then ensure that you delete the backslash character.

   The following snippet shows the sample response.

```
{
  "EMAIL": [
    {
      "classifier_name": "NLP Email Address",
```

```
      "end": 18,
      "score": 1,
      "start": 0
    },
    {
      "classifier_name": "Email Address",
      "col": "0",
      "col_idx": 0,
      "score": 1
    }
  ]
}
```

Chapter 12

Using Webhooks

Protegrity Discover enables a webhook functionality that is used to send the scan information in a POST response from the Protegrity Discover machine to an external URL every time a discover job identifies sensitive data.

To use the webhook functionality, you need to set up an API on an external server, which can receive the notifications from Protegrity Discover. You then need to provide the endpoint URL of the external API as an input to the webhook functionality. When a discover job identifies sensitive data in the scan results, a POST request is sent to the specified URL. The POST request contains the following response details related to the scan results:

- *scan_id* - ID of the scan result
- *coordinate* - Location of the sensitive data
- *classification* - The classification type that reported the sensitive data values
- *confidence_score* - Score that determines the confidence or severity level in the data classification findings. The *confidence_score* value is expressed in decimals.
- *observed* - Observed count of unique sensitive data values examined from the sampled data
- *est_total* - Estimated count of unique sensitive data values examined from the sampled data
- *esa_data_element* - The ESA data element associated with the classification type. This parameter is applicable only if you have associated an ESA data element with the classifier, in the **General** tab of the classifier.

  For more information about associating an ESA data element to a classifier, refer to *step 2* of the section *Updating the General Tab*.

  For more information about retrieving data elements from the ESA, refer to the section *Retrieving ESA Data Elements*.

You can then choose to use the response information based on your requirements.

For example, you can create a custom service that generates a support ticket every time a notification is received from Protegrity Discover.

12 Configuring a Webhook

You can configure the webhook in the following ways:

- *Global configuration* - You can configure a single webhook for all the discover jobs. In this case, a POST request will be sent to the specified URL whenever any discover job identifies a sensitive data in the scan results. You can configure the webhook globally by modifying the *sureloc.json* file.

  For more information about configuring a webhook globally, refer to the section *Configuring a Webhook Globally*.

- *Local configuration* - You can configure a separate webhook for each discover job. In this case, you need to specify the *webhook* configuration setting in the *Advanced Configuration* dialog box for each discover job.

For more information about the *webhook* configuration setting, refer to the section *Scan Job Advanced Configuration Settings*.

For more information about creating a discover job, refer to the section *Creating a Discover Job*.

You can specify the global and local configurations simultaneously. In this case, both the endpoints mentioned in the global and local configurations will receive a POST response from the Protegrity Discover machine.

12.1 Configuring a Webhook Globally

The following section describes the steps for configuring a global webhook, which is applicable to all the discover jobs.

Before you begin

If you want to mutually authenticate the Protegrity Discover system with the endpoint server that you want to create for accepting the notifications from the webhook, then ensure that you create a self-signed or a trusted client certificate for the endpoint server. You need to copy this certificate to the machine on which Protegrity Discover has been installed. This certificate is used to authenticate the Protegrity Discover application, when sends the POST request to the endpoint server, whenever sensitive data is found in the scan results.

Important: Protegrity recommends that you use a certificate-based authentication method for authenticating the Protegrity Discover application to the endpoint server.

To configure a webhook globally:

1. If you want to enable mutual TLS authentication between Protegrity Discover and the endpoint server, then perform the following steps to upload the client certificate to the certificate repository.
   a. Login to the Appliance Web UI and navigate to **Settings** > **Network Settings** > **Certificate Repository**.
      The Certificate Repository screen appears.
   b. Click **Upload new files** to upload the client certificate of the endpoint server to the certificate repository.
      When Protegrity Discover sends a POST request to the URL of the endpoint server with the details of the scan results, the endpoint server uses the uploaded certificate to authenticate Protegrity Discover.

      For more information about uploading a certificate, refer to the section *Uploading Certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*.

2. Perform the following steps to modify the *sureloc.json* file.
   a. On the Appliance Web UI, navigate to **Settings** > **System** > **Files**.
      The List of Product Files screen appears.
   b. In the **Discover - Settings** section, edit the *sureloc.json* file.
   c. Edit the following code block in the *sureloc.json* file.

```
"scanner": {
    "webhook": {
      "url": "<URL of the endpoint server>",
      "timeout": 0.1,
      "verify":"<Value for verifying the client certificate>",
      "cert": <ID of the client certificate used by the endpoint server to authenticate
```

```
Protegrity Discover>"
    }
```

Important: Before saving the *sureloc.json* file, ensure that you validate the JSON syntax using an online validator.

If you save the *sureloc.json* file with an invalid JSON syntax, then you will not be able to restart the Protegrity Discover service. As a result, you will not be able to login to the Protegrity Discover Web UI.

In case you are unable to login to the Protegrity Discover Web UI, then you can directly access the Appliance Web UI by navigating to the *https://{Management IP}/index.html* URL, and then edit the *sureloc.json* file to fix the invalid JSON syntax.

d. Modify the following fields in the webhook configuration.

Field	Description
url	Specify the URL of the endpoint server, which is used to receive the notifications sent by the Protegrity Discover webhook. For example, specify the value of the URL as *https:///<host:port>/path* The following parameters are specified in the URL: • *host* - IP address or host name of the machine where the endpoint server has been installed. • *port* - Port number that is used to receive the notifications sent by the Protegrity Discover webhook. In this example, *path* denotes a the endpoint of an API that is set to receive the notifications from the Protegrity Discover webhook. **Note:** Ensure that you specify the accurate end points for the API on the endpoint server. If you specify the accurate IP address and port number, but incorrect endpoints, then only an empty POST request, without any response, is sent to the endpoint URL. By default, the value of this parameter is set to *null*.
timeout	Specifies the time for which Protegrity Discover tries to communicate with the endpoint server, after which the connection times out. By default, this value is set to *0.1*. The unit of the *timeout* parameter is seconds.
cert	Specify the ID of the client certificate, which is mentioned on the **Certificate Repository** screen, used by the endpoint server to authenticate Protegrity Discover. If the endpoint server is not using a certificate-based authentication method, then you need to specify the value of the *cert* parameter as *null*. By default, the value of this parameter is set to *null*.

Field	Description
verify	Specify whether the self-signed or trusted client certificate that you have uploaded to establish secure communication between the endpoint server and Protegrity Discover, needs to be verified. You can either specify a boolean value or a string value for this parameter. If you specify a boolean value, then this parameter determines whether the client certificate is verified by the endpoint server. In this case, you can specify one of the following values for this parameter: • *true* - The client certificate is verified. • *false* - The client certificate is not verified. You can also specify the ID of the CA certificate. Protegrity Discover uses this value to internally identify the path where the Certificate Authority bundle is stored. The Certificate Authority bundle is used to verify the client certificate. For more information about the *verify* parameter, refer to the section *SSL Cert Verification* in the Python *Requests* module. By default, the value of this parameter is set to *null*.

After you have modified the *sureloc.json* file, you need to login to the Appliance Web UI, navigate to the **System** > **Services** screen, and restart the Discover service.

For more information about restarting the Discover service, refer to the *step 3* of the section *To manage system services*.

If you are unable to restart the Discover service from the **System** > **Services** screen, then you need to login to the Protegrity Discover CLI Manager and restart the Discover service.

3. If you are unable to restart the Discover service from the **System** > **Services** screen, then perform the following steps to restart the Discover service.
   a. Login to the Protegrity Discover CLI Manager.
      For more information about logging in to the Protegrity Discover CLI Manager, refer to the section *Access CLI Manager* in the *Protegrity Appliances Overview Guide 9.2.0.0*.
   b. Navigate to **Administration** > **Services**.
      The **Service Management** dialog box appears.

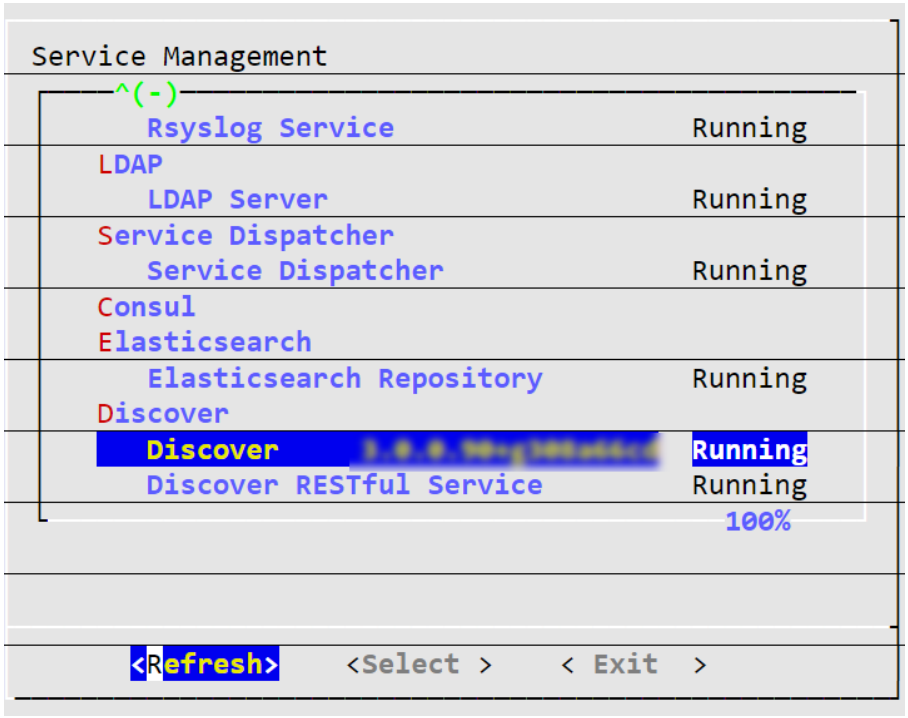


Figure 12-1: Service Management

c. In the **Discover** section, select the *Discover <version>* service, and then click **Select**.
The details of the Discover service appear.

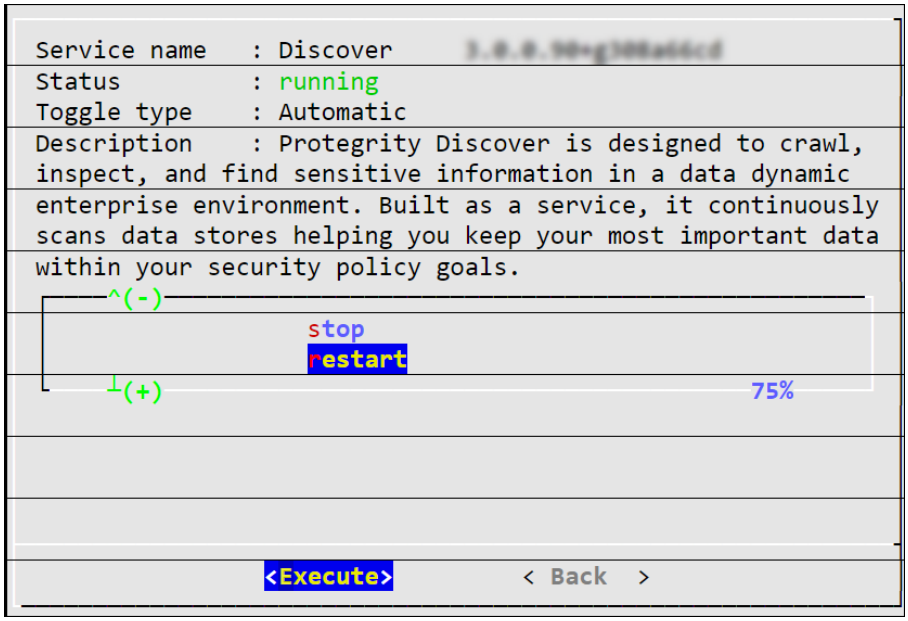


Figure 12-2: Protegrity Discover Service

d. Select **restart**, and then click **Execute**.
The Discover service restarts.

For more information about restarting the Discover service, refer to the section *Start and Stop Services* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

Chapter 13

Appendix A: Protegrity Discover-specific Term Definitions

Boost/Weight

A boost or weight, discussed in the context of a classifier, is the probability of a keyword getting qualified for the scan, as per the defined regular expression pattern.

Classification

The grouping of data that is done to represent a particular type of sensitive data after evaluation by the classifiers. For example, a collection of 16 digit values evaluated to be a potential credit card number is classified to be a Credit Card Number (CCN).

Classifier

The classifier defines the data identification rules to classify sensitive data from the sampled data.

Confidence Score

It is the score that determines the confidence or severity level in the data classification findings.

Coordinate

A coordinate represents the location of sensitive data, which can be any system, database, schema, table, column, or file path.

Datastore/ Node

The datastore base or nodes define the different systems that are supported for a data discovery scan.

Job

A task that you can create to scan a particular coordinate.

Reference/Referential Data

A pre-populated list of names and addresses containing the data dictionaries, such as, common names, city, postal codes, or states. By default, the list is pre-populated with Protegrity Discover to support commonly used data dictionaries.

Chapter 14

Appendix B: Usage of Regular Expressions

This section explains the usage of regex patterns with examples to help you define custom patterns, as per your data discovery search requirements. For any data discovery solution, the usage of regular expressions, which is referred to as *regex*, determines the elements used to search for and identify sensitive data.

Regular expressions are used for pattern matching in which a sequence of characters are used to define a search pattern. Using this pattern matching, you can define the rules required to match a *string* or *text*. In the context of data discovery requirements, you can define regex patterns to match the sensitive data that you must discover. The regex patterns can also identify keywords or metadata information. For example, a regex pattern for *Name* can look for possible *first name* and *last name* values if it matches the pattern '[a-zA-Z'`]{2,32}'.

The regex patterns include string literals or meta-characters. The following table explains the meaning of some of the commonly-used patterns.

Table 14-1: Protegrity Discover - Regex Patterns with Description

Regex Type	Character/String	Description	Example
String literal	abc	Matches any sequence with 'abc' string as a part of it.	www.abc.com
Character	.	Matches any character.	The expression ssn. will match 'ssndetails' but not 'ssn'
	[]	Matches anything inside the square brackets. The addition of ^ character, which is called as negated character class, means exception from the search.	The expression ssn [ab^c] will match 'ssna' and 'ssnb', but will not match 'ssnc'.
Anchors	^	Matches the specified characters at the beginning of a string.	The expression ^ssn will match 'ssndetails' but not 'detailsssn'.
	$	Matches the specified characters at the end of a string.	The expression ssn$ will match 'detailsssn' but not 'ssndetails'
Quantifiers	?	Matches the preceding element or element that follows this expression for one or zero time.	The expression behaviou?r will match both 'behavior' or 'behaviour'
	?i	Case insensitive search to match the preceding element or element that follows for one or zero time.	The expression (?i)(telephone|mobile) will match 'telephone' or 'mobile'
	*	Matches the preceding element for zero or more times.	The expression nea*t will match 'net', 'neat', 'neet', and so on
	+	Matches the preceding element for one or more times.	The expression se+t will match 'set' and 'seet' but not 'seat'
	|	Selects either the string before or after the | character.	The expression ssn | ccn will match either 'ssn' or 'ccn'
	{}	An expression {**x**} that checks if the element that precedes matches **x** times or not.	The expression s{3} will match 'sss', 'ssss', or 'sssn' but not 'ssn'
Escape Sequence	\	The element that follows the \ character is used as a literal. **Note:** Some special meta-characters beginning with \ are not escape sequences. The commonly	The expression schema\.ssncolumn will match 'schema.ssncolumn'

Regex Type	Character/String	Description	Example
		used special meta-characters are listed as follows in this table.	
Special meta-characters	\s	Matches any whitespace character, such as a space, tab, or a line break.	The expression Social\sSecurity\sNumber will match 'Social Security Number' but not 'Social<Security>Number'
	\S	Matches any non-whitespace character.	The expression \SSocialSecurityNumber will match 'ASocialSecurityNumber' or '1SocialSecurityNumber'
	\w	Matches any alphanumeric character.	The expression \w\w\w will match 'ssn' or 's1n'
	\W	Matches any non-alphanumeric character.	The expression ssn\W will match 'ssn!' or 'ssn?'
	\d	Matches any digit.	The expression ssn\d\d will match 'ssn12' or 'ssn56'
	\D	Matches any non-digit character.	The expression ssn\D\D will match 'ssnno' or 'ssndetails'
	\b	Matches a word boundary.	The expression \bssn will match 'ssndetails' but not 'thessndetails'
	\B	Matches a non-word boundary.	The expression \Bssn will match 'thessndetails' but not 'ssndetails'
Blocking and capturing	()	Used in conjunction with the | character to make a choice between elements.	The expression pr(a|e)y will match 'pray' or 'prey'
		In addition, \1 recalls the first sub-expression and \2 recalls the second sub-expression.	The expression [0-9]([-])[0-9]\1[0-9] will match 7-5-2 and 2-8-1 but not 13-2, 132, or 1-32

The following table lists some of the regex patterns with examples.

Table 14-2: Protegrity Discover - Classifier Regex Patterns with Examples

Classifier type	Regex Pattern	Example match
Phone Numbers	(?i)(^|[^a-z])(telephone|mobilephone)([^a-z]|$	telephone mobilephone
Password	(?i)pass[\\W_\\s]*?(word|code)	password passcode
Date of birth	(?i)d(ate)?[\\s_-]?o(f)?[\\s_-]?b(irth)?	DOB dob dateofbirth
Social Security Number	(?i)s(ocial)?s(ecurity)?n(umber)?	ssn socialsecuritynumber SSN
File Path	\\[^\\]+$	\\fs1\shared

Chapter 15

Appendix C: Scan Job Advanced Configuration Settings

This section describes the advanced configuration settings that you can set while creating or editing a job.

Each configuration setting is a key-value pair setting in JSON format. You can separate multiple key-value pairs using a comma separator. You can override the default values of the configuration settings from the **Config** field, which is visible if you click the **Advanced Settings** link on the *Add Discover Job* or *Edit Discover Job* dialog box.

The configuration settings that are displayed depend on the selected datastore.

The following table describes the advanced configuration settings and lists the datastores for which each setting is applicable. The table also lists the default values for each configuration setting.

Table 15-1: Protegrity Discover - Advanced Configuration Settings

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
port	Specify the port number for accessing the datastore.	`{ "port":<value> }`	• AWS S3 • Azure Storage (Blob) • EXAsol • Hadoop File System (HDFS) • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Network File System (NFS) • Oracle Database • PostgreSQL • SharePoint • Teradata	The default values change depending on the targeted system: • AWS S3 - ***443*** • Azure Storage (Blob) - ***443*** • EXAsol - ***8563*** • WebHDFS service for Hadoop File System (HDFS) - ***50070*** • Hive - ***10000*** • IBM DB2/UDB - ***50000*** • IBM DB2/zOS - ***5030*** • Microsoft SQL Server - ***1433*** • MySQL - ***3306*** • Network File System (NFS) - ***2049*** • Oracle Database - ***1521*** • PostgreSQL - ***5432*** • SharePoint - ***443*** • Teradata - ***1025***

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
			• Windows Share (CIFS)	• Windows Share (CIFS) - ***445*** **Note:** If you want to use the HttpFS service to access data from HDFS, then you must specify the required port number. The default port number used for the HttpFS service is *14000*.
protocol	Specify the protocol used to communicate with the datastore over the Internet. You can specify one of the following values: • https • http	`{ "protocol":"<value>" }`	• AWS S3 • Hadoop File System (HDFS)	***https***
max_file_size	Specify the maximum size of the files above which Protegrity Discover does not scan the file. If you want to scan a file whose size is more than the value defined in the *max_file_size* setting, then you must modify the value. However, if the file type is *CSV*, *TXT*, *Apache Avro*, or *Apache Parquet*, then Protegrity Discover partially scans the file even if its size exceeds the value specified in the *max_file_size* setting. In this case, the number of bytes that Protegrity Discover scans from the top of the file equals the value specified in the *max_file_size* setting. Protegrity Discover does not scan the remaining bytes in the file. **Note:** In SharePoint, Protegrity Discover scans only those *CSV*, *TXT*, *Apache Avro*, and *Apache Parquet* files whose maximum size does not exceed the value specified in the *max_file_size* setting.	`{ "max_file_size":<value> }`	• AWS S3 • Azure Storage (Blob) • Hadoop File System (HDFS) • SharePoint • Network File System (NFS) • Windows Share (CIFS)	***5242880***. The unit is in bytes.
no_cache	Specify whether Protegrity Discover should scan the files, irrespective of any change in the file. You can set one of the following values: • *1* - Protegrity Discover always scans the files even if the files have not been modified. • *0* - Protegrity Discover caches the files and checks the time stamp on these files. If the files have not changed, then Protegrity Discover skips these files while performing a scanning operation. It scans only those files that have been modified.	`{ "no_cache":<value> }`	• AWS S3 • Azure Storage (Blob) • Hadoop File System (HDFS) • SharePoint • Network File System (NFS)	***0***

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
			• Windows Share (CIFS)	
ocr_for_pdf	Specify whether Protegrity Discover should perform an Optical Character Recognition (OCR) procedure on the scanned text that is part of a PDF document. **Note:** When you scan text, it is converted to an image. If such an image is part of a PDF, then Protegrity Discover needs to perform an OCR procedure on the PDF to identify the text from the image. Conversely, Protegrity Discover does not require to perform an OCR procedure on a PDF to identify text that is not scanned. You can set one of the following values: • *1* - Protegrity Discover always performs an OCR procedure on PDFs to identify scanned text. **Note:** Due to the OCR procedure, the time taken to scan a PDF that contains scanned text is significantly higher than the time taken to scan a PDF that does not contain any scanned text. As a result, the discover job for scanning data from a PDF that contains scanned text can take a longer duration to complete. • *0* - Protegrity Discover does not perform an OCR procedure on PDFs to identify scanned text. **Important:** If a PDF contains both scanned text and text that is not scanned, and you have set *0* as the value of the *ocr_for_pdf* configuration setting, then Protegrity Discover will only identify the text that is not scanned.	`{ "ocr_for_pdf": <value> }`	• AWS S3 • Azure Storage (Blob) • Hadoop File System (HDFS) • SharePoint • Network File System (NFS) • Windows Share (CIFS)	***0***
schema	Specify the Kerberos schema.	`{ "schema":"<value>" }`	Hive	***default***
krbhostfqdn	Specify the host name of the Hive server as the fully qualified name for Kerberos authentication. The default value *_HOST* is used to indicate the host name of the Hive server.	`{ "krbhostfqdn": "<value>" }`	Hive	***_HOST***
aws_config	Specify the following AWS configuration parameters: • *connect_timeout* - Specify the time for which a Protegrity Discover scan job will try to connect with an AWS S3 datastore, after which the scan is aborted. • *read_timeout* - Specify the time for which a Protegrity Discover scan job will try to read the data	`{ "aws_config": { "connect_timeout":<value>`	AWS S3	The default values for the AWS configuration parameters are: • *connect_timeout* - ***60***. The unit is in seconds.

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	from an AWS S3 bucket, after which the scan is aborted. • *retries* - Specify the value of the *max_attempts* parameter. It indicates the maximum number of attempts a Protegrity Discover scan job will try to connect with an AWS S3 datastore, after which the scan is aborted.	`"read_timeout" :<value> "retries": { "max_attempts" :<value> } }`		• *read_timeout* - **`60`**. The unit is in seconds. • *max_attempts* - **`2`**
options	Specify the options that are passed as arguments to the *mount* command using the *-o* command-line option. You must specify the arguments as a comma-separated list. **Note:** For more information on the options that can be passed as arguments to the *mount* command, refer to the manual pages (man pages) for the *mount* command and NFS file format for Linux.	`{ "options":"<value_1>, <value_2>" }`	Network File System (NFS)	By default, the following values are specified: • **`nolock`** - Specify this option to ensure that multiple Protegrity Discover machines that have mounted a shared folder can simultaneously scan the shared folder. **Important:** It is recommended that you retain this default value to ensure that Protegrity Discover successfully scans the shared folder using NFS. • **`sec=krb5`** - Specify this option to use Kerberos authentication for accessing the shared folder on the NFS server. This option is available only if you have selected *Kerberos Authentication* as the authentication type for NFS.
extra_parameters	Specify the additional parameters that are passed as arguments to the *mount* command, other than those arguments that are passed using the *-o* option. **Note:** For more information on the additional parameters that can be passed as arguments to the *mount* command, refer to the manual pages (man pages) for the *mount* command and NFS file format for Linux.	`{ "extra_param": "<value_1> <value_2>" }`	Network File System (NFS)	By default, the value of this parameter is set to **`-v`**, which indicates the verbose mode.
filesystem_type	Specify the type of file system for NFS. You can set one of the following values: • *nfs* - Specify this value if the version of the NFS server is 2, 3, or 4.	`{ "filesystem_type":"<value>" }`	Network File System (NFS)	**`nfs`**

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	• *nfs4* - Specify this value if the version of the NFS server is 4. This value is only used for legacy purposes. **Note:** For more information on the NFS file system type, refer to the manual pages (man pages) for the *mount* command and NFS file format for Linux.			
sample_size	Defines the size of sample data that you intend to scan. You can specify the sample size as a fixed number (static value) or as a percentage. The sample size determines a chunk of data from the entire data container that must be considered for the scan. For example, you can set sample size of 1000 records for scanning a database table containing 10K records. If you do not specify any value for the *sample_size* configuration parameter, then a default value *1000* is considered as the sample data. If the *sample_size* configuration parameter is expressed as a percentage, then the GET_ROW_COUNT query is used to calculate the *%(sample_count)i* parameter, by multiplying the total number of rows with the percentage value. For more information about the GET_ROW_COUNT query and its format, refer to the table *System Queries with Context*. **Note:** Apache Avro does not support a percentage value for the *sample_size* parameter. **Important:** The *sample_size* parameter depends on the network and available memory of the system where Protegrity Discover has been deployed. If you increase the default sample size of 1000, then it can lead to performance issues if your system does not have sufficient network bandwidth or memory.	`{ "sample_size": <value> }`	• EXAsol • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Oracle Database • PostgreSQL • Teradata	`1000`
hdfs_config	Specify the following HDFS configuration parameters: • *timeout* - Specify the connection timeout and the read timeout. Connection timeout specifies the time for which a Protegrity Discover scan job will try to connect with an HDFS server, after which the scan is aborted. Read timeout specifies the time for which a Protegrity Discover scan job will try to read the data from an HDFS server, after which the scan is aborted.	`{ "hdfs_config": { "timeout":"<value 1>, <value 2>" "mutual_auth": "<value>" "max_concurrency":<value> "urls": ["<value 1>", "<value 2>"]`	Hadoop File System (HDFS)	The following are the default values for the HDFS configuration parameters: • *timeout* - `"15, 30"`, where the connection timeout is *15* seconds and the read timeout is *30* seconds. • *mutual_auth* - `OPTIONAL` • *max_concurrency* - `1` • *urls* - Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	If you want to specify a single value for the connection and read timeouts, then you can specify a float value for the timeout. For example, you can specify the timeout value as: `{"timeout":<timeout_value>}` However, if you want to specify different values for connection and read timeouts, then you can specify the timeout values as a tuple. For example, you can specify the timeout values as: `{"timeout":"<connect_timeout_value>, <read_timeout_value>"}` For more information about the *timeout* parameter, refer to the section *Timeouts* in the Python *Requests* module. • *mutual_auth* - Specify whether mutual authentication needs to be enforced between Protegrity Discover and the HDFS server. You can specify one of the following values: • *REQUIRED* - Mutual authentication must be enforced. • *OPTIONAL* - Mutual authentication is optional. • *DISABLED* - Mutual authentication is not enabled. This parameter is only applicable if you select *Kerberos Authentication* as the authentication mode. For more information about the *mutual_auth* parameter, refer to the section *Kerberos authentication* in the HdfsCLI documentation. • *max_concurrency* - Specify the maximum number of concurrent requests that are allowed to be sent to a Kerberos KDC service for authentication. This parameter is only applicable if you select *Kerberos Authentication* as the authentication mode. For more information about the *max_concurrency* parameter, refer to the section *Kerberos authentication* in the HdfsCLI documentation. For more information about the Kerberos KDC, refer to the section *Kerberos*. • *urls* - Specify multiple host names or IP addresses for the HDFS Name Node or Data Node, in case you are using a high availability cluster for the Name	} }		

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	Node or Data Node. You also need to specify the port number along with the host name or IP address. For example, consider that you want to use the WebHDFS service to access the HDFS data. The HDFS cluster has two Name Nodes, which have the host names as *hostname_1* and *hostname_2*, in a high availability configuration. In this case, you can specify the following value for the *urls* parameter: `{"urls": ["http://hostname_1:port", "http://hostname_2:port"]}` Protegrity Discover will establish a connection with the first Name Node. However, if the first Name Node fails, then Protegrity Discover will automatically establish a connection with the second Name Node.			
certs	Use this configuration parameter if you want to use a self-signed certificate to securely communicate between the HDFS server and Protegrity Discover. In this parameter, you need to specify the path where the custom certificate has been uploaded. For example, set the value of the *certs* attribute to `/etc/ksa/certs/<ID of the custom certificate>.pem`. You can obtain the ID of the custom certificate from the **Certificate Repository** screen. For more information about the Certificate Repository, refer to the section *Certificate Repository* in the *Protegrity Certificate Management Guide 9.2.0.0*. For more information about uploading a certificate, refer to the section *To upload certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*. If your certificate also contains the private key, then you need to specify the path where the certificate file has been uploaded. If you have separately uploaded the certificate file and the private key, then you must specify the paths where both the files have been uploaded, as a comma separated string.	• `{ "certs": "<Path where the certificate has been uploaded>" }` - If the private key is part of the certificate • `{ "certs": "<Path where the certificate has been uploaded, Path where the private key has been uploaded>" }` - If the private key has been separately uploaded	Hadoop File System (HDFS)	Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	For example, set the following value of the *certs* attribute. `{ "certs": "/etc/ksa/certs/<ID of the custom certificate>.pem, /etc/ksa/certs/<ID of the file containing the private key>.key" }` You can obtain the ID of the custom certificate and of the file containing the private key from the **Certificate Repository** screen. For more information about the Certificate Repository, refer to the section *Certificate Repository* in the *Protegrity Certificate Management Guide 9.2.0.0*. For more information about the *certs* parameter, refer to the *Client Side Certificates* section in the Python *Requests* module. **Note:** This parameter is only applicable if you have selected *Kerberos Authentication*, while creating a discover job.			
verify	Specify whether the self-signed certificate that you have uploaded to establish a secure communication between the HDFS server and Protegrity Discover, needs to be verified. You can either specify a boolean value or a string value for this parameter. If you specify a boolean value, then this parameter determines whether the TLS certificate of the server is verified. In this case, you can specify one of the following values for this parameter: • *true* - The server certificate is verified. • *false* - The server certificate is not verified. If you specify a string value for this parameter, then the value indicates the path where the Certificate Authority bundle is stored. For more information about the *verify* parameter, refer to the section *SSL Cert Verification* in the Python *Requests* module. **Note:** This parameter is only applicable if you have selected *Kerberos Authentication*, while creating a discover job.	`{ "verify":"<value>" }`	Hadoop File System (HDFS)	true

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
webhdfs_prefix	Specify the value of the prefix that is used in the URL to scan the data from the HDFS server. For example, Protegrity Discover uses the following URL for performing any operation on an HDFS server, where */webhdfs/v1* is the default value of the *webhdfs_prefix* parameter: ***http://<HOST>:<HTTP_PORT>/webhdfs/v1/<PATH>?op=...*** You can modify this default value, if required.	`{ "webhdfs_prefix":<"value"> }`	Hadoop File System (HDFS)	***/webhdfs/v1***
get_delegation_token_url	Specify the URL for retrieving the token from the HDFS server. For example, you can specify the following URL: ***http://<HOST>:<PORT>/webhdfs/v1/?op=GETDELEGATIONTOKEN[&renewer=<USER>][&service=<SERVICE>][&kind=<KIND>]*** You can specify the following query parameters in the URL: • *USER* - Specify the name of the user who has permissions to access the HDFS server. For more information about the *USER* parameter, refer to the section *Username* query parameter in the WebHDFS REST API documentation. • *SERVICE* - Specify the name of the service where you want to use the token. For example, you can specify the value of this parameter as ***<Hostname or IP address of the Name Node>:<port>***. For more information about the *SERVICE* parameter, refer to the section *Token Service* query parameter in the WebHDFS REST API documentation. • *KIND* - Specify the kind of the token requested. For example, you can specify the value of this parameter as *HDFS_DELEGATION_TOKEN*. For more information about the *KIND* parameter, refer to the section *Token Kind* query parameter in the WebHDFS REST API documentation.	`{ "get_delegation_token_url":< "value"> }`	Hadoop File System (HDFS)	Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	For more information about the query parameters used in the URL, refer to the *WebHDFS REST API* documentation. **Note:** This parameter is only applicable if you have selected Token Authentication, while creating a discover job.			
cancel_delegation_token_url	Specify the URL for canceling the delegation token from the HDFS server. For example, you can specify the following URL: `http://<HOST>:<PORT>/webhdfs/v1/?op=CANCELDELEGATIONTOKEN&token=` For more information about the query parameters used in the URL, refer to the *WebHDFS REST API* documentation. **Note:** This parameter is only applicable if you have selected Token Authentication, while creating a discover job.	`{ "cancel_delegation_token_url":<"value"> }`	Hadoop File System (HDFS)	Not applicable
renew_delegation_token_url	Specify the URL for renewing the delegation token from the HDFS server. For example, you can specify the following URL: ***`http://<HOST>:<PORT>/webhdfs/v1/?op=RENEWDELEGATIONTOKEN&token=`*** For more information about the query parameters used in the URL, refer to the *WebHDFS REST API* documentation. **Note:** This parameter is only applicable if you have selected Token Authentication, while creating a discover job.	`{ "renew_delegation_token_url":<"value"> }`	Hadoop File System (HDFS)	Not applicable
random_fetch	Determine the method in which the data is retrieved from a database table. You can set one of the following values for this parameter: • *true* - Retrieve the data from a database table in a random sequence. • *false* - Retrieve the data from a database table sequentially.	`{ "random_fetch":<value> }`	• EXAsol • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Oracle Database	true

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
			• PostgreSQL • Teradata	
fetch_count	Determine the number of records retrieved at a time from the database table, from a sequence list. A sequence list defines the specific positions in a database table, from where Protegrity Discover retrieves the data for scanning. For example, consider a scenario where the sample size is set to 1000 and the fetch count is set to 100. In this case, Protegrity Discover identifies 10 positions within the database table, from where it retrieves 100 records each for scanning. The positions are calculated internally. Each position is separated from each other by at least a value that is equal to the fetch count. For example, if the fetch count is 100 and the first position starts from the first row of the database table, then the second position will start at minimum from the 101st row, or 201st row, or 301st row, and so on. Similarly, if the second position starts from the 201st row, then the third position will start at minimum from the 301st row, or 401st row, or 501st row, and so on. The order in which the positions are determined depends on the *random_fetch* parameter. If the *random_fetch* parameter is set to *true*, then Protegrity Discover determines the positions randomly. If the *random_fetch* parameter is set to *false*, then Protegrity Discover determines the positions sequentially.	`{ "fetch_count": <value> }`	• EXAsol • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Oracle Database • PostgreSQL • Teradata	100
custom_auth	Specify the configuration for connecting to a third-party authentication system, such as, CyberArk, for retrieving the password of the datastore that you want to scan. You can specify the following parameters: • *name* - Specifies the name of the CyberArk configuration that determines the CyberArk system with which you want communicate. For example, specify the value as *cyberark*. For more information about the CyberArk configuration name, refer to the *CyberArk configuration name* entry in the table in *step 2d* of the section *Setting up Custom Authentication Using CyberArk*. • *Object* - Specifies the unique name of the account created in the CyberArk system for storing the password of the required datastore. For more information about the *Object* parameter, refer to the section *Call the Web Service using REST* in the CyberArk documentation.	`{"custom_auth" : {"name":"cyberark", "params": { "Object":"<Unique account name>" } } }`	For more information about the datastores supported by CyberArk, refer to the *CyberArk documentation*.	Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
webhook	Specify the configuration for creating a local webhook, which enables you to send the scan results for the specific discover job to an external URL. You can specify the following parameters: • *url* - Specify the URL of the endpoint server, which is used to receive the notifications sent by the Protegrity Discover webhook. This parameter is mandatory. For example, specify the value of the URL as *https://<host:port>/path* The following parameters are specified in the URL: • *host* - IP address or host name of the machine where the endpoint server has been installed. • *port* - Port number that is used to receive the notifications sent by the Protegrity Discover webhook. In this example, *path* denotes the endpoint of an API that is set to receive the notifications from the Protegrity Discover webhook. **Note:** Ensure that you specify the accurate end points for the API on the endpoint server. If you specify the accurate IP address and port number, but incorrect endpoints, then only an empty POST request, without any response, is sent to the endpoint URL. • *timeout* - Specifies the time for which Protegrity Discover tries to communicate with the endpoint server, after which the connection times out. The unit is in seconds. This parameter is optional. • *verify* - Specify whether the self-signed or trusted client certificate that you have uploaded to establish secure communication between the endpoint server and Protegrity Discover, needs to be verified. You can either specify a boolean value or a string value for this parameter. If you specify a boolean value, then this parameter determines whether the TLS certificate of the server is verified. In this case, you can specify one of the following values for this parameter: • *true* - The server certificate is verified. • *false* - The server certificate is not verified.	`{ "url": "value", "timeout": "value", "verify":"value", "cert": value" }`	• AWS S3 • Azure Storage (Blob) • EXAsol • Hadoop File System (HDFS) • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Network File System (NFS) • Oracle Database • PostgreSQL • SharePoint • Teradata • Windows Share (CIFS)	Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
	If you specify a string value for this parameter, then the value indicates the path where the Certificate Authority bundle is stored. This parameter is optional. For more information about the *verify* parameter, refer to the section *SSL Cert Verification* in the Python *Requests* module. • *cert* - Specify the ID of the client certificate, which is mentioned on the **Certificate Repository** screen, used by the endpoint server to authenticate Protegrity Discover. This parameter is optional.			
update_keytab	Specify whether you want to update an existing keytab file. You can set one of the following values for this parameter: • *true* - Overwrites an existing keytab file. • *false* - Does not overwrite an existing keytab. If you try to upload a new keytab file with the same name as that of an existing keytab file, then an error appears.	`{ "update_keytab ":<value> }`	• Hadoop File System (HDFS) • Hive • Microsoft SQL Server • Network File System (NFS)	false
login_timeout	Specify the time for which a Protegrity Discover scan job will try to login to a database, after which the scan is aborted. The value specified in the *login_timeout* parameter is used to override the timeout value that is specified in the *timeout* parameter of the pyodbc *connect()* function, which sets the *SQL_ATTR_LOGIN_TIMEOUT* parameter of the corresponding ODBC driver. The *SQL_ATTR_LOGIN_TIMEOUT* parameter, which is an attribute of the *SQLSetConnectAttr* function, determines the time that the ODBC driver waits for a login request, before the request times out. For more information about the *SQLSetConnectAttr* parameter and its attributes, refer to the *Microsoft SQL Documentation*. For more information about the pyodbc *connect()* function and its parameters, refer to the *pyodbc Documentation*. If you specify the value of the *login_timeout* parameter as *0*, then the Protegrity Discover uses the default timeout value of the targeted database, if such a value has been set. The *login_timeout* value is expressed in seconds.	`{ "login_timeout ":<value> }`	• EXAsol • Hive • IBM DB2/UDB • IBM DB2/zOS • Microsoft SQL Server • MySQL • Oracle Database • PostgreSQL • Teradata	The default values change depending on the targeted system: • EXAsol - **0** • Hive - **0** • IBM DB2/UDB - **0** • IBM DB2/zOS - **0** • Microsoft SQL Server - **0** • MySQL - **5** • Oracle Database - **0** • PostgreSQL - **5** • Teradata - **0** **Note:** Protegrity recommends a timeout value of *5* seconds for MySQL and PostGres. Ensure that the value is non-zero. If the *login_timeout* value is set to *0* for MySQL and PostGres, then you cannot save the job if the **System**

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
				setting in the job is set to *Auto Detect* and the database service is not available on the requested port. **Note:** If your connectivity to the database is slow, then you can choose to increase the value specified in the *login_timeout* parameter.
role_arn	Specify the Amazon Resource Name (ARN) of the IAM role that has permissions to access the AWS S3 buckets. The ARN is used to uniquely identify a resource in AWS. You need to specify this parameter if you want to scan the AWS S3 buckets associated with another AWS account or the same account as that of the AWS EC2 instance.	`{ "role_arn":<"value"> }`	AWS S3	Not applicable
role_session_name	Specify the session name for the IAM role, whose ARN you have provided in the *role_arn* parameter. If you do not specify the value for this parameter, then the value is automatically generated as shown in the following snippet. `Protegrity_Discover_<Instance_id>` *Instance_id* is dynamically generated. **Note:** This setting is applicable only if the user has provided the *role_arn* parameter.	`{ "role_session_name":<"value"> }`	AWS S3	Not applicable
aws_external_id	Specify the external ID for the IAM role, whose ARN you have provided in the *role_arn* parameter. This is an optional parameter. This parameter is required only if the use of an external ID has been set to required during the creation of the role. **Note:** This setting is applicable only if the user has provided the *role_arn* parameter.	`{ "aws_external_id":<"value"> }`	AWS S3	Not applicable

Configuration Setting	Description	Syntax	Applicable Datastores	Default Value
session_expiration_time	Specify the time after which the AWS session will expire. **Note:** This setting is applicable only if the user has provided the *role_arn* parameter.	`{ "session_expiration_time":<"value"> }`	AWS S3	3600. The unit is in seconds.
minimum_renew_session_time	Specify the time remaining in a session, after which the new session or temporary security credentials will be created. For example, if the *session_expiration_time* is 3600 seconds and the *minimum_renew_session_time* is 300 seconds, then the AWS session will be renewed when less than 300 seconds are remaining in the session. **Note:** This setting is applicable only if the user has provided the *role_arn* parameter.	`{ "minimum_renew_session_time" :<"value"> }`	AWS S3	300. The unit is in seconds.
aws_assume_role_config	Specify optional parameters for the *assume_role* method in the AWS SDK for Python (Boto3), if you have specified a value in the *role_arn* parameter. For more information about the parameters that you can specify in the *aws_assume_role_config* setting, refer to the parameters specified in the *assume_role()* method in the Boto3 documentation. **Caution:** Do not provide the following parameters, as they have already been specified as part of other configuration settings: • *RoleARN* • *RoleSessionName* • *DurationSeconds* • *ExternalId* **Note:** This setting is applicable only if the user has provided the *role_arn* parameter.	`{ "aws_assume_role_config": { "<parameter_1>":<value 1> "<parameter_2>":<value 2> "<parameter_3>":<value 3> } }`	AWS S3	Not applicable

Chapter 16

Appendix D: Supported File Formats

This section lists the file formats supported by Protegrity Discover.

Protegrity Discover can be used to identify sensitive data from the following unstructured and semi-structured file formats:

- Unstructured file formats
  - *doc*
  - *docx*
  - *xls*
  - *xlsx*
  - *txt*
  - *pdf*
  - *jpg*
  - *jpeg*
  - *bmp*
  - *png*
  - *gif*
  - *jfif*
- Semi-structured file formats
  - *json*
  - *xml*
  - *csv* (comma-delimited)
  - *sas7bdat*
  - *sd7*
  - *parquet* - Protegrity Discover supports the following compression algorithms to decompress the Apache Parquet data:
    - UNCOMPRESSED
    - SNAPPY
    - GZIP
    - LZO
    - BROTLI
    - LZ4
    - ZSTD
  - *avro* - Protegrity Discover supports the following compression algorithms to decompress the Apache Avro data:
    - Snappy
    - Deflate
    - Zstandard
    - Bzip2

- LZ4
- XZ

Important: Protegrity Discover scans only those Apache Avro files that contain the schema along with the data.

Important: If you have created a table in Hive and want to store the output file with the file format as Apache Avro, then ensure that you save the file with the extension as *.avro*. To generate an output file from the Hive table with the *.avro* extension, you must run the following command on the Hive database:

```
set hive.output.file.extension=.avro;
```

Chapter 17

Appendix E: ODBC INI File Structure

The ODBC ini (*odbcinst.ini*) file is the configuration file for the unixODBC driver manager, which is provided as part of the Protegrity Discover installation package.

For more information about the unixODBC driver manager, refer to the *unixODBC project home page*.

The *odbcinst.ini* is divided into multiple sections, as shown in the following snippet. Each section corresponds to a specific ODBC driver.

```
[Section Heading1]
Key 1=Value 1
Key 2=Value 2
Key 3=Value 3

[Section Heading2]
Key 1=Value 1
Key 2=Value 2
Key 3=Value 3
```

Each section consists of two entities:

- *Section heading* - Used to identify a specific ODBC driver. You must specify the same section heading in braces as the value of the *Driver* parameter in the connection string, while creating a new datastore.

  For more information about creating a new datastore, refer to the section *Adding a New Datastore*.

  If you have uploaded a new ODBC driver, then the section heading is created by concatenating the values of the database type, vendor name, and driver version that you have specified while uploading the driver.

  For more information about uploading a new ODBC driver, refer to the section *ODBC Setup and System Configuration Settings*.

- *Key-value pairs* - Used to specify the configuration settings for a specific ODBC driver.

  Out-of-the-box, Protegrity Discover supports the following ODBC drivers:

  - IBM ODBC driver for DB2
  - Microsoft ODBC driver for SQL Server
  - MySQL database ODBC driver
  - Oracle ODBC driver for Oracle
  - PostgreSQL database ODBC driver
  - Teradata database ODBC driver

  The default configuration settings for the out-of-the-box ODBC drivers are provided in the *odbcinst.ini* file. The following snippet lists the default configuration settings for the Oracle and Teradata ODBC drivers in the *odbcinst.in* file.

  ```
  [Oracle]
  Description=Oracle ODBC driver for Oracle 12c
  Driver=/opt/oracle/instantclient_12_2/libsqora.so.12.1

  [Teradata]
  ```

```
DriverManagerEncoding=UTF-16
Description=Teradata Database ODBC Driver 16.20
Driver=/opt/teradata/client/16.20/lib64/tdataodbc_sb64.so
APILevel=CORE
ConnectFunctions=YYY
DriverODBCVer=3.51
SQLLevel=1
```

If you have uploaded a new ODBC driver, then only the Driver-Path key value pair is added to the *odbcinst.ini* file by default.

Important: The configuration settings, or the key value pairs, are specific to each ODBC driver. For information on each setting and its default values for the out-of-the-box ODBC drivers, refer to the documentation provided by the vendor of the specific ODBC driver.

Similarly, if you are installing a new ODBC driver, then you must first refer to the documentation provided by the vendor of the ODBC driver for information on the driver settings that are applicable and their default values.

Chapter 18

Appendix F: Default Classifiers

This section lists the default classifiers provided by Protegrity Discover to identify sensitive data.

The following table lists the default classifiers that have been provided by Protegrity Discover out-of-the-box. It also lists the tabs on the **Classifiers** screen that are applicable for each classifier.

Table 18-1: Default Classifiers and Applicable Tabs

Default Classifiers	*General*	*Metadata*	*Qualification*	*Regex*	*Reference*	*Source Code*	*Spacy Patterns*	*Settings*
Codice Fiscale	**Yes**	**Yes**	**Yes**	**Yes**	No	No	No	No
Country ISO Alpha 2 Code	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
Country ISO Alpha 3 Code	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
Country Name	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
Credit Card	**Yes**	**Yes**	**Yes**	No	No	No	No	No
Date Of Birth	**Yes**	**Yes**	**Yes**	No	No	No	No	**Yes**
Dutch Tax ID	**Yes**	**Yes**	**Yes**	No	No	No	No	No
Email Address	**Yes**	**Yes**	**Yes**	No	No	No	No	**Yes**
IBAN	**Yes**	**Yes**	**Yes**	No	No	**Yes**	No	No
IMAGE RECOGNITION	**Yes**	**Yes**	No	No	No	No	No	No
IPv4	**Yes**	**Yes**	**Yes**	**Yes**	No	No	No	No
IPv6	**Yes**	**Yes**	**Yes**	**Yes**	No	No	No	No
MAC	**Yes**	**Yes**	**Yes**	**Yes**	No	No	No	No
NLP Credit Card	**Yes**	**Yes**	No	No	No	No	**Yes**	No
NLP Email Address	**Yes**	**Yes**	No	No	No	No	**Yes**	**Yes**
NLP Persons Name	**Yes**	**Yes**	No	No	No	No	**Yes**	No
NLP Phone Number	**Yes**	**Yes**	No	No	No	No	**Yes**	No
NLP US Address	**Yes**	**Yes**	No	No	No	No	**Yes**	No
NLP US Social Security Number	**Yes**	**Yes**	No	No	No	No	**Yes**	No
Password	**Yes**	**Yes**	**Yes**	**Yes**	No	No	No	No
Persons Name	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No

Default Classifiers	*General*	*Metadata*	*Qualification*	*Regex*	*Reference*	*Source Code*	*Spacy Patterns*	*Settings*
Phone Number	**Yes**	**Yes**	**Yes**	No	No	No	No	**Yes**
Place Name	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
Postal Code	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
State Code	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No
State Name	**Yes**	**Yes**	**Yes**	No	**Yes**	No	No	No

Chapter 19

Appendix G: Understanding Protegrity Discover-specific Permissions

This section lists the Protegrity Discover-specific permissions.

The following table lists the permissions that are required for using Protegrity Discover.

Table 19-1: Protegrity Discover-specific Permissions

Permission	Description
Discover Admin	Allows users to perform all operations available as part of the Protegrity Discover Web UI. You can also access the Protegrity Discover REST API documentation.
Discover Viewer	Allows users to log on to the Protegrity Discover Web UI as a viewer. You can also access the Protegrity Discover REST API documentation. However, you cannot edit or download any files from the Appliance Web UI.
Discover RestAPI	Allows users to access the Protegrity Discover REST APIs.

You need to add users to access Protegrity Discover, and assign them the required Protegrity Discover-specific permissions through roles.

For more information about adding users, refer to *step 23* in the section *Managing the Appliance Information*.

For more information about assigning permissions to user roles, refer to the section *Managing Roles* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

For more information about managing users, refer to the section *Managing Users* in the *Protegrity Appliances Overview Guide 9.2.0.0*.

Chapter 20

Appendix H: Integrating Protegrity Discover with CyberArk

The CyberArk Application Access Manager (AAM) provides custom authentication to access datastores that are scanned by Protegrity Discover. This enables Protegrity Discover to connect to the datastores without requiring to store the password.

> **Note:** The CyberArk Application Access Manager is deployed using an agentless configuration.
>
> For more information about the CyberArk Application Access Manager, refer to the *CyberArk documentation*.

The following figure illustrates the communication flow between Protegrity Discover and CyberArk for retrieving the password for the datastore that needs to be scanned.

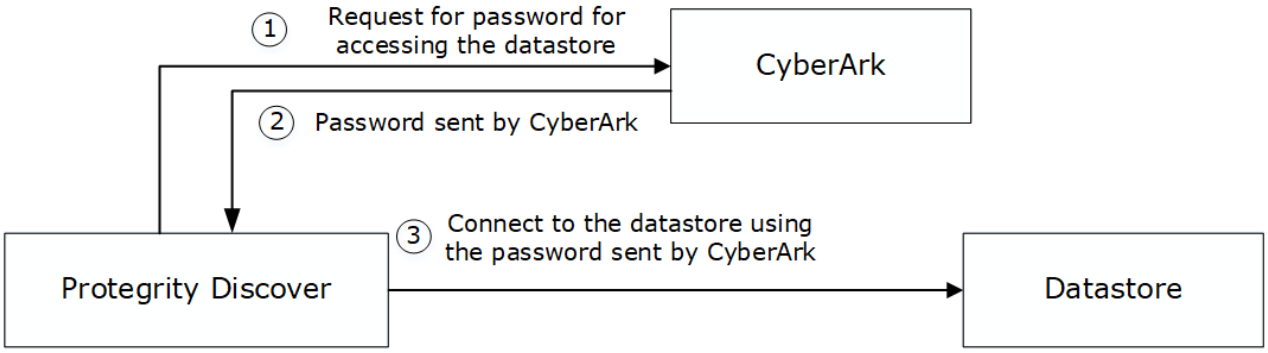


Figure 20-1: Communication Flow Between Protegrity Discover and CyberArk

The CyberArk system has been configured to store the password for the datastore that needs to be scanned using Protegrity Discover. The following steps specify the sequence that is followed for requesting a password from the CyberArk system.

1. Protegrity Discover sends a request to the CyberArk system for retrieving the password to connect to the datastore.
2. CyberArk authenticates the request from Protegrity Discover and returns the password for connecting to the datastore.
3. Protegrity Discover uses the password obtained from the CyberArk system to connect to the datastore.

> **Important:** Protegrity Discover does not save the password that is obtained from CyberArk.

20.1 Setting up Custom Authentication Using CyberArk

The following sections describe the prerequisites and tasks for setting up custom authentication for Protegrity Discover using CyberArk.

Before you begin

- Ensure that you define the Protegrity Discover application in the Password Vault Web Access (PVWA) Applications Management UI in the CyberArk system.

  For more information about defining an application, refer to the section *Add applications* in the CyberArk documentation.

- Ensure that you have created a self-signed or a trusted client certificate for the CyberArk system. You need to copy this certificate to the machine on which Protegrity Discover has been installed. This certificate is used to authenticate the Protegrity Discover application, when it sends a request to the CyberArk system for obtaining the password of the datastore that you want to scan.

  > **Important:** Protegrity recommends that you use a certificate-based authentication method for authenticating the Protegrity Discover application to the CyberArk system.

To set up custom authentication using CyberArk:

1. Perform the following steps to upload the client certificate to the certificate repository.
   a. Login to the Appliance Web UI and navigate to **Settings** > **Network Settings** > **Certificate Repository**.

      The Certificate Repository screen appears.
   b. Click **Upload new files** to upload the CyberArk client certificate to the certificate repository.

      When Protegrity Discover sends a request to the CyberArk system for retrieving the passwords for a specific datastore, CyberArk uses the uploaded certificate to authenticate Protegrity Discover.

      For more information about uploading a certificate, refer to the section *Uploading Certificates or CRL* in the *Protegrity Certificate Management Guide 9.2.0.0*.

2. Perform the following steps to modify the *sureloc.json* file.
   a. On the Appliance Web UI, navigate to **Settings** > **System** > **Files**.

      The List of Product Files screen appears.
   b. In the **Discover - Settings** section, edit the *sureloc.json* file.
   c. Edit the following code block in the *sureloc.json* file.

```
"custom_auth": {
    "<CyberArk configuration name>": {
      "type": "CyberArk",
      "url": "https://hostname:16945/AIMWebService/api/Accounts",
      "cert": "<ID of the client certificate used by the CyberArk to authenticate
Protegrity Discover>",
      "verify": "<Value for verifying the client certificate>",
      "params": {
        "AppId": "<Application ID defined in CyberArk>",
        "Safe": "<Safe name defined in CyberArk>"
      }
    }
  }
```

   d. Modify the following fields in the custom authentication section.

Field	Description
CyberArk configuration name	Specify a name for the CyberArk configuration. If you have multiple CyberArk systems, then you can create separate configurations for each of the systems. For example, if you have separate CyberArk systems for the development and testing environments, then you can create two entries for each configuration, You can specify the names for the individual configurations as *cyberark_dev* and *cyberark_qa*. In this example, you can add the following code block in the *sureloc.json* file: `"custom_auth": { "cyberark_dev": { "type": "CyberArk", "url": "<value>", "cert": "<value>", "verify": "<value>", "params": { "AppId": "<value>", "Safe": "<value>" } }, "cyberark_qa": { "type": "CyberArk", "url": "<value>", "cert": "<value>", "verify": "<value>", "params": { "AppId": "<value>", "Safe": "<value>" } } }` When you are creating a discover job, you need to specify the name of the CyberArk configuration as the value of the *name* parameter in the *custom_auth* configuration setting in the advanced configuration settings for that job. For more information about the *custom_auth* configuration setting, refer to the section *Scan Job Advanced Configuration Settings*.
type	Specifies the type of authentication. For example, specify the value as *CyberArk*.
url	Specify the URL of the Central Credential Provider Web Service, which is used by Protegrity Discover to retrieve the password of the datastore that you want to scan. For example, specify the value of the URL as *https://<host:port>/AIMWebService/api/Accounts* The following parameters are specified in the URL: • *host* - IP address or host name of the machine where the Central Credential Provider Web Service has been installed.

Field	Description
	• *port* - Port number that is used to access the Central Credential Provider Web Service AIMWebService is the name of the Central Credential Provider Web Service. For more information about the Central Credential Provider and the Central Credential Provider Web Service, refer to the section *Central Credential Provider* in the CyberArk documentation.
cert	Specify the ID of the client certificate, which is mentioned on the **Certificate Repository** screen, used by the CyberArk system to authenticate Protegrity Discover. If the CyberArk server is not using a certificate-based authentication method, then you need to specify the value of the *cert* parameter as *null*.
verify	Specify whether the self-signed or trusted client certificate that you have uploaded to establish secure communication between the CyberArk system and Protegrity Discover, needs to be verified. You can either specify a boolean value or a string value for this parameter. If you specify a boolean value, then this parameter determines whether the client certificate is verified by the CyberArk system. In this case, you can specify one of the following values for this parameter: • *true* - The client certificate is verified. • *false* - The client certificate is not verified. You can also specify the ID of the CA certificate. Protegrity Discover uses this value to internally identify the path where the Certificate Authority bundle is stored. The Certificate Authority bundle is used to verify the client certificate. For more information about the *verify* parameter, refer to the section *SSL Cert Verification* in the Python *Requests* module.
params/appID	Specifies the unique ID of the Protegrity Discover application that has been defined in the PVWA Applications Management UI.
params/Safe	Specifies the name of the safe where the password of the datastore to be scanned is stored.

For more information about the parameters defined in the CyberArk system, refer to the section *Call the Web Service Using REST* in the CyberArk documentation.

3. Perform the following steps to modify the advanced configurations while creating a discover job.
   For more information about creating a discover job, refer to the section *Creating a Discover Job*.
   a. In the **Add Discover Job** dialog box, click **Advanced Settings** to display the **Config** text area.

b. Specify the following advanced configuration settings for using the custom authentication.

```
{"custom_auth":{"name":"cyberark","params":{"Object":"<Unique account name"}}}
```

c. Modify the following fields in the custom authentication.

Field	Description
name	Specifies the name of the CyberArk configuration that you want to use for connecting to the CyberArk system. For example, specify the value as *cyberark*. For more information about the CyberArk configuration name, refer to the CyberArk configuration name entry in the table in *step 2d*.
params/Object	Specifies the unique name of the account created for storing the password of the required datastore.

For more information about the parameters defined in the CyberArk system, refer to the *Central Credential Provider Implementation Guide*.

You can use this configuration to connect to a CyberArk system for retrieving the password for the datastore that you want to scan.

Chapter 21

Appendix I: Percent-Encoding Special Characters

This section describes how you can percent-encode commonly used special characters using UTF-8 encoding. In percent-encoding, a character is replaced by a percent character and two hexadecimal digits, based on the UTF-encoding standards. Percent-encoding is also known as URL encoding.

For more information about percent-encoding, refer to the section *2.1 Percent-Encoding* in the *W3 Uniform Resource Identifier* specification.

The following table lists the percent-encoded values for the commonly used special characters.

Table 21-1: Percent-Encoding

Character	Percent-Encoding using UTF-8 Encoding Format
space	%20
!	%21
"	%22
#	%23
$	%24
%	%25
&	%26
'	%27
(	%28
)	%29
*	%2A
+	%2B
,	%2C
-	%2D
/	%2F
:	%3A
;	%3B
<	$3C
=	%3D
>	%3E
?	%3F
@	%40
[	%5B
\	%5C
]	%5D
^	%5E

Character	Percent-Encoding using UTF-8 Encoding Format
_	%5F
`	%60
{	%7B
|	%7C
}	%7D
~	%7E

Chapter 22

Appendix J: File Metadata Collected in Filestores

This section lists the metadata that is collected after scanning files in a filestore.

The following table describes the file metadata collected for each filestore.

Table 22-1: File Metadata

Filestore	File Metadata	Example
AWS S3	• *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *storage_class* - Name of the AWS S3 storage class where the file has been stored. For example, **Standard** or **Standard-IA**. • *file_type* - Format of the scanned file • *aws_metadata* - Metadata created by users in AWS. This parameter is set to *null*, if users have not created any custom key-value pairs in the **Metadata** field for the particular file in AWS S3. **Important:** Only user-defined metadata is supported. • *aws_tags* - Tags created by end users in AWS. This parameter is set to *null*, if users have not created any custom key-value pairs in the **Tags** field for the particular file in AWS S3.	The following snippet shows an example where users have created custom key-value pairs in the **Metadata** and **Tags** fields in AWS S3. `file_metadata: {"modified_time":"2021-03-10 13:00:09","file_size":805145,"storage_class":"STANDARD","file_type":"csv","aws_metadata": {"owner_name":"<Owner Name>"},"aws_tags": {"test1":"value1"}}` The following snippet shows an example where users have not created any custom key-value pairs in the **Metadata** and **Tags** fields in AWS S3. `file_metadata: {"modified_time":"2021-03-10 13:00:09","file_size":805145,"storage_class":"STANDARD","file_type":"csv","aws_metadata":null,"aws_tags":null}`
Azure Storage (Blob)	• *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *file_type* - Format of the scanned file • *created_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *azure_metadata* - Metadata created by users in Azure. This parameter is set to *null*, if users have not created any custom key-value pairs in the **Metadata** field for the particular blob in Azure. • *azure_tags* - Tags created by end users in Azure. This parameter is set to *null*, if users have not created any custom key-value pairs in the **Blog index tags** field for the particular blob in Azure.	The following snippet shows an example where users have created custom key-value pairs in the **Metadata** and **Blob index tags** fields in Azure. `file_metadata: {"modified_time":"2021-08-17 12:59:03","file_size":412,"file_type":"txt","created_time":"2020-09-24 17:53:43","azure_metadata": {"owner_name":"<Owner Name>"},"azure_tags": {"test1":"value1"}}` The following snippet shows an example where users have not created any custom key-value pairs in the **Metadata** and **Blob index tags** fields in Azure. `file_metadata: {"modified_time":"2020-09-24 17:54:07","file_size":412,"file_type":"txt","created_time":"2020-09-24 17:54:07","azure_metadata":null,"azure_tags":null}`

Filestore	File Metadata	Example
CIFS	• *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *created_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_type* - Format of the scanned file	`file_metadata: {"modified_time":"2021-06-28 07:02:00","file_size":190000,"created_time":"2021-06-28 12:53:49","file_type":"csv"}`
HDFS	• *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *file_permission* - Permissions applied on the file • *owner* - Name of the user who has created the file • *group* - Name of the group whose members have permissions to access the file. • *file_type* - Format of the scanned file	`file_metadata: {"modified_time":"2020-09-14 11:58:08","file_size":805145,"file_permission":"644","owner":"hdfs","group":"supergroup","file_type":"csv"},`
NFS	• *created_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *file_permission* - Permissions applied on the file • *file_type* - Format of the scanned file • *owner* - Name of the user who has created the file • *group* - Name of the group whose members have permissions to access the file	`file_metadata: {"created_time":"2020-06-02 08:07:11","modified_time":"2020-06-01 18:06:55","file_size":1266835,"file_permission":"100644","file_type":"txt","owner":"root","group":"root"}`
SharePoint	• *modified_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *file_size* - Size of the file in bytes • *file_type* - Format of the scanned file • *created_time* - Date and time when the file was modified. The format is YYYY-MM-DD HH:MM:SS. • *owner* - Name of the user who has created the file	`file_metadata: {"modified_time":"2019-09-13 05:40:30","file_size":1270027,"file_type":"csv","created_time":"2019-09-03 10:45:19","owner":"<Owner_name>"}`

Chapter 23

Appendix K: Using File Metadata to Create Custom Classifiers

This section lists sample use cases for creating custom classifiers that will classify the files using file metadata.

23.1 Example 1: Creating a Custom Classifier to Identify Large Video Files

This section describes how to create a custom classifier to identify video files that are larger than 1 GB in size.

To create a custom classifier to identify large files:

1. Create a classifier.
   For more information about creating a classifier, refer to the section *Creating Classifiers*.
2. Click + to add a new Python boolean expression in the **Keyword Patterns** area.
   The **Add New Metadata Keyword** dialog box appears.
3. Specify the following details in the **Add New Metadata Keyword** dialog box.
   For more information about specifying the metadata details, refer to the section *Updating the Metadata Tab*.

Field	Description
Name	Specify a unique name for the Python boolean expression.
Type	Select **File_type**.
Boost	Specify a value to boost the confidence score. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the boost value is applied to the confidence score.
Score	Specify the confidence score value. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the score value is added to the final confidence score. For example, specify the confidence score as *0.5*.
Expression	Specify the Python boolean expression for identifying the file metadata keyword. In this case, you want to search for video files. You can then specify the following expression. `"%(file_type)s" in ["mp3", "mp4"]`

Field	Description
Continue	Select this check box to include this pattern and all subsequent patterns in the **Keyword Patterns** area to identify the metadata. If you clear this check box, then the corresponding expression is included for identifying the metadata. However, all subsequent expressions are excluded from identifying the metadata.

Important: The metadata keyword appears on the **Classifications** screen in the Protegrity Discover Web UI. If the metadata keyword contains any sensitive data, then the sensitive data will appear on the **Classifications** screen.

4. Click **Add** to add the Python boolean expression.
   The Python boolean expression is added to the **Keyword Patterns** area.
5. Repeat steps *2* to *4* to add another Python boolean expression for identifying the files that are larger than 1 GB in size.
6. Specify the following details in the **Add New Metadata Keyword** dialog box.

Field	Description
Name	Specify a unique name for the Python boolean expression.
Type	Select **File_size**.
Boost	Specify a value to boost the confidence score. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the boost value is applied to the confidence score.
Score	Specify the confidence score value. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the score value is added to the final confidence score. For example, specify the confidence score as *0.5*. **Important:** If you want to accurately identify the file using the metadata, then ensure that the confidence score for both the metadata must equal 1.
Expression	Specify the Python boolean expression for identifying the file metadata keyword. In this case, you want to search for files that are larger than 1 GB in size. You can then specify the following expression. `%(file_size)s > 1024**3`
Continue	Select this check box to include this pattern and all subsequent patterns in the **Keyword Patterns** area to identify the metadata. If you clear this check box, then the corresponding expression is included for identifying the metadata. However, all subsequent expressions are excluded from identifying the metadata.

7. Click **Add** to add the Python boolean expression.
   The Python boolean expression is added to the **Keyword Patterns** area.
8. Navigate to the **Testbed** tab.
9. Specify a value in the **Test Metadata JSON** area.
   For more information about specifying a value in the **Test Metadata JSON** area, refer to *step 5* in the section *Testing the Classifier*.
10. Click **Run Test** to test the metadata.

The classifier evaluates the Python boolean expression specified in the **Metadata** tab using the values specified in the **Test Metadata JSON** area. If the evaluation is successful, then the results appear in the **Test Results** area and the classifier is activated.

> **Important:** If the classifier returns the confidence score as 1, then the classifier accurately identifies the file using the metadata. In this case, the classifier does not scan the contents of the file.

23.2 Example 2: Creating a Custom Classifier to Scan Sensitive Data within Legacy Files Created by a Specific User

This section describes how to create a custom classifier to identify sensitive data within files created by a user named *Jane Smith* before *31st December 2010*.

To create a custom classifier to identify large files:

1. Create a classifier.
   For more information about creating a classifier, refer to the section *Creating Classifiers*.
2. Click + to add a new Python boolean expression in the **Keyword Patterns** area.
   The **Add New Metadata Keyword** dialog box appears.
3. Specify the following details in the **Add New Metadata Keyword** dialog box.
   For more information about specifying the metadata details, refer to the section *Updating the Metadata Tab*.

Field	Description
Name	Specify a unique name for the Python boolean expression.
Type	Select **Owner**.
Boost	Specify a value to boost the confidence score. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the boost value is applied to the confidence score.
Score	Specify the confidence score value. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the score value is added to the final confidence score. For example, specify the confidence score as *0.4*.
Expression	Specify the Python boolean expression for identifying the file metadata keyword. In this case, you want to search for video files. You can then specify the following expression. `"%(owner)s" == "Jane Smith"` **Note:** The value depends on the format used for the owner name in your file system.
Continue	Select this check box to include this pattern and all subsequent patterns in the **Keyword Patterns** area to identify the metadata.

Field	Description
	If you clear this check box, then the corresponding expression is included for identifying the metadata. However, all subsequent expressions are excluded from identifying the metadata.

Important: The metadata keyword appears on the **Classifications** screen in the Protegrity Discover Web UI. If the metadata keyword contains any sensitive data, then the sensitive data will appear on the **Classifications** screen.

4. Click **Add** to add the Python boolean expression.

   The Python boolean expression is added to the **Keyword Patterns** area.

5. Repeat steps *2* to *4* to add another Python boolean expression for identifying the files that were created before 31st December 2010.

6. Specify the following details in the **Add New Metadata Keyword** dialog box.

Field	Description
Name	Specify a unique name for the Python boolean expression.
Type	Select **File_size**.
Boost	Specify a value to boost the confidence score. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the boost value is applied to the confidence score.
Score	Specify the confidence score value. The Python boolean expression is evaluated against the metadata and if the evaluation result is *true*, then the score value is added to the final confidence score. For example, specify the confidence score as *0.4*. **Important:** In this case, you want Protegrity Discover to search for sensitive data within only those files that were created by Jane Smith before 31st December 2010. However, you do not want the classifiers to skip these files. Therefore, in this case, you must ensure that the total confidence score of both the metadata must be less than 1. As a result, the classifiers will search all the files within the file system. But, they will assign a higher confidence score to the sensitive data present in the files created by Jane Smith before 31st December 2010.
Expression	Specify the Python boolean expression for identifying the file metadata keyword. In this case, you want to search for files that have been created before 31st December 2010. You can then specify the following expression. `%(created_time)s < datetime.datetime(2010, 12, 31, 23, 59, 59).timestamp()`
Continue	Select this check box to include this pattern and all subsequent patterns in the **Keyword Patterns** area to identify the metadata. If you clear this check box, then the corresponding expression is included for identifying the metadata. However, all subsequent expressions are excluded from identifying the metadata.

7. Click **Add** to add the Regex pattern.

   The Regex pattern is added to the **Keyword Patterns** area.

8. Navigate to the **Testbed** tab.

9. Add test data.

For more information about adding test data, refer to the *step 3* in the section *Testing the Classifier*.

10. Specify a value in the **Test Metadata JSON** area.

    For more information about specifying a value in the **Test Metadata JSON** area, refer to *step 5* in the section *Testing the Classifier*.

11. Click **Run Test** to test the metadata.

    The classifier evaluates the Python boolean expression specified in the **Metadata** tab using the values specified in the **Test Metadata JSON** area.

    Even if the evaluation is successful, as the combined confidence score of both the file metadata does not equal 1, the classifier tries to classify the data that you have either provided as free text or have specified in an uploaded file.

    If the evaluation is successful and the confidence score equals or exceeds 1, then the results appear in the **Test Results** area and the classifier is activated.