Discovery

Abstract

Easily find and assess the risk of non-compliance to GDPR, CCPA, HIPAA, and other regulations for data distributed across multiple cloud databases, analytics platforms, reporting systems, and geographies.

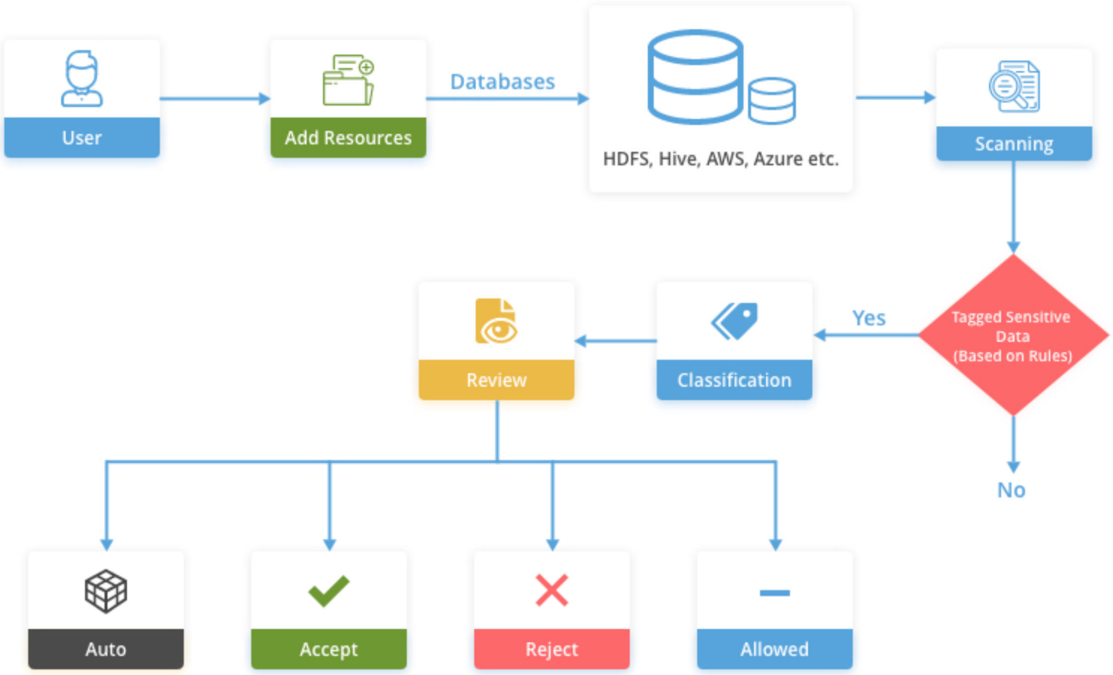
Table of Contents

Get started with Discovery

helps reveal information about your data and usage.

crawls computing assets such as databases or files, which are called *data sources*. It scans the data sources to identify sensitive information like credit card numbers, Social Security Numbers, and other personal, restricted, or confidential information.

classifies or labels this information to create a comprehensive catalog of your sensitive data. You can review these classifications to accept or reject them or refine the scanning via rules, dictionaries, models, and patterns.



The scans use a variety of techniques you can manipulate to identify and classify sensitive information:

- Pattern matching with regular expressions.
- Dictionaries to look up data from a whitelist or blacklist.
- Sophisticated heuristics that look at both the data content and the context in which the data is located, such as the table or column name.

By reviewing and updating the classifications generated by the scanners, you can further implement policies to protect sensitive data in conformance with your enterprise's requirements. For example, you might want a policy that allows non-privileged users to see only the masked or transformed versions of certain sensitive fields such as SSNs or credit card numbers.

By grouping data sources into administrative *data zones*, maintenance and control of the assets in these zones can be delegated to the owners of the data in your enterprise's organizational groups.

has a variety of reports to aggregate, summarize, drill down into the classification results across the entire collection of data assets.

Planning for

This is a general approach to setting up .

- Make sure that has been Install and Enable [11].
- Take an inventory of the data assets you want to monitor, including databases and tables and applications, such as or . Be sure to identify the data owners of those assets.
- Enable your data sources [36] for scanning in .
- Optionally, create data zones [93] of those data sources to delegate administration to your organizational groups.
- Define scans [75] to classify those data sources, including resource scoring, scanning schedules, which data sources to include, and which to exclude.
- Based on the system's classifying tags, refine the scans using the following techniques:
  - Dictionaries [58].
  - Regular expression patterns [62].
  - Models [64].
  - Rules [71].
- Establish organizational mechanisms to implement a compliance workflow, including the following:
  - Monitoring alerts [113].
  - Checking for movement of data sources across data zones [92].
  - Accepting or rejecting classifications [110] from scans.
- Enhance security by masking or encrypting database table, columns, rows, or other fields. For more information, see the Get started with Encryption.

Install and Enable

Use the Privacera Manager to install and enable .

See the following information:

- Set up Discovery on AWS for Privacera Platform [11].
- Set up Discovery on Azure for Privacera Platform [13].
- Set up Discovery on GCP for Privacera Platform [17].

Set up Discovery on Privacera Platform

Set up Discovery on AWS for Privacera Platform

This topic shows you how to set up the AWS configuration for installing Privacera Discovery in a Docker and Kubernetes (EKS) environment.

IAM policies for Discovery on AWS

To use the Privacera Discovery service, ensure the following IAM policies are attached to the `Privacera_PM_Role` role to access the AWS services.

The policy to create AWS resources is required only during installation or when Discovery is updated through Privacera Manager. This policy gives permissions to Privacera Manager to create AWS resources like DynamoDB, Kinesis, SQS, and S3 using terraform.

- ${AWS_REGION}: AWS region where the resources will get created.

```
 {
"Version":"2012-10-17",
"Statement":[
    {
        "Sid":"CreateDynamodb",
        "Effect":"Allow",
        "Action":[
            "dynamodb:CreateTable",
            "dynamodb:DescribeTable",
            "dynamodb:ListTables",
            "dynamodb:TagResource",
            "dynamodb:UntagResource",
            "dynamodb:UpdateTable",
            "dynamodb:UpdateTableReplicaAutoScaling",
            "dynamodb:UpdateTimeToLive",
            "dynamodb:DescribeTimeToLive",
            "dynamodb:ListTagsOfResource",
            "dynamodb:DescribeContinuousBackups"
        ],
        "Resource":"arn:aws:dynamodb:${AWS_REGION}:*:table/privacera*"
    },
    {
        "Sid":"CreateKinesis",
        "Effect":"Allow",
        "Action":[
            "kinesis:CreateStream",
            "kinesis:ListStreams",
            "kinesis:UpdateShardCount"
        ],
        "Resource":"arn:aws:kinesis:${AWS_REGION}:*:stream/privacera*"
    },
    {
        "Sid":"CreateS3Bucket",
        "Effect":"Allow",
        "Action":[
            "s3:CreateBucket",
            "s3:ListAllMyBuckets",
            "s3:GetBucketLocation"

        ],
        "Resource":[
            "arn:aws:s3:::*"
        ]
    },
    {
        "Sid":"CreateSQSMessages",
        "Effect":"Allow",
        "Action":[
            "sqs:CreateQueue",
            "sqs:ListQueues"
        ],
        "Resource":[
            "arn:aws:sqs:${AWS_REGION}:${ACCOUNNT_ID}:privacera*"
        ]
    }
]
}
```

CLI configuration for Discovery on AWS

1. SSH to the instance where Privacera is installed.
2. Run the following commands.

```
cd ~/privacera/privacera-manager
cp config/sample-vars/vars.discovery.aws.yml config/custom-vars/
vi config/custom-vars/vars.discovery.aws.yml
```

3. Edit the following properties. For property details and description, refer to the **Configuration Properties** below.

   ```
   DISCOVERY_BUCKET_NAME: "<PLEASE_CHANGE>"
   ```

   To configure a bucket, add the property as follows, where `bucket-1` is the name of the bucket:

   ```
   DISCOVERY_BUCKET_NAME: "bucket-1"
   ```

   To configure a bucket containing a folder, add the property as follows:

   ```
   DISCOVERY_BUCKET_NAME: "bucket-1/folder1"
   ```
4. Uncomment/Add the following variable to enable Autoscalability of Executor pods:

   ```
   DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_ENABLED: "true"
   ```
5. (Optional) If you want to customize Discovery configuration further, you can add custom Discovery properties. For more information, refer to Set custom Discovery properties on Privacera Platform [134].
   For example, by default, the username and password for the Discovery service is padmin/padmin. If you choose to change it, refer to Add custom properties using Privacera Manager on Privacera Platform.
6. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   ./privacera-manager.sh update
   ```

Configuration properties for Discovery on AWS

Property	Description	Example
`DISCOVERY_BUCKET_NAME`	Set the bucket name where Discovery will store its metadata files	container1
[Properties of Topic and Table names](../pm-ig/customize_topic_and_tables_names.md)	Topic and Table names are assigned by default in Privacera Discovery. To customize any topic or table name, refer to the link.	

Enable realtime scan

An AWS SQS queue is required, if you want to enable realtime scan on the S3 bucket.

After running the PM update command, an SQS queue will be created for you automatically with the name, `privacera_bucket_sqs_{{DEPLOYMENT_ENV_NAME}}`, where `{{DEPLOYMENT_ENV_NAME}}` is the environment name you set in the `vars.privacera.yml` file. This queue name will appear in the list of queues of your AWS SQS account.

If you have an SQS queue which you want to use, add the `DISCOVERY_BUCKET_SQS_NAME` property in the `vars.discovery.aws.yml` file and assign your SQS queue name.

If you want to enable realtime scan on the bucket, see Scan resources [24].

Set up Discovery on Azure for Privacera Platform

This topic allows you to setup the Azure configuration for installing Privacera Discovery.

Prerequisites

Ensure the following prerequisites are met:

Azure storage account

- Create an Azure storage account. For more information, refer to Microsoft's documentation Create a storage account.
- Create a private-access container. For more information, refer to Microsoft's documentation Create a container.
- Get the access key. For more information, refer to Microsoft's documentation: View account access keys.

Azure Cosmos DB account

- Create an Azure Cosmos DB.
- Get the URI from the **Overview** section.
- Get the Primary Key from the **Settings > Keys** section.
- Set the consistency to **Strong** in the **Settings > Default Consistency** section.

For Terraform

- Assign permissions to create Azure resources using **managed-identity**.

Procedure

1. SSH to the instance where Privacera is installed.
2. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   cp config/sample-vars/vars.kafka.yml config/custom-vars
   vi config/custom-vars/vars.kafka.yml
   ```
3. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   cp config/sample-vars/vars.discovery.azure.yml config/custom-vars
   vi config/custom-vars/vars.discovery.azure.yml
   ```
4. Edit the following properties. For property details and description, refer to the **Configuration Properties** below.

   ```
   DISCOVERY_FS_PREFIX: "<PLEASE_CHANGE>"
   DISCOVERY_AZURE_STORAGE_ACCOUNT_NAME: <PLEASE_CHANGE>"
   DISCOVERY_COSMOSDB_URL: <PLEASE_CHANGE>"
   DISCOVERY_COSMOSDB_KEY: "<PLEASE_CHANGE>"
   DISCOVERY_AZURE_STORAGE_ACCOUNT_KEY: "<PLEASE_CHANGE>"
   CREATE_AZURE_RESOURCES: "false"
   DISCOVERY_AZURE_RESOURCE_GROUP: "<PLEASE_CHANGE>"
   DISCOVERY_AZURE_COSMOS_DB_ACCOUNT: "<PLEASE_CHANGE>"
   DISCOVERY_AZURE_LOCATION: "<PLEASE_CHANGE>"
   ```
5. (Optional) If you want to customize Discovery configuration further, you can add custom Discovery properties. For more information, refer to Set custom Discovery properties on Privacera Platform [134].

   For example, by default, the username and password for the Discovery service is padmin/padmin. If you choose to change it, refer to Add custom properties using Privacera Manager on Privacera Platform.
6. To configure real-time scan for audits, refer to Enable Pkafka for real-time audits in Discovery on Privacera Platform [18].
7. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   ./privacera-manager.sh update
   ```

Configuration properties for Discovery on Azure

Property	Description	Example
`DISCOVERY_ENABLE`	In the **Basic** tab, enable/disable Privacera Discovery.	
`DISCOVERY_REALTIME_ENABLE`	In the **Basic** tab, enable/disable real-time scan in Privacera Discovery. For real-time scan to work, ensure the following: • If you want to scan the default ADLS app registered by the system at the time of installation, keep its app properties unchanged in Privacera Portal. • If you want to scan a user-registered app, the app properties in Privacera Portal and its corresponding discovery.yml should be the same. • At a time, only one app can be scanned.	
`DISCOVERY_FS_PREFIX`	Enter the container name. Get it from the **Prerequisites** section.	container1
`DISCOVERY_AZURE_STORAGE_ACCOUNT_NAME`	Enter the name of the Azure Storage account. Get it from the **Prerequisites** section.	azurestorage

`DISCOVERY_COSMOSDB_URL`
`DISCOVERY_COSMOSDB_KEY`
	Enter the Cosmos DB URL and Primary Key. Get it from the **Prerequisites** section.	`DISCOVERY_COSMOSDB_URL: "https://url1.documents.azure.com:443/"`
		`DISCOVERY_COSMOSDB_KEY: "xavosdocof"`

`DISCOVERY_AZURE_STORAGE_ACCOUNT_KEY`	Enter the Access Key of the storage account. Get it from the **Prerequisites** section.	GMi0xftgifp==
[Properties of Topic and Table names](../pm-ig/customize_topic_and_tables_names.md)	Topic and Table names are assigned by default in Privacera Discovery. To customize any topic or table name, refer to the link.	
`PKAFKA_EVENT_HUB`	In the **Advanced > Pkafka Configuration** section, enter the Event Hub name. Get it from the **Prerequisites** section.	eventhub1
`PKAFKA_EVENT_HUB_NAMESPACE`	In the **Advanced > Pkafka Configuration** section, enter the name of the Event Hub namespace. Get it from the **Prerequisites** section.	eventhubnamespace1
`PKAFKA_EVENT_HUB_CONSUMER_GROUP`	In the **Advanced > Pkafka Configuration** section, enter the name of the Consumer Group. Get it from the **Prerequisites** section.	congroup1

`PKAFKA_EVENT_HUB_CONNECTION_STRING`	In the **Advanced > Pkafka Configuration** section, enter the connection string. Get it from the **Prerequisites** section.	Endpoint=sb://eventhub1.servicebus.windows.net/;
		SharedAccessKeyName=RootManageSharedAccessKey;
		SharedAccessKey=sAmPLEP/8PytEsT=

`CREATE_AZURE_RESOURCES`	For terraform usage, assign the value as **true**. Its default value is false.	true
`DISCOVERY_AZURE_RESOURCE_GROUP`	Get the value from the Prerequisite section.	resource1
`DISCOVERY_AZURE_COSMOS_DB_ACCOUNT`	Get the value from the Prerequisite section.	database1

Set up Discovery on Databricks for Privacera Platform

This topic covers the installation of Privacera Discovery on Databricks.

1. SSH to the instance as USER.
2. Run the following commands.

```
cd ~/privacera/privacera-manager
cp config/sample-vars/vars.discovery.databricks.yml config/custom-vars/
vi custom-vars/vars.discovery.databricks.yml
```

3. Add and provide the following details in *custom-vars/vars.discovery.databricks.yml* file if the Databricks plugin is not enabled. To configure Databricks plugin, see Configure Databricks Spark Fine-Grained Access Control Plugin [FGAC] [Python, SQL].

   ```
   DATABRICKS_HOST_URL: "<PLEASE_UPDATE>"
   DATABRICKS_TOKEN: "<PLEASE_UPDATE>"

   DATABRICKS_WORKSPACES_LIST:
   - alias: DEFAULT
       databricks_host_url: "{{DATABRICKS_HOST_URL}}"
       token: "{{DATABRICKS_TOKEN}}"
   ```

4. Edit the following properties. For property details and description, refer to the **Configuration Properties** below.
   AWS

   ```
   DATABRICKS_DRIVER_INSTANCE_TYPE: "m5.xlarge"
   DATABRICKS_INSTANCE_TYPE: "m5.xlarge"
   DATABRICKS_DISCOVERY_MANAGE_INIT_SCRIPT: "true"
   DATABRICKS_DISCOVERY_SPARK_VERSION: "7.3.x-scala2.12"
   DATABRICKS_DISCOVERY_INSTANCE_PROFILE: "arn:aws:iam::<ACCOUNT_ID>:instance-profile/<D
   DISCOVERY_AWS_CLOUD_ASSUME_ROLE: "true"
   DISCOVERY_AWS_CLOUD_ASSUME_ROLE_ARN: "arn:aws:iam::<ACCOUNT_ID>:role/<DISCOVERY_IAM_R
   ```

   Azure

   ```
   DATABRICKS_DRIVER_INSTANCE_TYPE: "Standard_DS3_v2"
   DATABRICKS_INSTANCE_TYPE: "Standard_DS3_v2"
   DATABRICKS_DISCOVERY_MANAGE_INIT_SCRIPT: "true"
   DATABRICKS_DISCOVERY_SPARK_VERSION: "7.3.x-scala2.12"
   ```

NOTE

PRIVACERA_DISCOVERY_DATABRICKS_DOWNLOAD_URL is no longer in use. The Discovery Databricks packages will be downloaded from PRIVACERA_BASE_DOWNLOAD_URL.

Databricks Discovery configuration properties

Property	Description	Example
DATABRICKS_DRIVER_INSTANCE_TYPE	For AWS driver's instance type can be "m5.xlarge" or "m5.2xlarge" For Azure driver's instance type can be "Standard_DS3_v2"	m5.xlarge
DATABRICKS_INSTANCE_TYPE	For AWS driver's instance type can be "m5.xlarge" or "m5.2xlarge" For Azure driver's instance type can be "Standard_DS3_v2"	m5.xlarge
SETUP_DATABRICKS_JAR		
USE_DATABRICKS_SPARK		

Property	Description	Example
DATABRICKS_ELASTIC_DISK		
DATABRICKS_DISCOVERY_MANAGE_INIT_SCRIPT	Set to true if you want to create databricks init script.	false
DATABRICKS_DISCOVERY_WORKERS		
DATABRICKS_DISCOVERY_JOB_NAME		
DATABRICKS_DISCOVERY_SPARK_VERSION	Spark version can be as follows: • 6.4.x-scala2.11 (Spark 2.4) • 7.3.x-scala2.12 (Spark 3.0) • 7.4.x-scala2.12 (Spark 3.0) • 7.5.x-scala2.12 (Spark 3.0) • 7.6.x-scala2.12 (Spark 3.0)	7.3.x-scala2.12
DATABRICKS_DISCOVERY_INSTANCE_PROFILE	Property is used for the instance role, for the Databricks instance node where your discovery will be running	arn:aws:iam::1234564835:instance-profile/privacera_databricks_cluster_iam_role
DISCOVERY_AWS_CLOUD_ASSUME_ROLE	Property to grant Discovery access to AWS services to perform the scanning operation.	true
DISCOVERY_AWS_CLOUD_ASSUME_ROLE_ARN	ARN of the AWS IAM Role	arn:aws:iam::12345671758:role/DiscoveryCrossAccAssumeRole_k

Set up Discovery on GCP for Privacera Platform

This topic allows you to set up the GCP configuration for installing Privacera Discovery in a Docker and Kubernetes environment.

Prerequisites for setting up Discovery on GCP

Ensure the following prerequisites are met:

- Create a service account and add the following roles. For more information, refer to Creating a new service account.
  - Editor
  - Owner
  - Private Logs Viewer
  - Kubernetes Engine Admin (Required only for a Kubernetes environment)
- Create a Bigtable instance and get the Bigtable Instance ID. For more information, refer to Creating a Cloud Bigtable instance.

CLI configuration for Discovery on GCP

1. SSH to the instance where Privacera is installed.
2. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   cp config/sample-vars/vars.discovery.gcp.yml config/custom-vars/
   vi config/custom-vars/vars.discovery.gcp.yml
   ```
3. Edit the following properties. For property details and description, refer to the **Configuration Properties** below.

   ```
   BIGTABLE_INSTANCE_ID: "<PLEASE_CHANGE>"
   DISCOVERY_BUCKET_NAME: "<PLEASE_CHANGE>"
   ```
4. (Optional) If you want to customize Discovery configuration further, you can add custom Discovery properties. For more information, refer to Set custom Discovery properties on Privacera Platform [134].

   For example, by default, the username and password for the Discovery service is padmin/padmin. If you choose to change it, refer to Add custom properties using Privacera Manager on Privacera Platform.

5. For real-time scanning, run the following.

```
cd ~/privacera/privacera-manager
cp config/sample-vars/vars.pkafka.gcp.yml config/custom-vars/
```

> **NOTE**
>
> - Recommended: Use **Google Sink** based approach to enable real-time scan of applications on different projects, click here [86].
> - Optional: Use **Google Logging API** based approach to enable real-time scan of applications on different projects, click here [82].

6. Run the following commands.

```
cd ~/privacera/privacera-manager
./privacera-manager.sh update
```

Configuration properties for Discovery on GCP

Property	Description	Example
BIGTABLE_INSTANCE_ID	Get the value by navigating to **Navigation Menu > Databases > BigTable > Check the instance id column**.	BIGTABLE_INSTANCE_ID: "table_1"
DISCOVERY_BUCKET_NAME	Give a name where the Discovery will store it's metadata files.	DISCOVERY_BUCKET_NAME="bucket_1"

Enable Pkafka for real-time audits in Discovery on Privacera Platform

This topic shows you how to enable Pkafka for real-time audits in Privacera Discovery.

Prerequisites

- Create an Event Hub namespace with a region similar to the region of a Storage Account you want to monitor. For more information, refer to Microsoft's documentation: Create an Event Hubs namespace
- Create an Event Hub in the Event Hub namespace. For more information, refer to Microsoft's documentation: Create an event hub
- Create a consumer group in the Event Hub by going to Azure Portal > Event Hubs namespace > Event Hub > Consumer Groups > +Consumer Group. The **Consumer Groups** tab will be under **Entities** of the **Event Hub** page.
- Get the connection string of the Event Hub namespace. For more information, refer to Microsoft's documentation: Get connection string for a namespace
- Create an Event Subscription for the Event Hub namespace with the **Event Type** as **Blob Created** and **Blob Deleted**. For more information, refer to Microsoft's documentation: Integration with Event Grid

> **NOTE**
>
> When you create an event grid subscription, clear the checkbox **Enable subject filtering**.

Procedure

To enable Pkafka, follow these steps:

1. SSH to the instance where Privacera is installed.
2. Run the following commands.

```
cd ~/privacera/privacera-manager
cp config/sample-vars/vars.pkafka.azure.yml config/custom-vars/
vi config/custom-vars/vars.pkafka.azure.yml
```

3. Edit the following properties. For property details, see Pkafka configuration properties [19].

```
PKAFKA_EVENT_HUB: "<PLEASE_CHANGE>"
PKAFKA_EVENT_HUB_NAMESPACE: "<PLEASE_CHANGE>"
PKAFKA_EVENT_HUB_CONSUMER_GROUP: "<PLEASE_CHANGE>"
PKAFKA_EVENT_HUB_CONNECTION_STRING: "<PLEASE_CHANGE>"
DISCOVERY_REALTIME_ENABLE: "true"
```

4. Run the following commands.

```
cd ~/privacera/privacera-manager
./privacera-manager.sh update
```

Pkafka configuration properties

Property	Description	Example
PKAFKA_EVENT_HUB	Enter the Event Hub name. Get it from the **Prerequisites** section above.	eventhub1
PKAFKA_EVENT_HUB_NAMESPACE	Enter the name of the Event Hub namespace. Get it from the **Prerequisites** section above.	eventhubnamespace1
PKAFKA_EVENT_HUB_CONSUMER_GROUP	Enter the name of the Consumer Group. Get it from the **Prerequisites** section above.	congroup1

PKAFKA_EVENT_HUB_CONNECTION_STRING	Enter the connection string. Get it from the **Prerequisites** section above.	Endpoint=sb://eventhub1.servicebus.windows.net/;
		SharedAccessKeyName=RootManageSharedAccessKey;
		SharedAccessKey=sAmPLEP/8PytEsT=

DISCOVERY_REALTIME_ENABLE	Add this property to enable/disable real-time scan. By default, it is set to false. **NOTE** This is a custom property, and has to be added separately to the YAML file. For real-time scan to work, ensure the following: • If you want to scan the default ADLS app registered by the system at the time of installation, keep its app properties unchanged in Privacera Portal. • If you want to scan a user-registered app, the app properties in Privacera Portal and its corresponding `discovery.yml` should be the same. • At a time, only one app can be scanned.	true

Customize topic and table names on Privacera Platform

By default, topic and table names are assigned and managed internally by Privacera Discovery. Also, the deployment environment name is attached as a suffix to the topic and table names.

For example, the default name for a Classification Topic in Privacera Discovery is shown as below:

```
CLASSIFICATION_TOPIC: "privacera_classification_info_{{DEPLOYMENT_ENV_NAME}}"
```

To customize the name of a topic or table, you can do one of the following:

- Remove the {{DEPLOYMENT_ENV_NAME}} variable as suffix.
- Re-define a new topic/table name.

If you want to customize any topic or table name, refer to the property in the the following table.

- Uncomment the topic, and enter a name, along with the {{DEPLOYMENT_ENV_NAME}} as the suffix.
- To remove the {{DEPLOYMENT_ENV_NAME}} as the suffix, refer to the `DISCOVERY_DEPLOYMENT_SUFFIX_ID` property in this table.
- {{DEPLOYMENT_ENV_NAME}} is the name of the environment you have given in the `vars.privacera.yml`

Property	Description	Example customization
`PRIVACERA_PORTAL_TOPIC_DYNAMIC_PREFIX`	Uncomment and enter a custom name to add a prefix to the real-time topic for Data Sources in Privacera Portal.	`PRIVACERA_PORTAL_TOPIC_DYNAMIC_PREFIX="privacera_scan_worker"`
`CLASSIFICATION_TOPIC`	Streams Privacera Discovery classification information [110] generated after scanning for consumers to post-process, such as writing the data to Solr	`CLASSIFICATION_TOPIC: "privacera_classification_info_{{DEPLOYMENT_ENV_NAME}}"`
`ALERT_TOPIC`	Streams alert [123] data which consumers to post-process, such as writing the data to Solr	`ALERT_TOPIC: "privacera_alerts_{{DEPLOYMENT_ENV_NAME}}"`
`SPARK_EVENT_TOPIC`	Streams Spark events for debugging purpose	`SPARK_EVENT_TOPIC: "privacera_spark_events_{{DEPLOYMENT_ENV_NAME}}"`
`RESULT_TOPIC`	Streams error logs consumers to post-process, such as writing the data to Solr for displaying on the Privacera Portal diagnostic page	`RESULT_TOPIC: "privacera_results_{{DEPLOYMENT_ENV_NAME}}"`
`OFFLINE_SCAN_TOPIC`	Streams batch file events after listing, which is consumed by Privacera Discovery to initiate scanning of batch file [37]	`OFFLINE_SCAN_TOPIC: "privacera_offline_scan_{{DEPLOYMENT_ENV_NAME}}"`
`AUDITS_TOPIC`	Streams real-time audit events consumed by Privacera Discovery for real-time scanning [37]	`AUDITS_TOPIC: "privacera_audits_{{DEPLOYMENT_ENV_NAME}}"`
`SCAN_RESOURCE_INFO_TOPIC`	Streams data for scan summary information [116] reporting about scan request jobs	`SCAN_RESOURCE_INFO_TOPIC: "privacera_scan_resources_info_{{DEPLOYMENT_ENV_NAME}}"`
`RIGHT_TO_PRIVACY_TOPIC`	Streams events for triggering the Right to Privacy compliance policy [103]	`RIGHT_TO_PRIVACY_TOPIC: "privacera_right_to_privacy_{{DEPLOYMENT_ENV_NAME}}"`
`DELAY_QUEUE_TOPIC`	Streams real-time events to HDFS for delayed processing	`DELAY_QUEUE_TOPIC: "privacera_delay_queue_{{DEPLOYMENT_ENV_NAME}}"`
`APPLY_SCHEME_TOPIC`	Streams events for triggering the de-identification compliance policy [98]	`APPLY_SCHEME_TOPIC: "privacera_apply_scheme_{{DEPLOYMENT_ENV_NAME}}"`
`ML_CLASSIFY_TAG_TOPIC`	Streams events for triggering tag detection via Machine Learning Models [64]	`ML_CLASSIFY_TAG_TOPIC: "privacera_ml_classify_tag_{{DEPLOYMENT_ENV_NAME}}"`
`LINEAGE_TOPIC`	Streams lineage information for consumers for writing the data to Solr	`LINEAGE_TOPIC: "privacera_lineage_{{DEPLOYMENT_ENV_NAME}}"`

Property	Description	Example customization
RESOURCE_TABLE ALERT_TABLE SCAN_REQUEST_TABLE ACTIVE_SCANS_TABLE MLRESOURCE_TABLE LINEAGE_TABLE AUDIT_SUMMARY_TABLE STATE_TABLE SCAN_STATUS_TABLE	You can customize the table names. Uncomment the table, and enter a name, along with the {{DEPLOYMENT_ENV_NAME}} as the suffix. To remove the {{DEPLOYMENT_ENV_NAME}} as the suffix, refer to the DISCOVERY_DEPLOYMENT_SUFFIX_ID property in this table.	RESOURCE_TABLE: "privacera_resource_v2_{{DEPLOYMENT_ENV_NAME}}" ALERT_TABLE: "privacera_alert_{{DEPLOYMENT_ENV_NAME}}" SCAN_REQUEST_TABLE: "privacera_scan_request_{{DEPLOYMENT_ENV_NAME}}" ACTIVE_SCANS_TABLE: "privacera_active_scans_{{DEPLOYMENT_ENV_NAME}}" MLRESOURCE_TABLE: "privacera_mlresource_v2_{{DEPLOYMENT_ENV_NAME}}" LINEAGE_TABLE:"privacera_lineage_{{DEPLOYMENT_ENV_NAME}}" AUDIT_SUMMARY_TABLE: "privacera_audit_summary_{{DEPLOYMENT_ENV_NAME}}" STATE_TABLE: "privacera_state_{{DEPLOYMENT_ENV_NAME}}" SCAN_STATUS_TABLE: "privacera_scan_status_{{DEPLOYMENT_ENV_NAME}}"
DISCOVERY_DEPLOYMENT_SUFFIX_ID	Use this property to remove the {{DEPLOYMENT_ENV_NAME}} variable as suffix from the topic/table names. **NOTE** This is a custom property, and has to be added separately to the YAML file.	DISCOVERY_DEPLOYMENT_SUFFIX_ID: ""
DISCOVERY_BUCKET_SQS_NAME	You can customize the SQS bucket name. Uncomment the table, and enter a name, along with the {{DISCOVERY_DEPLOYMENT_SUFFIX_ID}} as the suffix.	DISCOVERY_BUCKET_SQS_NAME: "privacera_bucket_sqs_{{DISCOVERY_DEPLOYMENT_SUFFIX_ID}}"

Enable Discovery on PrivaceraCloud

To enable Discovery, click the **Discovery** toggle button.

Enable real-time scanning of an S3 bucket

To enable real-time scanning on an S3 bucket, do the following steps. This step assumes you have an existing setup of an SQS account with a queue created. If you do not have an SQS account, set up an account and then create a queue.

1. Click **ENABLE** button next to the Real-Time Scanning.
   The **Real-Time Scanning Info** dialog appears.
2. Get the following information from the SQS account and enter them in the given fields:
   - With **Use IAM Role** disabled:
     - SQS Endpoint
     - SQS Access Key

    - SQS Secret Key
    - SQS Region
    - SQS Queue Name
  - With **Use IAM Role** enabled:
    - SQS Endpoint
    - SQS IAM Role
    - SQS Region
    - SQS Queue Name
3. Click **Test Connection** to check if the connection is successful, and then click **Save Settings**.

Thereafter, use the toggle to either disable or enable real-time scanning, and use the pen icon to modify the existing configuration.

Configure real-time scanning for ADLS on PrivaceraCloud

For real-time scanning to be configured, you need to configure an Event Hub. It will process all the events sent from the storage container, whenever a new resource gets added.

Event Hub requires a storage account to store checkpoint information. Checkpointing is a process by which readers (i.e Pkakfa) mark or commit their position within a partition event sequence. In this case, blob storage container is used for storing checkpoints while processing events from Event Hubs.

1. Configure Event Hub:
   a. Create an Event Hub namespace with a region similar to the region of a Storage Account you want to monitor. Refer to Microsoft documentation on how to Create an Event Hubs namespace .
      Use this Event Hub namespace name in **Eventhub Namespace**.
   b. Create an Event Hub in the Event Hub namespace. Refer to Microsoft documentation on how to Create an event hub .
      Use this event hub name in **Eventhub Name**.
   c. Get Eventhub Sas Key Name and Eventhub Sas key:
      i. Navigate to **Event hub namespace** > **Event hub**.
      ii. Under **Settings**, click **Shared access policies**.
      iii. Click **+Add** to create a new Sas policy.
           The **Add SAS Policy** section is displayed on the right.
      iv. Enter a policy name and select appropriate claims.
      v. Click the new policy to populate keys.
         Use the policy name in **Eventhub Sas Key Name**, and use either the Primary key or Secondary key in **Eventhub Sas key**.
2. Create Consumer Group for Pkafka:
   a. Navigate to **Event Hubs namespace** > **Event Hub** > **Consumer Groups** > **+Consumer Group**. The Consumer Groups tab will be under **Entities** of the Event Hub page.
   b. Create a consumer group with name as **pkafkagroup1**.
3. Configure Checkpoint Storage for Pkafka:
   a. Get Eventhub Storage Account Name:
      Use an existing storage account or create a storage account to use with Eventhub. Refer to Microsoft documentation on how to Create a Storage Account.
      Use this storage account name in **Eventhub Storage Account Name**.
   b. Get Eventhub Storage Account Key:
      i. Navigate to the storage account.
      ii. Under **Security + networking**, click **Access keys**.
      iii. Click **Show Keys** for keys to be populated.
      iv. Use **Key1** value in **Eventhub Storage Account Key**.
   c. Get Eventhub Storage Container Name:

Use an existing container name or create a storage container to use with Eventhub. Refer to Microsoft documentation on how to Create a Container .

Use this container name in **Eventhub Storage Container Name**.

d. Get the Eventhub URL Prefix:

i. Navigate to the container.

ii. Open the container and click **Properties**, container property details are populated on the right.

iii. Use the URL prefix in **Eventhub Storage Url Prefix**.

4. Enable Real-Time Scan:

a. Click **ENABLE** button next to the Real-Time Scanning.

The **Real-Time Scanning Info** dialog appears.

b. Provide the following information:

- Eventhub Namespace
- Eventhub Name
- Eventhub Sas Key Name
- Eventhub Sas key
- Eventhub Storage Url Prefix
- Eventhub Storage Account Name
- Eventhub Storage Account Key
- Eventhub Storage Container Name

c. Click **Test Connection** to check if the connection is successful, and then click **Save Settings**.

Thereafter, use the toggle to either disable or enable real-time scanning, and use the pen icon to modify the existing configuration.

Scan resources

The key feature of Privacera Discovery is scanning your data for sensitive data. Scanning gives you tagged-classifications of your data resources that you can analyze and refine.

Supported file formats for Discovery Scans

can scan the following file formats:

- Structured data with taggable content and metadata:
  - .avro
  - .avro (nested)
  - .csv
  - .html
  - .json
  - .json (nested)
  - .orc
  - .parquet
  - .parquet (nested)
  - .sas
  - .tsv
  - .xls
  - .xlsx
  - .xml
- Compressed/archive data with taggable content and metadata:
  - .gzip (single or multiple files)
  - .gz (single or multiple files)
  - .lzo/.lzop
  - .jar (single or multiple files)
  - .tar.gz (single or multiple files)
  - .snappy.parquet
  - .snappy.orc
  - .snappy.avro
  - .zip (single or multiple files)
  - .zlib.orc
  - .zlib.parquet
  - .zlib.avro
- Unstructured data with taggable content and metadata:
  - .dat
  - .doc
  - .docx
  - .pdf
  - .txt
- Media data with taggable metadata:

  > **NOTE**
  >
  > For the following file formats, Discovery only supports metadata extraction.

  - .jpeg
  - .mp4

- .mpeg

scan targets

crawls targeted data sources to identify and applies metadata labels called *tags* to potentially sensitive data, such as credit card numbers or email addresses.

Access Manager *Tag Policies* can then be created so that user access can be controlled and monitored.

is enabled by default.

Disable or reenable on PrivaceraCloud

is enabled by default.

The account owner or the administrator can disable or reenable .

To disable or reenable :

1. Go to **Settings > Account**.
2. Under the **Discovery** heading, toggle **Enable Discovery** to **ON** or **OFF**.

Connect Applications

You can configure the applications supported by in **Settings > Applications** . For more information, see Data sources on Privacera Platform [26].

Discovery scan targets

After an application has been connected, subsets of that data can be configured as scan targets.

To define a scan target:

1. To add specific databases and tables to be scan targets, go to **Discovery > Data Source**.
2. Click the desired application in the **Applications** section.
3. Click **ADD** to define a subset of that Datasource for scanning.
   - Enter the database name (*database* or for Snowflake *database*.*schema*) or wildcard asterisk for all databases.
   - Enter one or more comma-separated table names or wildcard asterisk for all tables. Wildcard asterisks can also be used in table names as prefix, suffix, or inside the name.

Start a scan

After a scan target is established, it can be scanned. Each database/table set is listed by row, under **Scanning Details**. The columns are: **Database**, **Tables**, and **Actions**.

To start a scan, click **SCAN RESOURCE**.

The following message is displayed:

> Scan request is in the queue, please check after 2 minutes.

View a scan

Completion status and various statistics for all scan **Discovery > Scan Status**.

> **NOTE**
>
> - During JDBC database scanning, data for binary type columns is skipped and only non-binary columns data is scanned.
> - During JDBC database scanning, data for string or text column types is trimmed to avoid Out of Memory error.

Processing order of scan techniques

applies tags [41] to dataset attributes using defined rules [71] . This is done by comparing data against dictionaries [58] and models [64]. The application of tags depends on the order of relevant rules. After a rule is triggered, the rest of the relevant rules are not processed.

After creating rules, you can reorder them into the necessary sequence to ensure that your data is tagged appropriately. See Reorder Structured Rules [74] for more information.

Register data sources on Privacera Platform

Data sources on Privacera Platform

With Privacera Discovery on Privacera Platform, you connect your third-part application data to Privacera so you can scan it for snsitive data.

Add a system data source on Privacera Platform

1. From the Privacera main menu, scroll down to **Settings** and click **Data Source Registration**.
2. From the **Data Source Registration** page, click **+ Add System**.
3. Enter System Name in the **Name** field. (*Required*).
4. Enter a brief description in the **Description** field. (*Optional*)
5. Click **Save**.
   Your new entry appears upon page refresh.

Add a resource data source on Privacera Platform

1. Select the settings icon in a data source detail box to add resources to your system. Resources can be applications, tables, or filesystems.
2. Select an application from the drop-down menu.
3. Enter a **Name**, an optional **Description**, and an **Application Code** in the **Application Detail** dialog box.
4. Set the status toggle to **Enable**.
5. Click **Save**.

   > **NOTE**
   > You can optionally test your data source connection at this point by selecting **Test Connection**.

6. Select the **Application Properties** tab. You can import exsting application properties from a file using the **Import Properties** option. Open a browser window, select a **JSON** file, and click **Add**.
7. In the **Add New Properties** section, add the following properties for **Dataserver**. Add one property per line.
   SSL: If SSL is enabled for Dataserver, use the following properties.

```
explorer_proxy_enable=true
explorer_proxy_host=dataserver
explorer_proxy_port=8282
explorer_proxy_protocol=https
explorer_protocol=http
```

Non-SSL: If SSL is **not** enabled for Dataserver, use the following properties.

```
explorer_proxy_enable=true
explorer_proxy_host=dataserver
explorer_proxy_port=8181
explorer_proxy_protocol=http
explorer_protocol=http
```

8. Click **Test Connection**.
9. Click **Next**.

> **NOTE**
>
> To minimize the inflow of audits to Privacera, there is an option to add inclusion filter support for CDH (HDFS and Hive).

A success banner displays upon a successful addition.

Add AWS S3 application data source on Privacera Platform

The following steps show you how to add an AWS S3 application. You can allows users to access multiple S3 accounts using `AssumeRole`.

1. Create an AWS S3 application on the Privacera Platform Portal.
   a. Click **Setting** > **+ Add Application**.
   b. Select **AWS S3 Application**.
   c. Enter the **Application Name** and **Application Code**.
   d. Select the **Application Properties** tab.
   e. You can import existing application properties from a file using the **Import Properties** option. Browse and select the **JSON** file and click **Add**.
   f. Enable **Folder name tagging** toggle button to include folder names during scanning and to tag the folders based on dictionary values.
   g. Under **Add New Properties**, add the following for **Dataserver**. Add one property per line.
      - **SSL**: If SSL is enabled for Dataserver, use the following properties:

        ```
        explorer_proxy_enable=true
        explorer_proxy_host=dataserver
        explorer_proxy_port=8282
        explorer_proxy_protocol=https
        explorer_protocol=http
        ```
      - **Non-SSL**: If SSL is **not** enabled for Dataserver, use the following properties:

        ```
        explorer_proxy_enable=true
        explorer_proxy_host=dataserver
        explorer_proxy_port=8181
        explorer_proxy_protocol=http
        explorer_protocol=http
        ```
   h. Click **Test Connection**.
   i. Click **Next**.
      When the AWS S3 application is added successfully a success banner is displayed.

2. Create one more AWS S3 application following the above steps, and add the following custom property:
   `explorer_assume_role_arn=arn:aws:iam::${111111111111}:role/${s3_assume_role}`

TIP

To minimize the in-flow of audits to Privacera audits, there is an option to add inclusion filter support for CDH (HDFS and Hive).

Add Azure ADLS data source on Privacera Platform

The following steps show you how to add an Azure ADLS application:

1. Click **Setting** > **+ Add Application**.
2. Select **Azure ADLS**.
3. Enter the **Application Name** and **Application Code**.
4. Select the **Application Properties** tab. You can import existing application properties from a file using the **Import Properties** option. Browse and select the **JSON** file and click **Add**.
5. Enable **Folder name tagging** toggle button to include folder names during scanning and to tag the folders based on dictionary values.
6. Under **Add New Properties**, add the following for **Dataserver**. Add one property per line.
   - **SSL**: If SSL is enabled for Dataserver, use the following properties:

     ```
     explorer_proxy_enable=true
     explorer_proxy_host=dataserver
     explorer_proxy_port=8282
     explorer_proxy_protocol=https
     explorer_protocol=http storage_type=blob
     ```
   - **Non-SSL**: If SSL is **not** enabled for Dataserver, use the following properties.

     ```
     explorer_proxy_enable=true
     explorer_proxy_host=dataserver
     explorer_proxy_port=8181
     explorer_proxy_protocol=http
     explorer_protocol=http storage_type=blob
     ```
7. Click **Test Connection**.
8. Click **Next**.
   When the Azure ADLS application is added successfully a success banner is displayed.

Add Databricks Spark SQL data source on Privacera Platform

Prerequisites

Have the following details ready to enter into the data source definition in Privacera:

- A username and password in the target system that has read/write permission.
- The name of the JDBC driver you need.
- A JDBC connection string to communicate with the target data source.

To add Databricks Spark SQL data source in Privacera Platform:

Procedure

1. Navigate to **Settings** > **Data Source Registration**.
2. (Optional) Click **Add System** or modify an existing data source.
3. Enter a name and description for the data source and click **Save**.
4. Locate the new data source system name and from the wrench icon select **Add Data Source**.
5. In the **Add Data Source** dialog, on the **Choose** tab, select **Databricks Spark SQL**.
6. On the **Configure** tab:
   -
7. Enter a required **Application Name** of your choice.
8. Enter a required **Application Code** of your choice. This is an identifier for your own use.
9. If you have prepared a properties file in JSON format, click **Import Properties** and load the file.
10. Scroll to find the following properties and enter the values you prepared:
    a. `jdbc.username`: Enter the Email ID used to login to the Databricks account console.
    b. `jdbc.password`: On Databricks account console:
       i. Navigate to **Settings** > **User Settings** > **Access Tokens**.
       ii. Click **Generate New Token**.
       iii. Use the Token as your password.
    c. `jdbc.url`: On the Databricks account console:
       i. Click **Compute** and select the Cluster.
       ii. Navigate to **Advance Options** and click **JDBC/ODBC** tab.
       iii. Copy the URL from the **JBDC URL** section and update as shown in the following example:

```
Original URL:
jdbc:spark://<yourHostname>:443/default;transportMode=http;ssl=1;httpPath=sql/

New URL:
jdbc:hive2://<yourHostname>:443/default;transportMode=http;ssl=true;httpPath=s
```

11. Accept the default values for all other properties or modify them if needed.
12. At the bottom left, to verify the properties, click **Test Connection**.

> **NOTE**
> Your Databricks cluster should be up and running before clicking Test Connection.

13. At the bottom right, click **Next** to save the data source or **Back** to return to the **Choose** tab.

Add Google BigQuery (GBQ) data source on Privacera Platform

1. From the Privacera main menu, open **Settings**, and click **Data Source Registration**.
2. Add a **System** with the name **GBQ**.
3. Click the **Setting** icon of your added system, and click **+ Add Application**.
4. Choose **Google BigQuery** as the application.
5. Enter the following:
   - Name
   - Description
   - Application Code
6. (Optional) Enable **Status**.
7. Click **Save**.
8. In the **Application Properties** section, add the following properties:
   - Enter the **Google Project Id** (*Required*).
   - **Default Datasource for RealTime Scan** - This value is set to `false` by default. Set this value to `true` when adding the data source for a default project.

- **Enable Auto Scan Real-Time**: Enter `true` for Discovery to auto scan real-time and not check the included resources.
- Click **Save**.

Add Google Pub-Sub data source on Privacera Platform

1. From the navigation menu, select **Settings** > **Data Source Registration**.
2. Under your GCP system, select **+Add New Data Source** > **Google Cloud Storage**.
3. In the **Add Data Source** dialog, select and enter the following properties:
   - **Google Project Id**: `${PROJECT_ID}` (*Required*)
   - **scan.result.topic**: `${Scan_Topic_Name}` (*Required*). Use the same topic name you created as part of the prerequisite steps.
   - **scan.result.project.id**: `${Specify_ID_of_Cross_Project}`. If you do not specify a project ID, the system will consider applying a default project ID.
4. Click **Test Connection** to verify the configuration.
5. Click **Save**.

Add Google Cloud Storage data source on Privacera Platform

There are two ways to add a Google Cloud Storage (GCS) data source on Privacera Platform:

- Using a credential file [30]
- Using a Project ID [31]

Add Google Cloud Storage data source on Privacera Platform using a credential file

A credential type is a JSON file downloaded from the GCP that allows you to access the GCP service account from outside. Attaching this credential file will give access to the resources in the environment which can be used to run Discovery scans on GCP resources, such GCS or GBQ.

There are two ways to incorporate the credential file.

- **Local File Path**: Provide the path of the local file system to where the credential file is saved, and the system will read and copy internally to configuration location.
- **File**: Upload the credential file using a browser, and the system will copy internally to configuration location.

To add a GCS data source with credential file type, do the following:

1. Under GCP, add a new **Data Source**, then select **Google Cloud Storage**.
2. Enter the following:
   - **Name**: A name is provided by default. If required, enter a preferred name.
   - **Description**: Enter a suitable description
   - **Application Code**: An application code is an unique identifier for a data source. A code is provided by default. If required, enter a preferred code. No two data sources can have the same application code.
3. In the **Application Properties** section, add the following properties:
   - **Credential Type**: Select **Google Credentials Local File Path** from the drop-down list.
   - **Google Credentials Local File Path**: `/tmp`
   - **Google Project Id**: `${PROJECT_ID}`
   - **Default Datasource for RealTime Scan** : This value is set to **false** by default. Set this value to **true** if you have more than one data source. In such scenarios, it is recommended that you identify one data source as the default data source which will be used for real-time scanning.
   - Enable **Folder name tagging** toggle button to include folder names during scanning and to tag the folders based on dictionary values.

- **Fast Track Data Zones List**: Enter fast track data zones list. Ensure to enable `DISCOVERY_FASTTRACK_REALTIME_ENABLED` property, for more information see Discovery Properties [134].
- **Enable Auto Scan Real-Time**: Enter `true` for Discovery to auto scan real-time and not check the included resources.

> **NOTE**
>
> If **Enable Auto Scan Real-Time** is set to `true`, ensure that quarantine location, transfer location, and archive location are in the exclude resources to avoid re-scanning.

4. Scroll down to the bottom of the screen, and under **Add new properties** enter the following properties:
   - **SSL**: If SSL is enabled for Dataserver, use the following properties.

     ```
     explorer_proxy_enable=true
     explorer_proxy_host=dataserver
     explorer_proxy_port=8282
     explorer_proxy_protocol=https
     explorer_protocol=http
     ```
   - **Non-SSL**: If SSL is **not** enabled for Dataserver, use the following properties.

     ```
     explorer_proxy_enable=true
     explorer_proxy_host=dataserver
     explorer_proxy_port=8181
     explorer_proxy_protocol=http
     explorer_protocol=http
     ```
5. Click **Save**.

Add Google Cloud Storage data source on Privacera Platform using a Project ID

A project ID is a unique ID assigned to a GCP project. The project ID is required in order to interact with resources in the project. Using this project ID, you can access the resources defined in the project and run Discovery scans on those resources.

To add a GCS data source with project ID, do the following:

1. Under GCP, add a new **Data Source**, then select **Google Cloud Storage**.
2. Enter the following:
   - **Name**: A name is provided by default. if required, enter a preferred name.
   - **Description**: Enter a suitable description
   - **Application Code**: An application code is an unique identifier for a data source. A code is provided by default. if required, enter a preferred code. No two data sources can have the same application code.
3. In the **Application Properties** section, add the following properties:
   - **Credential Type**: Select **Google Credentials Local File Path** from the drop-down list.
   - **Google Credentials Local File Path**: `/tmp`
   - **Google Project Id**: `${PROJECT_ID}`
   - **Privacera Configuration Bucket**: `gcs`. Use the same bucket name you added in GCP Configuration [17].
   - **Default Datasource for RealTime Scan** - This value is set to **false** by default. Set this value to **true** if you have more than one data source. In such scenarios, it is recommended that you identify one data source as the default data source which will be used for real-time scanning.

- **Fast Track Data Zones List**: Enter fast track data zones list. Ensure to enable `DISCOVERY_FASTTRACK_REALTIME_ENABLED` property, for more information see Discovery Properties [134].
- **Enable Auto Scan Real-Time**: Enter `true` for Discovery to auto scan real-time and not check the included resources.

> **NOTE**
>
> If **Enable Auto Scan Real-Time** is set to `true`, ensure that quarantine location, transfer location, and archive location are in the exclude resources to avoid re-scanning.

4. Click **Save**.

If you want to scan multiple resources, or resources from a different project, see Set up cross-project scanning on Privacera Platform [32].

Set up cross-project scanning on Privacera Platform

Prerequisites

- Project should be created on GCP console.
- Cluster should have access to all cross projects.
- Create a topic on GCP console.

Procedure

1. From the navigation menu, select **Settings** > **Data Source Registration**.
2. Under GCP system, click **+Add New Data Source** > **Google Cloud Storage**.
3. From the **Add Data Source** dialogue box, select and enter the following properties:
   - **Project Id**: `${PROJECT_ID}` (Mandatory)
   - `scan.result.topic`: '${Scan_Topic_Name}` (Mandatory). Use the same topic name which you have created as part of prerequisites.
   - `scan.result.project.id`: `${Specify_ID_of_Cross_Project}`. If you do not specify the project ID of cross project, the system will consider it your default project ID.
4. To verify the configuration, click **Test Connection**.
5. Click **Save**.
6. Log on to the GCP console and click **Topic**.
7. Search for the topic name, then click **VIEW MESSAGES** on top panel.
8. Click **PULL**, and expand the respective message to view the details.

Google Pub-Sub Topic message scan on Privacera Platform

Prerequisites

Ensure that following prerequisites are met:

- Project should be created on GCP console.
- Cluster should have access to cross projects.
- Topics to be scanned, should be created under Google Project ID on GCP console.
- Pub-Sub result scan topic should be created on GCP console. Eg. *pub_sub_scan_result_topic_t1*

Procedure

1. From the navigation menu, select **Settings** > **Data Source Registration**.

2. Under GCP system, **+Add New Data Source**, and then select **Google Pub-Sub**.
3. On the **Add Data Source** dialogue, enter the following properties:
   - Google Project Id (Mandatory): `${PROJECT_ID}`
   - `pubsub.topic.request.user`
   - `pubsub.scan.result.topic.prefix`: By default, this field auto-populate `pub_sub_scan_result` as prefix.

     Example: **Topic to scan**: topic_t1

     Pubsub scan result: pub_sub_scan_result**_topic_t1**

     > **NOTE**
     > User is allowed to enter the custom prefix as well, as per the choice.

   - `scan.result.topic`: ${Scan_Topic_Name}

     Scan.result.topic should be created under Scan.result.project.id. If Scan.result.project.id is not specified, then Scan.result.topic will consider default project id.
   - `scan.result.project.id`: ${Specify_ID_of_Project}

     If you do not specify the ID of project then system will consider default project id.
4. Click **Save**.
5. Now, add a new Pub-Sub Topic which you want to submit for scan.

   a. Log on to GCP console, and navigate to **Project** > **Pub/Sub topics** and then click **CREATE TOPIC**.
6. Go back to Privacera Portal > **Data Source**, and then select **gcp-Google Pub-Sub** from Application list.

   a. Under Include Resource, click **+Add**, and then enter the pub sub topic name. Eg. *privacera_scan_topic*.
7. Publish a message on the topic which is added in Include Resource for gcp-Google Pub-Sub.

   a. Go to GCP console, and then navigate to **Topic** > **PUBLISH MESSAGE**

   b. Enter the message in the **Message body**.

   > **NOTE**
   > Only the text format is supported in the Message body.

   c. Click **PUBLISH**.
8. Now, on the Privacera home page, expand the **Data Inventory** menu, and then click on **Classification** from left menu.
9. On the Classification page, select the Pub-sub topic name from search, and then look for the tags which are tagged under **Tag** column.

   > **NOTE**
   > Classification is applied as soon as you publish message from GCP console i.e. only for the latest scanned message will be visible on the classification page.

10. Go to the GCP console, and then check the Pub-Sub scan result topic which was created to publish the scan result, it should have the scan result for all the messages.

Add JDBC-based systems as data sources for Discovery on Privacera Platform

The following systems can be connected to Privacera Discovery as data sources via Java Database Connectivity (JDBC):

- Amazon Aurora
- Microsoft SQL Server
- MySQL
- Oracle
- Postgres
- PrestoSQL

> **NOTE**
>
> Starburst PrestoSQL versions are supported through version 350-e.

- Redshift
- Snowflake
- Spark SQL
- Synapse
- Trino
- Starburst

The general process is as follows:

1. Create or identify a service user in the target system with read/write privileges.
2. Determine the JDBC connection string to the data and database name in that target.
3. Define these details as properties in the Privacera Platform.

Prerequisites

Have the following details ready to enter into the data source definition in Privacera:

- A username and password in the target system that has read/write permission.
- The name of the JDBC driver you need.
- A JDBC connection string to communicate with the target data source.

Required properties for JBDC data sources on Privacera Platform

Values for the following properties are described in Required Name of JDBC Driver per Target System [35], Username and Password [35], and Required JDBC Connection String [35].

> **NOTE**
>
> The format of the `jdbc.url` value varies by target system. Not all systems require `databaseName`.

```
jdbc.driver.class=<jdbc_driver_name>
jdbc.username=<user_with_readwrite_permission>
jdbc.password=<login_credentials_of_identified_user>
jdbc.url=jdbc:<protocol>://<hostname>:<port>;databaseName=<database_name>
```

Required name of JDBC Driver per target system

Depending on the target system, for the `jdbc.driver.class` definition you enter in the Privacera properties, use one of the JDBC drivers shown below.

- Amazon Aurora: `org.mariadb.jdbc.Driver`
- Microsoft SQL Server: `com.microsoft.sqlserver.jdbc.SQLServerDriver`
- MySQL: `com.mysql.jdbc.Driver`
- Oracle: `oracle.jdbc.driver.OracleDriver`
- Postgres: `org.postgresql.Driver`
- PrestoSQL: `org.apache.hive.jdbc.HiveDriver`
- Redshift: `com.amazon.redshift.jdbc.Driver`
- Snowflake: `net.snowflake.client.jdbc.SnowflakeDriver`
- Spark SQL (Databricks): `org.apache.hive.jdbc.HiveDriver`
- Synapse: `com.microsoft.sqlserver.jdbc.SQLServerDriver`
- Trino: `io.trino.jdbc.TrinoDriver`
- Starburst: `io.trino.jdbc.TrinoDriver`

Username and password

Identify the name of a user who must have read/write permission in your data source and the login credentials for that user. These values are needed for `jdbc.username` and `jdbc.password` properties in Privacera.

Required JDBC connection string

The `jdbc.url` value you enter in the Privacera properties must be one of the following, where `<domainName>, <port>`, and other variables are for your specific environment:

- Amazon Aurora: `jdbc:mysql://<domainName>:<port>/<dbname>`
- Microsoft SQL Server: `jdbc:sqlserver://<domainName>:<port>;databaseName=<db_name>`
- MySQL: `jdbc:mysql://<domainName>:<port>/<dbname>`
- Oracle: `jdbc:oracle:thin:@//<domainName>:<port>/<dbname>.localdomain`
- Postgres: `jdbc:postgresql://<domainName>:<port>/<dbname>`
- PrestoSQL: `jdbc:presto://<domainName>:<port>/<catalog_name>`
- Redshift: `jdbc:postgresql://<domainName>:<port>/<dbname>`
- Snowflake: `jdbc:snowflake://<domainName>:<port>/?warehouse=<name_of_policysync_warehouse>`
- Spark SQL (Databricks): `jdbc:hive2://<domainName>:<port>/default;transportMode=http;ssl=true;httpPath=sql/protocolv1/o/0/xxxx-xxxxxx-xxxxxxxx;AuthMech=3;`
- Synapse: `jdbc:sqlserver://<domainName>:<port>;databaseName=<dbname>`
- Trino: `jdbc:trino://<host>:<port>/<catalog>`
- Starburst: `jdbc:trino://<host>:<port>/<catalog>`

NOTE

The following three databases can be added as catalog on Trino and Starburst server: , ,

Add JDBC-Based data source on Privacera Platform

These are the setup and steps to add a JDBC-based data source. Have the details listed in the planning sections above ready to enter into the data source definition in Privacera

To add a JDBC-based data source in Privacera Platform:

1. From the navigation menu, select **Settings** > **Data Source Registration**.
2. (Optional) Click **Add System** or modify an existing data source.
3. Enter a name and description for this data source.
4. Click **Save**.
5. Locate the new data source system name and from the wrench icon on the right, select **Add Data Source**.
6. In the **Add Data Source** dialog, on the **Choose** tab, select **JDBC APPLICATION**.
7. On the **Configure** tab:
8. Enter a required **Application Name** of your choice.
9. Enter a required **Application Code** of your choice. This is an identifier for your own use.
10. If you have prepared a properties file in JSON format, click **Import Properties** and load the file.
11. Scroll to find the following properties and enter the values you prepared:
    - `jdbc.username`
    - `jdbc.password`
    - `jdbc.driver.class`
    - `jdbc.url`
12. Accept the default values for all other properties or modify them if needed.
13. To verify the properties, click **Test Connection**.
14. Click **Next** to save the data source or **Back** to return to the **Choose** tab.

Add and scan resources in a data source

The following example enables scanning on an **-Aurora DB** resource. It is recommended that you familiarize yourself with the names of the resources you want to enable before scanning as they will appear in a drop-down menu.

To enable scanning on an resource, do the following:

1. From the navigation menu, select **Discovery > Data Source**.
2. From the **Applications** list, select **-Aurora DB**.
3. Click **Add** to add a resource for scanning.
   a. Type the text of the resource and it will display the list of resources that matches the text.
   b. Select the scan type.
   c. Click **Save**.
4. Click the **Status** toggle to globally enable scanning.
   - For real-time scan, resources will be automatically scanned when they are added to the **Included Resources** list.
   - For offline scan, click **Scan Resource** button to initiate a scan.
5. Repeat these steps as needed for other data resources or applications you intend to enable for scanning.
   - The names of displayed fields will be different depending on the type of resource or application you are configuring (for example, **Include Resource** or **Include Database or Table**).
   - Resources in the landing zones are automatically scanned by Privacera. For more information on **Data Zones** see Data zones overview [93].

and

Using a single or data source, you can scan resources from multiple projects. You can search for projects to be added, and select resources from the project to be included for scanning. To retrieve the list of projects in or , configure the Google Cloud Manager API.

NOTE

Data Explorer does not support showing resources from multiple projects. It only shows resources for the project with which the data source is configured.

Prerequisites

To allow Privacera search for projects on your Google account, you need to enable the API services in the project you registered as a data source. Refer the Google documentation to enable API services.

Add resources to or data sources

Before you can add resources to a data source, your data source must be registered and the prerequisite requirements must be met in order to continue. For more information on registering a data source, see Register data sources on Privacera Platform [26].

To add resources to or data sources, do the following:

1. From the navigation menu, select **Discovery** > **Data Source**.
2. From the Applications section, select a or data source.
3. Click **Add**.
4. In the **Add Resource** dialog, enter the following:
   a. Enter the **Project ID** of the resource you want to scan. You can enter an asterisk (*) to get a list of projects.
      - For , the **Project ID** will be appended to the dataset or table name.
      - For , the **Project ID** will not append to the bucket name as they are unique across a project.

   b. Enter the **Resource** you are including in the project.

      NOTE

      Resources can be added from multiple projects. Existing resources will be updated with a project ID. If you have resources in a specific directory, you can add this location path so that all of the databases/tables in that location are scanned.

      - For , add the bucket resources.
      - For , add the datasets or tables.

   c. Select a scan type:
      - **Scan**: Select this option if you want to perform real-time/offline scan.
      - **Incremental**: Select this option if you want to scan the resource once. During a re-scan, the resource gets added in the **Excluded Resources** list.

   d. **Multi-input**: Turn on this button if you want to switch to a multiple input view and add multiple resources, one per line.

   e. Click **Save**.
5. To enable the real-time/offline scan for the or data source, click the **Status** toggle.

Start a scan

There are several ways to start scans in Privacera Discovery:

- From the **Data Sources** page, which is described here.
- For offline (re-scan) or realtime (continuous) scans. See Start offline and realtime scans [38].
- If you have set up datazones, starting a scan, called **reevaluation**, is discussed in Data zones overview [93].

To start a scan from the **Data Source** page, follow these steps:

1. From the **Applications** section, select the application that contains the resource you want to scan.
2. In the **Scanning Details** section, locate the resource you want to scan.
3. Click **SCAN RESOURCE**.

A message appears indicating that a scan has been initiated.

Start offline and realtime scans

There are two ways to scan resources in :

- Offline [38].
- Realtime [38].

Start offline scanning

You can manually scan resources (offline scanning) from the **Data Sources** page.

To start offline scanning, follow these steps:

1. From the navigation menu, select **Discovery** > **Data Sources**.
2. Select a resource from the **Applications** list.

   NOTE
   Ensure that the application is enabled.

3. Under **Include Resource** tab, check the **Rescan** checkbox of the resource to be scanned.
   The **Info and Success** dialog is displayed.

Start realtime scanning

By default, scans resources that you add to an application (realtime scanning). When a new file is added to the **Include Resource** tab of the **Data Source** page, realtime scanning occurs.

To scan the resource in realtime, the application should be enabled and resource should be added to the **Include Resource** tab in the application. For example, to copy a file from the cluster to HDFS, use the following command:

```
hdfs dfs -put -f <local-src> … <HDFS_dest_path>
```

For AWS S3, you can fetch S3 tags. for more information see Configure S3 for real-time scanning on Privacera Platform.

View classification results

You can view scan results on the **Classification** page. For more information, see Classifications overview [110].

Scan Status overview

After you trigger a manual scan, you can check the progress of the scan from the **Scan Status** page.

During manual or offline rescan, if a file under a specified directory does not exist, the scan marks that the data was deleted in Classification. This is applicable only when realtime scan is disabled. The deleted resources are stale and can be viewed under Stale Resources.

Scan IDs that have not resulted in any tag classification are periodically removed from the status page.

View scan status summaries

To check the status of your scans, select **Discovery** >**Scan Status** from the navigation menu.

Scans can have the following statuses:

- **Pending**: Number of scan requests in pending state.
- **Listing**: Number of scan requests in listing state.
- **Running**: Number of scan requests in running state.
- **Success** Number of successfully completed scan requests.
- **Failed**: Number of failed scan requests.
- **Killed**: Number of killed scan requests.
- **Cancelled**: Number of cancelled scan requests.
- **Retry**: Number of scan requests moved into retry state.

> **NOTE**
>
> Scanning durations are shown for data in different stages. For example, **Listing** shows the time taken to scan the existing data.

Individual scans on the Scan Status page

The **Scan Status** page displays a table of individual scans that includes the following information:

- **Scan Id**: The scan ID with a clickable link to view a summary of the scan.
  The **Scan Type** is shown as *Scan* (which is a full scan) or *Incremental*.
- **Status**: The status of the scan request.
- **Scan/Total Resource**: The number of files or tables scanned out of the total number present in the scan request.
- **Application**: The name of the application, such as Hadoop-Hive or Azure-ADLS.
- **Resource**: The name of resource. Click the resource to view the classification page for that resource.
- **Create Time**: The date and time that the scan was triggered.
- **Start Time**: The start time of the scan.
- **End Time**: The end time of scan.
- **Duration**: The scan duration.
- **Request User**: The name of the user who triggered the scan.
- **Type**: The type of scan, such as offline or realtime scan.
- **Policy**: The name of the policy.

View individual Summary Reports

Click **View Summary Report** in the **Scan Id** column to view **Summary Info** details for the selected scan ID such as **Tagged Resources, UnTagged Resources, Excluded, Failed Resources, Properties, Diagnostic Info, Logs, Scan Cleanup, and Stale Resources**.

Export Scan Summary

To download the scan summary, click **Export** and follow the leading prompts.

Cancel a scan

To cancel a scan, follow these steps:

1. Go the **Scan Status [38]** page.
2. Locate the scan that you want to cancel.

3. Click **Cancel**.

Trailing forward slash (/) in data source URLs/URIs

If a data source URL/URI has a trailing `/` (forward slash), then will scan the folders in the bucket individually. If the data source URL/URI does not have a trailing `/`, then the folders in the bucket will be scanned together.

For example, say the following three folders are in an S3 bucket:

- A
- A_1
- A_1_2

If these three folders need to be scanned individually, then the URL/URI in the data source should be listed as:

- `s3://bucket/A/`
- `s3://bucket/A_1/`
- `s3://bucket/A_1_2/`

If the three folders need to be scanned together, then the URL/URI in the data source should be listed as:

- `s3://bucket/A`

Or, if you want to scan A_1 and A_1_2, then the URL/URI should be listed as:

- `s3://bucket/A_1`

This will scan both `s3://bucket/A_1` and `s3://bucket/A_1_2`.

Configure Discovery scans

Tags

Tags are an important part of and access control. In addition to security policies for resources and roles, you can create policies based on tags. Using tag-based policies, you can manage access to sensitive data regardless of where the data is stored.

scans data sources and tags all sensitive information across the enterprise. Example tags include `PERSON_NAME`, `PII, ADDR`, or `EMAIL_ADDR`. A dataset attribute, such as a column, table, or file, can be tagged with metadata information that can be used to classify the data asset. For example, a column titled "Email" or "Phone_Number" can be tagged as `PII`.

Tags enrich existing information about your data. Data administrators can create access control policies based on the tags created by . You can view your tags from the **Tags Information** page.

If you have defined rules, the generation of tags depends on the order of the rules. For more information, see Processing order of scan techniques [26] and Rules [71].

Add Tags

You can add tags in from the **Tags Information** page.

To add a tag, folow these steps:

1. From the Privacera home page, expand the **Discovery** menu and select **Tags Information**.
2. Click the **+** icon.
   The **Add Tag** dialog is displayed.
3. In the **Tag Name** field, enter a name for the tag.
4. In the **Description** field, enter a description of the tag (optional).
5. Click **Save**.
   The tag is added.

Import Tags

To import a tag file in JSON format, follow these steps:

1. Click the **Import** icon.
   The **Import** dialog displays.
2. Select the JSON file you want to export.
3. Click **Save**.
   The tag file is imported.

Add, edit, or delete Tag attributes

The **Attributes** section displays a list of attributes associated with a tag. You can search the list of attributes using the search box. The **Attributes** section also displays the total number of records with this tag.

To add an attribute for a specific tag, follow these steps:

1. In the **Tags Information** page, select the tag from the **Tags** list.
2. Click **Add Attribute**
   The **Add Attribute** dialog displays.
3. In the **Name** field, enter the name of the attribute.
4. In the **Value** field, enter the value of the attribute.

5. Click **Save**.
   The attribute is added to the selected tag.

NOTE
You can delete or edit the attribute from the **Actions** column.

Edit Tag descriptions

You can edit the descriptions of tags in from the **Tags Information** page.

NOTE
You cannot change a tag name after the tag is created.

To edit the description of a tag, follow these steps:

1. In the **Tags Information** page, select the tag you want to edit from the **Tags** list and click **Edit**.
   The **Edit Tag** dialog is displayed.
2. Update the **Description** field.
3. Click **Save**.
   The tag is updated.

Delete Tags

You can delete tags in from the **Tags Information** page.

To delete a tag, follow these steps:

1. In the **Tags Information** page, select the tag you want to edit from the **Tags** list and click **Delete**.
   The following message is displayed: “Are you sure you want to delete this tag?”
2. Click **Yes** to delete the tag or **No** to return to the **Tags Information** page.

Export Tags

To export the tag file in JSON format, follow these steps:

1. Click **Export**.
2. Check the checkbox of the required tag and click the **Export**. You can select multiple tags.
   The tag file is exported.

Search for Tags

You can search for tags in from the **Tags Information** page.

To search for a tag, enter the name of the tag into the **Search Tag** field.

Fetch AWS S3 Tags

allows you to fetch AWS S3 tags. There are two types of tags that can be fetched:

- **Object Tags**: Tags associated with the AWS S3 object or files in buckets.

- **Bucket Tags:** Tags associated with the S3 bucket.

To fetch AWS S3 tags, follow these steps:

1. Navigate to **Discovery** > **Tags Information** and create a tag named `AWS_S3_TAG`.
2. Navigate to **Settings** > **Data Source Registration** and add or update the application properties as below:
   a. Set `"Fetch S3 Object Tags": true`
   b. Set `"Fetch S3 Bucket Tags": true`

   > **NOTE**
   > By default these properties are disabled and set to false.
3. Go to **Data Inventory** > **Classifications** and click `AWS_S3_TAG` under the **Tag** column, then click on **View attributes** link.
4. Click **View attributes** .
   AWS S3 tags will be displayed in the **Data Info** grid.

> **NOTE**
>
> - If the `AWS_S3_TAG` tag is not created, then AWS S3 tags will not be fetched and the tag will not be displayed in **Classification** page.
> - If both the Object and Bucket tags are enabled and have a common tag, then the Object tag will override the Bucket tag. For example: If the Bucket tag is `owner=user1` and the Object tag is *owner=user2*, then the `AWS_S3_TAG` tag will have `owner=user2` as its attribute.
> - Tags fetched from AWS S3 will be added as attributes of the `AWS_S3_TAG`. This tag with attributes will be synced to Apache Ranger. Verify using the following URL: `https://<EC2_Instance_IP>:6182/service/tags/tags`.

Propagate Tags to Ranger

Privacera Discovery allows you to classify information in files as tags when you scan files in a application. The tags can be used in access policies to configure access control for the application.

Apache Ranger requires the tagged information while applying a policy. This topic describes how you can propagate the tag details from Discovery to Apache Ranger.

This feature is supported for the following applications:

-
-
-
-
-
-
-

General process for configuring an application

You need to configure some advanced properties for the application where all the data to be scanned are stored.

Determine which of the supported applications you want to configure.

For each application:

1. Enable **Access Management** and **Data Discovery** for the application.
2. For Data Discovery, on the **ADVANCED** tab, enter the following properties in the **Add New Custom Properties** text box, where
   - `ranger.writer.enable=true`
   - `cluster_name=privacera`
   - `service_name=privacera_<name_of_application>`

where `service_name=privacera_<name_of_application>` depends on the application you are configuring:

- : `service_name=privacera_hive`
- : `service_name=privacera_mssql`
- : `service_name=privacera_hive`
- : `service_name=privacera_redshift`
- : `service_name=privacera_s3`
- : `service_name=privacera_snowflake`
- : `service_name=privacera_gcs`

Validate the configuration

To validate the configuration, you run a scan to create the classification tags and then use curl with the Ranger API to see the results.

To create the tags, perform an offline or online scan. For more information, see "Start a scan" in scan targets [25].

You can use the following Ranger API to retrieve the pushed tagged information:

```
curl -i -L -k -u <username>:<password> -H "Content-type: application/json" -X GET <hostn
```

where:

- `<username>`:`<password>`: Credentials of a PrivaceraCloud user. See Create user [44].
- `<hostname-of-ranger>`: Ranger Admin URL. See Get Ranger Admin URL [44].

Create user

1. Go to **Settings > User Management** > and click **Add**.
2. Enter the required details. Select role as **Admin** from the dropdown.
3. Click **Save**.

Get Ranger Admin URL

1. Go to **Settings > Api Key** and click the API Key info icon. The **Api Key Info** dialog appears.
2. For the **Ranger Admin URL**, click **Copy URL**. This is the endpoint to connect to Ranger.

TagSync using Apache Ranger on Privacera Platform

allows you to classify your data using tags [41]. Tags can be used in access policies to manage access to sensitive data.

requires the tagged information while applying a policy. This topic describes how you can propagate the tag details from Discovery to .

Enable TagSync

You need to enable TagSync in the Privacera Portal by configuring the following properties in the Application Properties UI:

```
ranger.writer.enable=true
send.inherited.table.tags.to.ranger=true
```

Properties to add based on service type

Apart from above properties, you need to add the additional properties based on service type in Application Properties UI. These properties will help to verify TagSync in using the Ranger utility script.

For example:

```
service_name=privacera_s3
cluster_name=privacera
```

The value of `service_name` depends on the application that you want to apply TagSync to. The following is a list of services and values for each application:

S3

```
service_name=privacera_s3
cluster_name=privacera
```

Redshift

```
service_name=privacera_redshift
cluster_name=privacera
```

PostgreSQL

```
service_name=privacera_postgres
cluster_name=privacera
```

Snowflake

```
service_name=privacera_snowflake
cluster_name=privacera
```

DynamoDB

```
service_name=privacera_dynamodb
cluster_name=privacera
```

MSSQL/Synapse

```
service_name=privacera_mssql
cluster_name=privacera
```

MySql/MariaDB/AuroraDB/Databricks Spark SQL

```
service_name=privacera_hive
cluster_name=privacera
```

TagSync validation scenarios

TagSync can be validated in the following scenarios:

- Auto scanning [46]
- Meta tagging [47]

> **NOTE**
> Allowed and rejected tags will not be synced to .

Auto scanning

On the **Classifications** page, files are classified with system classified tags. After classification, all system-classified and manually accepted tags are synced to .

Parent-Child Level TagSync in :

Based on database applications or file systems, the following is the criteria to sync parent and child tags:

Database applications

Example 1. Scenario

If the resource is a database, then the database gets classified as:

- Database, tag1, tag2, etc.

In Ranger, child entries are created as below:

- (Database): tag1, tag2, etc.

Example 2. Scenario

If the resource is a table, the classification is as shown as below:

- (Database, table), tag1, tag2, etc. then in Ranger child level entry can be seen as below:

In Ranger, child level entry can be seen as below:

- (Database, table): tag1, tag2, etc.

Example 3. Scenario

If the resource is a column, on the UI the classification is as shown below:

- (Database, table, column), tag1, tag2, etc.

In Ranger, only column level tags will be synced:

- (Database, table, column), tag1, tag2. etc.

File System

- For a folder or file, all the tag levels are allowed.
- For a field, only the same tag level is allowed.

Meta tagging

Meta tags are applied at the table, file or folder level. They are also synced to at the table, file or folder level. Only system classified and manually classified tags are synced to .

Folder tagging

By default folder tagging feature is not enabled, you can enable folder tagging at the application settings using **Folder name tagging** toggle button. Folder tagging includes folder names during scanning and tags the folders based on dictionary values.

Example 4. Scenario

Create a new dictionary with following fields:

- **Name**: Enter the dictionary name.
- **Type**: Select the tagging type from the dropdown menu.
- **Apply For**: Select **metaname**.
- **Tags**: Add existing or new tag names.

Save and add the folder names that you wish to tag. The names should match either folder, file, or field name in the scanned files.

Add S3 resources on any file or folder, system will add a tag on the folder with values that are matching from the dictionary and that are present in the path.

On the Classification page, you can see folder resource along with tags.

Open scan summary, under tagged resource tab you will see all tagged folders with scan reason as **Resource is folder**.

Check for tags in ranger using tag sync tool, you need to add all necessary fields in application s3 settings to enable ranger tag sync.

Post-processing tags

System classified and manually classified tags that are applied using post processing rules are synced to .

Re-evaluate

In the case of re-evaluation, system classified and manually classified datazone tags are synced to . Resources that are deleted through datazone policies will be removed from as well.

Add or edit tags

You can add or edit tags manually on the original classified resources from following pages:

- **Classifications**: From the navigation menu, select **Data Inventory** > **Classifications**.
- **Resource Detail**: From the navigation menu, select **Data Inventory** > **Classifications**. Select a resource and click **Resource Detail**.
- **Data Explorer**: From the navigation menu, select **Data Inventory** > **Data Explorer**.
- **Data Zone Dashboard**: From the navigation menu, select **Compliance Workflow** > **Data Zone Dashboard**.

When a user adds tags manually from the pages listed above, the tag status is set by default to “Accepted : Manually classified” and it will be synced to .

Add a resource

You can manually add tags to unclassified resources. When you add such resources and add a tag to them, the tag status is set by default to “Accepted : Manually classified” and it will be synced to .

To add resource, select **Data Inventory** > **Classifications** from the navigation menu and click **Add Resource**.

Tag status changes

Tag status changes will affect TagSync. Only system classified and manually accepted tags will be synced to . The following are few scenarios for tag status changes:

- If the status of a tag is changed from system classified to rejected or allowed, then the tag will be removed from .
- If the status of the tag is changed from manually accepted to allowed or rejected, then the tag will be removed from .
- If the tag status resets to system classified from rejected or allowed, then the tag be synced .
- If the tag status is changed to manually classified from rejected or allowed, then the tag will be synced to .
- If the tag status is changed from system classified to manually classified, then the synced tags in will remain unchanged.

Remove tags

You can manually remove added tags if you have rejected them. If you remove a tag from a resource using the **Add/Edit** option, then the tag will be removed from as soon as you reject it.

Remove resources

If a resource is added manually and has only manually classified tags, then after your reject the last tag the resource will be removed from .

If a resource has system classified tags and you reject the last tag, the resource will be removed from as last TagSync for the same resource will get removed.

Rescan of same file

- If you rescan a resource that is already synced with and no changes were made to rules or datazone policies, then TagSync will remain unchanged.
- If post-processing rules are disabled, then rescanning a file will remove post-processing tags.
- If a datazone tag is disabled or a resource removed from a datazone, then the datazone tag will be removed from upon rescan.
- If a meta tag rule or a meta tag is disabled, then the meta tag will be removed from upon rescan.
- If a status change is applied before a rescan of a file, as per status change TagSync will also affect.

Validate TagSync in

You can view tags that are getting pushed to using curl commands as well as using the Ranger tag utility script.

Validate TagSync using curl command

```
curl -i -L -k -u admin:${PRIVACERA_PASSWORD} -H "Content-type: application/json" -X GET
https://${PRIVACERA_HOST}:6182/service/tags/resources/service/privacera_postgres
```

The above curl command will give the list of resources that are synced to , but the response of this curl command is not in a readable format. Therefore , it is recommended to use the Ranger tag utility to check TagSync.

Validate TagSync using the Ranger Tag Utility

The following is a Python script created to communicate with all Ranger API methods. This will return the response in a readable format:

- Run the following command to download required files:

  ```
  wget https://privacera.s3.amazonaws.com/public/pm-demo-data/ranger_tag_utility.py -O ra
  ```
- Download the file on your local system and execute the following command to view the TagSync response.

  SSL instance

  ```
  python3 ranger_tag_utility.py     --operation list_tags     --host ${PRIVACERA_HOST}
  ${RANGER_USERNAME}     --password ${RANGER_PASSWORD}     --servicename privacera_redsh
  ```

  Non-SSL instance

  ```
  python3 ranger_tag_utility.py     --operation list_tags     --host ${PRIVACERA_HOST}
  ${RANGER_USERNAME}     --password ${RANGER_PASSWORD}     --servicename privacera_maprf
  ```
- (Optional) Change the service name as per the application.

  Output

  ```
  Received Tag Data for path : ['/testdir/sample_files/file_format/avro/test.avro'] => ta
  Received Tag Data for path : ['/testdir/sample_files/file_format/avro/test.snappy.avro
  Received Tag Data for path : ['/testdir/sample_files/file_format/avro/test1.avro'] => 
  Received Tag Data for path : ['/testdir/sample_files/file_format/avro/twitter.avro'] =>
  Received Tag Data for path : ['/testdir/sample_files/file_format/avro/twitter.snappy.av
  ```

Ranger TagSync custom properties on Privacera Platform

The following table contains the list of custom properties that can be configured for Ranger Tagsync. To use a custom property from the table, just add it to the following YML file in the custom-vars folder configured as per your environment:

- vars.ranger.tagsync.yml

Property	Description	Values	Default Value
RANGER_TAGSYNC_INSTALL	To enable Tagsync, set this property to true.		false
RANGER_TAGSYNC_IMAGE_NAME	Privacera Tagsync image name		{{privacera_hub_url}}/ranger-tagsync
RANGER_TAGSYNC_IMAGE_TAG	Privacera Tagsync image tag name		PRIVACERA_IMAGE_TAG
TAGSYNC_RANGER_URL	Ranger URL for the Tagsync to sync the tags.		http://ranger:6080
TAGSYNC_TAG_SOURCE_ATLASREST_ENDPOINT	Required only when you set the SOURCE as REST.		${ATLAS_HOST}:21000
TAGSYNC_RANGERTAGSYNC_PASSWORD	Password for Tagsync user to use an API to Ranger.		welcome1
TAGSYNC_TAG_DEST_RANGER_ENDPOINT	Ranger URL for the Tagsync to sync the tags.		http://ranger:6080
TAGSYNC_TAG_DEST_RANGER_SSL_CONFIG_FILENAME	SSL config file name is used by Tagsync to push tags to SSL-enabled Ranger and PolicyMgr files. It is required to be modified only when custom changes are made to the file.		/opt/ranger/ranger-tagsync/conf.dist/ranger-policymgr-ssl.xml
TAGSYNC_TAG_SOURCE_ATLAS_ENABLED	Enable Kafka as a SOURCE for Tagsync.		true
TAGSYNC_TAG_SOURCE_ATLAS_KAFKA_SERVICE_NAME	Service Name to be used while communicating with Kafka.		kafka
TAGSYNC_TAG_SOURCE_ATLAS_KAFKA_SECURITY_PROTOCOL	Protocol to be used to communicate to Kafka.		PLAINTEXTSASL

Property	Description	Values	Default Value
TAGSYNC_TAG_SOURCE_ATLAS_KERBEROS_PRINCIPAL	If Kafka is kerberos-enabled, then set the value to the principal name used by Tagsync to sync the tags.		
TAGSYNC_TAG_SOURCE_ATLAS_KERBEROS_KEYTAB	If Kafka is kerberos-enabled, then set the value to the keytab location used by Tagsync to sync the tags.		
TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED	Enable REST-based Tagsync to Ranger. This is not recommended as REST has limitation for number of tags it can push to Ranger.		false
TAGSYNC_TAG_SOURCE_ATLASREST_DOWNLOAD_INTERVAL_IN_MILLIS	Tagsync interval required only when TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED is set to true.		900000
TAGSYNC_TAG_SOURCE_ATLASREST_USERNAME	Atlas user name required only when TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED is set to true.		
TAGSYNC_TAG_SOURCE_ATLASREST_PASSWORD	Atlas password required only when TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED is set to true.		
TAGSYNC_TAG_SOURCE_FILE_ENABLED	To enable file-based TagSync.		false
TAGSYNC_TAG_SOURCE_FILE_FILENAME	Location of the file required only when TAGSYNC_TAG_SOURCE_FILE_ENABLED is set to true.		/etc/ranger/data/tags.json
TAGSYNC_TAG_SOURCE_FILE_CHECK_INTERVAL_IN_MILLIS	Tagsync interval, required only when TAGSYNC_TAG_SOURCE_FILE_ENABLED is set to true.		60000
TAGSYNC_TAGSYNC_ATLAS_CUSTOM_RESOURCE_MAPPERS	Any custom mappers to be configured in Tagsync for mapping Atlas entities to Ranger type definitions.		org.apache.ranger.tagsync.source.atlas.AtlasS3ResourceMapper
TAGSYNC_TAGSYNC_KEYSTORE_FILENAME	File will be generated to store the credentials for Ranger password for rangerTagsync user.		/etc/ranger/tagsync/conf/rangertagsync.jceks
TAGSYNC_TAG_SOURCE_ATLASREST_KEYSTORE_FILENAME	File will be generated to store the password for Atlas when TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED is set to true.		/etc/ranger/tagsync/conf/atlasuser.jceks
TAGSYNC_TAG_SOURCE_ATLASREST_SSL_CONFIG_FILENAME	SSL config file name to communicate to Atlas required when TAGSYNC_TAG_SOURCE_ATLASREST_ENABLED is set to true.		
TAGSYNC_UNIX_USER	User to run the process.		ranger
TAGSYNC_UNIX_GROUP	File permission group.		ranger
TAGSYNC_LOGDIR	Log location for Tagsync application.		log
TAGSYNC_PID_DIR_PATH	Location to store the PID file for the Java process.		/var/run/ranger
TAGSYNC_IS_SECURE	Property to check whether Tagsync Is secure (kerberos-enabled).		false
TAGSYNC_PRINCIPAL	Tagsync principal required only when the TAGSYNC_IS_SECURE is set to true.		
TAGSYNC_KEYTAB	Tagsync keytab location required only when the TAGSYNC_IS_SECURE is set to true.		
TAGSYNC_HADOOP_CONF	Hadoop Conf location.		/etc/hadoop/conf
TAGSYNC_FILE_PERMISSION	File permission on the PM host for the templates generated by PM. For example, file permissions on the file, install.properties.		700
TAGSYNC_K8S_SERVICE_ACCOUNT	Service Account Name to be used during installation in a Kubernetes environment.		privacera-sa
TAGSYNC_ROOT_LOG_LEVEL	Log-level for the root.		info

Property	Description	Values	Default Value
TAGSYNC_RANGER_LOG_LEVEL	Log-level for the org.apache.ranger.tagsync package.		info
Memory Variables			
TAGSYNC_SMALL_MEMORY_MB	TAGSYNC MEMORY in MB for Java process if deployment size is set to SMALL.		1024
TAGSYNC_MEDIUM_MEMORY_MB	TAGSYNC MEMORY in MB for Java process if deployment size is set to MEDIUM.		4096
TAGSYNC_LARGE_MEMORY_MB	TAGSYNC MEMORY in MB for Java process if deployment size is set to LARGE.		8192
TAGSYNC_HEAP_MIN_MEMORY_MB	Depending upon the DEPLOYMENT SIZE the value will be calculated above properties.		1024
TAGSYNC_HEAP_MIN_MEMORY	Minimum Java Heap memory used by Ranger Tagsync. Setting this value will override TAGSYNC_HEAP_MIN_MEMORY_MB. For example, TAGSYNC_HEAP_MIN_MEMORY: "1g"		1024M
TAGSYNC_HEAP_MAX_MEMORY_MB	Maximum Java Heap memory in MB used by Ranger Tagsync. For example, TAGSYNC_HEAP_MAX_MEMORY_MB: "1024"		1024
TAGSYNC_HEAP_MAX_MEMORY	Maximum Java Heap memory used by Ranger Tagsync. Setting this value will override TAGSYNC_HEAP_MAX_MEMORY_MB. For example, TAGSYNC_HEAP_MAX_MEMORY: "1g"		1024M
TAGSYNC_K8S_MEM_REQUESTS_MB	Minimum amount of Kubernetes memory in MB to be requested by Ranger Tagsync. For example, TAGSYNC_K8S_MEM_REQUESTS_MB: "1024"		1024
TAGSYNC_K8S_MEM_REQUESTS	Minimum amount of Kubernetes memory to be used by Ranger Tagsync. Setting this value will override TAGSYNC_K8S_MEM_REQUESTS_MB. For example, TAGSYNC_K8S_MEM_REQUESTS: "1G"		1024M
TAGSYNC_K8S_MEM_LIMITS_MB	Maximum amount of Kubernetes memory in MB to be requested by Ranger Tagsync. For example, TAGSYNC_K8S_MEM_LIMITS_MB: "1024"		1024
TAGSYNC_K8S_MEM_LIMITS	Maximum amount of Kubernetes memory to be used by Ranger Tagsync. Setting this value will override TAGSYNC_K8S_MEM_LIMITS_MB. For example, TAGSYNC_K8S_MEM_LIMITS: "1G"		1024M
TAGSYNC_CPU_MIN	Minimum amount of Kubernetes CPU to be requested by Ranger Tagsync. For example, TAGSYNC_CPU_MIN: "0.5"		0.5
TAGSYNC_CPU_MAX	Maximum amount of Kubernetes CPU to be used by Ranger Tagsync. For example, TAGSYNC_CPU_MAX: "0.5"		0.5
TAGSYNC_K8S_CPU_REQUESTS	Minimum amount of Kubernetes CPU to be requested by Ranger Tagsync. For example, TAGSYNC_CPU_MIN: "0.5"		0.5
TAGSYNC_K8S_CPU_LIMITS	Maximum amount of Kubernetes CPU to be used by Ranger Tagsync. For example, TAGSYNC_CPU_MAX: "0.5"		0.5
TAGSYNC_HELM_CHART_VERSION	Tagsync Helm Chart Version		4.3.0

Sync Ranger tag store with Atlas using TagSync on Privacera Platform

This topic shows how you can configure Ranger TagSync to synchronize the Ranger tag store with Atlas.

Configuration

1. Run the following commands.

```
cd ~/privacera/privacera-manager
cp config/sample-vars/vars.ranger-tagsync.yml config/custom-vars/
vi config/custom-vars/vars.ranger-tagsync.yml
```

2. Edit the following properties.

Property	Description	Example
`RANGER_TAGSYNC_ENABLE`	Property to enable/disable the Ranger TagSync.	true
`TAGSYNC_TAG_SOURCE_ATLAS_KAFKA_BOOTSTRAP_SERVERS`	Kakfa bootstrap server where Atlas publishes the entities. Tagsync listens and pushes the mapping of Atlas entities and tags to Ranger.	kafka:9092
`TAGSYNC_TAG_SOURCE_ATLAS_KAFKA_ZOOKEEPER_CONNECT`	Zookeeper URL for Kafka.	zoo-1:2181
`TAGSYNC_ATLAS_CLUSTER_NAME`	Atlas cluster name.	privacera

`TAGSYNC_TAGSYNC_ATLAS_TO_RANGER_SERVICE_MAPPING`	(Optional) To map from Atlas Hive cluster-name to Ranger service-name, the following format is used:
	`clusterName,componentType,serviceName;clusterName2,componentType2,serviceName2`
	Note: There are no spaces in the above format.
	For Hive, the notifications from Atlas include the name of the entities in the following format:
	`dbName@clusterName dbName.tblName@clusterName dbName.tblName.colName@clusterName`
	Ranger Tagsync needs to derive the name of the Hive service (in Ranger) from the above entity names. By default, Ranger computes Hive service name as: clusterName + "_hive".
	If the name of the Hive service (in Ranger) is different in your environment, use following property to enable Ranger Tagsync to derive the correct Hive service name.
	`TAGSYNC_ATLAS_TO_RANGER_SERVICE_MAPPING = clusterName,hive,rangerServiceName`
		{{TAGSYNC_ATLAS_CLUSTER_NAME}},hive,privacera_hive;{{TAGSYNC_ATLAS_CLUSTER_NAME}},s3,privacera_s3

`TAGSYNC_TAGSYNC_ATLAS_DEFAULT_CLUSTER_NAME`	(Optional) Default cluster name configured for Atlas.	{{TAGSYNC_ATLAS_CLUSTER_NAME}}
`TAGSYNC_TAG_SOURCE_ATLAS_KAFKA_ENTITIES_GROUP_ID`	(Optional) Consumer Group Name to be used to consume Kafka events.	privacera_ranger_entities_consumer

NOTE

You can also add custom properties that are not included by default. See Ranger TagSync custom properties on Privacera Platform [49].

3. Run the following command.

```
cd ~/privacera/privacera-manager
./privacera-manager.sh update
```

Add Tags with Ranger REST API

Prerequisite: Make sure the repo is created on Ranger for tags and Hive has the same tag service selected.

To add a tag using Rest API in Ranger, use the following steps:

1. Create privacera_tags in the Ranger Tag Based Policy.
2. Associate the privacera_tags to Hive service.

   ```
   vi atlas_tag_test.json
   ```
3. Edit the JSON file shown below based on your specific table/tag information.

```
{
  "op": "add_or_update",
  "serviceName": "dublin_hive",
  "tagVersion": 0,
  "tagDefinitions": {
    "0": {
      "name": "TEST_TAG",
      "source": "Atlas",
      "attributeDefs": [],
      "id": 0,
      "isEnabled": true
    }
  },
  "tags": {
    "0": {
      "type": "TEST_TAG",
      "owner": 0,
      "attributes": {},
      "id": 0,
      "isEnabled": true
    }
  },
  "serviceResources": [
    {
      "serviceName": "dublin_hive",
      "resourceElements": {
        "database": {
          "values": [
            "db_name"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "column": {
          "values": [
            "column_name"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "table": {
          "values": [
            "table_name"
          ],
```

```
                "isExcludes": false,
                "isRecursive": false
              }
            },
            "id": 0,
            "isEnabled": true
          }
        ],
        "resourceToTagIds": {
          "0": [
            0
          ]
        }
      }
```

Update the following variables

- serviceName
- tagDefinitions[‘0’].name
- tags[‘0’].type
- serviceResources[0].serviceName
- serviceResources[0].resourceElements[‘database’].values[0]
- serviceResources[0].resourceElements[‘column’].values[0]
- serviceResources[0].resourceElements[‘table’].values[0]

```
curl -i -L -k -u admin:${RANGER_ADMIN_PASSWORD} \
-H "Content-type: application/json" \
-d @atlas_tag_test.json \
-X PUT http://<RANGER_HOST>:6080/service/tags/importservicetags
```

- Wait for a couple of minutes and run the following:

```
select * from <database_name>.<table_name>
```

Hive

1. Create privacera_tags in the Ranger Tag Based Policy.
2. Associate the privacera_tags to Hive service.
3. Create a JSON file where you can add tags.

   ```
   vi hive_tag.json
   ```
4. Edit the JSON file shown below based on your specific table/tag information.

```
      {
        "op": "add_or_update",
        "serviceName": "${Hive_Service_Name}",
        "tagVersion": 0,
        "tagDefinitions": {
          "0": {
            "name": "${Tag_Name}",
            "source": "Atlas",
            "attributeDefs": [],
            "id": 0,
            "isEnabled": true
          }
        },
```

```
  "tags": {
    "0": {
      "type": "${Tag_Type}",
      "owner": 0,
      "attributes": {},
      "id": 0,
      "isEnabled": true
    }
  },
  "serviceResources": [
    {
      "serviceName": "${Hive_Service_Name}",
      "resourceElements": {
        "database": {
          "values": [
            "${Database}"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "table": {
          "values": [
            "${Table}"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "column": {
          "values": [
            "${Column}"
          ],
          "isExcludes": false,
          "isRecursive": false
        }
      },
      "id": 0,
      "isEnabled": true
    }
  ],
  "resourceToTagIds": {
    "0": [
      0
    ]
  }
}
```

Sample hive_tag.json

```
 {
  "op": "add_or_update",
  "serviceName": "privacera_hive",
  "tagVersion": 0,
  "tagDefinitions": {
    "0": {
      "name": "SSN",
      "source": "Atlas",
      "attributeDefs": [],
      "id": 0,
      "isEnabled": true
    }
  },
  "tags": {
    "0": {
      "type": "SSN",
      "owner": 0,
      "attributes": {},
      "id": 0,
      "isEnabled": true
    }
  },
  "serviceResources": [
    {
      "serviceName": "privacera_hive",
      "resourceElements": {
        "database": {
          "values": [
            "finance"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "table": {
          "values": [
            "ssn_finance_us"
          ],
          "isExcludes": false,
          "isRecursive": false
        },
        "column": {
          "values": [
            "SocialSecurity"
          ],
          "isExcludes": false,
          "isRecursive": false
        }
      },
      "id": 0,
      "isEnabled": true
    }
  ],
  "resourceToTagIds": {
    "0": [
      0
    ]
  }
}
```

5. Push the tag to Ranger.

Add Tag

```
curl -i -L -k -u admin:<RANGER_ADMIN_PASSWORD> -H "Content-type: application/json" -d @h
```

Get Tagged Resource

```
curl -i -L -k -u admin:<RANGER_ADMIN_PASSWORD> -H "Content-type: application/json" -X GE
```

S3

1. Create privacera_tags in the Ranger Tag Based Policy
2. Associate the privacera_tags to S3 Service.
3. Create a JSON file where you can add tags.

   ```
   vi s3_tag.json
   ```

   ```
   {"op":"add_or_update","serviceName":"${S3_Service_Name}","tagVersion":0,"tagDefinitio
   ```

   Sample JSON:

   ```
   {"op":"add_or_update","serviceName":"privacera_s3","tagVersion":0,"tagDefinitions":{"
   ```
4. Push the tag to Ranger.

   ```
   curl -i -L -k -u admin:welcome1 -H "Content-type: application/json" -d @s3_tag.json -
   ```

   Response:

   ```
   HTTP/1.1 204 No Content
   Set-Cookie: RANGERADMINSESSIONID=517FD2032481415D188C6925FA96E7E3; Path=/; HttpOnly
   X-Frame-Options: DENY
   X-XSS-Protection: 1; mode=block
   Strict-Transport-Security: max-age=31536000; includeSubDomains
   Content-Security-Policy: default-src 'none'; script-src 'self' 'unsafe-inline' 'unsaf
   Cache-Control: no-cache, no-store, max-age=0, must-revalidate
   Pragma: no-cache
   Expires: 0
   X-Content-Type-Options: nosniff
   Content-Type: application/json
   Date: Sun, 08 Mar 2020 18:55:44 GMT
   Server: Apache Ranger
   ```

   To get the tagged resources list.

   ```
   curl -i -L -k -u admin:welcome1 -H "Content-type: application/json" -X GET http://${R
   ```

   Response:

   ```
   [{"id":5,"guid":"6b9234f1-69d9-40b0-9865-fe5bec45b469","isEnabled":true,"createdBy":"
   ```

Test the Tag-Based Policies for S3 with the sample given above:

1. Create user <kate> in EC2 and add permissions read, metaread, write, metawrite to the S3 bucket ${Bucket_Name} in privacera_s3 service.
2. Create a deny tag-based policy for user <kate> - tag = SSN, Component = S3, permissions = read, write.
3. Now try to access the ${Bucket_Name} with user <kate>.
4. Denied audit is seen with ${SSN} tag in the audits.

Dictionaries

Dictionaries are lists of values used to identify data elements. Privacera Discovery matches dictionaries against your resources and data and can be applied to either content or metanames.

Example dictionaries include:

- A dictionary of US person names used to identify names in a database.
- A dictionary of common column name patterns used to identify a column of account IDs.

Dictionaries support multiple include/exclude patterns. This helps enable a longer transition from conventional patterns for pattern matching. For example, the 'email' conventional pattern and its associated structured and unstructured rules can be disabled and the same pattern value can be added as part of a new dictionary lookup. The resulting rules can then be configured just as conventional patterns.

Types of dictionaries

There are three types of dictionaries in Privacera Discovery:

- **Exact match:** the value of the data must exactly match the value in the dictionary.
- **Fuzzy match:** the matching is based on fuzzy logic instead of exact match.
- **Pattern match:** the values in the dictionary are regular expressions.

Dictionary Keys

The key is used by Discovery Rules [71] to associate a tag with a resource element. Because a dictionary can be applied to either content or metaname, a naming convention is used for the key:

- **Content dictionary:** LOOKUP suffix.
- **Metaname dictionary:** KEYWORD suffix.

Manage dictionaries

Privacera Discovery comes pre-loaded with a set of useful dictionaries. You can also create your own custom dictionaries and configure rules [71] to use them.

The values in a dictionary can come from a text file that can be uploaded through the portal or directly copied into your installation. For smaller dictionaries, you can add values using the Privacera portal either one by one or with the bulk input interface. For dictionaries that are file-based, you can add additional values or exclude existing values using the Privacera portal.

When a dictionary is created or modified, the updated dictionary becomes available for use within a few minutes.

Default dictionaries

The following is a list of the default Privacera-supplied dictionaries. The name of a dictionary in general describes the purpose of the dictionary. For precise details, look at the dictionary itself in the Platform UI.

- AU_BSB_LOOKUP
- BINARY_MIME_KEYWORD
- CC_KEYWORD
- CC_PROTECTED_KEYWORDDisabled
- CITY_KEYWORD
- COUNTY_KEYWORD
- CRIMINAL_RECORD_LOOKUP
- DISALLOW_DOB_KEYWORDDisabled
- DISALLOW_NAME_KEYWORDDisabled
- DISALLOW_ZIP_KEYWORDDisabled

- DOB_KEYWORD
- ETHNICITY_LOOKUP
- EXEC_MIME_KEYWORD
- GEO_KEYWORD
- GPS_KEYWORD
- IMAGE_MIME_KEYWORD
- ISO3166_CC_LOOKUP
- MEDICAL_RECORD_LOOKUP
- ORG_LOOKUP
- PASSPORT_KEYWORD
- PASSWORD_KEYWORD
- PERSON_NAME_KEYWORD
- PERSON_NAME_LOOKUP
- PII_ID_KEYWORD
- SSN_KEYWORD
- STATE_KEYWORD
- SWIFT_BIC_KEYWORDDisabled
- SWIFT_BIC_LOOKUPDisabled
- TAX_ID_KEYWORD
- UK_ELECTORAL_ROLL_KEYWORDDisabled
- UK_NHS_KEYWORDDisabled
- UK_NINO_KEYWORDDisabled
- UK_POSTAL_TOWN_LOOKUPDisabled
- US_ABA_NUMBER_KEYWORDDisabled
- US_ADDRESS_KEYWORD
- US_CITY_KEYWORD
- US_CITY_LOOKUP
- US_COUNTY_KEYWORDDisabled
- US_COUNTY_LOOKUPDisabled
- US_DLICENSE_KEYWORD
- US_DLICENSE_LOOKUP
- US_STATE_KEYWORD
- US_STATE_LOOKUP
- US_ZIP_KEYWORD
- US_ZIP_LOOKUP

Add a dictionary

To add a dictionary, follow these steps:

1. On the **Dictionaries** page, click the **+** sign.
   The **Add Dictionary** dialog is displayed
2. Enter the following details:
   - The **Name** of the dictionary (required)
   - The **Description** of the dictionary.
   - The **Key** field is not editable because it is populated by the system. You have the option to add IPv4 and IPv6 address regexes as an option under Key description for regexes and used to lookup dictionary content.
   - The required **File name**.
3. Select the required **Type**: Exact, Pattern, or Fuzzy match.

NOTE

For pattern dictionaries, see Pattern Validation [62]. .

4. Select **Apply For**. The choices are **content or metaname.** If you select metaname, for pattern type dictionaries, you have the choice to apply the input tags directly to the resource. See Add Meta Tags Directly to Dictionary [60].
5. Select the **Status** (enabled by default).
6. Click **Save**.
   The dictionary is added.

Add meta tags directly to a dictionary

When you create a new dictionary of type pattern, you can apply meta tags directly to a data source. The option appears after you select the combination of pattern and metaname.

Import a dictionary

To import a dictionary in JSON format, follow these steps:

1. On the Discovery page, click **Import**.
   The **Import** dialog is displayed.
2. Select the JSON file of the dictionary you want to import and click **Save**.
   The dictionary configuration file is imported.

Upload a dictionary

To upload a dictionary, follow these steps:

1. In the **Dictionaries** page, click **Upload Dictionary**.
   The **Upload Dictionary** dialog is displayed.
2. Select the .txt file of the dictionary you want to upload.
3. Click **Save**.
   The dictionary file is uploaded.

Enable or disable a dictionary

To enable or disable a dictionary, follow these steps:

1. On the **Dictionaries** page, select a dictionary from the **Dictionary** list
2. Click the **Status** toggle to enable or disable the dictionary.

Include a Dictionary

You can filter the list of included dictionaries using the search included dictionary option. This tab also displays the current count of records relying on the dictionary.

The **Include Dictionary** tab displays the following:

- **Name**: Name of the dictionary.
- **Description**: The lookup/keyword description.
- **Actions**: Edit or delete dictionaries.
- **Bulk Edit/Delete:** Select this to edit or delete the dictionary values in bulk. After selecting, click **x** to delete the values.

Exclude a dictionary

You can filter the list of excluded dictionaries using the search excluded dictionary option. This tab also displays the total record count.

The **Exclude Dictionary** tab displays the following information:

- **Name**: Indicates name of the dictionary.
- **Actions**: Allows you to edit and delete the dictionary.

To add a lookup in the **Exclude Dictionary** tab, follow these steps:

1. On the **Dictionaries** page, select a dictionary from the **Dictionary** list.
2. Select the **Exclude Dictionary** tab and click **+Add**.
   The **Add Dictionary** dialog displays.
3. In the **Name** field, enter the names of the dictionaries, one name per line.
4. In the **Description** field, enter a description for the dictionary.
5. Click **Save**.
   The lookup is added to the selected dictionary.

Add keywords to an included dictionary

To add a keyword or lookup under **Include Dictionary**, follow these steps:

1. On the **Dictionaries** page, select a dictionary from the dictionary list.
2. In the **Include Dictionary** tab, click **ADD**.
   The **Add Dictionary** dialog is displayed.
3. Enter the name of the keyword or lookup, one name per line.
4. Add a **Description** for the dictionary name.
5. Click **Save**.
   The keyword or lookup is added to the selected dictionary in the **Include Dictionary** tab.

Edit a dictionary

To edit a dictionary, follow these steps:

1. In the **Dictionaries** page, select a dictionary from the dictionary list and click **Edit**.
   The **Edit Dictionary Info** dialog is displayed.
2. Update the required fields.
3. Click **Save**.
   The dictionary is updated.

Copy a dictionary

To make a copy of a dictionary, follow these steps:

1. On the **Dictionary** page, select a dictionary from the dictionary list and click **Create Copy**.
2. The **Copy Dictionary** Info **dialog is displayed with selected** Type **and** Apply For** values.
3. Enter the following details:
   - Enter the Name dictionary (required).
   - Enter the Description of dictionary.
   - Enter the File name (required).
   - Select the Type (required).
   - Select the Apply For (required).
   - Select the Status (enabled by default)
4. Click **Save**.
   A copy of the dictionary is created.

Export a dictionary

To export a dictionary in JSON format, follow these steps:

1. On the **Dictionaries** page, click **Export** .
2. Check the checkbox of the required dictionary and click **Export**.

> **NOTE**
> You can select multiple dictionaries.

The dictionary file is exported.

Search for a dictionary

To search for a dictionary, navigate to the **Dictionaries** page and enter the dictionary name into the search bar.

Test dictionaries

Pattern validation

If the dictionary is of type pattern, you can validate its regexes.

To validate a pattern, follow these steps:

1. In the Dictionaries page, add a new dictionary of type '**Patterns'**.
   The **Add Dictionary** field for the pattern type is displayed.
2. Enter a complex **Expression** (regex).
3. Enter the **Description** for the expression.
4. Enter the **Input Test Data.**
5. Click **Test Expression**.

The message "Passed" or "Failed" appears in the **Test Output** field.

Test against a data source

To test changes to a dictionary, follow these steps:

1. Perform an offline scan [38] of the data source that has sensitive fields you want to test.
2. Check the **Scan Status**.
3. After the scan is completed, open the resource to verify if the scan classified the tags correctly.

The tags are classified under **Data Inventory** > **Classification**.

Dictionary tour

To see an explanation of the different components of a dictionary, click **Tour** on the **Dictionaries** page.

Patterns

Patterns are deprecated. Embed patterns in Dictionaries instead.

> **NOTE**
>
> In a future release Discovery patterns will be removed from the left nav, because they are not used frequently. Instead, customers should now embed patterns in Dictionaries [58]. If you have any patterns in use, you should move them to Dictionaries [58] now.
>
> Patterns are regular expressions (regexes) that match specific data elements in your data resources.
>
> Privacera-supplied regular expressions can match common patterns like email addresses and URLs.
>
> You can also define your own regexes to isolate patterns in your data to augment Privacera's patterns.

Default patterns

The following is a list of the default patterns in Privacera. You can view details about each of the patterns from the **Patterns** page.

- ACCOUNT
- CREDIT_CARD
- EMAIL
- FINANCIAL
- IPV4
- IPV6
- MAC_ADDRESS
- STREET_ADDRESS
- UK_DRIVER_LICENSE
- UK_ELECTORAL
- UK_NINO
- UK_POSTAL_CODE
- UK_US_PASSPORT
- URL
- ZIPCODE

Add patterns

To add a pattern, do the following:

1. From the navigation menu, select **Discovery** > **Patterns**.
2. Click **Add Pattern**.
   The **Add Pattern** dialog is displayed.
3. In the **Pattern Name** field, enter a pattern name.
4. From the **Applied On** dropdown menu, select one of the following options:
   - **All**: Pattern matching is applied at the file level (default).
   - **File Content**: Pattern matching is applied to the content of the file.
   - **File Name**: Pattern matching is applied based on file name.
   - **Table/Column Name**: Pattern matching is applied based on table or column name.
5. Using the **Regex Status** toggle, enable (default) or disable regexes.
6. In the **Expression** field, enter an expression. For example, an expression for a bank account number might be `b(d{9}\|d{12})b`.

7. In the **Description** field, enter a description.
8. In the **Input Test Data** field, enter your test data.
9. Click **Test Expression** to verify the expression you entered into the **Input Test Data** field.
   The test result is displayed in **Test Output** field.
10. Click **Save**.

The pattern is added.

Edit patterns

To edit a pattern, do the following:

1. On the **Patterns** page, locate the pattern you want to edit and click the **Edit** icon in the **Actions** column.
2. Update the required fields.
3. Click **Save**.

The pattern is updated.

Delete patterns

To delete a pattern, do the following:

1. On the **Patterns** page, locate the pattern you want to delete and click the **Delete** icon in the **Actions** column.
   You are prompted with a message to confirm the deletion.
2. Click **Yes** to delete the pattern.

The pattern is deleted.

Export JSON pattern files

To export patterns to a file in JSON format, do the following:

1. On the **Patterns** page, click **Export**.
2. Select the patterns you want to export and click **Export**.
   The pattern file is exported.

Import JSON pattern files

To import a pattern file in JSON format, do the following:

1. On the Patterns page, click **Import**.
2. Select the JSON file you want to import and click **Save**.

The pattern file is imported.

Search for patterns

To search for a pattern, enter the pattern name in the search bar on the **Patterns** page and click **Enter**.

The search results are displayed.

Models

Models detect specific data elements in your data resources. The detection is done with various algorithms and heuristics.

Types of models

Privacera supports different types of models. You can filter the list of models using the search model option. This tab also displays the present number of record count.

Generic models

These are various general model parameters you can use to tailor matching of data.

Parameter	Data Type	Default	Description
`INCLUDE_PATTERN_<#>`	String	None	Patterns to be matched. Can contain more than one pattern by changing the value of the <#> variable. For example: `INCLUDE_PATTERN_1`, `INCLUDE_PATTERN_2`, `INCLUDE_PATTERN_3`.
`EXCLUDE_PATTERN_<#>`	String	None	Patterns to be excluded from matching. Can contain more than one pattern by changing the value of the <#> variable. For example, `EXCLUDE_PATTERN_1`, `EXCLUDE_PATTERN_2`, `EXCLUDE_PATTERN_3`.
`ONLY_DIGITS`	Boolean	FALSE	Indicates whether matching should use only the digits. Setting this parameter TRUE removes all non-numeric characters in the string before matching. For example, `1234-5` is treated as `12345`.
`CHECK_DIGIT_CODE_VALIDATE`	String	None	Indicates whether to evaluate a checksum digit based on the last digit. Valid values: • `LUHN` • `ABA` • `CUSIP` • `DIHEDRAL` • `IBAN` • `UK_NHS` • `MOD11` • `ISBN10`
`DO_LOOKUP`	Boolean	FALSE	Indicates whether to use patterns specified by the `LOOKUP_PATTERN` parameter. If this parameter is set to TRUE, the patterns specified in `LOOKUP_PATTERN` are used.
`LOOKUP_DICT`	String	None	A dictionary name or key. See Dictionaries [58].
`LOOKUP_PATTERN`	String	None	Pattern for matching. See Patterns [62].
`ISO3166_CC_VALIDATE_FLAG`	Boolean	FALSE	Indicates whether to use Privacera-defined matching to validate an ISO two-character country code. If this parameter is set to TRUE, `ISO3166_CC_PATTERN` is used.
`ISO3166_CC_PATTERN`		None	A valid pattern for matching country codes. See Patterns [62].
`ISO3166_CC_LOOKUP_KEY`		None	Name of a defined dictionary. See Dictionaries [58].

Credit card model

The credit card model detects credit card numbers. It validates numbers based on the issuing network, length, and Luhn checksum.

Parameter	Type	Default	Meaning
`CC_PATTERN`	String	Privacera-supplied pattern for credit card numbers with range of digits, space or hyphen separated.	Credit card pattern, if you want to override the supplied pattern.
`DEFAULT_TYPES`	Boolean	True	Validate against known issuing network prefixes.
`LUHN_CHECK`	Boolean	True	Validate the Luhn checksum on the credit card number.

Supported credit card types

Credit Card Type	Description	Examples
American Express (AMEX) Card	Starts with starting with 34 or 37 and having 15 digits.	34xxxxxxxxxxxxx 37xxxxxxxxxxxxx
JCB	• Starts with 2131 or 1800 and followed by 11 digits. • Starts with 35 followed by 14 digit.	2131xxxxxxxxxxx 35xxxxxxxxxxxxxx

Credit Card Type	Description	Examples
Maestro	Starts with 5018, 5020, 5038, 6304, 6759, 6761, 6763 followed by 8 to 15 digits	6761xxxxxxxx
Master Card	• Starts with 51 to 55 and having 14 digits • Starts with 2221 and having 12 digits • Starts with 27 and followed by 13 digits.	51xxxxxxxxxxxx 2221xxxxxxxx 27xxxxxxxxxxx
Visa Card	Starts with 4 and followed by 13 or 16 digits.	4xxxxxxxxxxxx 4xxxxxxxxxxxxxxx
Diners Club Card	Starts with 300 to 305 or 3095 or 36 or 38 or 39 and followed by 14 digits.	300xxxxxxxxxxx 3095xxxxxxxxxx
VPay (Visa) Card	Starts with and followed by 13 or 19 digits.	4xxxxxxxxxxxx 4xxxxxxxxxxxxxxxxxx

Regular expressions to match credit card numbers

Models for credit cards can define additional custom regular expressions to match against credit card types and numbers not explicitly supported by this model. Data that matches these regexes and passes the Luhn check is tagged as CC.

These additional regular expressions are entered into the **Properties** field when you create your model, as described in Create models [69].

Some examples of regexes for credit cards:

- Match JCB credit card numbers:
  `ADDITIONAL_REGEX_JCB: ^((?:2131|1800|35\d{3})\d{11})$`
- Match Maestro credit card numbers:
  `ADDITIONAL_REGEX_MAESTRO: ^((?:5018|5020|5038|6304|6759|6761|6763)\d{8,15})$`

Regex property name

The property name in Privacera must have the following prefix:

`ADDITIONAL_REGEX`.

This can be followed by some identifying string for your needs.

Regex property value

- The regex value must indicate the beginning and end of the regexes by following this structure, as shown in the examples:
  `^(`*`your_regexes_here`*`)$`
- You should thoroughly test *`your_regexes_here`* before you put them into a Privacera Discovery model to verify that they return the desired results.

Interaction of regexes and Luhn checksum

If a regex matches but the Luhn checksum fails, the matched credit card number might not be tagged as CC. Verifying the Luhn checksum is enabled by default. So if the data is not tagged as CC as expected, you can disable verifying the Luhn checksum by setting the following property:

```
LUHN_CHECK:false
```

NOTE
Disabling the Luhn checksum is not recommended, because the credit card numbers should be checked for compliance to the number formats and algorithms.

Date of birth model

The Date of Birth model detects various date formats.

Parameter	Type	Default	Meaning
`MIN_AGE_YEARS`	Integer	5	Age lower threshold.
`MAX_AGE_YEARS`	Integer	100	Age upper threshold.
`USE_ALGO`	Boolean	True	Tagging is done based on an algorithm to detect random distribution.
`DATE_REGEX_`*`var1`*	String	–	Pattern that matches a custom date format *var1*.
`DATE_FORMAT_`*`var1`*	String	–	Date Format that matches the pattern for *var1*.

Pre-configured date formats are:

- International YYYYMD format with 4 digit year
- US MDY with 4 digit or 2 digit year
- Month abbreviated MDY

Additional formats can be configured. For example, configure a regex and a Java date format:

Parameter	Type
`DATE_REGEX_1`	`\d{4} \d{2} \d{2}`
`DATE_FORMAT_1`	`yyyy MM dd`

EIN model

The EIN model detects Employer Identification Number using patterns and digit validation.

Parameter	Type	Default	Meaning
`EIN_PATTERN`	String	Default	EIN digit pattern if you want to override the default pattern.
`VALIDATIONS`	Boolean	True	Age upper threshold.
`STRICT_PATTERN`	Boolean	True	Allow match only if EIN has exact format.

Geo latitude and longitude model

The geo model detects latitude and longitude coordinates. It can validate these values based on a geographical area.

Parameter	Type	Default	Meaning
`MIN_LAT`	Double	US min latitude	Lower limit (southern) on latitude.
`MAX_LAT`	Double	US max latitude	Upper limit (northern) on latitude.
`MIN_LONG`	Double	US min longitude	Lower limit (west) on longitude.
`MAX_LONG`	Double	US max longitude	Upper limit (east) on longitude.
`MIN_FRACTIONAL_DIGITS`	Integer	3	Minimum number of digits after the decimal point.

IMEI model

The IMEI model detects International Mobile Equipment Identity numbers that are used to identify mobile phones. It validates the Luhn checksum and the length of the IMEI.

ITIN model

The ITIN model detects Individual Tax Identifier Numbers (identifiers of individual taxpayers). It validates the format and digits of the ITIN.

Parameter	Type	Default	Meaning
`ITIN_PATTERN`	String	Default	ITIN digit pattern if you want to override the default pattern.
`STRICT_PATTERN`	Boolean	True	Allow match only if ITIN has exact format.

MIME model

The MIME model detects a file based on its Multipurpose Internet Mail Extensions type. The MIME type is detected using a combination of file extension and magic bytes in the header of the file. The detected MIME type is then looked up in a dictionary of MIME types.

Parameter	Type	Default	Meaning
`LOOKUP_DICT`	String	–	Identifier of dictionary of MIME types.

There are two pre-configured MIME model instances.

- For detecting executable files: `LOOKUP_DICT=EXEC_MIME_KEYWORD`.
- For detecting image files: `LOOKUP_DICT=IMAGE_MIME_KEYWORD`.

Phone number model

The Phone Number model detects phone numbers. It validates the format of the phone numbers based on the country for which it is configured.

Parameter	Type	Default	Meaning
`COUNTRY_CODE`	String	US	Two-character country code.

SSN model

The SSN model detects US Social Security Numbers. It validates the format and checks against a blacklist of SSN numbers.

Parameter	Type	Default	Meaning
`SSN_PATTERN`	String	Default	Override the default SSN pattern.
`VALIDATIONS`	Boolean	True	Validate against known blacklist of SSNs.
`STRICT_PATTERN`	Boolean	False	Allow match only if SSN has exact format.
`USE_9_DIGIT_PATTERN`	Boolean	False	Match against any nine digit number without format.
`USE_4_DIGIT_PATTERN`	Boolean	False	Match against any four digit number without format. Disables validation with blacklist of SSN.
`STRICT_EXT_PATTERN`	Boolean	True	Allow match only if SSN has exact format that is hyphen-, dot-, or space-separated.

Examples of Invalid SSNs

The SSN model would determine that the following SSNs are invalid.

- SSN starting with 9 or 666 or 000 or 98765432.
- SSN with 00 as the 4th and 5th digits.
- SSN with 0000 as the sixth through ninth digits.
- Any SSN like these:
  - 123456789
  - 111111111
  - 222222222
  - 333333333

- 444444444
- 555555555
- 666666666
- 777777777
- 888888888
- 999999999

VIN model

The VIN model detects Vehicle Identification Numbers. It validates the length and the VIN checksum.

Zip model

The Zip model detects US Zip codes. It detects both 5 digit and 5+4 digit variations and validates against a dictionary of US Zip codes.

Parameter	Type	Default	Meaning
`ZIP_DICT_KEY`	String	`US_ZIP_LOOKUP`	Key of the US Zip dictionary.
`ZIP_PATTERN`	String	Default	Validates content regular expression for list of ZIP codes.
`STRICT_PATTERN`	Boolean	False	Allow match only if Zip code has exact format. If set to true then only nine digits containing '-' and starting with five digits are considered a Zip code.

Default models

The following is a list of the default models in Privacera. For precise details, look at the model itself in the Platform UI.

- `DOB_ML_MODEL`
- `CC_ML_MODEL`
- `ZIP_ML_MODEL`
- `IMEI_ML_MODEL`
- `SSN_ML_MODEL`
- `EXEC_ML_MODEL`
- `MIME_ML_MODEL`
- `PHONE_NUMBER_ML_MODEL`
- `GEO_LAT_LONG_ML_MODEL`
- `CC_ML_MODEL_PROTECTED`
- `EIN_ML_MODEL`
- `ITIN_ML_MODEL`
- `VIN_ML_MODEL`
- `SSN_9_DIGIT_ML_MODEL`
- `SSN_4_DIGIT_ML_MODEL`
- `IMAGE_FILE_ML_MODEL`
- `IMAGE_ML_MODEL`

Create models

To create a model, follow these steps:

1. From the navigation menu, select **Discovery** > **Models**.
2. Click **Add Model**.
   The **Add Model** dialog is displayed.
3. In the **Name** field, enter a name for the model.
4. In the **Description** field, enter a description of the model.
5. In the **Key** field, enter a model key.
6. From the **Type** dropdown menu, select a model type.

> **NOTE**
> See Types of models [64] for more information.

7. From the **Apply For** dropdown menu, select **File content**.

> **NOTE**
> File content is resource content.

8. Enable or disable the model using the **Model Status** toggle.
9. Add model properties by clicking **+**.
10. Enter a key and value into the **Key** and **Value** field. For example: Key: MIN_FRACTIONAL_DIGITS, Value: 2. You can add multiple model properties.

> **NOTE**
> For example: Key: `MIN_FRACTIONAL_DIGITS`, Value: 2. You can add multiple model properties.

11. Click **Save**.
    The model is created.

Edit models

You can edit a model by clicking the **Edit** icon in the **Actions** column.

To edit a model, follow these steps:

1. Click the **Edit** icon in the **Actions** column.
   The **Edit Model** dialog displays.
2. Make your desired changes.
3. Click **Save**.
   The model is updated.

Delete models

You can edit a model by clicking the **Delete** icon in the **Actions** column.

To delete a model, follow these steps:

1. Click the **Delete** icon in the **Actions** column.
   The **Confirm Delete** dialog displays.
2. Select **Delete** to confirm the deletion.
   The model is deleted.

Import models

To import a model file in JSON format, follow these steps:

1. In the **Models** home page, click the **Import** option.
   The **Import** dialog is displayed.
2. Browse and select the JSON file and click **Import**.

The model file is imported.

Export models

To export a model file in JSON format, follow these steps:

1. In the **Models** page, click **Export**.
2. From the drop-down menu, select one of the following options:
   - **All Records**: Export the entire set of models.
   - **Select Records**: Select the specific model to export. You can select multiple models.
3. Click **Export**.
   The JSON file is exported.

Rules

You can create and manage custom and system-provided rules in Privacera Discovery. By executing the conditions in each rule, Discovery applies classifications to your data. The output tag associated with the processed rule is applied to the resource as the final tag.

The generation of tags [41] depends on the order of the rules. See Processing Order of Scan Techniques [26].

You can also create rule mappings.

Types of rules

There are three types of rules in Privacera Discovery:

- Structured
- Unstructured
- Post-processing

Example rules and classifications

Based on the tags found in a structured or unstructured rule or a table in various columns, we can assign a tag to the file or the table. This is an AND conditions of output tags. For example, you can set multiple rules as follows:

1. If a file has `PERSON_NAME`**AND**`EMAIL`**AND**`SSN` , tag as `PII`.
2. If a file has `USER_ID`**AND**`GEO`, tag as `SENSITIVE` .
3. If a file has `USER_ID`**AND**`IP` , tag as `SENSITIVE` .

Import rules and mappings

To import a JSON rule file for a structured rule, follow these steps:

1. From the navigation menu, select **Discovery** > **Rules**.
2. On the **Rules** page, click **Import**.
   The **Import** dialog is displayed.
3. Click **Choose File** and select the JSON file.

   > **NOTE**
   >
   > Selecting **Clean Previous** deletes all existing rules.

4. Click **Save**.

The rule file is imported.

Export rules and mappings

To export a rule file in JSON format for a structured rule, follow these steps:

1. From the navigation menu, select **Discovery** > **Rules**.
2. Click **Export**.

3. Select the files you wish to export.
4. Click **Export**

The rule file is exported.

Unstructured rules

Create an unstructured rule

To create an unstructured rule, follow these steps:

1. From the navigation menu, select **Discovery** > **Rules**.
2. On the **Rules** page, click **Unstructured** > **Create Rule**.
   The **Create Rule** dialog is displayed.
3. Enter the following details:
   - **Rule Name**: Name of the rule.
   - **Description**: Description of the rule (optional).
   - **Must Have**: From the dropdown menu, select dictionaries, patterns, or models to be included in the rule.
   - **Must Not Have**: From the dropdown menu, select dictionaries, patterns, or models to be excluded from the rule (optional).
   - **Word Proximity**: Name of a pattern to identify sensitive information within the specified number of words.
   - **Key order strict**: Using the toggle, indicate whether key order is strictly followed.
   - **Enable rule**: Using the toggle, enable or disable the rule.
4. Review the information in the **Rule preview** section.
5. Click **Save**.

The unstructured rule is created.

Structured rules

Default structured rules

The following is a list of the default structured rules in Privacera.

- Australia Bank Account Number
- Australia Bank BSB code
- Australia Driver License
- IBAN Rule
- rule_auto_1P
- rule_auto_2P
- rule_auto_3P
- rule_auto_4P
- rule_auto_5M
- rule_auto_6M
- rule_auto_7M
- rule_auto_8M
- rule_auto_9M
- rule_biometric
- rule_biometric_keyword
- rule_cc
- rule_city_name
- rule_criminal_keyword
- rule_dob

- rule_email
- rule_ethnicity_keyword
- rule_gps
- rule_gps_6_digit
- rule_medical_keyword
- rule_national_id
- rule_password
- rule_person_name
- rule_phonenumber
- rule_pii_id_keyword
- rule_political_keyword
- rule_religion_keyword
- rule_sexual_orientation_keyword
- rule_ssn_4_digit
- rule_ssn_9_digit
- rule_ssn_strict
- rule_ssn_strict_fallback
- rule_state_name
- rule_street_address
- rule_tax_id_9_digit
- rule_tax_id_strict
- rule_trade_union_keyword
- Rule US ABA Routing Number
- Rule US ABA Routing Number 2
- rule_us_dlicense_keyword
- rule_us_zip
- rule_viewership_keyword
- rule_web_keyword
- SWIFT BIC Bank ID rule
- SWIFT BIC Bank ID Rule 2
- UK Driver License Rule
- UK Electoral Roll number
- UK NHS Rule
- UK NHS Rule 2
- UK NINO Rule
- UK NINO RULE 2
- UK Phone Number Rule
- UK Postal Code
- UK Postal Town
- UK US Passport

Create a structured rule

To create a structured rule, follow these steps:

1. From the navigation menu, select **Discovery** > **Rules**.
2. On the **Rules** page, click **Structured** > **Create Rule**.
   The **Create Rule** dialog is displayed.
3. In the **Create Rule** dialog, enter the following details:
   - **Name**: The name of the rule.
   - **Description**: A description of the rule (optional).
   - **Must Have**: From the dropdown menu, select dictionaries, patterns, or models to be included in the rule.

- **Must Not Have**: From the dropdown menu, select dictionaries, patterns, or models to be included in the rule.
- **Score Type**: From the dropdown menu, select one of the following options:
  - **Auto**: If the rule is applied, the resource is classified as **System**.
  - **Review:** If the rule is applied, the resource is classified as **Pending Review**.
- **Output Tags**: The tags associated with the rule.
- **Key For Samples**: The keys from the objects in the **Must Have** dropdown menu.
- **Enable rule**: The rule is enabled or disabled.

4. Review the information in **Rule preview** section.
5. Click **Save**.
   The structured rule is created.

Reorder structured rules

Rule order decides the priority of the rules applied during classification.

To reorder rules, follow these steps:

1. On the **Rules** page, click **Reorder**.
2. Drag the rules up or down to change the order.
3. Click **Save Order**.
   The new order is saved.

Post-processing

With post-processing, the data is scanned and then the rules are applied on the tagged data in multiple passes. Post-processing can be used with both real-time and offline scans. Based on the output tags of the rules applied after the initial scan, with post-processing you can add additional tags on the parent or child data resources.

Post-processing rules should be applied after datazone and tag propagation is done.

For example, after the initial scan of a structured or unstructured file or columns within a table, will identify the data and classify them with tags based on the rules. After the initial scan has tagged various columns within a table or a file, you can use post-processing rules to assign additional tags to the file or the parent table.

Enable post-processing

To enable post-processing, follow these steps:

1. Navigate to **Setting** > **System Configuration**.
2. Search for the property `privacera.portal.rules.post_process.enable=false`.

   > **NOTE**
   > The default setting is false.

3. Set the property to true.

Example of post-processing rules on tags

1. From the navigation menu, select **Discovery** > **Rules**.
2. On the **Rules** page, select **Post-Processing**.
3. Create a new rule with the following condition: If `PERSON_NAME` and `SSN` are found, apply the `SENSITIVE` tag.
4. Rescan the file to apply the post-processing rules.
   The fields are now classified as `SENSITIVE` and the tag is applied in the unformatted view.

Rule mappings

Create a rule mapping

To create a rule mapping, follow these steps:

1. From the navigation menu, select **Discovery** > **Rules**.
2. On the **Rules** page, click **Rule Mapping** > **Add Mapping**.
   The **Add Key Tag Mapping** dialog is displayed.
3. From the **Key** dropdown menu, select a dictionary, pattern, or model.
4. From the **Tag** dropdown, select a tag.

   > **NOTE**
   > You can add multiple keys and tags by clicking **+**.

5. Click **Save**
   The rule mapping is created.

Configure scans

You can configure Privacera Discovery scans to suit your needs.

Scan setup

Using , you can configure scans and set threshold scores to determine if a resource should be reviewed for non-compliance. This is done from the **Scan Setup** page.

To view the **Scan Setup** page, select **Discovery** > **Scan Setup** from the navigation menu.

The **Scan Setup** page displays the following information:

- **Application Status**: The total number of enabled and disabled applications.
- **System Classification**: This allows you to set the global value at what percentage match will cause the scanned resource to be classified. To automatically classify the associated tags, enable the auto classification feature using the enable/disable toggle.
- **Minimum Review**: This allows you to set the global minimum value that will send the tagged resources to the Pending Review status under classification for manual verification. Tag scores falling below the review score are ignored.
- **Reduce Score**: If a column has empty data but is meta-tagged with 100% score, this reduces the score with the value that is set here. For example: If it is configured to 50, then the final score set for that column tag will be 50 and it will be re-evaluated based on the auto-classification and review score threshold.
  If you toggle the reduce score enable, it will reduce. If you toggle the reduce score enable, it will reduce the score of the associated meta tag. If you disable the reduce score feature, the meta tags will not be auto-classified.
- **Rescan Type**: For file system and database applications, scanning options include:
  - **Incremental**: Only scans resources that have been modified since the previous scan.
  - **Scan**: Rescans the resource completely regardless of previous scans.

Adjust default scan depth on Privacera Platform

operations are computationally intensive. Therefore, Discovery defaults to scanning only a sample of targeted data in order to determine whether sensitive information is present.

Individual customers are responsible for determining what level of scanning is necessary to meet their regulatory requirements. You can adjust the sampling size by setting the `DISCOVERY*MAX*` variables detailed in Set custom Discovery properties on Privacera Platform [134].

Classifications using random sampling on PrivaceraCloud

By default PrivaceraCloud scans at "shallow depth" of a database. That is, for performance, the system examines the records of the database to derive the classifications.

This default assumes that the database itself is uniform with normalized data and the records are accurately represented.

However, with some unnormalized databases, this uniformity might be lacking.

If you suspect that your database values are not uniform, you can configure PrivaceraCloud to take a *random sample* from the entire database for analysis in classification.

One purpose of random sampling is to help isolate these data variations to eliminate them.

Supported JDBC applications for random sampling

Random sampling is supported for the following applications:

- MySQL
- Oracle
- Trino

If you configure random sampling for any other database, it is ignored.

Prerequisites for random sampling

- Know the names of the applications you want to randomly sample.
- Be sure to have the JDBC connection details for those applications.
- To minimize performance impact, determine if your database can be considered "large". By default, PrivaceraCloud considers any database with 10,000 records or more to be large. In this case, the random sampling is based on a subset of the data.

Define datasource (application) and configure random sampling

Random sampling is part of configuring a datasource. For details on setup, see Applications.

Enable random sampling

To enable random sampling for a database:

1. Go to **Settings > Applications**.
2. Under **Connected Applications**, click the name of the application.
3. On the **BASIC** tab, Click the toggle *jdbc.random.record.fetching*.
4. If your database has more than 10,000 records, specify the approximate number in the *rows.as.small.dataset* field.

Effects of random sampling

Random sampling has some visible effects.

Performance impact

You might perceive a delay in the running of random samples. Performance times can increase depending on the size of the sample.

Variations in classifications

You should not expect the same classification results for the same database from random sample to random sample.

Each random sample operates on a subset of the data. Depending on variations in the sampling of values in the database, the results of classification can vary.

Each random sample is unique. The records are selected randomly and so results vary from sample to sample.

For example, suppose an **EMAIL** column does not have consistent values:

1. Sometimes, a delimiter that distinguishes a first and a last name with @-sign indicating Internet domain.
   A random sampling of such records can result in a consistent classification as **PERSON NAME**.
2. Sometimes a bare username with no delimited last name and with no @-sign at all.
   The inconsistent variation in the data makes a concrete classification difficult to derive.

This same inherent inconsistency in the column values can result in variations of classification from run to run, each with its own unique random sampling.

Enable Discovery Realtime Scanning Using IAM Role on PrivaceraCloud

In this topic, you will learn how to use IAM roles to configure AWS S3 service for realtime scanning.

Create an IAM role with AWS S3 permissions

1. Log in to the AWS console.
2. Go to **Identity and Access Management (IAM)** and navigate to **Access management > Users/Groups/Roles**.
3. Create a role or edit an existing AWS IAM role. Refer to AWS documentation on how to create an IAM Role.
4. Navigate to the role created or the role you are editing.
   a. Open the role.
      The role **Summary** page is displayed.
   b. Copy the **Role ARN**.
      Use the ARN in **IAM Role ARN** field when providing **Application Properties** details for the data source.
5. Add a policy to AWS IAM role.
   a. Open the role you are editing.
   b. Click **Permissions** tab.
   c. On the **Permissions Policies** section, click **Attach Policies** or **Add inline policy**.
      The **Create policy** page is displayed.
   d. Click the **JSON** tab to add the policy and permissions.
      Refer to the following sample permission JSON for the role on S3 bucket. Ensure your have **Get** and **List** actions in the permissions policy of the role and enter the bucket name in `bucket-name`.

      NOTE

      You can scan multiple buckets in multiple regions or same region from a single IAM role that is configured as part of data source. This single IAM role should have access permission to access these buckets.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAccountLevelS3Actions",
            "Effect": "Allow",
            "Action": [
                "s3:ListAllMyBuckets",
                "s3:Get*"
            ],
            "Resource": "*"
        },
        {
            "Sid": "AllowListAndReadS3ActionOnMyBucket",
            "Effect": "Allow",
            "Action": [
                "s3:Get*",
                "s3:List*"
            ],
            "Resource": [
                "arn:aws:s3:::bucket-name/*",
                "arn:aws:s3:::bucket-name",
                "arn:aws:s3:::bucket1-name/*",
                "arn:aws:s3:::bucket1-name",
                "arn:aws:s3:::bucket2-name/*",
                "arn:aws:s3:::bucket2-name",
                "arn:aws:s3:::bucket3-name/*",
                "arn:aws:s3:::bucket3-name",
                "arn:aws:s3:::bucket4-name/*",
                "arn:aws:s3:::bucket4-name",
                "arn:aws:s3:::bucket5-name/*",
                "arn:aws:s3:::bucket5-name",
                "arn:aws:s3:::bucket6-name/*",
                "arn:aws:s3:::bucket6-name"
            ]
        },
        {
            "Sid": "AllowReadS3ActionOnMyQueue",
            "Effect": "Allow",
            "Action": [
                "sqs:ReceiveMessage",
                "sqs:DeleteMessage",
                "sqs:GetQueueUrl"
            ],
            "Resource": [
                "<ARN of SQS queue>"
            ]
        }
    ]
}
```

NOTE

Multiple buckets of the same region can be configured to a single SQS queue. Bucket should be mapped to the configured SQS queue in the above policy.

e. Click **Review policy**.
   The **Review policy** section is displayed.
f. Enter the policy **Name** and click **Create policy**.

6. Establish IAM Role Trust Relationship with Discovery Data Access Role.
   a. Open the role role you are editing.
   b. Click the **Trust relationships** tab.

c. Click **Edit trust relationship**.

d. Refer to the following JSON to add a new policy document.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::870790086151:role/DISCOVERY_PROD_DATA_ACCESS_
                "Service": "s3.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }

    ]

}
```

e. Click **Update Trust Policy** to save this revision.

Configure AWS S3 access using IAM role

Connect application

1. Go to **Settings** > **Applications**.
2. On the **Applications** screen, select .
3. Enter the application **Name** and **Description**, and then click **Save**.

You can see and with the toggle buttons.

NOTE

If you don't see in your application, enable it in **Settings** > **Account** > **Discovery**.

Enable

1. Click the toggle button to enable for your application.
2. On the **BASIC** tab, enter values in the following fields.
   - With **Use IAM Role** disabled:
     a. AWS Access Key: AWS data repository host account Access Key.
     b. AWS Secret Key: AWS data repository host account Secret Key.
     c. AWS Region: AWS S3 bucket region.
        For the first time, real-time discovery is disabled.
   - With **Use IAM Role** enabled:
     a. IAM Role ARN: Enter the actual IAM Role using a full AWS ARN.
     b. AWS Region: AWS S3 bucket region.
3. On the **ADVANCED** tab, you can add custom properties.
4. Using the **IMPORT PROPERTIES** button, you can browse and import application properties.
5. Click the **TEST CONNECTION** button to check if the connection is successful, and then click **Save**.

Go to **Data Source** to add a resources using this connection as Discovery targets.

Enable Real-time Scanning on ADLS Gen 2 on PrivaceraCloud

Prerequisites

Ensure the following prerequisites are met. To configure them, see About the Account page on PrivaceraCloud.

- Select **Enable Real-Time Scanning**.
- Configure Event Hub for scanning.
- Create Consumer Group for Pkafka.
- Configure Checkpoint Storage for Pkafka.

Create a Storage Account and Event Subscription for Scanning

1. Log in to Azure Portal.
2. Use an existing storage account or create a new one. For more information, see Create a storage account.

   Use this storage account name in **Storage Account Name** when providing **Application Properties** details for the datasource.
3. Get Storage Account Key:
   a. Navigate to the storage account.
   b. Under **Security + networking**, click **Access keys**.
   c. Click **Show Keys** for keys to be populated.
   d. Use appropriate key value in **Storage Account Key** when providing **Application Properties** details for the datasource.
4. Use an existing container or create a new one. Refer to Microsoft documentation on how to create a container.
5. Get URL Prefix:
   a. Navigate to the container and click **Properties**.

      Container property details are populated on the right.
   b. Use the URL prefix in the **Application Properties** details for the datasource.
6. Create a event subscription. Refer to Microsoft documentation on how to Create an Event Grid subscription.
   a. Navigate to the storage account.
   b. On the left menu, select **Events** and click **+ Event Subscription**.

      Create Event Subscription page is displayed.
   c. On the **Create Event Subscription** page within the **Basic** tab, provide the following values:
      i. Enter the **Event Name** and **Event Schema**.
      ii. Topics Details are auto populated.
      iii. Choose Event Type as **Blob Created** and **Blob Deleted**.
      iv. Choose Endpoint type as **Event Hubs**.
      v. Select an Endpoint from **Select Event Hub** dialog.
         A. From the **Event Hub Namespace** dropdown, choose the Event Hub Namespace you created.
         B. From the **Event Hub** dropdown choose the Event Hub you created.
         C. Click **Select Confirmation**.
      vi. Click **Create**.

   NOTE

   It is recommended to disable *soft delete* on blob storage account as ORC and Parquet file scanning is not supported when *soft delete* is enabled.

Connect ADLS Gen2 Application for Data Discovery on PrivaceraCloud

1. Go the **Setting > Applications**.
2. In the **Applications** screen, select **ADLS Gen2**.
3. Enter the application **Name** and **Description**, and then click **Save**.
4. Click the toggle button to enable the **Data Discovery** for ADLS Gen2.
5. In the **BASIC** tab, enter the values in the following fields:
   - **JDBC URL**
   - **JDBC Username**
   - **JDBC Password**
6. In the **ADVANCED** tab, you can add custom properties.
7. Click **TEST CONNECTION** to check if the connection is successful, and then click **Save**.
8. To add a resource to be scanned in real-time, navigate to **Discovery > Data Source**. See Data sources on Privacera Platform [26].
9. To verify the scan status, navigate to **Discovery > Scan Status**.
10. To see the scan results, navigate to **Data Inventory > Classifications**.

Enable Real-time Scanning of S3 Buckets on PrivaceraCloud

To enable realtime scanning of S3 buckets:

1. To **Enable Real-Time Scanning** for AWS S3, see About the Account page on PrivaceraCloud.
2. To connect a new AWS S3 application, see Connect S3 to PrivaceraCloud. Alternatively, to edit an existing AWS S3 application:
   a. Go the **Setting > Applications**.
   b. In the **Applications** screen, select S3.
   c. Click the pen icon next to the **Account Name**.
   d. Disable and enable the toggle button to see the configuration screen.
   e. Click the **Real-Time Enable** toggle button.
   f. Click the clipboard icon to copy the **Real-Time Event Name**, which will be used to configure event notifications from S3 buckets in the AWS account.
   g. Click **SAVE**.
3. Apply access policy in the SQS Queue to allow the S3 bucket to send events. Refer to the AWS documentation for detailed information on configuring access policies - Click here
   a. Navigate to SQS Queue and select the queue (test_queue).
   b. Provide the correct Access Policy to SQS queue, so that S3 is allowed to put events into SQS queue. Refer to the following example to apply access policy:

```
{"Version":"2008-10-17","Id":"__default_policy_ID","Statement":[{"Sid":"__owner_s
```

4. Configure event notifications from S3 buckets to the SQS Queue. See the AWS documentation for detailed information.
   a. Go to the S3 bucket you want to link with the SQS queue.
   b. On the **Properties** tab, navigate to the **Event Notifications** section and choose **Create event notification**.
   c. In the event name, paste the **Real-Time Event Name** copied from the step 2.e. Enter a bucket name, for example, `test-bucket`.
   d. Select the event type as required from **Event types**.
   e. Select **Destination** type as **SQS Queue**, and then choose the SQS queue (test_queue) from the dropdown list.
   f. Click **Save Changes**.
5. Include and scan resources from datasource.
   a. Navigate to **Discovery > Data Source**.
   b. On the **Data Source** page, click the S3 application that needs to be set up for realtime scanning. The selected S3 application details are displayed.

c. Click **Include Resources** tab and ensure that the check mark is displayed when the realtime scanning is enabled.

d. Click **Add** to add a resource.

Connect ADLS Gen2 Application for Data Discovery on PrivaceraCloud

1. Go the **Setting > Applications**.
2. In the **Applications** screen, select **ADLS Gen2**.
3. Enter the application **Name** and **Description**, and then click **Save**.
4. Click the toggle button to enable the **Data Discovery** for ADLS Gen2.
5. In the **BASIC** tab, enter the values in the following fields:
   - **JDBC URL**
   - **JDBC Username**
   - **JDBC Password**
6. In the **ADVANCED** tab, you can add custom properties.
7. Click **TEST CONNECTION** to check if the connection is successful, and then click **Save**.
8. To add a resource to be scanned in real-time, navigate to **Discovery > Data Source**. See Data sources on Privacera Platform [26].
9. To verify the scan status, navigate to **Discovery > Scan Status**.
10. To see the scan results, navigate to **Data Inventory > Classifications**.

Include and exclude resources in GCS

Add a bucket in the inclusion

1. From the navigation menu, select **Discovery > Data Source**.
2. From the **Applications** menu, select **GCS-Google Cloud Storage**.
3. In the **Include Resource** tab, add the resources that need to be scanned and discovered by .
   You can use the following patterns to add resources:
   - gs://bucket_name
   - gs://bucket_name/folder_child/
   - gs://bucket_name/*
   - gs://bucket_name/filename.format
   - gs://<bucket_name>/<folder_name>/*

Add a bucket in the exclusion

1. From the navigation menu, select **Discovery** > **Data Sources**.
2. From the **Applications** menu, select **GCS-Google Cloud Storage**.
3. In the **Exclude Resource** tab, add the resources that need to be ignored by .
   You can use the following patterns to add resources:
   - *.format
   - */file_path/*
   - * folder_name/file*
   - * file_name *
   - file_name*
   - * file_name
   - gs://<bucketname>/<folder_name>/*.file_extension

Configure real-time scan across projects in GCP

You can enable real-time scan for applications in different projects in GCP. An application in GCP can be Google Cloud Storage (GCS) or Google BigQuery (GBQ).

By default, only one application of GCS is created at the time of installation. If you have multiple projects containing resources in GCP and want to scan them in real-time, then do the following:

Prerequisites

Ensure the following prerequisites are met:

- Get the project IDs of each project:
  - Project where the instance is configured
  - Cross project(s) containing the resources to be scanned
- Give permissions to the project instance to access the cross project resources (GCS buckets, GBQ datasets).

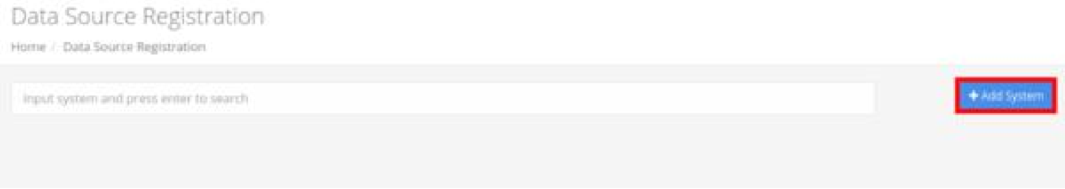
  1. Get the service account name of the project where the instance is configured.
  2. Navigate to the cross project > **IAM & Admin** > **IAM** > **Add**.
  3. Enter the service account name, and add the following roles:
     - Editor
     - Private Logs Viewer

Configuration

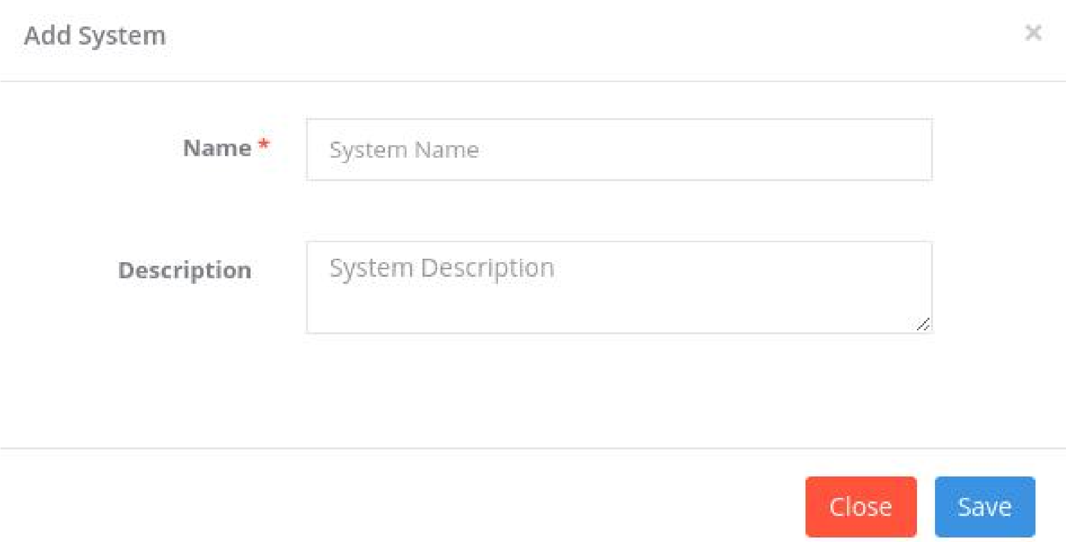
1. Add the following property to the `vars.discovery.gcp.yml` YML file, and assign the projects IDs.

   ```
   PKAFKA_CROSS_PROJECT_IDS=project_id_2,project_id_3
   ```
2. Run the following commands.

   ```
   cd ~/privacera/privacera-manager
   ./privacera-manager.sh update
   ```
3. After installing/updating Privacera Manager, add the GCP projects in Privacera Portal.
   a. In Privacera Portal, add new GCS and GBQ with the project ID.
      i. On the Privacera home page, expand the **Settings** menu and click on **Data Source Registration** from left menu.
      ii. On the Data Source Registration page, click **+Add System**.


Data Source Registration
Home / Data Source Registration
Input system and press enter to search
+ Add System

The Add System pop-up displays.



iii. Enter System Name in the **Name** field. (Mandatory) Example: Azure

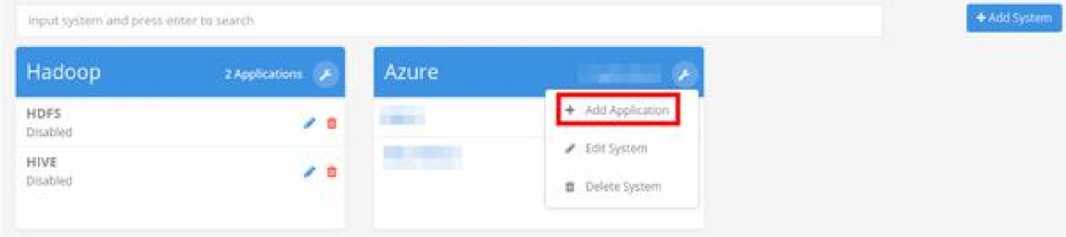
iv. Enter the description in the **Description** field. (Optional)

v. Click **Save**.

The Application page displays with newly added system.

Now, let's add the application in system, use the following steps:

i. Click on the **Setting icon** of the system and then click **+Add Application**.

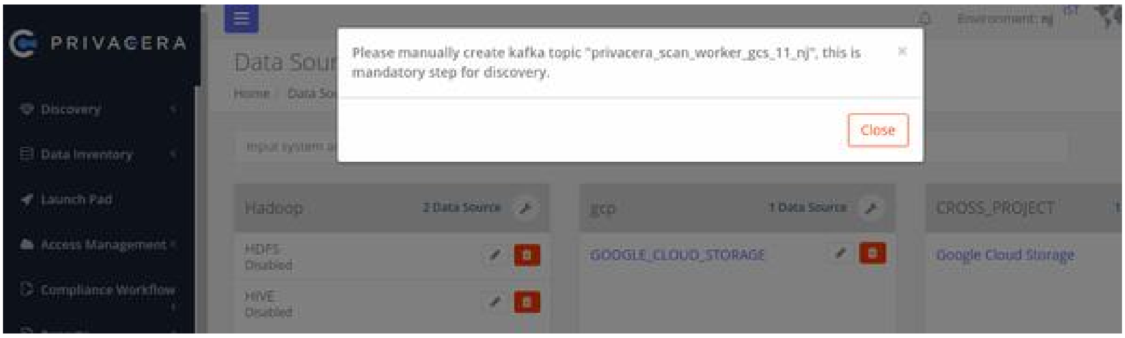


ii. Select the Application. Example: Google Cloud Storage

iii. Enter the **Application Name**, **Application Code**, and **Project ID**. (Mandatory)

iv. Click **Save**.

b. After adding the application, you will be instructed to manually create a topic in the GCP Console as shown in the image below.



Please manually create kafka topic "privacera_scan_worker_gcs_11_nj", this is mandatory step for discovery.

In the image, the topic name is **privacera_scan_worker_gcs_11_nj**. Use this name to create a topic on the instance where Privacera is installed. For more information on creating a topic in GCP, see Create and manage topics.

Enable offline scanning on ADLS Gen 2 on PrivaceraCloud

Get Azure Storage account name, account key, and URL prefix

1. Login to Azure portal.

2. Use an existing storage account or create a new storage account. Use this storage account name in the **Storage Account Name** field when providing **Application Properties** details for the datasource. For more information, see Create a Storage Account.
3. Use an existing storage container or create a new storage container. For more information, see Create a Container.
4. Get an Azure Storage account key:
   a. Navigate to the storage account.
   b. Under **Security + networking**, click **Access keys**.
      The key details are populated on the right.
   c. Click **Show Keys** for keys to be populated.
   d. Use **Key1** value in **Storage Account Key** when providing **Application Properties** details for the datasource.
5. Get the URL prefix:
   a. Navigate to the container, and then click **Properties**.
      The container property details are populated on the right.
   b. Use the URL prefix in the **Application Properties** details for the datasource.

Connect ADLS Gen2 Application for Data Discovery on PrivaceraCloud

1. Go the **Setting > Applications**.
2. In the **Applications** screen, select **ADLS Gen2**.
3. Enter the application **Name** and **Description**, and then click **Save**.
4. Click the toggle button to enable the **Data Discovery** for ADLS Gen2.
5. In the **BASIC** tab, enter the values in the following fields:
   - **JDBC URL**
   - **JDBC Username**
   - **JDBC Password**
6. In the **ADVANCED** tab, you can add custom properties.
7. Click **TEST CONNECTION** to check if the connection is successful, and then click **Save**.
8. To add a resource to be scanned in real-time, navigate to **Discovery > Data Source**. See Data sources on Privacera Platform [26].
9. To verify the scan status, navigate to **Discovery > Scan Status**.
10. To see the scan results, navigate to **Data Inventory > Classifications**.

Include and exclude datasets and tables in GBQ

1. From the navigation menu, select **Discovery > Data Source**.
2. From the **Applications** menu, select **GBQ-Google BigQuery**, and then click **Include Dataset or Table**.
3. In the **Include Dataset or Table** tab, add the dataset or table that needs to be scanned and discovered by .
   You can use the following patterns:
   - */sales_2018
   - */sales_2020
   - *finance_sandbox1/*
   - eCommerce/sales_2020
   - finance*/sales*

Exclude a dataset or table in GBQ

1. From the navigation menu, select **Discovery > Data Source**.
2. From the **Applications** menu, select **GBQ-Google BigQuery**, and then click **Exclude Dataset or Table**.

3. In the **Exclude Dataset or Table** tab, add the dataset or table that needs to be ignored by .
   You can use the following patterns:
   - */sales_2018
   - */sales_2020
   - *finance_sandbox1/*
   - eCommerce/sales_2020
   - finance*/sales*

Google Sink to Pub/Sub

This topic covers how to use a Sink based approach to read the real time audit logs for real time scanning in **Pkafka** for Discovery, instead of using the Cloud logging API. The following are key advantages of Sink based approach:

- All the logs will be synchronized to a Sink.
- Sinks are exported to a destination Pub/Sub topic.
- **Pkafka** subscribes to the Pub/Sub topic and it will read the audit data from the topic and will pass on the Privacera topic and a real time scan will be triggered.

You need to create following resources on Google Cloud Console:

1. Destination to write logs from Sink: Following destination are available to write logs from Sink:
   a. Cloud Storage
   b. Pub/Sub Topic
   c. Big Query
   In this document, Pub/Sub Topic [86] is considered as a destination for a Sink.
2. Create a Sink [87]

Create Pub/Sub topic

1. Log on to Google Cloud Console and navigate to Pub/Sub topics page.
2. Click the **+ CREATE TOPIC**.
3. In the **Create a topic** dialog, enter the following details:
   - Enter the unique topic name in the **Topic ID** field. For example, **DiscoverySinkTopic**.
   - Select **Add a default subscription** checkbox.
4. Click **CREATE TOPIC**.

   > **NOTE**
   > If required, you can create a subscription in a later stage, after creating the topic, by navigating to **Topic** > **Create Subscription** > **Create a simple subscription**.
   >
   > Note down the subscription name as it will be used inside a property in Discovery.

5. If you created a default subscription, or created a new subscription, you need to change the following properties:
   - **Acknowledgement deadline**: Set as 600.
   - **Retry policy**: Select as **Retry after exponential backoff delay** and enter the following values:
     - Minimum backoff(seconds): `10`
     - Maximum backoff (seconds): `600`
6. Click **Update**.

> **NOTICE**
>
> You can configure GCS lineage time using custom properties, that are not readily apparent by default. See Set custom Discovery properties on Privacera Platform [134].

Create a Sink

1. Login to the Google Cloud Console and navigate to the **Logs Router** page. You can perform the above action using the Logs Explorer page as well by navigating to **Actions** > **Create Sink**.
2. Click **CREATE SINK**.
3. Enter Sink details:

   a. **Sink name** (*Required*: Enter the identifier for Sink.

   b. **Sink description** (*Optional*): Describe the purpose, or use case for the Sink.

   c. Click **NEXT**.
4. Now, enter Sink destination:

   a. Select Sink service.

   b. Select the service where you want your logs routed. The following services and destinations are available:

   - Cloud Logging logs bucket: Select or create a Logs Bucket.
   - BigQuery: Select or create the particular dataset to receive the exported logs. You also have the option to use partitioned tables.
   - Cloud Storage: Select or create the particular Cloud Storage bucket to receive the exported logs.
   - Pub/Sub: Select or create the particular topic to receive the exported logs.
   - Splunk: Select the Pub/Sub topic for your Splunk service.
   - Select as **Other Project**: Enter the Google Cloud service and destination in the following format:

     ```
     SERVICE.googleapis.com/projects/PROJECT_ID/DESTINATION/DESTINATION_ID
     ```

     For example, if your export destination is a Pub/Sub topic, then the Sink destination will be as following:

     ```
     pubsub.googleapis.com/projects/google_sample_project/topics/sink_new
     ```
5. Choose which logs to include in the Sink:

   Build an inclusion filter: Enter a filter to select the logs that you want to be routed to the Sink's destination. For example:

   ```
   (resource.type="gcs_bucket" AND
   resource.labels.bucket_name="bucket-to-be-scanned" AND
   (protoPayload.methodName="storage.objects.create" OR protoPayload.methodName="storage
   protoPayload.methodName="storage.objects.get")) OR
   resource.type="bigquery_resource"
   ```

   Add all of the bucket names you want to scan in the above filter as resources in Discovery.

   ```
   bucket_name="bucket-to-be-scanned" AND
   ```

   In case of multiple buckets, you will need to specify it as an “OR” condition, for example:

   ```
   (resource.type="gcs_bucket" AND resource.labels.bucket_name="bucket_1" OR resource.la
   ```

   In above example, three buckets are identified to be scanned - `bucket_1`, `bucket_2`, `bucket_3`.
6. Click **DONE**.

Cross-project scanning

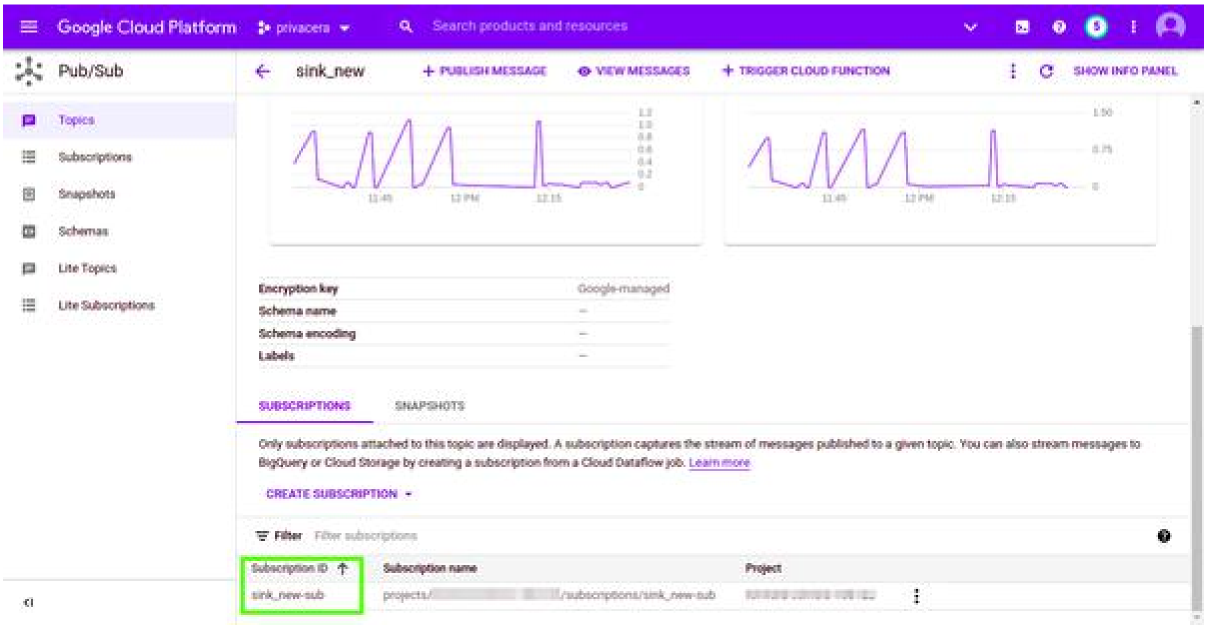
- For cross project scanning of GCS & GBQ resources, you need to create a Sink in another project and add the destination as a Pub/Sub topic of project one.
- You can refer to the same step as mentioned above for creating the Sink in the destination by navigating to **Destination** > Select as **Other project** and enter the **Pub/Sub topic name** in the following format: 'pubsub.googleapis.com/projects/google_sample_project/topics/sink_new'
- To access the Sink created in another project, you need to add the Sink writer identity service account in the IAM administration page of the project where you have the **Pub/Sub** topic and the VM instance present.
- To get the Sink Writer Identity, perform the following steps:
  - Go to the **Logs Router** page > select the **Sink** > select the dots icon > select **Edit Sink Details** > **Writer Identity** section, copy the service account.
  - Go to the **IAM Administration** page of the project where you have the **Pub/Sub Topic** and the VM instance > select **Add member** > **Add the service account** of the **Writer Identity** of the Sink created above.
  - Choose the role **Owner and Editor**
  - Click **Save**. Verify whether the service account which you added is present as a member on the **IAM Administration** page.

Google Sink configuration properties

- Add the following properties to the file: `vars.pkafka.gcp.yml`

  ```
  PKAFKA_USE_GCP_LOG_SINK_API: "true"
  PKAFKA_GCP_SINK_DESTINATION_PUBSUB_SUBSCRIPTION_NAME: ""
  ```
- For the above property, add the **Subscription name** as the value created in the **Pub/Sub Topic**.

Note that **Subscription ID** can be used as the value of the above property. Refer to the following screenshot for more information.



Encryption key	Google-managed
Schema name	–
Schema encoding	–
Labels	–

Only subscriptions attached to this topic are displayed. A subscription captures the stream of messages published to a given topic. You can also stream messages to BigQuery or Cloud Storage by creating a subscription from a Cloud Dataflow job. Learn more

Subscription ID ↑	Subscription name	Project
sink_new-sub	projects/[redacted]/subscriptions/sink_new-sub	[redacted]

Data Zones on Privacera Platform

The **Data Zones** page on Privacera Platform displays information about your data zones. This information is displayed in five different tabs:

- **Resources**: This tab allows you to add files and folders for scanning so that you can apply policy to them. You can filter the list of resources using the search bar. The **Resources** tab displays the following information:
  - **Application**: The name of an application.
  - **Resource**: The name of a resource.
  - **Re-evaluate**: Allows you to re-validate resource files. Before selecting **Re-evaluate** , the resource file must already be scanned. This option is only available in the *Right to Privacy* policy and *Expunge* policies because these policies do not work with real-time and offline scans.
  - **Actions**: Allows you to edit or delete a resource.
- **Delegated Admin**: A delegated admin has permission to scan data zone resources. By default, the delegated admin is *privacera*. Click the edit icon to change the delegated admin name.
- **Owners**: A list of owners. You can filter the list using the search bar. The **Owners** tab displays the following information:
  - **Owner**: The name of the owner.
  - **Description**: The description of the owner.
  - **Actions**: Allows you to edit or delete an owner.
- **Policies**: A list of policies. You can filter the policy list using the search bar. The **Policies** tab displays the following information:
  - **Policy**: The name of the policy.
  - **Type**: The type of policy.
  - **Conditions**: The conditions pertaining to the policy.
  - **Alert Level**: The alert levels: High, Medium, or Low.
  - **Actions**: The actions related to policy.
  - **Enabled**: The status of policy: Enabled or Disabled.
  - **Settings**: This allows you to edit the policy as well as you can delete the policy on clicking on respective icon under Settings column.
- **Tags**: This tab displays the tags associated with the data zone. You can modify the tags by clicking the **Edit**.

Planing data zones on Privacera Platform

Before you create a data zone, you should:

- Identify the data owners and data governors for the data zone. Make sure these people have been added to Privacera as users.
- Identify the resources, data sources and applications that should be included in the data zone.
- Decide on a useful name and explanatory description for the data zone
- Study the types of data zone policies [89] to determine the kinds of policies you want to enforce in the data zone.

Data Zone Dashboard

Data zones are used to group and label areas within your data lake to serve specific, well defined purposes. You can apply different policies and workflows to the resources in your data zones for tailored control over your data.

Datazone - Information page

Information about individual data zones can be viewed on the **Datazone - Information** page.

To view the **Datazone - Information** page, do the following:

1. From the navigation menu, select **Compliance Workflow** > **Data Zone Dashboard**.
2. Select the data zone you want to view.
   The **Datazone - Information** page displays.

The **Datazone - Information** page displays the following information:

- **Resource**: The list of resources.
- **Tag**: The list of tags.
  - **Show All Tag**: View all of the tags. By default, this is disabled.
  - **Add / Edit:** Add or edit the existing tags.

Datazone - Information page search filters

You can apply the following search filters on the **Datazone - Information** page:

- **Search by Resource**: Search using resource names.
- **Search by Application**: Filter results by selecting an application from the dropdown menu.
- **Search by Tags**: Filter results by selecting tags from the dropdown menu.

Enable data zones on Privacera Platform

To enable a data zone, do the following:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. On the Data Zones page, select the created data zone and enable it using the **Status** toggle.
   The data zone is enabled.

Add resources to a data zone on Privacera Platform

You can add two types of resources to a data zone:

- Files
- Database table names

To add resources to an existing data zone, do the following:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. Select a data zone from the **Data Zones** menu and click **ADD RESOURCE**.
   The **Add Resource** dialog is displayed.
3. Select an application from the **Application** dropdown menu (required).
4. In the **Resource** field, enter a resource name.

   > **NOTE**
   > You can add * wildcard entries for the table name.

5. Click **Save**.
   The File Format resource is added.

   > **NOTE**
   > Similarly, you can add the Table format resource. i.e. DB Name and Table Name.

6. Click **Save** to create the Resource.

Create a data zone on Privacera Platform

To create a data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. In the Data Zones page, click **+**.
   The **Add Data Zone** dialog is displayed.
3. In the **Data Zone Name** field, enter a name for the data zone.
4. In the **Description** field, enter a description (optional).
5. Click **Save**.
   The data zone is created.

Edit data zones on Privacera Platform

To edit an existing data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. In the **Data Zones** page, select the data zone to edit and click **Edit**.
   The **Edit Data Zone** dialog is displayed.
3. In the **Data Zone Name** field, enter a name for the data zone (required).
4. In the **Description** field, enter description of the data zone.
5. Click **Save**.

The data zone is updated.

Delete data zones on Privacera Platform

To delete a data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. On the Data Zones page, select the created data zone and click **Delete**.
   The **Confirm Delete** dialog displays.
3. Click **Delete**.
   The data zone is deleted.

Import data zones on Privacera Platform

To import a data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. In the Data Zones page, click the **Import** icon.
   The **Import Data Zone** dialog is displayed.
3. Browse and select the JSON file you want to import.

   > **NOTE**
   > Only JSON format is allowed.

4. Click **Import**.
   The data zone is imported.

Export data zones on Privacera Platform

To export a data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. On the Data Zones page, click the **Export** icon.

3. Select the Data Zone(s) you want to export and click **Export**.
   The **Export Data Zone** dialog displays.
4. Select either **JSON** or **CSV** as the export format.
5. Click **Export**.
   The data zone is downloaded to your computer.

You can filter the data zone list using the Search Data Zone option. Also, the refresh feature allows you to view the updated datazone list.

Disable data zones on Privacera Platform

To disable a data zone, do the following:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. On the Data Zones page, select the created data zone disable it using the **Status** toggle.
   The data zone is disabled.

Create tags for data zones on Privacera Platform

To create a tag for data zone, do the following:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. In the **Data Zones** page, select an existing data zone and click the **Tags** tab.
3. Click **Edit** and select the **Tag(s)**.
4. Select the tag(s) from the **Tags** dropdown menu.
5. Click **Save**.

The tags are created.

Data zone movement

To view a summary of data zone movement, select **Compliance Workflow** > **Data Zone Movement** from the navigation menu.

View undefined data zone movements

On the **Data Zone Movement** page, click **Show Undefined Zone Movements** to view undefined zone movements.

Filter data zone movements

You can filter the data zone list using the Filter Data Zone option. You can also filter data zone movements by date range, including:

- Today
- Yesterday
- Last 30 Days
- This Month
- Last Month
- Custom Range

> **NOTE**
> By default, the date range is set to Last 7 Days.

Click **Refresh** to refresh the list of data zones.

Data zones overview

On Privacera Platform, data zones on are distinct areas in a data lake that serve specific and well-defined purposes.

Data owners and data governors can create data zones based on domains, business functional ownership, or other logical groupings. Some examples of data zones:

- A data zone to manage customer data under the guardianship of a customer data steward.
- A data zone to manage finance data assets under the guardianship of a data administrator from the finance organization.

Data zones simplify data access management and relieve IT of the burden of managing policies for the entire enterprise. The administrative function for a data zone can be delegated to specific data owners who have the proper permissions/roles to administer the zone. Administrators can apply selective workflow policies [105] to their data zones.

Configure data zone policies on Privacera Platform

Data zone policies are configured to monitor resources in a particular data zone or data lake. Alerts can be raised based on restricted users, user groups, subnets, subnet-range, tags, and restricted zones.

To create a policy for data zone, follow these steps:

1. From the navigation menu, select **Compliance Workflow** > **Data Zones**.
2. In the **Data Zones** page, select the data zone and click the **Policies** tab.
3. Click **Add Policy**.
   The **Add Policy** dialog is displayed.
4. In the **Name** field, enter a name for the policy (required).
5. Select an alert level from the **Alert Level** dropdown menu.
6. Select a policy type from the **Type** dropdown menu (required).

   > **NOTE**
   > This will change the Source label as needed. By default, **Disallowed Movement** policy is selected.

7. Enter a description into the **Description** field.
8. Using the **Status** toggle, set the status of the policy. By default, it is set to **Enable**.
9. Select the required Application.
10. Click **Save**.
    The policy is created.

Encryption for Right to Privacy (RTP) on Privacera Platform

These are details for creating a data zone in Privacera Discovery **Compliance Workflow** with the Right To Privacy (RTP) policy, which encrypts the resources specified in the data zone.

For background, see the following:

- Data zones overview [93]
- Right to Privacy policy [103]

Create a New Data Zone Called "RTP"

1. Expand the **Discovery** > **Compliance Workflow** menu.
2. Click **Data Zones**.

3. On the **Data Zones** page, click the **+** icon.
4. For **Data Zone Name**, enter RTP.
5. Enter a **Description**.
6. Click **Save**.

The Data Zone is added.

Add RTP Policy to Data Zone

1. In the **Data Zones* page, select the created datazone.
2. Click the **Policies** tab.
3. Click **+Add Policy**.
4. Enter a required **Name**.
5. Enter a **Description**.
6. From the **Type** of policy menu, select **Right to Privacy**.
7. Select the **Alert Level**.
8. Enable or disable the **Status** of the policy.
9. Select the required **Application**.
10. Select Lookups: **Application** and **Lookup File Location**.
11. If you want literal masking, select the **Use LITERAL** checkbox.
12. Click **Save**.

The RTP policy is created.

Add Resource to be Encrypted

1. On the **Data Zones** page, select the datazone.
2. Click **+Add Resource**.
3. Enter the required **Application** name.
4. Enter the required **Resource** name.
5. Click **Save**.
   The file format resource is added. You can also add table format resources, such as DB name and table name.
6. Click the **Re-evaluate** checkbox for the resource to be encrypted.
7. Click the soft refresh button to update completion status.
   After the checkbox is cleared, the encryption has completed.

Verifying the Encryption

To verify the encryption:

1. Navigate to **Data Inventory** > **File Explorer**.
2. Select the resource to which you applied the RTP policy.
3. Verify the the data has been encrypted.

Original data can be viewed under the archive folder.

Workflow policy use case example

Workflow policy without encryption

Add the workflow policy without encryption

Follow the steps above to add a workflow policy. In the policy, clear the **Encrypt Data** checkbox, if selected.

Add a resource

1. Select a datazone that you want to apply the workflow policy to.
2. Select the **Resources** tab.
3. Click **Add Resource.**.

   > **NOTE**
   > You can add a folder or file as a resource. Resource files must be in CSV, Parquet, orc, JSON, or avro format.

4. Click **Save**.

When you run the scan on the datazone, the policy will now be applied and the data in the file will not be encrypted.

Workflow policy with encryption

Add the workflow policy with encryption

Follow the steps above to add a workflow policy. In the policy, select the **Encrypt Data** checkbox, and select an **Encryption Scheme** to the tag you want to encrypt.

Add a resource

1. Select a datazone that you want to apply the workflow policy to.
2. Select the **Resources** tab.
3. Click **Add Resource** button. You can add a folder or file as a resource.

   > **NOTE**
   > Resource files must be in CSV, Parquet, orc, JSON, or avro fomat.

4. Click **Save**.

Now, when you run the scan on datazone, the policy will be applied and the data in the file will be encrypted, for those tags that were marked to be encrypted.

Workflow Expunge policy

Enable Workflow Expunge policy

By default, the Workflow Expunge policy is not visible in the dropdown list of policies. To configure the Workflow Expunge policy, do the following in Discovery of Privacera Manager and Privacera Portal:

Privacera Manager

1. Run the following commands:

   ```
   cd ~/privacera/privacera-manager
   cp config/sample-vars/vars.aws.discovery.yml config/custom-vars/
   vi config/custom-vars/vars.aws.discovery.yml
   ```
2. Add the following property:

   ```
   DISCOVERY_WORKFLOW_EXPUNGE_POLICY_ENABLED=true
   ```
3. Run the update:

```
cd ~/privacera/privacera-manager
./privacera-manager.sh update
```

Privacera Portal

Go to **System configuration** in the portal and add the following custom properties:

```
privacera.portal.datazone.policy.workflow.expunge.enable=true
```

Add the workflow policy

Follow the steps above to add a workflow policy. In the policy, select the **Encrypt Data** checkbox, and select an **Encryption Scheme** to the tag you want to encrypt.

Add a resource

1. Select a datazone that you want to apply the workflow policy to.
2. Select the **Resources** tab.
3. Click **Add Resource** button. You can add a folder or file as a resource.

   NOTE

   Resource files must be in JSON format.

4. Click **Save**.

When you run the scan on the datazone, the policy will now be applied and the data in the file will be encrypted for those tags that were marked to be encrypted.

Define Discovery policies on Privacera Platform

Privacera Discovery comes with policies that can control various aspects of scanning datasources, such as the following:

- Disallow specific users or groups from accessing a datasource.
- "De-identify" (that is, encrypt or mask) data that has been tagged.
- Detect when a data source has moved from one subnet to another or from a defined data zone to a different zone.

Disallowed Groups policy

The Disallowed Groups Policy policy raises an alert if a user belonging to a restricted user group moves data into a specified data zone. You can add multiple user groups by clicking enter after each value. For example: safari, HDFS, superusers, and admin.

The Disallowed Groups Policy has the following fields:

- **Name**: The name of the Disallowed Groups Policy.
- **Type**: The type of policy.
- **Alert Level**: The alert level: high, medium, or low.
- **Description**: The description of the Disallowed Groups Policy.
- **Disallowed Users**: Allows you to add multiple user groups to be disallowed to move the data into a specific data zone.

Disallowed Movement Policy

This policy helps to monitor and raise alert if a user moves data to a restricted zone from any selected zones. You can add multiple source data zones by pressing enter after each value.

The Disallowed Movement Policy has the following fields:

- **Name**: The name of the Disallowed Movement Policy.
- **Type**: The type of policy.
- **Alert Level** :The alert level: high, medium, or low.
- **Description**: The description of the Disallowed Movement Policy.
- **Source**: Allows you to add multiple data zones to be disallowed.

Compliance Workflow policies on Privacera Platform

Privacera has the following types of Compliance Workflow policies:

- Disallowed Movement Policy [97]
- Disallowed Tags policy [100]
- Disallowed Subnets Policy [99]
- Disallowed Users Policy [102]
- Disallowed Groups policy [97]
- Disallowed Subnet Range Policy [99]
- Workflow policy [105]
- De-identification policy [98]
- Right to Privacy policy [103]
- Expunge policy [101]
- Workflow Expunge Policy [104]

NOTE

If you want to use encryption for Compliance Workflow policies (i.e., De-Identification, Right to Privacy, and Workflow Encryption), you have to add the `privacera_service_discovery` user. See Add Discovery user for encryption service.

NOTE

The following Compliance Workflow policies are not supported on the GCP platform:

- Workflow Policy
- De-identification Policy
- Right to Privacy Policy
- Expunge Policy
- Workflow Expunge Policy

Supported file formats by workflow policy type

The following table shows the supported file formats for each policy type.

Policies	csv	avro	parquet	json	orc
Workflow with Encryption	Yes	Yes	Yes	Yes	Yes
Workflow without Encryption	Yes	Yes	Yes	Yes	Yes
Workflow Expunge	-	-	-	Yes	-
De-identification	Yes	Yes	Yes	Yes	Yes
RTP	Yes	Yes	Yes	Yes	-
Expunge	Yes	Yes	Yes	Yes	-

De-identification policy

The De-identification policy encrypts sensitive data from resources based on specified tags.

Supported data sources

The following data sources are supported in the cloud for the De-identification policy:

-
-
-
- AuroraDB Postgres
- AuroraDB
-

Supported file formats

For a list of supported file formats that the De-identification policy can be applied to, see Compliance Workflow policies on Privacera Platform [97].

De-identification policy fields

The De-identification policy has the following fields:

- **Name**: The name of the De-identification policy .

- **Type**: The type of policy.
- **Alert Level (Optional)** : The alert level: high, medium, or low.
- **Description (Optional)**: A description of the De-identification policy.
- **Status**: A toggle used to enable or disable the policy. It is enabled by default.
- **Application**: The data source from which the scanned resources can be accessed and where the De-identification policy will be applied.
- **Destination Location**: The location where the encrypted sensitive data will be transferred.

  > **NOTE**
  > Some applications such as Snowflake and Presto SQL follow the `[Db].[Schema].[Table]` hierarchy. You need to provide the destination location in the correct format `[Db].[Schema]` for these applications.

- **Archive Location**: This field specifies the location where a copy of the input file is stored before any tagged records are encrypted.

  > **NOTE**
  > Some applications such as Snowflake and Presto SQL follow the `[Db].[Schema].[Table]` hierarchy. You need to provide the archive location in the correct format `[Db].[Schema]` for these applications.

- **Search for tags**: The tags used to identify or classify the data to be encrypted.
- **Apply Encryption Schemes**: A list of scheme names that have been added to the **Schemes** page. To view the schemes, select **Encryption & Masking** > **Schemes** from the navigation menu.

Add a resource to a data zone

To add a resource to a data zone, see Add resources to a data zone on Privacera Platform [90].

When you run a scan on a data zone, the policy will be applied and the data will be encrypted and moved to the destination location. The source file will be moved to the archive location.

If the destination location is not provided, the data will be encrypted in the resource file itself.

Disallowed Subnets Policy

This policy helps to monitor and raise alerts if users moving the data into a specific data zone belong to restricted IP addresses. You can add multiple IP addresses by clicking enter after each value.

The Disallowed Subnets Policy has the following fields:

- **Name**: The name of Disallowed Subnets Policy.
- **Type**: The type of policy.
- **Alert Level** : The alert level: high, medium, or low.
- **Description**: The description for Disallowed Subnets Policy.
- **Disallowed Subnets**: Allows you to add multiple IP Addresses to be disallowed.

Disallowed Subnet Range Policy

The Disallowed Subnet Range Policy monitors and raises an alert if data is moved into a data zone that belongs to a restricted IP address range. The UI is similar to the `disallowed_subnets policy`, with the addition of a pair of IP addresses. Add a pair of IP addresses to specify the range by clicking enter after each single IP address.

The Disallowed Subnet Range Policy has the following fields:

- **Name**: The name of the Disallowed Subnet Range Policy.
- **Type**: The type of policy.
- **Alert Level** : The alert level: high, medium, or low.
- **Description**: The description of the Disallowed Subnet Range Policy.
- **Disallowed Subnet Range**: Allows you to add IP address ranges to be restricted. Restricted IP addresses are unable to move data into the specified data zones.

Disallowed Tags policy

This policy helps to monitor and raises an alert if any PII tags are identified. You can add multiple tags by clicking enter after each value.

The Disallowed Tags policy has the following fields:

- **Name**: The name of the Disallowed Movement policy.
- **Type**: The type of policy.
- **Alert Level** : The alert level: high, medium, or low.
- **Description**: The description of the Disallowed Movement policy.
- **Disallowed Tags**: Allows you to add multiple tags to be disallowed.

Add Disallowed Tags policy

If you are creating Disallowed Movement and Disallowed Tags policies, then you can capture data zone movement using Spark. Data Zone movement can be captured in HDFS to S3.

To capture Data Zone movement using Spark, follow these steps:

> **NOTE**
> These data zones are examples. You should create your own.

1. Create directories in HDFS and add the file in one of the HDFS locations:

   ```
   hdfs dfs -mkdir /colour/purple
   hdfs dfs -mkdir /colour/pink
   hdfs dfs -put /finance_us.csv /colour/purple/
   ```
2. Add both the created directories in Include resource of HDFS.
3. Create two Data Zones and add the two folders in those two Data Zones' Resources.
   - **SourceDz**: It should have resource e.g. /colour/purple/ and also the Data Zone tag.
   - **DestinationDz**: It should have resource e.g. /colour/pink/ and also the policies configured for disallowed movement and disallowed tags.
4. Set the **Application property** as follows:

   ```
   Generate Alert All Part Files = false
   ```

   > **NOTE**
   > If you set Generate Alert All Part Files to false, the system generates an alert *for the first two* part files. If you set this property to true, the system generates an alert for all part files.
5. Go to the terminal and log into Spark shell as follows:

```
spark-shell --packages com.databricks:spark-csv_2.10:1.5.0  scala> val df = sqlContex
```

The following output is displayed:

- **Kafka Topics**: Check the Kafka topics audit consumption for **Alerts** and **Lineage**.
- **Alerts Details**: Check the **Alerts Details** tab on the resource details for this resource.
- **Lineage**: Check the **Lineage** for this resource.
- **Alerts Generated for part file** : Check the Data Zone Graph for the alerts generation for the part files in DestinationDz.

Expunge policy

The Expunge policy removes sensitive information such as usernames and email addresses from your data. This information is moved into a quarantine folder.

The fields in the lookup file are compared to the records in the resource files. If the tag is found (the value in the lookup file matches the value in the resource file for the specified tag (Search for tags)), then the field value in the resource file will be deleted. Ensure that the header of the lookup file matches the header of the tag to be searched.

> **NOTE**
> The resource file should be scanned before applying the Expunge policy. The Expunge policy does not work on real-time or offline scans.

Expunge policy supported data sources

Thr Expunge policy supports the following data sources. Click the tab to display the data sources that are supported in the cloud.

-
    -
    -
    -
    - AuroraDB Postgres
    - AuroraDB
    -
- Microsoft
    - MSSQL Server Synapse
-
    -

Expunge policy supported file formats

For a list of supported file formats that the Expunge policy can be applied to, see Supported file formats by workflow policy type [98]

Expunge policy fields

The following fields are included in the Expunge policy:

- **Name**: The name of the Expunge policy.
- **Type**: The type of policy.
- **Alert Level**: The level of alert: high, medium or low.

- **Description**: The description of the Expunge policy.
- **Status**: A toggle to enable or disable the policy. It is enabled by default.
- **Application**: The data source from which the scanned resources can be accessed and where the Expunge policy will be applied.
- **Lookup Application**: The name of the data source containing lookup file. The lookup file should be in `.csv` format, with tag names in the header columns.
- **Lookup File Location**: The location of the lookup file.
- **Quarantine Location**: The location of the data removed from the input file.

> **NOTE**
> Some applications such as Snowflake and Presto SQL follow the `[Db].[Schema].[Table]` hierarchy. You need to provide the Quarantine location in the correct format `[Db].[Schema]` for these applications.

- **Archive Location (Optional)**: The location of a copy of the original file.

> **NOTE**
> Some applications such as Snowflake and Presto SQL follow the `[Db].[Schema].[Table]` hierarchy. You need to provide the Archive location in the correct format `[Db].[Schema]` for these applications.

- **Search for tags**: Tags that identify and classify the data to be removed.
- **Auto Run**: If this feature is enabled, the Expunge policy is applied after a specified time interval.

Example 5. Expunge policy example

- **Lookup File Location**: Add a `.csv` file to the **Lookup File Location** field, and it should specify which sensitive data needs to be removed from resources based on tags. For example: File name is *input.csv* file with EMAIL tag (sample@gmail.com).
- When the file is being scanned, if “sample@gmail.com” tagged with EMAIL is matched, then this row will be removed.

Consider the following example:

1. A file, **test_file.csv**, is added to a data zone. **Search for** as **EMAIL** tag is added.
2. The scheduler is triggered and the system applies the **Expunge** policy to the resource (**test_file.csv**).
3. After applying the **Expunge** policy, a row in **test_file.csv** that contains sensitive information is removed from the file and moved to the specified quarantine location.

Disallowed Users Policy

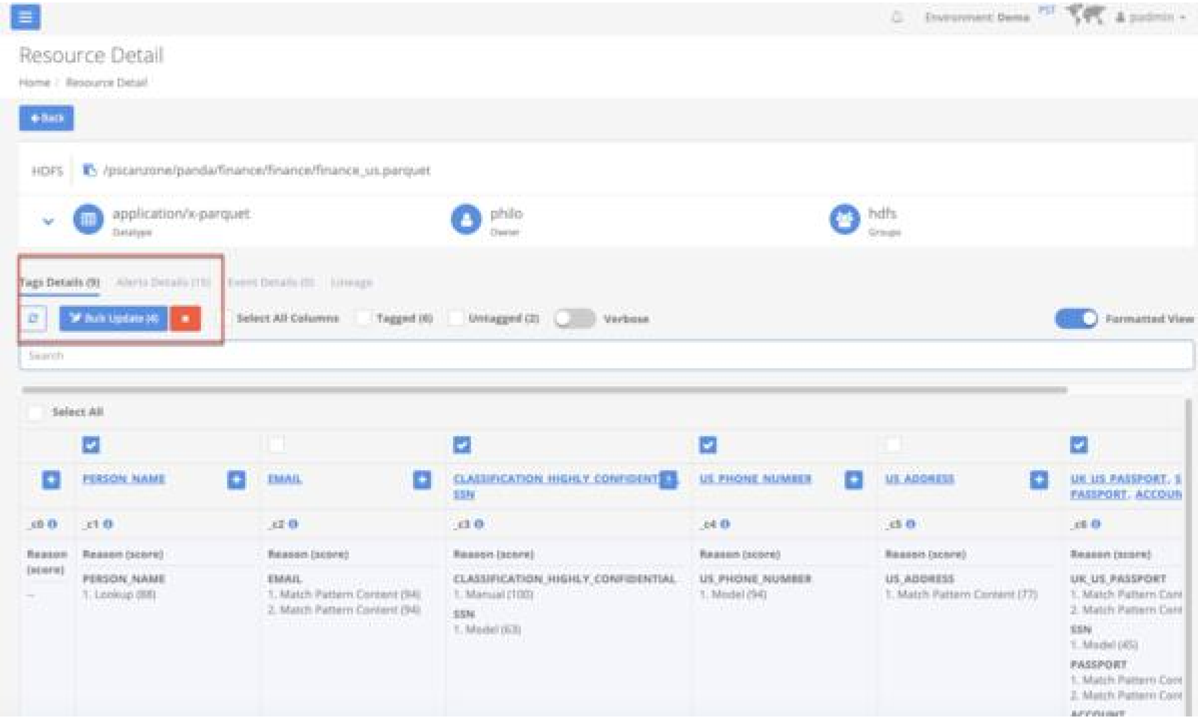
This policy helps to monitor and raise alert if restricted users move the data into a specific data zone. You can add multiple users by clicking enter after each value for e.g. sally, mark, jason as shown in the image.

Add Disallowed Users Policy

The Disallowed Users Policy has the following fields:

- **Name**: This field indicates name of Disallowed Users Policy.
- **Type**: This field indicates type of policy.
- **Alert Level** : This field indicates alert level: High, Medium, or Low.
- **Description**: This field indicates description for Disallowed Users Policy.

- **Disallowed Users**: This field allows you to add multiple users to be disallowed to move the data into specific data zone..



	☑	□	☑	☑	□	☑
	PERSON_NAME	EMAIL	CLASSIFICATION_HIGHLY_CONFIDENTIAL, SSN	US_PHONE_NUMBER	US_ADDRESS	UK_US_PASSPORT, SSN, PASSPORT, ACCOUN
_c0	_c1	_c2	_c3	_c4	_c5	_c6
Reason (score)	Reason (score)	Reason (score)	Reason (score)	Reason (score)	Reason (score)	Reason (score)
--	PERSON_NAME 1. Lookup (88)	EMAIL 1. Match Pattern Content (94) 2. Match Pattern Content (94)	CLASSIFICATION_HIGHLY_CONFIDENTIAL 1. Manual (100) SSN 1. Model (63)	US_PHONE_NUMBER 1. Model (94)	US_ADDRESS 1. Match Pattern Content (77)	UK_US_PASSPORT 1. Match Pattern Con 2. Match Pattern Con SSN 1. Model (45) PASSPORT 1. Match Pattern Con 2. Match Pattern Con ACCOUNT

Right to Privacy policy

With lookup data and static masking algorithms, sensitive information such as email addresses, phone numbers, and street addresses are encrypted in the source folder and subject to the Right to Privacy (RTP).

Lookup files must be in `.csv` format. The fields in the lookup file are compared to the records in the resource files. If the tag is found (the value in the lookup file matches the value in the resource file for the specified tag (Search for tags)), then the field value in the resource file will be encrypted. Ensure that the header of the lookup file matches the header of the tag to be searched.

> **NOTE**
>
> The resource file should be scanned before applying the RTP policy. The RTP policy does not work on real-time or offline scans.

Right to Privacy policy supported data sources

The following data sources are supported by the RTP policy. Click the tab to display the data sources that are supported in the cloud.

-
  -
  -
  -
  - AuroraDB Postgres
  - AuroraDB
  -

- Microsoft
  - ADLS
  - MSSQL Server Synapse
-
  -

Right to Privacy policy supported file formats

For a list of supported file formats that the Right to Privacy policy can be applied to, see Supported file formats by workflow policy type [98]

Right to Privacy policy fields

The following fields are included in the RTP policy:

- **Name**: The name of the RTP policy.
- **Type**: The type of policy.
- **Alert Level**: The level of alert: high, medium, or low.
- **Description**: A description of the RTP policy.
- **Status**: A toggle to enable or disable the RTP policy. It is enabled by default.
- **Application**: The data source from which the scanned resources can be accessed and where the RTP policy will be applied.
- **Lookup Application**: The name of the data source containing the lookup file. The lookup file must be in `.csv` format, with tag names in the header columns.
- **Lookup File Location**: The location of the lookup file.
- **Archive Location (Optional)**: This field specifies the location where a copy of the input file is stored before any tagged records are encrypted.

> **NOTE**
>
> Some applications such as Snowflake and Presto SQL follow the `[Db].[Schema].[Table]` hierarchy. You need to provide the archive location in the correct format `[Db].[Schema]` for these applications.

- **Search for tags**: Tags used to identify or classify data to be encrypted.
- **Apply Encryption Schemes**: A list of scheme names that have been added to the **Schemes** page. To view the schemes, select **Encryption & Masking** > **Schemes** from the navigation menu.
- **Use LITERAL**: If this feature is enabled, the sensitive values in the resource file are replaced with literals for scheme. For more information about LITERAL, see about LITERAL.
- **Auto Run**: If this feature is enabled, the RTP policy is applied after a specified time interval.

Example 6. Right to Privacy policy example

- Add a `.csv` file to the **Lookup File Location** field, and it should specify which sensitive data needs to be removed from resources based on tags. For example: File name is *input.csv* with EMAIL tag (sample@gmail.com), PERSON_NAME tag (Alex).
- Now, when the resource file is being scanned, if *sample@gmail.com* tagged with EMAIL and *Alex* tagged with PERSON_NAME are matched, then this row will be considered for RTP.

Workflow Expunge Policy

The Workflow Expunge policy removes sensitive data from resources based on specified tags. This policy accepts only newline-delimited JSON records format. For nested files, the Workflow Expunge policy is not supported.

Workflow Expunge policy supported data sources

The Workflow Expunge policy can be applied to the following data sources:

- AWS S3
- ADLS

Workflow Expunge policy supported file formats

For a list of supported file formats that the Workflow Expunge policy can be applied to, see Supported file formats by workflow policy type [98]

Workflow Expunge policy fields

The Workflow Expunge policy has the following fields:

- **Name**: The name of the Workflow Expunge policy.
- **Type**: The type of policy.

> **NOTE**
> The Workflow Expunge policy is not visible in the dropdown of policies by default.

- **Alert Level**: The level of alert: high, medium or low.
- **Description**: A description of the Workflow Expunge policy.
- **Status**: A toggle to enable or disable the Workflow Expunge policy. It is enabled by default.
- **Application**: The data source from which the scanned resources can be accessed and where the Workflow Expunge policy will be applied.
- **Transfer Location**: The location that the input file is transferred to if no tagged records match the tags specified in the policy.
- **Quarantine Location**: The location to which the input file is moved after the sensitive data is removed.
- **Archive Location (Optional)**: The location of a copy of the original file.
- **Search for tags**: Tags that help in identifying or classifying the data to be tagged and then expunged.

Add a resource to a data zone

To add a resource in the data zone, see Add resources to a data zone on Privacera Platform [90].

If the policy conditions are met (matching sensitive tags, file size exceeds the maximum limit, or excluded data type) when you run a scan on a data zone, then sensitive data is deleted from the file and moved to a quarantine location. Non-sensitive data will be moved to a transfer location.

Workflow policy

This policy includes conditions such as sensitive tags, maximum file size (for example, 1 MB), and excluded data types (for example, images). If any of the alert conditions are met, the file is moved to a quarantine location. If encryption is enabled and a sensitive tag is found, then the column with the sensitive tag is encrypted.

> **NOTE**
> For nested files, encryption is only supported for primitive data types, not complex data types.

Workflow policy supported data sources

The Workflow without encryption policy supports the following data sources:

- AWS S3
- ADLS
- GCP GCS

The Workflow with encryption policy supports the following data sources:

- AWS S3
- ADLS

Supported file formats

For a list of supported file formats that the Workflow policy can be applied to, see Supported file formats by workflow policy type [98].

Workflow policy fields

The following fields are included in the Workflow policy:

- **Name**: The name of Workflow policy.
- **Type**: The Workflow policy type.
- **Alert Level (Optional)**: The level of alert: high, medium, or low.
- **Description (Optional)**: A description of the Workflow policy.
- **Status**: A toggle to enable or disable the policy. It is enabled by default.
- **Application**: The data source from which the scanned resources can be accessed and where the Workflow policy will be applied.
- **Transfer Location (Optional)**: The location to which the input file is transferred if any of the alert conditions are not met.
- **Quarantine Location**: The location where the input file is moved if any of the alert conditions are met.
- **Archive Location (Optional)**: The location where a copy of the original file is moved before any tagged records are removed from it.
- **Search for tags**: The tags that help in identifying and classifying records that will be tagged and then expunged.
- **Apply Encryption Schemes**: This field appears when you select the **Encrypt Data** checkbox. This field is populated with the names of the schemes that have been added to the application's Scheme section. To view the schemes, click and expand the **Encryption & Masking** from left menu, and then select the **Schemes**.
- **Max File Size (MB)**: This field excludes files based on file size and raises an alert if the condition is met.
- **Exclude File Types**: This field excludes the files based on file type and raises an alert if the condition is met.

The workflow policy provides two options:

- Workflow policy without encryption
- Workflow policy with encryption

Workflow policy without encryption

The status of the workflow policy is enabled by default. If you do not want to encrypt your data, clear the **Encrypt Data** checkbox.

Add a resource to a data zone

To add a resource to a data zone, see Add resources to a data zone on Privacera Platform [90]

When you run a scan on a data zone, and if any of the alert conditions are met (matching sensitive tags, file size exceeds the maximum limit, or excluded data type), the file is moved to a quarantine location.

If none of the conditions are met and you have specified a transfer location, the file will be moved to the transfer location.

Workflow policy with encryption

If you want to encrypt data, select the **Encrypt Data** checkbox.

Add a resource to a data zone

To add a resource to a data zone, see Add resources to a data zone on Privacera Platform [90].

When you run a scan on a data zone, and if any of the alert conditions are met (matching sensitive tags, file size exceeding the maximum limit, or excluded data type), the column with the sensitive tag is encrypted and the file is moved to a quarantine location.

If none of the alert conditions are met and you have specified a transfer location, the file will be moved there.

If you have specified an archive location, the file will be moved to the archive location before being encrypted.

View scanned resources

You can examine the results of Discovery scans in several different ways. See also Discovery reports and dashboards [113].

Data Explorer on Privacera Platform

To access the **Data Explorer** screen, select **Data Inventory** > **Data Explorer** from the navigation menu.

Add or edit tags

On the **Data Explorer** screen, you can add and remove tags to and from resources.

To update a resource in Data Explorer, follow these steps:

1. In the **Data Explorer** screen, click **ADD / EDIT**.
2. Modify the **Tags**.
3. Click **Save**.
   The resource is updated.

The **Update Resource** dialog is displayed.

Filter the Data Explorer page

There are two ways to filter data on the **Data Explorer** screen:

- **Search by Resource**: View results by resource name.
- **Search by Tags**: View results by tags.

To view the resource details, click the folder link and the file name.

Add or remove or scan a resource from Data Explorer

To add, remove or scan a resource:

1. Enable the following property under **Settings** > **System Configurations** > **Portal Properties** > **Discovery Scan** section:
   `Enable Scan Resource on Data Explorer`
2. In the **Data Explorer** screen, select the resource and click the ellipsis under **Actions** column.
   The following options are displayed:

   Table 1.

Options	Description
Add / Edit Tags	You can add and remove tags to and from resources
Add to Include Resource	You can add a resource to Include Resource list
Remove from Include Resource	You can remove a resource from Include Resource list
Scan Resource	You can scan a resource

3. Select the required option to add or remove or scan a resource.

File Explorer overview

Using the File Explorer, you can browse object resources (buckets, files and folders) in AWS, ADLS and GCS. You can upload and retrieve the data to and from the bucket. With resource policies you associate with the bucket, you can control access by granting or revoking permissions for users, groups, or roles.

Prerequisites

- Be sure you have added the Dataserver data resource

Configure resource policy

After the Dataserver is added, configure the default policy to add an **Allow Condition** as follows:

- **AWS**: see Configure AWS S3 resource policies. For wide-open access, use * for bucket name and object path, with READ permission for the public group.
- **ADLS**: see Configure ADLS resource policies. For wide-open access, use * for storage account, container, name, and object path, with READ permission for the public group.

For general steps, see Create resource policies: general steps.

View files and folders using the File Explorer

To view the list of files and folders of an S3 bucket, do the following:

1. Navigate to **Data Inventory** > **File Explorer**.
   Since the policy is enabled, all the S3 buckets are displayed on the **File Explorer** page.
2. On the **File Explorer** page, you can do the following actions:

Action	Description
Refresh	Refreshes the S3 buckets
Search	Allows to search for a particular bucket
Filter	Hides or shows columns
Create Folder	Allows to create a folder
Upload	Allows to upload a file. You can upload all those files formats that are supported by AWS S3.
Delete	Allows to delete file(s) or folder(s)
Calculate	Calculates folder size
Copy to Clipboard	Copies the object path

Use case: AWS S3

By managing the permissions in the policy, you can provide access control on the S3 bucket. In this use case, you will see how you can allow/restrict a user from uploading files to a S3 bucket, and view the activity in an audit log.

1. Create a policy with *Read* and *Write* permissions.
2. Upload a file by performing the following steps:
   a. Navigate to **Data Inventory** > **File Explorer**.
   b. Go to the S3 bucket where you want to upload the file.
   c. Click **Upload**.
      Add File popup is displayed.
   d. Choose the file to upload and click **Upload** button.
      The file is uploaded with a success message and seen in the listing.
   e. Navigate to **Access Manager** > **Audits** to view the audit log for the upload action.
3. Edit the policy and remove the *Write* permission.
4. Upload a file by performing the following steps:
   a. Navigate to **Data Inventory** > **File Explorer**.
   b. Go to the S3 bucket where you want to upload the file.
   c. Click **Upload**, **Add File** popup is displayed.
   d. Choose the file to upload and click **Upload** button.
      "Access Denied" error message is displayed.
   e. Navigate to **Access Manager** > **Audits** to view the audit log for the denied upload action.

Classifications overview

The **Classifications** page displays all of the resources that has tagged.

Resources are displayed in the **Classifications** page if their score falls within the range defined in Scan setup [75]. The default score range is 40 to 70. If a resource falls within this range, it is also tagged.

If the score is below the minimum value (default 40), the tag is ignored and not applied to the resource.

Bulk update tags from the Classifications page

You can add new tags and do a bulk update for the resource directly from the **Classification** page.

1. In the Classification page, click the **Resource** that you want to add a tag to (for example, an HDFS resource).
   The **Resource Detail** page is displayed.
2. Click **+** next to the keyword that you want to add additional tags to.
   The **Update Resource** dialog displays.
3. Add the tags to the Resource using the **Tags** dropdown and click **Save**.
   The tags are added to the resource.
4. If you have added multiple tags, you can do a **Bulk Update** of the tags that you have added to the various columns.
5. Click **Bulk Update**.
6. Answer the prompts to Accept/Reject/Allow the tags and add a reason if required.
7. Click **Save** or **Close**.

Manually accept, reject, or allow a tag from the Classications page

To manually accept, reject, or allow a system-classified tag:

1. In the **Classification** page, select a **Tag.**
   The **Data Info** dialog displays.
2. In the **Status** column, select one of the following options:
   - **Accepted**: The resource and tag are under the **Classification** tab.
   - **Rejected**: The resource and tag are under **Show Rejected**.
   - **Allowed**: The resource and tag are under **Show Allowed**.

   ```
   privacera.portal.manual.whitelisted.tagging.enable=true(default=false)
   privacera.portal.manual.whitelisted.tagging.text=Allowed
   ```
3. Select **Accepted**, **Rejected**, or **Allowed** to set the status of the tag.
4. Add a reason and click **Save**.
   The resource is updated.

Similarly, you can set the status of the tags that are system-classified with pending review.

Filter the Classifications page

You can filter the results on the **Classifications** page using the following filters:

- **Partial or Exact Match:** Select the type of match you want to filter the resource data.
- **Search by Input resource**: This search filter allows you to view the result of resource.
- **Date Filter**: You can sort and filter the resource data by **Resource Create Time**, **Classification time**, or **Turn Off** time. You can specify a date range, which is 7 days by default.
- **Search by Application**: View by application name.
- **Search by Datazone**: View by datazone.
- **Search by Tags**: View by tags.
- **Search by Input ScanID**: View by scan ID.
- **Query Filters**: Apply conditional filters and view the result. For details on how to use the Query Filters, see Reports with the Query Builder [129].

- **More Filter**: This lists additional available filters. As you edit the search filters, the results are dynamically updated to match.
  - **Group Tables**: View the higher level classification for tables.
  - **Group Folders**: View the higher level classification for folders.
  - **Search by Location**: View the result by location (database/table name) of the resource. Additionally, you can search by HDFS location for Hive tables.
  - **System Classified**: View the results where tag is marked as system-classified.
  - **Show Allowed**: View the results where tag is marked as show allowed.
  - **Deleted:** View the results that have been deleted.
  - **Reviewed**: View the results that has been reviewed.

Export classifications by scan ID

To export the classifications from a particular scan ID, click **Export** for that scan ID and follow the leading prompts.

View classifications

To view the **Classifications** page, select **Data Inventory** > **Classifications** from the navigation menu.

The **Classifications** page displays classifications in a table. Structured data shows all meta-tags, including table-level and file-level, applied on the structured data columns along with their content and meta-tags. This helps to view all the parent tags that were distributed to the children classifications.

The **Classifications** page displays the following information:

- **Datazone**: The name of the datazone.
- **Application**: The name of the application.
- **Resource**: The name of the resource. Clicking on the resource displays the **Resource Detail** page. This page is divided into three tabs: **Tag Details**, **Alerts Details**, and **Lineage**, along with their count of records in each tab.
- **Updated On**: The date and time of the last update.
- **Tag**: A list of tags associated with the particular resource. Clicking on the tag name displays the **Data Info** dialog, which displays the following information:
  - **Field Name**: The full file location of the field.
  - **Sample Values**
  - **Reason**
  - **Score**
  - **Status**: **Accepted**, **Rejected**, and **Allowed**.
  - **Status Change Reason**

View rejected tags from the Classications page

To view rejected tags:

1. From the navigation menu, select **Data Inventory** > **Review**.
2. On the **Review** page, select **MORE FILTER**.
3. Select **Rejected**.
   The review page displays tags marked as **Rejected**.

Edit resources from the Classifications page

To edit a resource:

1. In the Classification page, click **Add / Edit**.
   The **Update Resource** dialog is displayed.
2. Add or modify the **Tags**.
   The resource is updated.

NOTE
You can add tags in bulk when you add or edit tags for a resource.

Discovery reports and dashboards

There are three categories of reports in the **Reports** menu.

- Built-in reports [115]: Reports provided by Privacera for a wide variety of information, such as data inventory, scan summaries, tag trends and summary, and more. Also available with Built-in reports [115] is the **Query Builder** [129] with which you can create your own custom reports.
- Saved Reports [128]: The results of running a report saved for later download.
- Offline reports [128]: Reports that result in a large number of rows and that would require much time to export are moved to Offline reports [128].

There are two dashboards for Discovery:

- Discovery Dashboard [114]
- Alerts Dashboard [113]

Alerts Dashboard

The **Alerts Dashboard** provides a brief overview of anomalies in data zone scans. If a data zone has a policy for a disallowed tag or disallowed movement (when a file is incorrectly copied from one data zone to another), then an alert is generated.

For more detail about disallowed tags and disallowed data zone movement, see Data zone movement [92].

View the Alerts Dashboard

To view the Alerts Dashboard, select **Compliance Workflow** > **Alerts Dashboard** from the navigation menu.

The **Alerts Dashboard** displays the following information:

- **Alert Time**: The time that the alert was triggered.
- **Alert Level**: The level of alert: high, medium, or low.
- **User**: The name of the user.
- **Policy**: The name of the policy.
- **Alert For**: The details of the alert.
- **Reason**: The reason for the alert.
- **Export**: See Export alert details [113].

Alerts Dashboard search filters

You can filter the alerts displayed on the **Alerts Dashboard** using the following methods:

- **Search by Category**: Allows you to view alerts by category.
- **Include Policies**: Allows you to view alerts that are marked under Include Policy.
- **Exclude Policies**: Allows you to view alerts that are marked under Exclude Policy.

Export alert details

You can export details about alerts from the **Alerts Dashboard** .

To export alert details, do the following:

1. On the **Alerts Dashboard**, click **Export**.
2. From the dropdown menu, select an export format: **CSV** or **JSON**.
   The **Export** dialog displays.

Discovery Dashboard

The Discovery Dashboard provides an at-a-glance view of key performance indicators (KPIs) relevant to a particular objective or business process.

To view the Discovery Dashboard, expand the **Discovery** menu and click **Dashboard**.

The Discovery Dashboard is divided into three major sections:

- Summary
- Graphs
- Trends

Summary

For each summary, hover over the info icon to view a description. The following information is displayed:

- **Files Scanned**: The number of files scanned by Discovery across all file systems.
- **Columns Scanned**: The number of columns scanned by Discovery including all tabular data.
- **Total Scanned**: The number of total resources scanned by Discovery.
- **Unique Tags Found**: The number of uniquely identified tags found by Discovery.
- **Tags Applied**: The number of tags applied automatically by Discovery.
- **Tags Pending Review**: The number of tags requiring review.

Graphs

The following information is displayed in the Graphs section of the Discovery Dashboard:

- **Data Map**: A ring chart displaying the percentages of data sources that have been scanned.
- **Datazones**: A bar chart displaying the number of datazones that have been scanned.

Trends

Trends is a categorization of tags:

- **Discovered**: The total number of occurrences discovered for the specified tag.
- **Applied Classification**: The total number of applied classifications for the specified tag.

Hover over the trend to see the exact number of discovered tags and applied classification with date and time.

To navigate to the Classification page, click **View Classification**.

Discovery Dashboard controls

Alerts and notifications appear in the uppermost bar. For a related dashboard, see Alerts Dashboard [113].

Click the **Refresh icon** to refresh the lists in the Discovery Dashboard.

The Discovery Dashboard has the following search filters:

- **Search by Application**: View scan results by application name.
- **Search by Datazone**: View scan results by datazone.
- **Search by Tags**: View scan results by tags.

You can also search by a specific date range. To use a specific date range filter, keep the **ON/OFF** toggle set to **ON**. You can search by the following date ranges:

- Today.

- Yesterday.
- Last 7 Days (default).
- Last 30 Days.
- This Month.
- Last Month.
- Custom Range.

Built-in reports

features the following **Built-in Reports**:

- Data Discovery Overview report [122]: The number of tags identified in a specific application.
- Data Inventory report [118]: Details of tags at the file or table/column level.
- Discovery Alerts report [123]: Alerts generated during scans.
- Scan Summary report [116]: Discovery scan summary.
- Tags Trends reports [124]: Tag trends identified during scans.
- Data Zone report [120]: Comprehensive details about data zones that have sensitive information or PII.
- Tag Review Summary report [126]: The number of tags that have been reviewed.
- Tag Review Inventory Summary report [125]: Details of tags that have been reviewed.
- Compliance Summary report [115]: A summary of top ten accesses to your data by type of access (service, user, IP address, or resource). Includes a heat map of total accesses by type and audit details.

Compliance Summary report

The Compliance Summary report is a summary of all the "components" (for example, users and IP addresses) that have access to your data.

The Compliance Summary report has the following sections:

- **Service**
- **Access By Tags**
- **Users**
- **IP Address**
- **Resources**
- **Heat Map**
- **Audits**

The **Heat Map** and **Audits** sections include:

- Data Access.
- Policy ID.
- Result.
- Event Time.
- Application.
- User.
- Service Name/Type.
- Service Type.
- Resource Name/Type.
- Access Type.
- Access.

View the Compliance Summary report

To view the Compliance Summary report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Compliance Summary** section, click **View or Edit Report**.
   The **Compliance Summary** page is displayed.

Apply filters to the Compliance Summary report

You can filter the data that appears on the Compliance Summary report using the following filters:

- **Search by Service**: View the result by service.
- **Search by Tags**: View the result by tag.
- **Search by Users**: View the result by user.
- **Search by IP Address**: View the result by IP Address.
- **Search by Resources**: View the result by resources.
- **Date**: View the results by date.
- **Exclude Service Users**: Exclude service users from the report.

To apply filters to the Compliance Summary report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Compliance Summary** section, click **View or Edit Report**.
3. Select the filters you want to apply.

Save the Compliance Summary report

To save the Compliance Summary report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Compliance Summary** section, click **View or Edit Report**.
3. Apply filters [116] (optional).
4. Click **Save**.
   The **Save Report** dialog displays.
5. Enter the following details:
   - Report Name (required)
   - Description
   - Select the Time
6. Click **Save**.
   The report is saved to **Saved Reports [128]**.

Download the Compliance Summary report

To download the Compliance Summary report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Compliance Summary** section, click **View or Edit Report**.
3. Apply filters [116] (optional).
4. Click **Download Report**.
   The report downloads in PDF format.

Scan Summary report

The Scan Summary report is a summary of Discovery scans.

The Scan Summary report includes the following information:

- **Time**: The date and time of the scan.
- **Application**: The name of the application.
- **Scan Id**: The scan identifier.
- **Resource**: The name of the resource with the full path.

- **Offline Scan Type**: Provides the offline scan type.
- **Scan Reason**: Provides scan reason for a resource.
- **Resource Type**: The status of the tag, such as tagged or untagged.

View the Scan Summary report

To view the Scan Summary Report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Scan Summary** section, click **View or Edit Report**.
   The **Scan Summary** report page is displayed.

Save the Scan Summary report

To save the Scan Summary report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Scan Summary** section, click **View or Edit Report**.
3. Apply filters [117] (optional).
4. Click **Save**.
   The **Save Report** dialog displays.
5. Enter the following details:
   - Report Name (mandatory).
   - Description.
   - Time.
6. Click **Save**.
   The report is saved to **Saved Reports [128]**.

Export the Scan Summary report to CSV

To export the Scan Summary report in CSV format, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Scan Summary** section, click **View or Edit Report**.
3. Click **Export to CSV**.
   The **Export** dialog displays.
4. You can select **All** or you can remove the default columns displayed under **Columns**.
5. Click **Export**.
   The **Scan Summary** report is downloaded in CSV format.

The exported CSV file includes the following information:

- **Time**: The creation date and time of classification.
- **Application**: The name of the application.
- **ScanID**: The scan identifier.
- **Resource**: The name of the resource.
- **Status**: The status of the tag, such as tagged or untagged.

Apply filters to the Scan Summary report

You can filter the data that appears in the Scan Summary report. The Scan Summary report includes the following filters:

- **Search by Resource**: View the results by the resource name.
- **Search by Application**: View the result by the application name.
- **Search by ScanID**: View the result by the scan id.
- **Scan ID**
- **Resource Status**

- **Date**: View the results by date.

When you select a **Failed Resource** and click **Show Logs**, the reason for the failure is shown in the report and can be exported

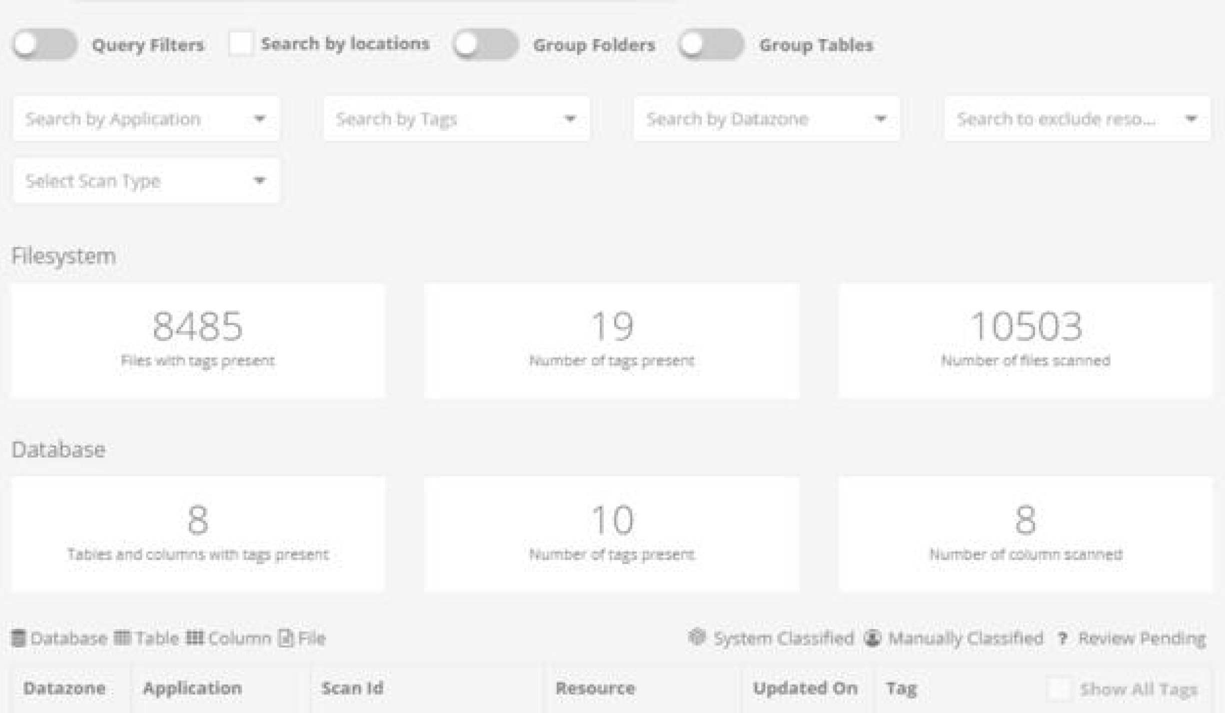
To apply filters to the Scan Summary report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Scan Summary** section, click **View or Edit Report**.
3. Select the filter you want to apply.

Data Inventory report

This report provides details of tags at the file- or the table/column-level.

- **Filesystem** shows a count of files with tags, number of tags, and number of files scanned.
- **Database** shows a count of tables and columns with tags, number of tags, and number of columns scanned.



Query Filters □ Search by locations Group Folders Group Tables

Search by Application ▾ | Search by Tags ▾ | Search by Datazone ▾ | Search to exclude reso... ▾

Select Scan Type ▾

Filesystem

8485 Files with tags present

19 Number of tags present

10503 Number of files scanned

Database

8 Tables and columns with tags present

10 Number of tags present

8 Number of column scanned

Database Table Column File — System Classified Manually Classified ? Review Pending

Datazone	Application	Scan Id	Resource	Updated On	Tag	□ Show All Tags

The following information is shown under the classifications:

- **Datazone**: Name of datazone.
- **Application**: Name of application.
- **Scan Id**: The scan identifier indicates the type of scan.
- **Resource**: Name of resource. Click the name of a resource to view the **Resource Detail** page. This page is divided into three tabs with counts of records in each tab: **Tag Details**, **Alerts Details**, and **Lineage**.
- **Updated On**: Date and time of most recent classification.
- **Tag**: List of tags associated with the particular resource. Click a tag name to view detailed information about the tag.
- The following information is shown under **Data Info**:
  - Field Name.
  - Sample Values.
  - Reason.

- Score.
- Status.
- Status Change Reason.

View/Edit Report

To view/edit a report, use the following steps:

1. On the Privacera home page, on the left, expand the **Reports** menu and click **Built-in Reports**.
2. Under **Data Inventory**, click **View or Edit Report**.
   The **Data Inventory** report page is displayed.
3. To edit/view the report, you can apply different search filters such as search by application, tag, and datazone. In addition, you can apply date ranges and other advanced filters.

Save Report

To save the report, use the following steps:

1. You can apply different search filters such as search by application, tag, and datazone. In addition, you can apply date ranges and other advanced filters.
2. Click **Save**.
3. Enter the following details:
   - Report Name (mandatory)
   - Description
   - Time
   - Exclude Resource
   - Exclude Description
4. Click **Save**.

The report is saved. You can view the saved report under **Saved Reports**.

Search Filters

The following filters are available. Also, whenever you apply the search filter, it is saved with the report.

- Basic search filter
- Query filter. See Reports with the Query Builder [129].

To apply various search filters, use the following steps:

1. On the Privacera home page, on the left, expand the **Reports** menu and click **Built-in Reports**.
2. Under **Data Inventory**, click **View or Edit Report**.
   The following search filters are available:
   - **Search by Application**: This search filter allows you to view the result by the application name.
   - **Search by Tags**: This search filter allows you to view the result by the tags.
   - **Search by Datazone**: This search filter allows you to view the result by the datazone.
3. On the **Reports** page, click **Search by Application**, **Search by Tags**, or **Search by Datazone**.
   You can select multiple items.
4. Based on selected search criteria, the report will be displayed. Similarly, you can search by **Tags** and **Datazone**.
5. Optionally, you can **Group Folders** and **Group Tables**.

Export to CSV

To export the report in CSV file, use the following steps:

1. In the **Built-in Reports** page, click **Export to CSV**.
2. The Export to CSV pop-up is displayed.

3. Select the required fields to include in the report:
   - **No of Resource or All**.
   - **Any of the following columns or All**:
     - **Application**: Name of application.
     - **Datazone**: Name of datazone.
     - **DB**: Name of database.
     - **Table**: Name of table.
     - **Column**: Name of column in the table.
     - **NonNullCount**: Non-null count.
     - **Score**: Score value of the tag.
     - **Tags**: Name of tag.
     - **TagReason**: The basis tag for applying the tag, such as lookup, model, match pattern content.
     - **TagStatus**: Status of tag.
     - **TagStatusReason**: Reason for change in tag status.
     - **Owner**: Name of resource owner.
     - **Location**: Path of resource.
     - **Encrypted**: True or false.
     - **CreatDate(mm/dd/yyyy)**: Creation date and time of resource.
     - **UpdateDate**: Update date and time of classification.
     - **Size**: Size of resource.
4. Check any of the following options:
   - Include Empty metaname tagged columns.
   - Include empty table.
   - Show only column tags on column.
   - Exclude external table location.
   - Exclude Reviewed Tags.
5. Click **Export**.

The report is downloaded.

Data Zone report

This report provides details about data zones that have sensitive data or PII.

- **Filesystem** shows a count of files with tags, number of tags, and number of files scanned.
- **Database** shows a count of tables and columns with tags, number of tags, and number of columns scanned.

The Data Zone Report displays the following information:

- Datazone name
- Datazone Create Date
- Each Tag Applied
- Latest Date Tag applied
- Latest Tag Rejected/Accepted
- Latest Tag Rejected/Accepted Date

View the Data Zone report

To view the Data Zone report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Zone Report** section, click **View or Edit Report**.
   The **Data Zone Report** page displays.

Save the Data Zone report

To save the Data Zone report, follow these steps:

1. On the **Data Zone Report** page, click **Save**.
2. Enter the following details:
   - Report Name (mandatory)
   - Description
   - Time
   - Exclude Resource
   - Exclude Description
3. Click **Save**.
   The report is saved to **Saved Reports**.

Apply filters to the Data Zone report

You apply the following filters to the Data Zone report:

- **Search by Datazone**: View results by datazone name.
- **Search by Tags**: View results by tag name.
- **Search by Tag attributes**: View results by tag attributes.
- **Search by Application**: View results by application name.

To apply filters to the Data Zone report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Zone Report** section, click **View or Edit Report**.
3. Select the filter you want to apply.

Export the Data Zone report to CSV

To export the Data Zone report in CSV format, do the following:

1. On the **Data Zone Report** page, click **Export to CSV**.
   The **Export** dialog displays.
2. Select the required fields to include in the report:
   - **No. of Resource or All**.
   - **Any or all of the following columns**:
     - **Application**: Name of application.
     - **Datazone**: Name of datazone.
     - **DB**: Name of database.
     - **Table**: Name of table.
     - **Column**: Name of column in the table.
     - **NonNullCount**: Non-null count.
     - **Score**: Score value of the tag.
     - **Tags**: Name of tag.
     - **TagReason**: The basis tag for applying the tag, such as lookup, model, match pattern content.
     - **TagStatus**: Status of tag.
     - **TagStatusReason**: Reason for change in tag status.
     - **Owner**: Name of resource owner.
     - **Location**: Path of resource.
     - **Encrypted**: True or false.
     - **CreatDate(mm/dd/yyyy)**: Creation date and time of resource.
     - **UpdateDate**: Update date and time of classification.
     - **Size**: Size of resource.
3. Check any of the following options:

  - Include Empty metaname tagged columns.
  - Include empty table.
  - Show only column tags on column.
  - Exclude external table location.
  - Exclude Reviewed Tags.
4. Click **Export**.
   The **Data Zone** report downloads in CSV format.

Data Discovery Overview report

The Data Discovery Overview report is a built-in report that provides counts of tags associated with specific applications.

- **Filesystem** shows a count of files with tags, number of tags, and number of files scanned.
- **Database** shows a count of tables and columns with tags, number of tags, and number of column scanned.

View the Data Discovery Overview report

To view the Data Discovery Overview report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Discovery Overview** section, click **View or Edit Report**.
   The **Data Discovery Overview** page is displayed.

Save the Data Discovery Overview report

To save the Data Discovery Overview report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Discovery Overview** section, click **View or Edit Report**.
3. Apply filters [122] (optional).
4. Click **Save**.
   The **Save Report** dialog is displayed.
5. In the **Report Name** field, enter a name for the report.
6. In the **Description** field, enter a description of the report (optional).
7. Select a time from the **Time** section.
8. Review the applied filters in the **Filter Parameter** section.
9. Click **Save**.
   The report is saved to **Saved Reports [128]**.

Download the Data Discovery Overview report

The Data Discovery Overview report can be exported in PDF format.

To download the Data Discovery Overview report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Discovery Overview** section, click **View or Edit Report**.
3. Click **Download Report**.
   The report downloads in PDF format.

Apply filters to the Data Discovery Overview report

You can filter the information displayed on the Data Discovery Overview report. The filters you apply are saved with the report.

You can apply the following filters to the Data Discovery Overview report:

- **Partial Match** or **Exact Match**
- **Filter by Application**
- **Filter by Tags**
- **Filter by Datazone**
- **Select Scan Type**
- **Search by Scan ID**
- **Date**

To apply filters to the Data Discovery Overview report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Data Discovery Overview** section, click **View or Edit Report**.
3. Select the filter you want to apply.

You can also apply Query Filters to the Data Discovery Overview report [123].

Apply Query Filters to the Data Discovery Overview report

You can apply Query filters to the Data Discovery Overview report by enabling the **Query Filters** toggle. See Reports with the Query Builder [129] for more information.

Data Discovery Overview graphs

The Data Discovery Overview report includes three graphs:

- **Tag Applied (Top 10)**
- **Tags by Application (Top 10)**
- **Tags by Datazone (Top 10)**

Discovery Alerts report

The Discovery Alerts report lists non-compliance alerts generated during scans. The Discovery Alerts report page displays the number of high, medium, and low level alerts as well as graphs that display information about the alerts.

View the Discovery Alerts report

To view the Discovery Alerts report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Discovery Alerts** section, click **View or Edit Report**.
   The **Discovery Alerts** report page is displayed.

Save the Discovery Alerts report

To save the Discovery Alerts report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Discovery Alerts** section, click **View or Edit Report**.
3. Apply filters [124] (optional).
4. Click **Save**.
   The **Save Report** dialog displays.
5. In the **Report Name** field, enter a name for the report.
6. In the **Description** field, enter a description of the report (optional).
7. Select a time from the **Time** section.
8. Review the applied filters in the **Filter Parameter** section.
9. Click **Save**.

Download the Discovery Alerts report

You can download the Discovery Alerts report in PDF format.

To download the Discovery Alerts report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Discovery Alerts** section, click **View or Edit Report**.
3. Click **Download Report**.
   The report downloads in PDF format.

Apply filters to the Discovery Alerts report

You can filter the information displayed on the Discovery Alerts report. The filters you apply are saved with the report.

You can apply the following filters to the Discovery Alerts report:

- **Search by Application**: View the results by application name.
- **Search by Datazone**: View the results by datazone.
- **Search by Alert Level**: View the results based on rating.
- **Search by Policy**: View the results by policy.
- **Search by Alert For**: View the results by alerts.
- **Date**: Search by specific date range.

To apply filters to the Discovery Alerts report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Discovery Alerts** section, click **View or Edit Report**.
3. Select the filter you want to apply.

You can also apply Query Filters [124] to the Discovery Alerts report.

Apply Query Filters to the Discovery Alerts report

You can apply Query filters to the Discovery Alerts report by enabling the **Query Filters** toggle. See Reports with the Query Builder [129] for more information.

Tags Trends reports

The Tag Trends report shows trends in tagging during Discovery scans.

View the Tag Trends report

To view the Tag Trends report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tags Trends** section, click **View or Edit Report**.
   The **Tags Trends** report is displayed.

Save the Tags Trends report

To save the Tags Trends report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tags Trends** section, click **View or Edit Report**.
3. Apply filters [125] (optional).
4. Click **Save**.
   The **Save Report** dialog displays.
5. Enter the following details:
   - Report Name (required).
   - Description.
   - Time.
6. Click **Save**.

The report is saved to **Saved Reports**.

Download the Tags Trends report

To download the Tags Trends report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tags Trends** section, click **View or Edit Report**.
3. Click **Download Report**.
   The report downloads in PDF format.

Apply filters to the Tags Trends report

You can filter the data displayed in on the Tags Trends report page. The filters you apply are saved with the report.

You can apply the following filters to the Tags Trends report:

- **Search by Application**: Filter the results by application name.
- **Search by Tags**: Filter the results by tag.
- **Search by Datazone**: Filter the results by datazone.
- **Date**: Filter the results by date and time.

To apply filters to the Tags Trends report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tags Trends** section, click **View or Edit Report**.
3. Select the filter you want to apply.

Tag Review Inventory Summary report

The Tag Review Inventory Summary report provides details about tags that have been reviewed.

The following information is included in the **Tag Review Inventory Summary** report:

- **User**: The name of the user. Click the username to view detailed information about the tag.
- **Datazone**: The name of the datazone.
- **Application**: The name of the application.
- **Resource**: The name of the resource. Click on the resource to view the Resource Detail page, which includes tabs such as Tag Details, Alerts Details, and Lineage along with the count of records in each tab.

View the Tag Review Inventory Summary report

To view the Tag Review Inventory Summary report, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tag Review Inventory Summary** section, click **View or Edit Report**.
   The **Tag Review Inventory Summary** report page displays.

Save the Tag Review Inventory Summary report

To save the Tag Review Inventory Summary report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tag Review Inventory Summary**, click **View or Edit Report**.
   The **Tag Review Inventory Summary** report page is displayed.
3. Apply filters [126] (optional).
4. Click **Save**.
   The **Save Report** dialog displays.

5. Enter the following details:
   - Report Name (mandatory).
   - Description.
   - Time.
6. Click **Save**.
   The report is saved to Saved Reports [128].

Export the Tag Review Inventory Summary report to CSV

To export the report in CSV file, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tag Review Inventory Summary** section, click **View or Edit Report**.
3. Click **Export to CSV**.
   The **Export** dialog displays.
4. You can select **All** columns or you can deselect any of the default columns.
5. Click **Export**.
   The **Tag Review Inventory Summary** downloads in CSV format.

The following are columns in the exported CSV file:

- **ReviewedBy**: The name of the reviewer.
- **Application**: The name of the application.
- **Datazone**: The name of the datazone.
- **DB**: The name of the database.
- **Table**: The name of the table.
- **Column**: The name of the column in the table.
- **Score**: The score value of the tag.
- **Tags**: The name of tag, such as `US_PHONE_NUMBER or EMAIL.`
- **TagReason**: Why the tag has been applied, such as lookup, model, or match pattern content.
- **TagStatus**: The status of the tag.
- **TagStatusReason**: The reason for the tag status change.
- **Owner**: The name of the resource owner.
- **Location**: The location (path) of the resource.
- **Size**: The size of the resource.
- **ReviewedOn(mm/dd/yyyy)**: The review date and time of classification in MM/DD/YYYY format. E.g. 04/16/2020 11:32:35

Apply filters to the Tag Review Inventory Summary report

You can filter the data that appears on the Tag Review Inventory Summary report using the following filters:

- **Search by Resource**: Allows you to view the result by the resource name.
- **Search by User**: Allows you to view the result by the user name.
- **Search by Application**: Allows you to view the result by application name.
- **Search by Tags**: Allows you to view the result by the tags.
- **Search by Datazone**: Allows you to view the result by the datazone.

To apply filters to the Tag Review Inventory Summary report, do the following:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. In the **Tag Review Inventory Summary** section, click **View or Edit Report**.
3. Select the filter you want to apply.

Tag Review Summary report

The following information is included in the **Tag Review Summary** report.

- **Review counts**: The number of reviewed resources, user accepted tags, and user rejected tags.
- **Charts**: section displays the below charts.
  - Tag Summary by all Users
  - Most accepted and rejected tags (Top 5)
  - Datazones

View/ddit report

To view/edit an existing report, use the following steps:

1. On the Privacera home page, expand the **Reports** menu and click on **Built-in Reports** from left menu.
2. Under Audit Summary>**Tag Review Summary**, click View or Edit Report.
   The **Tag Review Summary** report page is displayed.
3. To view/edit the report, you can apply the different search filters such as search by User, Application, Tag, and Datazone. You can also apply filter by using specific date range.

Save report

To save the report, use the following steps:

1. On the Privacera home page, on the left, expand the **Reports** menu and click **Built-in Reports**.
2. Under **Tag Review Summary**, click **View or Edit Report**.
3. You can apply different search filters such as search by application, tag, and datazone. In addition, you can apply date ranges and other advanced filters.
4. Click **Save**.
   The Save Report pop-up is displayed.
5. Enter the following details:
   - Report Name (mandatory).
   - Description.
   - Time.
6. Click **Save**.

The report is saved and is viewable under **Saved Reports**.

Download report

To download an existing report, use the following steps:

1. On the Privacera home page, on the left, expand the **Reports** menu and click **Built-in Reports** from left menu.
2. Under **Tag Review Summary**, click **View or Edit Report**.
3. On the **Built-in Reports** page, click **Download Report**.

The report is downloaded in PDF format.

Search filters

To apply various search filters, use the following steps:

1. On the Privacera home page, on the left, expand the **Reports** menu and click **Built-in Reports**.
2. Under **Tag Review Summary**, click **View or Edit Report**.
   The following search filters are available:
   - **Search by Users**: View the result by the Users name.
   - **Search by Application**: View the result by the application name.
   - **Search by Tags**: View the result by the tags.
   - **Search by Datazone**: View the result by the datazone.

   You can also search by date range.

Offline reports

Reports that result in a large number of rows and that would require much time to export are moved to **Offline Reports**. To view offline reports, select **Reports** > **Offline Reports** from the navigation menu. The **Offline Reports** page shows the status of every report export operation.

The **Offline Reports** page displays the following information:

- **Job Id**: The job ID of the offline report.
- **Job Status**: The job status. For example: success or failed.
- **Report Name**: The name of the report.
- **Created By**: The name of the user who created the report.
- **Create Time**: The report creation date and time.
- **End Time**: The report creation end date and time.
- **File Path**: The location of the file.
- **File Size**: The file size of the report.
- **Action**: Download the offline report.

Reports with more than million rows are exported in zip format.

Download offline reports

If the **Job Status** of a report is **Success**, you can download the report.

To download offline reports, follow these steps:

1. From the navigation menu, select **Reports** > **Offline Reports**.
2. Click the **Download Report** icon in the **Actions** column.
   The report is downloaded.

Saved Reports

The Saved Reports page displays the following information:

- **Report Name**: The name of the report.
- **Report Type**: The type of report. For example: Data Discovery Overview, Data Inventory, or Discovery Alerts.
- **Schedule Type**: The schedule type. For example: monthly, weekly, or hourly.
- **Start Time**: The start time of the report.
- **Day**: The day of the report.
- **Month**: The month of the report.
- **Owner**: The owner of the report. Click the column name to sort in ascending or descending order.
- **Updated By**: The name of the user who updated the report. Click the column name to sort in ascending or descending order.
- **Updated On**: The date and time the report was most recently updated. Click the column name to sort in ascending or descending order.
- **Actions**: Preview, edit, or delete the report.

Export saved reports to CSV

You can select multiple saved reports for exporting in CSV format.

To export saved reports in CSV format, follow these steps:

1. From the navigation menu, select **Reports** > **Saved Reports**.
   The **Saved Reports** page displays.
2. Click **Export CSV**.
3. From the dropdown menu, select the reports you want to export.

4. Click **Export**.
   The reports are exported in CSV format.

Customize report fields

You can select specific fields you want for the following reports:

- Data inventory
- Tag Review Inventory Summary
- Scan Summary
- Compliance Summary

To customize report fields, do the following:

1. Select the names of the reports that you want to export.
   The **Export to CSV** dialog displays.
2. Customize the fields you want to export:
   - **No. of Resource, All checkbox**: Clear this checkbox if you want only a specific number of resources.
   - **Columns, All checkbox**: Clear this checkbox to include only specific columns.
   - **Include empty metaname tagged columns**: Select this checkbox to include empty columns (columns that do not have data) but have been tagged.
   - **Include empty table**: Select this checkbox to include empty tables (tables that do not have data) but have been tagged at the table and column level.
   - **Include empty table**: Select this checkbox to include empty tables (tables that do not have data) but have been tagged at the table and column level.
   - **Show only column tags on column**: Select this checkbox to show only the tags at the column level but not the tags at the table level.
   - **Exclude External Table Location**: Select this checkbox to exclude a table for HDFS resource.
   - **Exclude Reviewed Tags**: Select this checkbox to exclude already reviewed tags from the report.
3. Click **Export**.
   The report is exported in CSV format.

> **NOTE**
> If the reports have a large number of rows and the export requires much time, they run as an offline job and the following message is displayed:
>
> "This request has been submitted as an Offline job. The Job Id is *<some_job_number>*."
>
> See **Offline Reports** [128] for more information.

Reports with the Query Builder

With the Query Builder, you can create reports using Privacera-supplied and custom filters.

There are two types of queries:

- **Privacera-supplied**: Dropdown menus to specify values of fields that must be included in the search results.
- **Query Filters** : Allow you specify fields that *must or must not* be included in the search results.

About the Scan Type filter

The **Scan Type** dropdown menu has the following selections:

- **Offline**: The report is based on the most recently stored data collected from the assets by Privacera.
- **Realtime**: The report is based on data collected from the assets at the time the report is run. Depending on the number of assets you have, network latency, and other considerations, generating a realtime report might require that you be patient for the results.

Access the Query Builder

To access the Query Builder, follow these steps:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. Select **View or Edit Report** to open either **Data Discovery Overview** or **Discovery Alerts**. The Query Builder page is displayed.

The Privacera-supplied filters are shown at the top. Search results are displayed at the bottom.

From left to right across the top:

- Partial or full match of the data
- Dropdown menus to search by **Application**, **Tag**, **Datazone**, or to exclude a certain resource.
- Search by location.
- Search only the last seven days.
- **Scan type**: either **Offline** or **Realtime**.
- **Group Folders** in the output.
- **Group Tables** in the output.

Use the Query Builder

This is a general approach to using the Query Builder. As you become more familiar with the tool, you will refine this general approach for your needs.

Prerequisites:

- Be sure you have a good understanding of the types of information you have associated with your cloud-based assets: the applications, tags, datazones, and other categories. See Available Query filters [130].
- Decide on the fields that will display the information you want to see. For instance, are you interested in the applications or the datazones or both?

To use the Query Builder, follow these steps:

1. From the dropdown menus on the Query Builder page, select the fields you want to see.
2. As you make selections, total counts of the assets in various categories based on your selection criteria and the complete results of the search are dynamically displayed.
3. Look at the output to determine if you have the data you want to see. You can adjust the menu selections to further refine the results. You can also use Available Query filters [130].to tailor the output to your needs.

Available Query filters

The following filters are the basis for both the Privacera-supplied search and custom Query Filters [130].

Filter	Description
Application	The application running on the asset, such as a web server or database.
Tags	User-defined information fields you have applied to assets.

Filter	Description
Tag Attributes	Additional criteria on those tags to further refine the information.
Datazone	Your logical grouping of the assets by type or security level of the data
Datazone Owner	The registered user or group that manages that datazone

Create custom reports with Query Filters

Using Query Filters, you can find and view precise information about your cloud-based assets.

Conditions, filters, “Are”, and “Are Not”

You can define certain conditions to filter the data to select the assets you are interested in. These filters are listed in Available Query filters [130]. You set these conditions as either **Are** applicable (the filter must match) or **Are Not** applicable (the filter must not match) to the assets you are interested in.

You can combine multiple conditions to further refine your search.

Enable Query Filters

To enable Query Filters:

1. From the navigation menu, select **Reports** > **Built-in Reports**.
2. Click **View or Edit Report** to open either **Data Discovery Overview** or **Discovery Alerts**.
3. Enable Query Filters using the **Query Filters** toggle.

Apply Query Filters

This is a general approach to using Query Filters. As you become more familiar with the tool, you will refine this general approach for your precise needs.

Prerequisites:

- Be sure you have a good understanding of the types of information you have associated with your cloud-based assets: the applications, tags, datazones, and other categories. See Available Query filters [130].
- Decide the condition or conditions that will display the information you want to see. For instance, are you interested in the applications or the datazones or both?

To apply Query Filters, follow these steps:

1. Enable Query Filters [131] using the **Query Filters** toggle.
2. Select **Add Condition**.
3. From the dropdown menus for the condition:
   a. On the left, select the needed filter.
   b. In the middle, specify whether that filter is applicable (**Are**) or not applicable (**Are Not**) to the search.
   c. On the right, enter a string that matches what you are interested in.

   A list of matching data is displayed from which you can select the exact field. The **Query Preview** on the right shows a text representation of the query.
4. Add more conditions, if required. To add another condition, specify **AND** to indicate that the condition is required or **OR** to indicate that the condition is optional.
5. After you have added all required conditions, click **Search**.

   The total counts of the assets in various categories based on your filter criteria and the complete results of the search are displayed.

Save or export a report

To reuse a search at a later time, click **Save** and enter a name for the search.

To download a copy of the resulting data in Comma-Separated Value (CSV) format, click **Export to CSV** and follow the leading prompts.

Discovery Health Check

The **Discovery Health Check** creates notifications about the status of .

To view the notifications, click the bell icon in the Privacera portal header. The **Critical Issues or Warnings** dialog displays.

Critical Issues or Warnings dialog

The **Critical Issues or Warnings** dialog displays the following information:

- **State**: The state of the notification: Ok , Warning, or Critical.
- **Version**: The software version number.
- **Service - Component**: The name of the service or component, such as Discovery.
- **Type**: The type of issue or warning, such as process or general.
- **Last Updated**: The last time the issue or warning was updated. You can hover over the column to view the exact date and time.
- **Description**: A description of the issue or warning.
- **Expiry Time**: The expiry date of the Keytab. The state of the notification is changed depending on the expiry date.

Add Keytab expiry date

You need to always keep the Keytab expiry date current so you can be notified whenever the Keytab expires.

To add a Keytab expiry date, follow these steps:

1. Click the notification icon in the Privacera portal header.
2. Under the **Expiry Time** column, click the date picker and select an expiry date.
3. Click **Close**.
   The Keytab expiry date is updated.

Keytab notification states

There are three notification states: Normal, Warning, and Critical.

Normal

The Normal state displays the bell icon in gray. This means that no action is needed.

Warning

The Warning state displays the bell icon in orange. This happens 10 days before the Keytab expires. This means that you update the Keytab before it expires to keep Privacera services running.

Critical

The Critical state displays the bell icon in red. This happens when Keytab has expired. This indicates that the Keytab needs to be renewed to keep Privacera services up and running.

Set custom Discovery properties on Privacera Platform

This topic provides the list of custom properties that can be configured for the Discovery service. It covers how you can configure the custom properties in Privacera Manager (PM) CLI.

To use a custom property from the properties table:

1. Add the property to the following YML file in the `custom-vars` folder configured as per your environment.
   - `vars.discovery.aws.yml`
   - `vars.discovery.azure.yml`
   - `vars.discovery.gcp.yml`
2. Run the following command:

```
cd ~/privacera/privacera-manager
./privacera-manager.sh update
```

Discovery properties

Property	Description	Values	Default Value
`DISCOVERY_IMAGE_NAME`			
`DISCOVERY_IMAGE_TAG`			
`DISCOVERY_ENABLE`	Set it true to enable Discovery.	true,false	
`USE_DATABRICKS_SPARK`	Enable to use Databricks Spark instead of Apache Spark.	true,false	
`DISCOVERY_INSTALL`			
`DISCOVERY_FS_PREFIX`	For accessing the filesytem of the cloud storage service, do the following: • For AWS and GCP, set the filesystem prefix. **s3a://** is the prefix for AWS, and **gs://** for GCP. • For Azure, set the container name. A container name is associated with your Azure storage account and where the blobs are organized containing the data to be scanned.	• s3a:// • StorageContainerName • gs://	
`DISCOVERY_CLOUD_TYPE`	Set the cloud type used for the Discovery setup.	• AWS • AZURE • GCP	
`DISCOVERY_TRUSTSTORE_PASSWORD`			
`AUTO_START_DATABRICKS_JOB`			
`DISCOVERY_REALTIME_ENABLE`	Set to true to enable real-time scan in Discovery.	true,false	false
`DISCOVERY_MENU_ENABLE`	Set to true to enable Discovery menu on Privacera Portal.	true,false	false
`DISCOVERY_LOG_LEVEL`			

Property	Description	Values	Default Value
`DISCOVERY_FOLDER_TAGGER_ENABLE` **NOTE** This property is supported only on AWS and Azure platforms.	Set this property to true to enable folder tagger job. Set this property to false to disable folder tagger job.	true,false	false
`DISCOVERY_STORE_SAMPLE_VALUES`	Whether any sample values should be stored for a column or field	true,false	false
`DISCOVERY_MAX_SAMPLE_VALUES`	Maximum sample values stored for a column or field.		
`DISCOVERY_ENCRYPT_SAMPLE_VALUES`	Whether the samples should be stored encrypted.	true,false;	false
`DISCOVERY_STREAM_SUFFIX`			
`DISCOVERY_STREAM_TAGS`			
`DISCOVERY_TABLE_SUFFIX`			
`DISCOVERY_TABLE_TAGS`			
`DISCOVERY_BUCKET_NAME`			
`DISCOVERY_BUCKET_TAGS`			
`DISCOVERY_CREATE_NOSQL_TABLES`			
`DISCOVERY_GEN_TERRAFORM_NOSQL_TABLES`	Set to true if you want to create Dynamodb tables using terraform. Set to false to disable terraform and create the resource manually.		true
`DISCOVERY_CREATE_STREAMS`			
`DISCOVERY_GEN_TERRAFORM_STREAMS`	Set to true if you want to create Kinesis streams using terraform. Set to false to disable terraform and create the resource manually.		true
`DISCOVERY_CREATE_BUCKET`			
`DISCOVERY_GEN_TERRAFORM_BUCKET`	Set to true if you want to create S3 bucket using terraform. Set to false to disable terraform and create the resource manually.		true
`DISCOVERY_GEN_TERRAFORM_AZURE_ACCOUNT`			
`DISCOVERY_SPARK_DRIVER_MEMORY`			
`DISCOVERY_SPARK_EXECUTOR_MEMORY`			
`DISCOVERY_SPARK_DRIVER_CORES`			
`DISCOVERY_SPARK_EXECUTOR_CORES`			
`DISCOVERY_SPARK_EXECUTOR_INSTANCES`			
`DISCOVERY_CREATE_DEFAULT_APP_IN_PORTAL`			
`DISCOVERY_COSMOSDB_FILE_REPOSITORY_PATH`			
`DISCOVERY_COSMOSDB_DOCUMENT_SIZE_LIMIT`			
`DISCOVERY_COSMOSDB_OFFER_THROUGHPUT`			
`DISCOVERY_AWS_CLOUD_ASSUME_ROLE`	Property to enable/disable to grant Discovery access to AWS services to perform the scanning operation.		true

Property	Description	Values	Default Value
DISCOVERY_AWS_CLOUD_ASSUME_ROLE_ARN			
DISCOVERY_BUCKET_SQS_NAME	Set this property if you want to set a custom name for a SQS queue.		privacera_bucket_sqs_{{DEPLOYMENT_ENV_NAME}}
DISCOVERY_SQS_TAGS			
DISCOVERY_CREATE_SQS			
DISCOVERY_GEN_TERRAFORM_SQS	Set to true if you want to create SQS resource using terraform. Set to false to disable terraform and create the resource manually.		true
DATABRICKS_INIT_DBFS_FOLDER			
DATABRICKS_DISCOVERY_CUST_CONF_ZIP_NAME			
DATABRICKS_DISCOVERY_INIT_SCRIPT_PATH			
DATABRICKS_DISCOVERY_SPARK_VERSION	The version of Spark used in a Databricks cluster.	• 6.4.x-scala2.11 (Spark 2.4) • 7.3.x-scala2.12 (Spark 3.0) • 7.4.x-scala2.12 (Spark 3.0) • 7.5.x-scala2.12 (Spark 3.0) • 7.6.x-scala2.12 (Spark 3.0) • 9.1.x-scala2.12	
DISCOVERY_SPARK_TASK_SCHEDULER_ENABLE			
DISCOVERY_RANGER_REST_ENABLED			
DISCOVERY_K8S_IMAGE_NAME			
DISCOVERY_K8S_IMAGE_TAG			
DISCOVERY_K8S_IMAGE_PULL_POLICY			
DISCOVERY_K8S_PVC_NAME			
DISCOVERY_K8S_PVC_STORAGE_SIZE_MB			
DISCOVERY_K8S_PVC_STORAGE_SIZE			
DISCOVERY_K8S_STORAGE_PROVISIONER			
DISCOVERY_K8S_SC_NAME			
DISCOVERY_K8S_PV_ENCRYPTED			
DISCOVERY_K8S_PV_KEY			
DISCOVERY_K8S_LOADBALANCER_EXTERNAL			
DISCOVERY_K8S_ANNOTATION_LOADBALANCER_ANNOTATION			

Property	Description	Values	Default Value
DISCOVERY_K8S_SPARK_UI_PORT			
DISCOVERY_K8S_SPARK_UI_PORT_EXTERNAL	Property to change the default port number for Discovery.		4040
DISCOVERY_K8S_SPARK_EVENT_LOG_ENABLED			
DISCOVERY_K8S_SPARK_DRIVER_PORT			
DISCOVERY_K8S_SPARK_BLOCKMANAGER_PORT			
DISCOVERY_K8S_SPARK_PORT_MAX_RETRIES			
DISCOVERY_K8S_SPARK_SERVICE_AC_NAME			
DISCOVERY_K8S_SPARK_DRIVER_MEMORY	Minimum amount of Kubernetes memory to be used by Discovery Driver. For example, DISCOVERY_K8S_SPARK_DRIVER_MEMORY: "1G"		
DISCOVERY_K8S_SPARK_EXECUTOR_MEMORY	Minimum amount of Kubernetes memory in MB to be requested by Discovery Executor. For example, DISCOVERY_K8S_SPARK_EXECUTOR_MEMORY: "1024".		
DISCOVERY_K8S_SPARK_DRIVER_CORES	Minimum amount of Kubernetes CPU to be requested by Discovery Driver. For example DISCOVERY_K8S_SPARK_DRIVER_CORES: “1”.		
DISCOVERY_K8S_SPARK_EXECUTOR_CORES	Minimum amount of Kubernetes CPU to be requested by Discovery Executor. For example DISCOVERY_K8S_SPARK_EXECUTOR_CORES: “1”.		
DISCOVERY_K8S_SPARK_EXECUTOR_INSTANCES			
DISCOVERY_K8S_SPARK_DRIVER_LIMIT_CORES	Maximum amount of Kubernetes CPU to be used by Discovery Driver. For example, DISCOVERY_K8S_SPARK_DRIVER_LIMIT_CORES: "0.5".		
DISCOVERY_K8S_SPARK_EXECUTOR_LIMIT_CORES	Maximum amount of Kubernetes CPU to be used by Discovery Executor. For example, DISCOVERY_K8S_SPARK_EXECUTOR_LIMIT_CORES: “0.5”.		
DISCOVERY_K8S_SPARK_EXECUTOR_REQUEST_CORES	Minimum amount of Kubernetes CPU to be used by Discovery Executor. For example, DISCOVERY_K8S_SPARK_EXECUTOR_REQUEST_CORES: “0.5”.		
DISCOVERY_K8S_SPARK_MASTER			
DISCOVERY_K8S_MEM_LIMITS			
DISCOVERY_K8S_MEM_REQUESTS			
DISCOVERY_K8S_CPU_LIMITS			
DISCOVERY_K8S_CPU_REQUESTS			
DISCOVERY_AZURE_APP_CLIENT_ID			
DISCOVERY_AZURE_STORAGE_ACCOUNT_NAME			
DISCOVERY_AZURE_URL_PREFIX			
DISCOVERY_AZURE_AUDIT_TYPE			
DISCOVERY_AZURE_LOCATION			

Property	Description	Values	Default Value
CREATE_AZURE_RESOURCES			
DISCOVERY_AZURE_RESOURCE_GROUP			
DISCOVERY_AZURE_APPLICATION_ID			
DISCOVERY_AZURE_TENANTID			
DISCOVERY_AZURE_APP_CLIENT_SECRET_BASE64			
DISCOVERY_AZURE_SUBSCRIPTION_ID			
DISCOVERY_AZURE_COSMOS_DB_ACCOUNT			
DISCOVERY_PORTAL_SERVICE_USERNAME			
DISCOVERY_PORTAL_SERVICE_PASSWORD			
DISCOVERY_CLOUD_MODE			
DISCOVERY_AWS_ENDPOINT_ENABLE			
DISCOVERY_KINESIS_ENDPOINT_URL			
DISCOVERY_DYNAMODB_ENDPOINT_URL			
DISCOVERY_SOLR_BASIC_AUTH_ENABLED			
DISCOVERY_SOLR_BASIC_AUTH_USER			
DISCOVERY_SOLR_BASIC_AUTH_PASSWORD			
PRIVACERA_DISCOVERY_SECRETS_FILE			
DISCOVERY_ENCRYPT_SECRETS			
PRIVACERA_DISCOVERY_SECRETS_KEYSTORE_PASSWORD			
DISCOVERY_ENCRYPT_PROPS_LIST			
DISCOVERY_PORTAL_SERVICE_PASSWORD			
PRIVACERA_DISCOVERY_DATASOURCE_PASSWORD			
RANGER_TAGSYNC_PASSWORD			
DISCOVERY_SOLR_BASIC_AUTH_PASSWORD			
PRIVACERA_DISCOVERY_DATASOURCE_PASSWORD			
DISCOVERY_FS_S3A_ACCCESS_KEY			
DISCOVERY_FS_S3A_SECRET_KEY			
DISCOVERY_CLUSTER_NAME			
DISCOVERY_AGENT_MODE			
DISCOVERY_LOGS_SOLR_ENABLE			
DISCOVERY_RANGER_HOOK_ENABLED			
DISCOVERY_SPARK_DOCKER_DRIVER_MEMORY			
DISCOVERY_SPARK_DOCKER_EXECUTOR_MEMORY			
DISCOVERY_SPARK_DOCKER_DRIVER_CORES			
DISCOVERY_SPARK_DOCKER_EXECUTOR_CORES			
DISCOVERY_SPARK_DOCKER_EXECUTOR_INSTANCES			
DISCOVERY_DOCKER_SPARK_MASTER			
DISCOVERY_OFFLINE_SCAN_DEBUG_ENABLED			
DISCOVERY_SCAN_BACKUP_CLEANER_INTERVAL_HR			
DISCOVERY_RTBF_POLICY_ENABLED			
DISCOVERY_WORKFLOW_POLICY_ENABLED			

Property	Description	Values	Default Value
DISCOVERY_WORKFLOW_EXPUNGE_POLICY_ENABLED			
DISCOVERY_DEIDENTIFICATION_POLICY_ENABLED			
DISCOVERY_CONTENT_SCANNING_ENABLED			
DISCOVERY_SCAN_OFFICE_MIME_TYPES_AS_ARCHIVE_ENABLED			
DISCOVERY_OFFLINE_SCAN_BACKUP_FOLDER			
DISCOVERY_DICT_BASE_PATH			
DISCOVERY_ML_BASE_PATH			
DISCOVERY_ML_TAG_ACTION_MODEL_PATH			
DISCOVERY_SCAN_REQUEST_FILES_DIR			
PARTIAL_MATCH_ENABLE			
DISCOVERY_COSMOSDB_URL			
DISCOVERY_COSMOSDB_KEY			
DISCOVERY_GEN_TERRAFORM_WITH_MSI_ROLE			
DISCOVERY_AZURE_HNS_ENALBED			
DISCOVERY_AZURE_ACCOUNT_REPLICATION_TYPE			
DISCOVERY_AZURE_ACCOUNT_KIND			
DISCOVERY_SAMPLE_VALUES_MAX_LENGTH	Maximum length of a sample that is stored for a column or field		
DISCOVERY_S3_AUDITS_ENABLE			
DISCOVERY_ADLS_AUDITS_ENABLE			
DISCOVERY_GCS_AUDITS_ENABLE			
DISCOVERY_GBQ_AUDITS_ENABLE			
DISCOVERY_DEPLOYMENT_SUFFIX_ID			
DISCOVERY_SERVICE_USER			
DISCOVERY_VERSION_FILE_NAME			
DISCOVERY_HEARTBEAT_UPDATE_INTERVAL_SEC			
DISCOVERY_SCAN_BACKUP_CLEANER_THRESHOLD_HR			
DISCOVERY_LOOKUP_COPY_TO_HDFS_INTERVAL_SEC			
DISCOVERY_GENERATE_SRC_ALERT_INTERVAL_MIN			
DISCOVERY_LOOKUP_COPY_TO_HDFS_FROM_AGENT			
DISCOVERY_RETRY_ON_FAILURE_INTERVAL_SEC			
DISCOVERY_SCAN_DELAY_RETRY_INTERVAL			
DISCOVERY_SCAN_DELAY_RETRY_COUNT			
DISCOVERY_HOST			
DISCOVERY_KAFKA_HEARTBEAT_INTERVAL_MS			
DISCOVERY_KAFKA_REQUEST_TIMEOUT_MS			
DISCOVERY_KAFKA_SESSION_TIMEOUT_MS			

Property	Description	Values	Default Value
DISCOVERY_KAFKA_CONNECTIONS_MAX_IDLE_MS			
DISCOVERY_KAFKA_ENABLE_AUTO_COMMIT			
DISCOVERY_KAFKA_AUTO_OFFSET_RESET			
DISCOVERY_KERBEROS_ENABLE			
DISCOVERY_SOLR_KERBEROS_ENABLE			
DISCOVERY_HBASE_KERBEROS_ENABLE			
DISCOVERY_KAFKA_KERBEROS_ENABLE			
DISCOVERY_KERBEROS_RELOGIN_INTERVAL_SECS			
DISCOVERY_PORTAL_KERBEROS_ENABLE			
DISCOVERY_SCAN_WORKER_KAFKA_SEND_BUFFER_MEMORY			
DISCOVERY_SCAN_WORKER_KAFKA_SEND_LINGERMS			
DISCOVERY_SCAN_WORKER_KAFKA_SEND_BATCHSIZE			
DISCOVERY_SCAN_WORKER_KAFKA_SEND_RETRIES			
DISCOVERY_SOLR_COLLECTION			
DISCOVERY_SOLR_LINEAGE_COLLECTION			
DISCOVERY_SOLR_ALERT_COLLECTION			
DISCOVERY_SOLR_RESOURCE_COLLECTION			
DISCOVERY_SOLR_OFFLINE_SCAN_SUMMARY_COLLECTION			
DISCOVERY_SOLR_RESOURCE_META_INFO_COLLECTION			
DISCOVERY_SOLR_RESOURCE_AUDIT_COLLECTION			
DISCOVERY_SOLR_SPARK_EVENT_COLLECTION			
DISCOVERY_SOLR_OFFLINE_SCAN_CLEANUP_COLLECTION			
DISCOVERY_UNSTRUCTURED_VALUE_CHECKING_ENABLED			
DISCOVERY_NUM_TOKENS_FOR_UNSTRUCTURED_DATA_DETECTION			
DISCOVERY_SCAN_INCLUDE_PART_FILES_MAX_INDEX			
DISCOVERY_ACTIVE_SCAN_ENABLE			
DISCOVERY_SPARK_JOB_SCHEDULER_SLEEP_TIME_MS			
DISCOVERY_AMOUNT_ARRAYVALUES_EXTRACTED			
DISCOVERY_RECOVERY_SPARK_DEFAULT_POOL_NAME			
DISCOVERY_CONSUMER_RECORD_WAIT_TIMEOUT_MS			
DISCOVERY_CONSUMER_RECORD_BATCH_SIZE			
DISCOVERY_RECOVERY_RETRY_MAX			
DISCOVERY_GENERAL_CONSUMER_QUEUE_SIZE			
DISCOVERY_OFFLINE_CONSUMER_QUEUE_SIZE			

Property	Description	Values	Default Value
DISCOVERY_CONSUMER_RECORD_DB_PATHS			
DISCOVERY_CONSUMER_RECORD_HANDLER_THREAD_POOL_SIZE	Property to configure the thread pool size for handling the consumer records. The property determines how many data source applications can be handled by the scheduler, so the property value should be more than the data source applications that are registered in an installation.		100
DISCOVERY_SEND_CHILD_TO_EXCLUDE_RESOURCE_INFO_ENABLE			
DISCOVERY_DYNAMODB_WRITE_ITEM_MAX_SIZE			
DISCOVERY_DYNAMODB_WRITE_BATCH_SIZE			
DISCOVERY_DYNAMODB_READ_BATCH_SIZE			
DISCOVERY_DYNAMODB_CHILD_COLUMN_LIMIT			
DISCOVERY_AZURE_PAYLOAD_LIMIT			
DISCOVERY_METASTORE_PAYLOAD_TABLE			
DISCOVERY_METANAME_LEAF_ONLY			
DISCOVERY_SEND_SPARK_JOB_EVENT			
DISCOVERY_RESTART_ON_STUCK_JOBS			
DISCOVERY_START_SCRIPT			
DISCOVERY_DB_MAX_STATEMENTS			
DISCOVERY_DB_MAX_POOL_SIZE			
DISCOVERY_DB_ACQUIRE_INCREMENT			
DISCOVERY_DB_MIN_POOL_SIZE			
DISCOVERY_COSMOSDB_MAX_POOL_SIZE			
DISCOVERY_COSMOSDB_RETRY_INTERVAL_SEC			
DISCOVERY_COSMOSDB_MAX_RETRY			
DISCOVERY_COSMOSDB_DATABASE_NAME			
DISCOVERY_SAVE_ARCHIVE_FILES			
DISCOVERY_RTBF_USE_ENCRYPTION			
DISCOVERY_DATAZONE_MONITOR_OFF_PREMISE_SRC_ENABLE			
DISCOVERY_DATAZONE_RESOURCE_REEVALUATE_ENABLED			
DISCOVERY_SCAN_NEW_SCANNER_ENABLE			
DISCOVERY_RIGHT_TO_PRIVACY_THREAD_POOL_SIZE			
DISCOVERY_OFFLINE_SCAN_RETRY_COUNT			
DISCOVERY_OFFLINE_SCAN_AUTO_RETRY_ENABLE			
DISCOVERY_OFFLINE_FILE_AND_FOLDER_COUNTING_TASK_POLL_TIME_MS			
DISCOVERY_OFFLINE_FILE_AND_FOLDER_COUNTING_TASK_TIMEOUT_MS			
DISCOVERY_OFFLINE_SCAN_PARTITION_ENABLE			
DISCOVERY_MAX_DICT_WORD_TO_SENTENCE_RATIO			

Property	Description	Values	Default Value
DISCOVERY_APPLY_METANAME_DICT_TO_UNSTRUCT			
DISCOVERY_MAX_BYTES_FOR_WORKFLOW			
DISCOVERY_PRECORDS_PARQUET_VERSION			
DISCOVERY_UNSTRUCT_TAGS_FILENAME			
DISCOVERY_WORKFLOW_DUPLICATE_FILE_RETRY_MAX_ATTEMPTS			
DISCOVERY_WORKFLOW_EXPUNGE_SPARKDF_SINGLE_FILE			
DISCOVERY_WORKFLOW_EXPUNGE_SPARKDF_ENABLE			
DISCOVERY_CLOUD_USE_ASSUMEROLE			
DISCOVERY_GCP_CLOUD_OUTPUTWRITERS_ENABLE			
DISCOVERY_DROOLS_POOL_SIZE			
DISCOVERY_DROOLS_USE_POOL			
DISCOVERY_INVALID_HEADER_CHARS_PAT			
DISCOVERY_MAX_HEADER_LEN			
DISCOVERY_STRUCT_VALUE_FULL_MATCH_ENABLED			
DISCOVERY_CLASSIFIER_AUTO_CREATE_MANUAL_TAG			
DISCOVERY_HBASE_BACKUP_TTL_MS			
DISCOVERY_HBASE_BACKUP_TTL_ENABLE			
DISCOVERY_HBASE_CLIENT_SCANNER_TIMEOUT_MS			
DISCOVERY_EXCLUSION_CLEANER_SLEEP_MIN			
DISCOVERY_EXCLUSION_CLEANER_BATCH_SIZE			
DISCOVERY_EXCLUSION_CLEANER_ENABLE			
DISCOVERY_FOLDER_TAGGER_BATCH_SIZE **NOTE** This property is effective only when DISCOVERY_FOLDER_TAGGER_ENABLE is enabled.	Represents how many records to fetch in each iteration.	100, 200, 300...	100

Property	Description	Values	Default Value
DISCOVERY_FOLDER_TAGGER_BACKOFF_TIME_SEC **NOTE** This property is effective only when DISCOVERY_FOLDER_TAGGER_ENABLE is enabled.	Fetches classification records from solr that are older than the time specified.	Time in seconds	120 seconds
DISCOVERY_FOLDER_TAGGER_SLEEP_TIME_MS **NOTE** This property is effective only when DISCOVERY_FOLDER_TAGGER_ENABLE is enabled.	Sleep time after each iteration.	Time in seconds	1000 mseconds
DISCOVERY_CMD_SERVER_ENABLED			
DISCOVERY_CMD_SERVER_PORT			
DISCOVERY_RULE_ENGINE_ADJUST_SCORES			
DISCOVERY_NOUN_LIST_FILE			
DISCOVERY_SPARK_JOB_MAX_TIME_MS			
DISCOVERY_ClASSIFY_RECORD_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_ClASSIFY_RECORD_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_ATLAS_HOOK_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_ATLAS_HOOK_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_NAV_TO_PRIVACERA_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_NAV_TO_PRIVACERA_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_SCAN_DELAY_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_SCAN_DELAY_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_ADLS_AUDITS_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_ADLS_AUDITS_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_S3_AUDITS_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_S3_AUDITS_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_DYNAMODB_AUDITS_MAPPER_TASK_POLL_TIME_MS			

Property	Description	Values	Default Value
DISCOVERY_DYNAMODB_AUDITS_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_HIVE_AUDITS_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_HIVE_AUDITS_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_CONTENT_CLASSIFIER_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_CONTENT_ClASSIFIER_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_CONTENT_SCAN_WORKER_TOPIC_PARTITION			
DISCOVERY_CONTENT_SCAN_COLLECTOR_CYCLE_TIME_MS			
DISCOVERY_DEFAULT_SPARK_PARTITION_PERCENT			
DISCOVERY_USE_SPARK_PARTITION_CALC			
DISCOVERY_HIVE_PROXY_USER_FEATURE			
DISCOVERY_KERBEROS_LOGIN_RETRY_INTERVAL_MS			
DISCOVERY_KERBEROS_LOGIN_NUM_RETRIES			
DISCOVERY_LFS_USE_FILE_MONITOR			
DISCOVERY_LFS_USE_FILE_WATCHER			
DISCOVERY_OFFLINE_SCAN_CLEANUP_THREAD_POOL_SIZE			
DISCOVERY_OFFLINE_SCAN_THREAD_POOL_SIZE			
DISCOVERY_QUICK_SCAN_LIMIT			
DISCOVERY_QUICK_SCAN_ENABLE			
DISCOVERY_DO_HDFS_SCHEMA_MAPPING			
DISCOVERY_ALLOW_FUZZY_MATCH_TAGS			
DISCOVERY_EXEC_MIMETYPE_REMOVE_DEFAULTS			
DISCOVERY_DEV_TEST_MODE			
DISCOVERY_TRIGGER_FILE_PATH			
DISCOVERY_POST_PROCESS_DROOLS_RULES_FILENAME			
DISCOVERY_CLASSIFIER_RULES_UNSTRUCT_FILENAME			
DISCOVERY_CLASSIFIER_RULES_FILENAME			
DISCOVERY_CLASSIFIER_DROOLS_RULES_FILENAME			
DISCOVERY_CHAT_SCAN_SKIP_INVALID_JSON_OUTPUT			
DISCOVERY_UNSTRUCT_AS_SINGLE_LINE			
DISCOVERY_POST_PROCESS_DATA_KEYSCORE_THRESHOLD			
DISCOVERY_UNSTRUCTURED_DATA_KEYSCORE_THRESHOLD			
DISCOVERY_STRUCTURED_DATA_KEYSCORE_THRESHOLD			
DISCOVERY_USE_KEYSCORE_THRESHOLD			
DISCOVERY__ML_PYTHON_FILE			
DISCOVERY_ML_CONDA_ENV_PATH			

Property	Description	Values	Default Value
DISCOVERY_ML_NLP_ENABLED			
DISCOVERY_POST_PROCESS_RULE_ENGINE_ENABLED			
DISCOVERY_RULE_ENGINE_DO_FALLBACK			
DISCOVERY_RULE_DATABASE_ENABLED			
DISCOVERY_RULE_ENGINE_ENABLED			
DISCOVERY_RULE_ENGINE_DROOLS_ENABLED			
DISCOVERY_RESOURCE_META_SCAN_MAPPER_CHECK_TASK_ACTIVE_INTERVAL_TIME_MS			
DISCOVERY_RESOURCE_META_SCAN_MAPPER_TASK_POLL_TIME_MS			
DISCOVERY_RESOURCE_META_SCAN_MAPPER_TASK_TIMEOUT_MS			
DISCOVERY_SCHEMA_MAP_BASE_PATH			
DISCOVERY_OFFLINE_SCAN_KAFKA_ENABLE			
DISCOVERY_ML_ENABLE			
DISCOVERY_SAS_SUFFIXES			
DISCOVERY_ENABLE_SIMPLE_KAFKA_CONSUMER_FOR_AUDIT_PARSING			
DISCOVERY_ENABLE_KAFKA_CONSUMER_FOR_MAPR_AUDIT_PARSING			
DISCOVERY_ENABLE_KAFKA_CONSUMER_FOR_AUDIT_PARSING			
DISCOVERY_ZIP_LOOKUP_KEY			
DISCOVERY_GENERIC_ML_TYPE			
DISCOVERY_CORE_NLP_ML_TYPE			
DISCOVERY_PHONE_NUMBER_ML_TYPE			
DISCOVERY_GEO_LAT_LONG_ML_TYPE			
DISCOVERY_DOB_ML_TYPE			
DISCOVERY_VIN_ML_TYPE			
DISCOVERY_ITIN_ML_TYPE			
DISCOVERY_EIN_ML_TYPE			
DISCOVERY_SSN_ML_TYPE			
DISCOVERY_IMEI_ML_TYPE			
DISCOVERY_CC_ML_TYPE			
DISCOVERY_ZIP_ML_TYPE			
DISCOVERY_LFS_WATCHER_POLLTIME_MS			
DISCOVERY_LFS_CREATE_MAX_TIME_MS			
DISCOVERY_LFS_WATCHER_CACHE_SIZE			
DISCOVERY_LFS_WATCHER_ENABLE			
DISCOVERY_LFS_APP_TOPIC			
DISCOVERY_LFS_APP			
DISCOVERY_GOOGLE_BIGQUERY_PARSE_CTAS			
DISCOVERY_DYNAMODB_ENABLE			
DISCOVERY_FUZZY_SCORING_SENSE_CHECK_ENABLE			
DISCOVERY_FUZZY_SCORING_MIN_CUTOFF_SCORE			
DISCOVERY_ML_SRC_DETECT_MODEL_PATH			
DISCOVERY_ML_MODEL_PATH			

Property	Description	Values	Default Value
DISCOVERY_ML_CLASSIFY_TAG_ACTION_ENABLE			
DISCOVERY_ML_CLASSIFY_SRC_CODE_ENABLE			
DISCOVERY_ML_CLASSIFY_TAG_ENABLE			
DISCOVERY_ML_STORE_SCAN_RESULTS			
DISCOVERY_OUTPUTWRITERS_ENABLE			
DISCOVERY_DATABRICKS_SPARK_ENABLE			
DISCOVERY_KAFKA_PRODUCER_COMPRESSION_CODEC			
DISCOVERY_SET_REMOTE_USER			
DISCOVERY_STALE_DATA_RETRY_COUNT			
DISCOVERY_AUDITS_TO_SOLR_ENABLED			
DISCOVERY_ATLAS_HOOK_SIMPLE			
DISCOVERY_ATLAS_HOOK_ENABLED			
DISCOVERY_SPLUNK_ENABLE			
DISCOVERY_SPLUNK_PORT			
DISCOVERY_SPLUNK_ALERT_INDEX			
DISCOVERY_SPLUNK_SCHEME			
DISCOVERY_SPLUNK_HEC_SOURCE			
DISCOVERY_ANOMALY_SCHEDULAR_ENABLE			
DISCOVERY_MONITORING_SCHEDULAR_ENABLE			
DISCOVERY_METRICS_JVM			
DISCOVERY_METRICS_KAFKA_TOPIC			
DISCOVERY_METRICS_KAFKA_INTERVAL_SEC			
DISCOVERY_METRICS_ENABLE_KAFKA			
DISCOVERY_METRICS_GRAPHITE_INTERVAL_SEC			
DISCOVERY_METRICS_GRAPHITE_ENABLE			
DISCOVERY_METRICS_CONSOLE_INTERVAL_SEC			
DISCOVERY_METRICS_ENABLE_CONSOLE			
DISCOVERY_METRICS_CSV_INTERVAL_SEC			
DISCOVERY_METRICS_ENABLE_CSV			
DISCOVERY_METRICS_CSVPATH			
DISCOVERY_SOLR_LOGS_COLLECTION			
DISCOVERY_SOLR_METRICS_COLLECTION			
DISCOVERY_DB_CPDS_TEST_ONCHECKIN			
DISCOVERY_DB_CPDS_TEST_ONCHECKOUT			
DISCOVERY_DB_CPDS_IDLECONN_TEST_PERIOD_SEC			
DISCOVERY_DB_CPDS_TESTQUERY			
DISCOVERY_COMMON_EXCLUDE_RESOURCE_LIST			
DISCOVERY_CSV_USE_HEADER			
DISCOVERY_SCAN_MARK_LIMIT_BYTES			
DISCOVERY_SCAN_MIN_CSV_FIELDS			

Property	Description	Values	Default Value
DISCOVERY_SCAN_HIVE_MAX_COLS	Maximum number of columns in a database table or fields in a structured file to be scanned. This can be overriden by using `record.max.fields` property at data source level.		2000
DISCOVERY_SCAN_HIVE_MAX_ROWS	Maximum number of rows of a database table to be scanned.		500
DISCOVERY_SCAN_MAX_LINES	Maximum number of records of a structured file to be scanned.		500
DISCOVERY_CONTENT_MAX_CHARACTER	Maximum number of bytes in a column cell or field cell to be scanned.		1000
DISCOVERY_TIKA_MAX_BYTES	Maximum number of bytes of an unstructured file to be scanned.		102400
DISCOVERY_MAX_TAG_SNIPPET_SAMPLE_VALUES	Maximum number of samples to be captured for display in a tag.		3
DISCOVERY_QUICK_COUNT_THRESHOLD			
DISCOVERY_KAFKA_CLASSIFIEDINFO_MAX_POLL_RECORDS			
DISCOVERY_KAFKA_CLASSIFIEDINFO_SESSION_TIMEOUT_MS			
DISCOVERY_KAFKA_CLASSIFIEDINFO_REQUEST_TIMEOUT_MS			
DISCOVERY_META_SCANNING_ENABLE			
DISCOVERY_OFFLINE_SCAN_SUMMARY_SOLR_ENABLE			
DISCOVERY_METRICS_SOLR_ENABLE			
DISCOVERY_NON_NULL_REPORT_OUTPUT_PATH			
DISCOVERY_CLASSIFICATION_NON_NULL_COUNT_ENABLE			
DISCOVERY_KAFKA_TOPIC_ENCRYPTION			
DISCOVERY_KAFKA_TOPIC_DISCOVERY			
DISCOVERY_KAFKA_DISCOVERY			
DISCOVERY_KAFKA_DISCOVERY_REQUEST_TIMEOUT_MS			
DISCOVERY_KAFKA_DISCOVERY_BOOSTRAP_SERVERS			
DISCOVERY_KAFKA_DISCOVERY_USE_SSL			
DISCOVERY_KAFKA_DISCOVERY_USE_KERBEROS			
DISCOVERY_KAFKA_DISCOVERY_NAME			
DISCOVERY_KAFKA_DISCOVERY_GROUP_ID			
DISCOVERY_KAFKA_DISCOVERY_POLL_TIME_MS			
DISCOVERY_KAFKA_DISCOVERY_ENABLE			
DISCOVERY_IS_ATLAS_TAG_ENABLE			
DISCOVERY_ATLAS_HOOK_VERSION			
DISCOVERY_SCAN_RESOURCE_META_INFO_SOLR			
DISCOVERY_IS_ATLAS_ENABLE			
DISCOVERY_SPARK_STREAMING_RECEIVER_MAXRATE			
DISCOVERY_SPARK_STREAMING_CHECKPOINT			
DISCOVERY_SPARK_ENABLE_HIVE_SUPPORT			

Property	Description	Values	Default Value
DISCOVERY_SPARK_LOCAL_MASTER			
DISCOVERY_SPARK_APPLICATION_NAME			
DISCOVERY_PORTAL_API_SCORE_THRESHOLD			
DISCOVERY_PORTAL_API_APP_LIST			
DISCOVERY_PORTAL_API_SYSTEM_LIST			
DISCOVERY_KERBEROS_PRINCIPAL			
DISCOVERY_KAFKA_ALERT_REPLICATION			
DISCOVERY_KAFKA_GROUP_ID			
DISCOVERY_GRAPHITE_HOST			
DISCOVERY_KAFKA_CLASSFICATION_INFO_REPLICATION			
DISCOVERY_MONITORING_HDFS_INPUT_PATH			
DISCOVERY_KERBEROS_KEYTAB			
DISCOVERY_SCAN_WORKER_KAFKA_GROUP_ID			
DISCOVERY_SOLR_ALERTS_COLLECTION			
DISCOVERY_SOLR_CLASSIFICATION_COLLECTION			
DISCOVERY_GRAPHITE_PORT			
DISCOVERY_HIVE_METASTORE_USEJDBC			
DISCOVERY_INIT_CONTAINER_COMMAND_LIST	You can provide a list of commands to download custom jars to a specified location inside the Discovery container. For example: `DISCOVERY_INIT_CONTAINER_COMMAND_LIST:-wget https://privacera/public/custom-1`		
DISCOVERY_SCAN_PARQUET_ORC_FROM_ARCHIVE_ENABLE	Property to enable/disable the scanning of ORC/Parquet files within a ZIP file.	true, false	false
DISCOVERY_SCAN_PARQUET_ORC_STREAM_FILE_SIZE_LIMIT	Property to set the file size limit in megabytes (MB) on the ORC/Parquet files being scanned from the archive location.		5242880
DISCOVERY_SCAN_PARQUET_TEMP_FILE_FROM_ARCHIVE_ENABLE	By default, Parquet files are stored in a temporary file within a zip file. Set to true to scan the Parquet files from a temporary file. Set to false to scan the Parquet files from a zip file stream.	true, false	true
DISCOVERY_SCAN_ORC_TEMP_FILE_FROM_ARCHIVE_ENABLE	By default, ORC files are stored in a temporary file within a zip file. Set to true to scan the ORC files from a temporary file. Set to false to scan the ORC files from a zip file stream.	true, false	false
DISCOVERY_GOOGLE_CLOUD_STORAGE_LINEAGE_LOOPBACK_TIME_MS	This property indicates time for GCS lineage loopback.	-	3000
DISCOVERY_GOOGLE_CLOUD_STORAGE_LINEAGE_CUTOFF_TIME_MS	This property indicates cut off time to wait for GCS log event for lineage.	-	300000
DISCOVERY_GOOGLE_CLOUD_STORAGE_LINEAGE_CUTOFF_TIME_CHECK_INTERVAL_MS	This property indicates fixed interval at which to check for delayed GCS lineage pending realtime file.	-	30000

Property	Description	Values	Default Value
`DISCOVERY_CONTENT_SCAN_THREAD_POOL_SIZE`	If you are scanning more than 2 datasource with different projects, then set this property as the number of projects you will be scanning in discovery.	-	2
`DISCOVERY_CONNECTION_TEST_INTERVAL_SEC`	The fixed interval in seconds at which all key Privacera internal components are checked. Status of the connection is sent to Portal. See Health Check	Allowable value is non-zero integer number of seconds. Recommended short duration and not to exceed 900 seconds (15 minutes).	60
`DISCOVERY_TELEMETRY_UPDATE_TO_SOLR`	Set to true to send telemetry to Apache Solr. Set to false to not send telemetry to the Apache Solr. The following telemetry is sent to Apache Solr: • Count of tags. • Count of resource scanned based on application and application type. • Scan amount based on application and application type. • Total compliance count and compliance count for individual policy.	true, false	true
`DISCOVERY_RTBF_SUMMARY_ENABLED`	Set this property to true to view the summary for RTP policy and Expunge policy on the UI for Auto Run jobs. Set this property to false to not view the summary. Although this property string contains "RTBF", the property relates to RTP.	true, false	false
`DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_ENABLED`	Whether to use dynamic resource allocation, which scales the number of executors registered with this application up and down based on the workload.	true, false	false
`DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_SHUFFLE_TRACKING_ENABLED`	Enables shuffle file tracking for executors, which allows dynamic allocation without the need for an external shuffle service. This option will try to keep alive executors that are storing shuffle data for active jobs.	true, false	true
`DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_EXECUTOR_IDLE_TIMEOUT`	If dynamic allocation is enabled and an executor has been idle for more than this duration, the executor will be removed.	-	60s
`DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_CACHED_EXECUTOR_IDLE_TIMEOUT`	If dynamic allocation is enabled and an executor which has cached data blocks has been idle for more than this duration, the executor will be removed.	-	120s

Property	Description	Values	Default Value
DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_MAX_EXECUTORS	Upper bound for the number of executors if dynamic allocation is enabled.	-	4
DISCOVERY_K8S_SPARK_MEMORY_OVERHEAD_FACTOR	This sets the Memory Overhead Factor that will allocate memory to non-JVM memory, which includes off-heap memory allocations, non-JVM tasks, and various systems processes.	-	0.1
DISCOVERY_HBASE_RETRY_ON_FAILURE_COUNT	Number of retries for Hbase connection.	-	2
DISCOVERY_HBASE_WAIT_BETWEEN_RETRY_MS	Wait time before retrying Hbase connection.	-	100 ms (milliseconds)
DISCOVERY_CONSUMER_ENABLE	Set this property to true if you want to start a separate consumer pod, which will be used for writing Classification and Scan Summary Data in Solr. Set this property to false if you do not require a separate consumer pod. **NOTE** This property is enabled only for AWS Kubernetes Spark.		
DISCOVERY_SPARK_JOB_MAX_TIME_MS	How long to wait (in milliseconds) before stopping a long running spark job.		14400000
DISCOVERY_K8S_SPARK_DYNAMIC_ALLOCATION_SHUFFLE_TRACKING_TIMEOUT	When enabled, shuffle tracking controls the timeout for executors that are holding shuffle data. The default value means that Spark will rely on the shuffles being garbage collected to be able to release executors. If garbage collection is slow to clean up shuffles, you can control when to time out executors, even when they are storing shuffle data.		300s
DISCOVERY_REALTIME_LINEAGE_ENABLED **NOTE** This property is supported only on Google Cloud Platform (GCP) for Google Cloud Storage (GCS) application.	Set to true to enable Discovery lineage tracking. Set to false to disable Discovery lineage tracking.	true, false	false

Property	Description	Values	Default Value
`DISCOVERY_FASTTRACK_REALTIME_ENABLE` **NOTE** This property is supported only on GCP for GCS application.	Set this property to true to enable fast track for real time scanning event. Set this property to false to disable fast track for real time scanning event. **NOTE** Add `DISCOVERY_FASTTRACK_REALTIME_ENABLE` ansible variable in any yml file in the following path: ~/privacera-manager/config/custom-vars/	true, false	false
`DISCOVERY_LINEAGE_PROCESSING_ENABLED` **NOTE** This property is supported only on GCP for GCS application.	Set this property to true to enable new lineage flow. Set this property to false to disable new lineage flow. **NOTE** Add `DISCOVERY_LINEAGE_PROCESSING_ENABLED` ansible variable in any yml file in the following path: ~/privacera-manager/config/custom-vars/	true, false	false
`DISCOVERY_DYNAMODB_TF_BILLING_MODE` **NOTE** This property is supported only on AWS	This property refers to `billing_mode` parameter of DynamoDB in AWS. The `billing_mode` parameter controls how you are charged for read and write throughout and how you manage capacity.	PAY_PER_REQUEST, PROVISIONED	PROVISIONED

Property	Description	Values	Default Value
`DISCOVERY_KINESIS_TF_STREAM_MODE` **NOTE** This property is supported only on AWS	This property refer to `stream_mode` parameter of Kinesis Stream in AWS. The `stream_mode` parameter controls how you are charged for read and write throughout and how you manage capacity.	ON_DEMAND, PROVISIONED	PROVISIONED
Memory Variables **NOTE** Memory variables are used only for Discovery on Kubernetes Spark.			
`DISCOVERY_DRIVER_HEAP_MIN_MEMORY_MB`	Minimum Java Heap memory in MB used by Discovery Driver. For example, DISCOVERY_DRIVER_HEAP_MIN_MEMORY_MB: "1024".		
`DISCOVERY_DRIVER_HEAP_MIN_MEMORY`	Minimum Java Heap memory used by Discovery Driver. Setting this value will override DISCOVERY_DRIVER_HEAP_MIN_MEMORY_MB. For example, DISCOVERY_DRIVER_HEAP_MIN_MEMORY: "1g".		
`DISCOVERY_DRIVER_HEAP_MAX_MEMORY_MB`	Maximum Java Heap memory in MB used by Discovery Driver. For example, DISCOVERY_DRIVER_HEAP_MAX_MEMORY_MB: "1024".		
`DISCOVERY_DRIVER_HEAP_MAX_MEMORY`	Maximum Java Heap memory used by Discovery Driver. Setting this value will override DISCOVERY_DRIVER_HEAP_MAX_MEMORY_MB. For example, DISCOVERY_DRIVER_HEAP_MAX_MEMORY: "1g".		
`DISCOVERY_DRIVER_K8S_MEM_REQUESTS_MB`	Minimum amount of Kubernetes memory in MB to be requested by Discovery Driver. For example, DISCOVERY_DRIVER_K8S_MEM_REQUESTS_MB: "1024".		
`DISCOVERY_DRIVER_K8S_MEM_REQUESTS`	Minimum amount of Kubernetes memory to be used by Discovery Driver. Setting this value will override DISCOVERY_DRIVER_K8S_MEM_REQUESTS_MB. For example, DISCOVERY_DRIVER_K8S_MEM_REQUESTS: "1G".		
`DISCOVERY_DRIVER_K8S_MEM_LIMITS_MB`	Maximum amount of Kubernetes memory to be requested by Discovery Driver. The value set in in this field will be considered as megabytes. For example, DISCOVERY_DRIVER_K8S_MEM_LIMITS_MB: "1024".		

Property	Description	Values	Default Value
DISCOVERY_DRIVER_K8S_MEM_LIMITS	Maximum amount of Kubernetes memory to be used by Discovery Driver. Setting this value will override DISCOVERY_DRIVER_K8S_MEM_LIMITS_MB. For example, DISCOVERY_DRIVER_K8S_MEM_LIMITS: "1G".		
DISCOVERY_DRIVER_CPU_MIN	Minimum amount of Kubernetes CPU to be requested by Discovery Driver. For example, DISCOVERY_DRIVER_CPU_MIN: "0.5".		
DISCOVERY_DRIVER_CPU_MAX	Maximum amount of Kubernetes CPU to be used by Discovery Driver. For example, DISCOVERY_DRIVER_CPU_MAX: "0.5".		
DISCOVERY_EXECUTOR_HEAP_MIN_MEMORY_MB	Minimum Java Heap memory in MB used by Discovery Executor. For example, DISCOVERY_EXECUTOR_HEAP_MIN_MEMORY_MB: "1024".		
DISCOVERY_EXECUTOR_HEAP_MIN_MEMORY	Minimum Java Heap memory used by Discovery Executor. Setting this value will override DISCOVERY_EXECUTOR_HEAP_MIN_MEMORY_MB. For example, DISCOVERY_EXECUTOR_HEAP_MIN_MEMORY: "1g".		
DISCOVERY_EXECUTOR_HEAP_MAX_MEMORY_MB	Maximum Java Heap memory in MB used by Discovery Executor. For example, DISCOVERY_EXECUTOR_HEAP_MAX_MEMORY_MB: "1024".		
DISCOVERY_EXECUTOR_HEAP_MAX_MEMORY	Maximum Java Heap memory used by Discovery Executor. Setting this value will override DISCOVERY_EXECUTOR_HEAP_MAX_MEMORY_MB. For example, DISCOVERY_EXECUTOR_HEAP_MAX_MEMORY: "1g".		
DISCOVERY_EXECUTOR_K8S_MEM_REQUESTS_MB	Minimum amount of Kubernetes memory in MB to be requested by Discovery Executor. For example, DISCOVERY_EXECUTOR_K8S_MEM_REQUESTS_MB: "1024".		
DISCOVERY_EXECUTOR_K8S_MEM_REQUESTS	Minimum amount of Kubernetes memory to be used by Discovery Executor. Setting this value will override DISCOVERY_EXECUTOR_K8S_MEM_REQUESTS_MB. For example, DISCOVERY_EXECUTOR_K8S_MEM_REQUESTS: "1G".		
DISCOVERY_EXECUTOR_K8S_MEM_LIMITS_MB	Maximum amount of Kubernetes memory in MB to be requested by Discovery Executor. For example, DISCOVERY_EXECUTOR_K8S_MEM_LIMITS_MB: "1024".		
DISCOVERY_EXECUTOR_K8S_MEM_LIMITS	Maximum amount of Kubernetes memory to be used by Discovery Executor. Setting this value will override DISCOVERY_EXECUTOR_K8S_MEM_LIMITS_MB. For example, DISCOVERY_EXECUTOR_K8S_MEM_LIMITS: "1G".		

Property	Description	Values	Default Value
DISCOVERY_EXECUTOR_CPU_MIN	Minimum amount of Kubernetes CPU to be requested by Discovery Executor. For example, DISCOVERY_EXECUTOR_CPU_MIN: "0.5".		
DISCOVERY_EXECUTOR_CPU_MAX	Maximum amount of Kubernetes CPU to be used by Discovery Executor. For example, DISCOVERY_EXECUTOR_CPU_MAX: "0.5".		
DISCOVERY_DRIVER_K8S_CPU_LIMITS	Maximum amount of Kubernetes CPU to be used by Discovery Driver. For example, DISCOVERY_DRIVER_K8S_CPU_LIMITS: "0.5".	true, false	false
DISCOVERY_DRIVER_K8S_CPU_REQUESTS	Minimum amount of Kubernetes CPU to be requested by Discovery Driver. For example, DISCOVERY_DRIVER_K8S_CPU_REQUESTS: "0.5".		
DISCOVERY_EXECUTOR_K8S_CPU_LIMITS	Maximum amount of Kubernetes CPU to be used by Discovery Executor. For example, DISCOVERY_EXECUTOR_K8S_CPU_LIMITS: “0.5”.		
DISCOVERY_EXECUTOR_K8S_CPU_REQUESTS	Minimum amount of Kubernetes memory to be used by Discovery Executor. For example, DISCOVERY_EXECUTOR_K8S_CPU_REQUESTS: “0.5”.		
DISCOVERY_CONSUMER_K8S_MEM_LIMITS	Maximum amount of Kubernetes memory to be used by Discovery Consumer. For example, DISCOVERY_CONSUMER_K8S_MEM_LIMITS: "1G".		
DISCOVERY_CONSUMER_K8S_MEM_REQUESTS	Minimum amount of Kubernetes memory to be used by Discovery Consumer. For example, DISCOVERY_CONSUMER_K8S_MEM_REQUESTS: "1G".		
DISCOVERY_CONSUMER_K8S_CPU_LIMITS	Maximum amount of Kubernetes CPU to be used by Discovery Consumer. For example, DISCOVERY_CONSUMER_K8S_CPU_LIMITS: "0.5".		
DISCOVERY_CONSUMER_K8S_CPU_REQUESTS	Minimum amount of Kubernetes CPU to be requested by Discovery Consumer. For example, DISCOVERY_CONSUMER_K8S_CPU_REQUESTS: "0.5".		
DISCOVERY_SKIP_JDBC_BINARY_COLUMN_LIST	During database scanning, binary columns data scanning is skipped and only non-binary columns data is scanned.		BINARY, VARBINARY, MEDIUMBLOB, LONGBLOB, IMAGE, BLOB, BFILE, RAW, LONG RAW, ROWID, UROWID, BYTEA, VARBYTE
DISCOVERY_ORC_ROW_BATCH_SIZE_LIMIT_IN_BYTES	When configuring MAX_LINES, the batch size should be smaller than the default value. This will decrease memory consumption and lower the chances for Out of Memory error.		1024 rows
DISCOVERY_AWS_S3_TAG_SYNC_ENABLE	Set this property to true if you want to enable AWS S3 tag sync. Set this property to false if you want to disable AWS S3 tag sync.	true, false	false

Enabling Multithreading for Different Consumers

For enabling multithreading for different consumers in the Discovery driver pod or Discovery consumer pod, refer to Configure system properties and follow these steps:

NOTE

This feature is supported only for AWS and Azure Kubernetes Spark.

1. For the Discovery driver, create the property file `discovery-custom.properties`.
2. For the Discovery consumer, create the property file `discovery-consumer-custom.properties`.
3. Add all of the following properties in both of the above files.

   NOTE

   The values in the following properties are recommended values.

AWS Properties

```
#privacera_offline_scan_topic privacera.discovery.cloud.consumer.config.offline.scan.sum
privacera.discovery.cloud.consumer.config.offline.scan.max.poll.records=1

#this is the timeout for offline scan job for each batch file
privacera.discovery.cloud.consumer.config.offline.scan.summary.task.timeout.ms=172800000
privacera.discovery.cloud.consumer.config.offline.scan.task.timeout.ms=172800000

#privacera_scan_resource_info_topic privacera.discovery.cloud.consumer.config.ow.solr.sc
privacera.discovery.cloud.consumer.config.ow.solr.resource.max.poll.records=10000
privacera.discovery.cloud.consumer.config.ow.solr.scan.resource.meta.max.poll.records=10

privacera.discovery.cloud.consumer.config.ow.solr.scan.resource.info.task.timeout.ms=172
privacera.discovery.cloud.consumer.config.ow.solr.resource.task.timeout.ms=172800000
privacera.discovery.cloud.consumer.config.ow.solr.scan.resource.meta.task.timeout.ms=172

privacera.discovery.cloud.consumer.config.ow.solr.scan.resource.info.parallel.size=50
privacera.discovery.cloud.consumer.config.ow.solr.scan.resource.meta.parallel.size=50
privacera.discovery.cloud.consumer.config.ow.solr.resource.parallel.size=50

#privacera_classification_topic privacera.discovery.cloud.consumer.config.ow.solr.classi
privacera.discovery.cloud.consumer.config.ow.resource.workflow.max.poll.records=10000
privacera.discovery.cloud.consumer.ow.ranger.rest.classifications.max.poll.records=10000

privacera.discovery.cloud.consumer.config.ow.solr.classifications.task.timeout.ms=864000
privacera.discovery.cloud.consumer.config.ow.resource.workflow.task.timeout.ms=86400000
privacera.discovery.cloud.consumer.ow.ranger.rest.classifications.task.timeout.ms=864000

privacera.discovery.cloud.consumer.config.ow.solr.classifications.parallel.size=50
privacera.discovery.cloud.consumer.ow.ranger.rest.classifications.parallel.size=50
```

Azure Properties

```
OFFLINE SCAN
privacera.discovery.kafka.consumer.config.offline.scan.max.poll.records=1
privacera.discovery.kafka.consumer.config.offline.scan.task.timeout.ms=432000000

OFFLINE SCAN SUMMARY
privacera.discovery.kafka.consumer.config.offline.scan.summary.max.poll.records=1
privacera.discovery.kafka.consumer.config.offline.scan.summary.task.timeout.ms=432000000

OUTPUT WRITER - SOLR CLASSIFICATION
privacera.discovery.kafka.consumer.config.ow.solr.classifications.max.poll.records=10000
privacera.discovery.kafka.consumer.config.ow.solr.classifications.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.solr.classifications.task.timeout.ms=864000

OUTPUT WRITER - SOLR SCAN RESOURCE INFO
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.info.max.poll.records=10
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.info.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.info.task.timeout.ms=432

OUTPUT WRITER - SOLR SCAN RESOURCE META
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.meta.max.poll.records=10
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.meta.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.solr.scan.resource.meta.task.timeout.ms=432

OUTPUT WRITER - SOLR SCAN RESOURCE
privacera.discovery.kafka.consumer.config.ow.solr.resource.max.poll.records=10000
privacera.discovery.kafka.consumer.config.ow.solr.resource.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.solr.resource.task.timeout.ms=432000000

OUTPUT WRITER - SOLR LINEAGE
privacera.discovery.kafka.consumer.config.ow.solr.lineage.max.poll.records=10000
privacera.discovery.kafka.consumer.config.ow.solr.lineage.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.solr.lineage.task.timeout.ms=86400000

OUTPUT WRITER - RESOURCE WORKFLOW
privacera.discovery.kafka.consumer.config.ow.resource.workflow.max.poll.records=10000
privacera.discovery.kafka.consumer.config.ow.resource.workflow.parallel.size=50
privacera.discovery.kafka.consumer.config.ow.resource.workflow.task.timeout.ms=86400000

OUTPUT WRITER - RANGER REST CLASSIFICATION
privacera.discovery.kafka.consumer.config.ow.ranger.rest.classifications.max.poll.record
privacera.discovery.kafka.consumer.config.ow.ranger.rest.classifications.parallel.size=5
privacera.discovery.kafka.consumer.config.ow.ranger.rest.classifications.task.timeout.ms
```