opentext™

OpenText Core
Data Discovery & Risk Insights

Application Version 24.4.0

Processing Agent Version 24.4.100

Frequently Asked Questions

Document Release Date: October 2024
Software Release Date: October 2024

Legal notices

Contents

Get started

This document addresses frequently asked questions about OpenText Core Data Discovery & Risk Insights.

About OpenText Core Data Discovery & Risk Insights

OpenText Core Data Discovery & Risk Insights lets you find, protect, and secure sensitive and high-value data within on-premises and cloud data platforms across your enterprise. Identify, collect, and organize content to ensure discovery of sensitive data. Configure how structured and unstructured sources and datasets are processed and categorized with Connect. Analyze your data under management with Analyze. Organize, review, and take action on documents and unstructured data items with Manage.

24.4.0.20241016

General

The following FAQs address questions about OpenText Core Data Discovery & Risk Insights in general.

What ports are used by the application?

Ports	Location	Usage
7957	available on agent host; does not need to be open for communication outside this box	processing agent
7432	available on agent host; does not need to be open for communication outside this box	PostgreSQL; will bind to the local loopback addresses of 127.0.0.1 (IPv4) and ::1 (IPv6)
7500, 7502	available on agent host; does not need to be open for communication outside this box	CFS (Connector Framework Server)
9312	network needs to be open from the agent host to the OpenText Core Data Discovery & Risk Insights service (gateway).	gateway for the processing agent
9025, 9310, 9320, 9390, 9400, 9420	open on client machines of all users accessing the OpenText Core Data Discovery & Risk Insights user interfaces	browser access to OpenText Core Data Discovery & Risk Insights user interfaces (Administration, Analyze, Connect, Manage) and Help Centers

The platform uses additional ports on the various server role hosts. For detailed information, see the OpenText Core Data Discovery & Risk Insights Implementation Help Center.

Security

The following FAQs address security questions.

How is the OpenText Core Data Discovery & Risk Insights environment secured in AWS (Amazon Web Services)?

Back office implementations follow industry standard security practices. including but not limited to:

- **All** incoming communication (from web users, on-premises agents, FTP clients, and so on) is transmitted exclusively via TLS 1.2+ using only high strength cipher suites (Encryption in Transit).
- Object and volume storage containing non-ephemeral data is encrypted using the industry standard AES-256 algorithm (Encryption at Rest).
- Industry standard Principle of least privilege (PoLP) is consistently applied. This is applied within the application, as well as within the back office (governing access to infrastructure resources, limiting intra-back office communications, and so on).

Consult the available Service Description documentation for more information.

OpenText Core Data Discovery & Risk Insights is multi tenanted. What protects my data from being seen by others?

Individual tenant data is stored in separate indexes and object storage locations.

Are the indexes backed up, and if so, for how long are the indexes kept?

Yes, the indexes are backed up. Consult the available Service Description documentation for more information.

What metadata is copied to the cloud ?

When objects are captured by the processing agent, the extracted content and relevant metadata is transmitted to the back office for further enrichment. The end results are then held in index storage within the back office.

Optionally, you may elect to also collect data (either by choice or to enforce a hold). In this case, a copy of the original data object is then also transmitted to the back office, which will be held in object storage within the back office.

What security is in place for transferring data from cloud data types? From agent data types?

OpenText Core Data Discovery & Risk Insights processing agents performing data capture and collection follow the same security procedures, regardless of the data type.

1. During configuration, the tenant provides information about the repository type, path to the dataset, and access credentials. This information makes up a dataset's definition, which is

used to reach the data type. Dataset definitions are encrypted and held securely within the back office.

2. An agent system connects to the back office and retrieves any pending tasks that have been delegated to it.
3. The dataset definition is provided to the authorized agent for use only when performing the specific task.
4. Data captured/collected is then transferred by the agent to the back office.

This applies to both individual private tenant on-premises agents and datasets assigned to be managed by cloud to cloud (C2C) agents operated from within the back office.

On-premises agents can operate on datasets that may only be available to the particular private customer system (such as, file system, private Exchange) or public locations (such as, SharePoint Online, Office 365) if the customer permits access to those.

C2C agents are restricted to exclusively operate on those data types that can be reached through public locations (such as, SharePoint Online, Office 365).

File access

The following FAQs address questions about accessing files under management.

What are the file path restrictions for file system sources and datasets?

Files whose full accessed share path exceeds 255 characters will not be accessible.

For file system **sources**, the following limitations exist.

- The source path cannot be more than a single directory beyond the host. For example, `\\server01.domain.com\folderA`. Further path refinement is defined by datasets.
- The hostname portion of the source path can contain only the following characters.
  - upper and lowercase alpha-numeric characters
  - `.` (period)
  - `-` (dash)
  - `_` (underscore)
- The source path cannot contain any of the following special characters.
  - `<` (less than)
  - `>` (greater than)
  - `:` (colon)
  - `"` (double quote)
  - `|` (vertical bar or pipe)
  - `?` (question mark)
  - `*` (asterisk)
  - `/` (forward slash)
- Files whose full accessed share path exceeds 255 characters will not be accessible.
- The path cannot contain `.` or `..` before, after, or in between slashes (`\`) with no other characters and cannot end with `.` .
  - Not valid:

    `\\company.domain.com\..` `\\company.domain.com\.`

  - Valid:

    `\\company.domain.com\ab..c` `\\company.domain.com\.abc`

For file system **datasets**, the following limitations exist.

- The sub-directory path cannot contain any of the following special characters.
  - < (less than)
  - > (greater than)
  - : (colon)
  - " (double quote)
  - | (vertical bar or pipe)
  - ? (question mark)
  - * (asterisk) / (slash)
- Files whose full accessed share path exceeds 255 characters will not be accessible.
- The path cannot contain **.** or **..** before, after, or in between slashes (\) with no other characters and cannot end with **.** .
  - Not valid:

    ```
    \..\        \.\abc        \abc\..        \abc\.\def
    ```

  - Valid:

    ```
    \ab..c        \.abc        \abc\d.e.f        \abc\.def\gh
    ```

What are the file path restrictions for file system targets and destinations?

Avoid hidden or system level CIFS share (such as, `\\server01\c$\folderA`).

Files whose full accessed share target and destination path exceeds 255 characters will not be accessible on the destination.

Processing and processing agent

The following FAQs address questions about file processing and the processing agent.

The task status on the Agent Activity page in Connect is listed as "waiting". What does this mean?

Some tasks require multiple processes, or steps, to complete and the task requested is in between steps and waiting to be picked up for the next step. For example, you want to send the items in a workbook to a target. For the items in this workbook, only the metadata was indexed and the source and destination are managed by different agent clusters. In this scenario, the items must be collected before they can be sent to the defined target. This task may show a "waiting" task status after the items are collected as the task waits to be picked up to send the items to the target.

The assigned agent may not be reachable. If the task status remains "waiting", ensure that the agents in the agent cluster assigned to the task are running and accessible. Specifically, verify that the agentAPI service is running on the agent host assigned to perform the task.

How are deleted items tracked?

OpenText Core Data Discovery & Risk Insights tracks file deletions when a processing job runs against a dataset. A job run occurs when a dataset is updated, either run on a schedule or manually updated from the Manage Datasets page in Connect (click the inline update icon for the dataset or the Update button in the dataset detail panel).

File systems

OpenText Core Data Discovery & Risk Insights tracks file deletions by directly comparing with the original file system location identified by the dataset path. Items are removed from the application index seven days after the deletion fro the source location is detected. If an item within a container file (such as ZIP) is deleted in the original file system location, the item is removed from the index as part of updating the container file when the job run occurs. In this case, the item may be removed from the index sooner than seven days after deletion is detected.

Exchange

No deletion detection from Exchange. OpenText Core Data Discovery & Risk Insights retains items it has already processed until a delete action is initiated from the application.

SharePoint
(missing or bad snippet)
Content Manager

OpenText Core Data Discovery & Risk Insights tracks the deletion of managed Content Manager items using the Content Manager delete events. Each time processing is run on a dataset—on a schedule, or on demand—the application checks the delete events. For each managed item that is deleted in Content Manager, the application deletes that item from the index. If an item within a container file (such as ZIP) is deleted from Content Manager, the item is removed from the index as part of updating the container file when the job run occurs.

To ensure accurate tracking of items deleted from Content Manager, ensure that the Content Manager datasets in OpenText Core Data Discovery & Risk Insights are updated more often than Content Manager administrator purges delete events. For example, if your Content Manager administrator purges delete events every 60 days, verify that your Content Manager datasets are updated at least every 59 days.

Google Drive

OpenText Core Data Discovery & Risk Insights tracks the deletion of managed Google Drive items using the change log for the Google drive defined by the source in Connect. Each time processing is run on a dataset—on a schedule, or on demand—the application checks the change logs for deleted items. For each managed item that is deleted in Google Drive, the application deletes that item from the application index. If an item within a container file (such as ZIP) is deleted in Google Drive, the item is removed from the index as part of updating the container file when the job run occurs.

To ensure accurate tracking of items deleted from Google Drive, ensure that the Google Drive datasets in Connect are updated more often than the maximum number of days Google Drive change logs are kept. For example, the default retention for change logs is 30 days. Verify that your Google Drive datasets are updated at least every 29 days.

How are duplicate items identified?

As the processing agent reads each file, it generates a SHA384 (Secure Hash Algorithm) checksum of the file. This is a hash function which takes an input, in this case the file, and produces an item hash value that is stored in the index. This means that OpenText Core Data Discovery & Risk Insights generates a fingerprint that can identify a file, excluding the name of the file.

Using the function of the deduplication task workbook, you define a dataset to represent official records to compare against, or you define rules to identify master items to compare against. The identified duplicate items (based on the metadata and content associated with the hash) and all family members of those items (such as attachments or parent item) are added to the workbook.
