Cloudera Octopai Data Lineage 1.0.0

Getting Started

Date published: 2025-10-09
Date modified: 2025-10-20

CLOUDERA

https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2026. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

What is Cloudera Octopai Data Lineage?

Cloudera Octopai Data Lineage is a comprehensive active metadata management platform designed to navigate complex data landscapes. This platform automates data lineage, discovery, and cataloging, providing users with an extensive understanding of their data ecosystem.

Cloudera Octopai Data Lineage covers the entire data ecosystem, including on-premises and cloud-based systems, and it integrates seamlessly with various ETLs, databases, analysis, and reporting tools.

Metadata, often described as data about data, is information that defines various aspects of data, helping data teams to understand and use data effectively. Active metadata management is a dynamic process that keeps metadata up to date and readily accessible for various data operations.

The following key capabilities are valid for Cloudera Octopai:

1. **Automated Data Lineage** – Cloudera Octopai provides detailed and comprehensive data lineage, showing how data moves and transforms across various systems. This cross-system, intra-system, and end-to-end column lineage helps data teams understand the entire journey of their data.
2. **Automated Discovery** – Cloudera Octopai uses machine learning to analyze metadata from data systems, enabling automatic discovery of data sources and their relationships.
3. **Knowledge Hub** – Cloudera Octopai creates an automatic catalog that is continuously updated, providing an organized view of all data assets.
4. **Ease of Setup and Use** – The platform can be set up in less than 24 hours, without the need for professional services. The setup process involves downloading the Cloudera Octopai client, extracting the metadata, and uploading the metadata files to a secured environment. The platform is not directly connected to the user environment, ensuring data security.
5. **Automatic Metadata Collection** – Cloudera Octopai automates the metadata collection process. The platform runs according to a schedule set by the user, creating metadata files, encrypting and uploading them to a secure vault. The platform continuously checks for new information and updates the metadata repository accordingly.
6. **Scalability and Flexibility** – Cloudera Octopai is designed to be scalable, with an elastic pricing model based on the number of source systems. Cloudera Octopai is also flexible, allowing organizations to focus on interpreting the data rather than maintaining it.

The value of Cloudera Octopai lies in its ability to provide visibility and trust in complex data environments. Cloudera Octopai can drastically reduce the time spent on tracking, finding, and understanding data, which can otherwise consume a significant portion of data teams' time. By automating these tasks, Cloudera Octopai allows data teams to manage their data more quickly, easily, and accurately, and to focus more on deriving insights from the data.

Data lineage

Data lineage involves tracing and visualizing the lifecycle of data within a single system or across multiple systems, providing insights into its origin, transformations, and usage.

Cross-systems lineage

Cross-systems lineage is the process of tracing and visualizing the data lifecycle across multiple systems and platforms within an organization. It allows you to see where data originates, how it changes and is used, and where it ultimately ends up.

Understanding cross-systems lineage can provide the following benefits:

- **Improved Decision Making** – By providing a clear view of data sources, transformations, and usage, decision-makers can have increased confidence in their data-driven insights. It validates that the data used in analysis and decision-making processes is accurate, trustworthy, and reliable.
- **Risk Management and Compliance** – For regulated industries, understanding data lineage can be crucial for compliance. Cross-systems lineage can demonstrate to regulators that data has been handled correctly.

Furthermore, it can help manage risk by identifying where sensitive data resides and ensuring appropriate security measures are in place.

- **Data Quality** – Cross-systems lineage helps identify data quality issues. By tracking data from its source, through transformations, and to its endpoint, inconsistencies, errors, or anomalies can be traced back to their origin for resolution.
- **System Migration or Consolidation** – When merging systems or migrating data from one system to another, understanding the lineage can help identify potential issues, dependencies, or impacts to downstream systems or processes.
- **Operational Efficiency** – Understanding cross-systems lineage can increase operational efficiency by eliminating redundant processes and identifying areas for automation or optimization.

Tip: Remember, the goal is to create a complete, end-to-end picture of your data's journey through the organization systems. By doing so, you can ensure data trustworthiness and integrity, improve compliance, and enhance operational efficiency.

Inner-system lineage

Inner-system lineage is the process of tracking and visualizing data as it moves and transforms within a single system or platform. It provides a detailed understanding of data origin, transformations, and usage, but within the boundaries of one system.

In contrast to cross-systems lineage, which is about understanding data across multiple systems, inner-systems lineage is more focused on a single system's data journey. Both are essential components of comprehensive data governance, but their use cases differ. While inner-systems lineage is ideal for system-specific data quality, efficiency, and security considerations, cross-systems lineage is beneficial for broader, organization-wide views of data flow, particularly in understanding dependencies and impacts across systems.

Inner-systems lineage has the following benefits and implementations:

- **Understanding Data Flow** – Inner-systems lineage provides a clear understanding of how data is created, transformed, and consumed within a specific system. This is particularly useful in complex environments where data undergoes numerous transformations or is used by multiple applications within the system.
- **Improving Data Quality** – If a data quality issue arises, inner-systems lineage allows you to trace the problem back to its source within the system. This can be instrumental in correcting data errors and improving overall data quality.
- **Streamlining System-Specific Processes** – By mapping out the data journey within a system, organizations can identify inefficiencies or bottlenecks in their processes. This can lead to better system-specific performance and efficiency.
- **Safeguarding Sensitive Data** – Within a system, sensitive data might be transformed or moved. Understanding the lineage of this data helps ensure that it is handled appropriately within that system, mitigating potential security risks.
- **System Enhancements and Migrations** – When updating system features or migrating to a new version, understanding the data lineage can help identify potential impacts or dependencies.

End-to-end column lineage (E2E)

End-to-end column lineage involves tracking the lifecycle of a specific data column or attribute from its origin, through all transformations, to its final form. This type of lineage gives a granular view of data handling and movement in your organization. It helps you understand how a specific data element changes, the dependencies it has, and the impact it might create throughout its lifecycle.

End-to-end column lineage provides the following values:

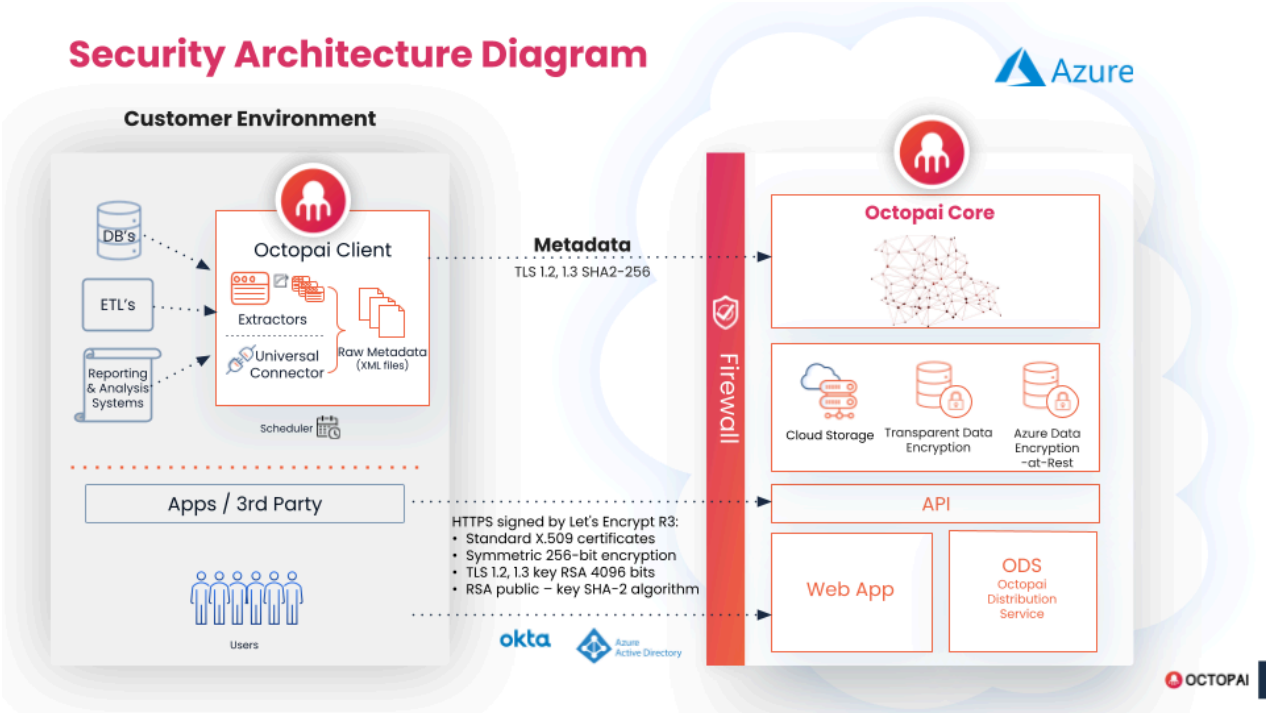
- **Data Provenance** – It helps understand the complete history of a data element. This includes the source system, any transformations or processing it has undergone, and where it is used in downstream systems and reports.
- **Data Quality Assurance** – If a data quality issue is identified in a column, tracing its lineage can help find the source of the issue. This might include identifying transformation errors, incorrect data mappings, or source system issues.

- **Change Impact Analysis** – If a change is planned in the source system or a transformation process, tracing the column lineage helps identify all the downstream systems, processes, or reports that might be impacted. This can help mitigate risks associated with system changes.
- **Regulatory Compliance** – In regulated industries, it is often necessary to demonstrate where specific data comes from and how it is transformed. Detailed column lineage can provide this information for audit or compliance purposes.
- **Data Trust** – For end users, understanding the lineage of a data column can increase trust in the data. If users can see where the data comes from and how it is handled, they might have more confidence in using it for decision-making.

Cloudera Octopai Data Lineage security architecture

The Cloudera Octopai Data Lineage security architecture ensures secure metadata extraction, storage, and transmission using advanced encryption methods.

Figure 1: Cloudera Octopai security architecture



The Cloudera Octopai security architecture consists of the following components:

1. **Metadata extractions**

   Cloudera Octopai sends to the customer a client called the Extractor to be installed on its private VM or Local Server. The configuration of this Client creates a Batch file for each source system, such as SQL server or SSIS, which is scheduled to run automatically on a regular basis using an automation scheduler, for example Control-M, UC-4, or Job of SQL server. The Extractor creates a readable metadata file in XML format for each source system.

2. **Encrypted Customer Portal**

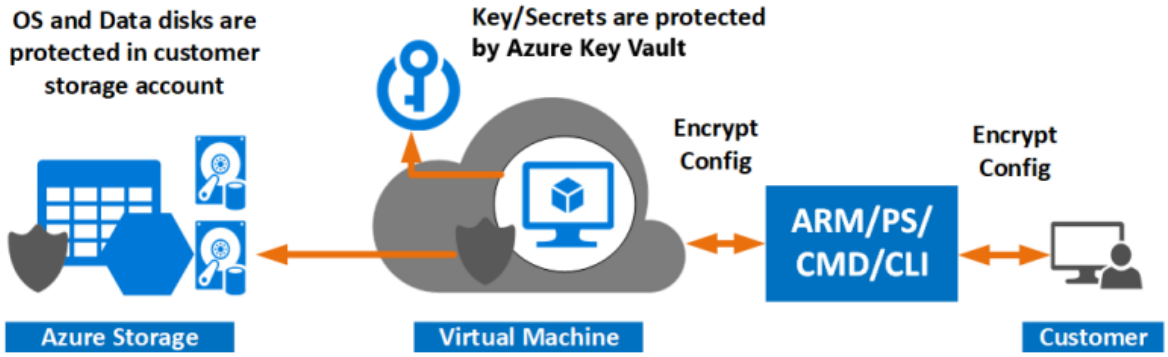
   The readable metadata files are uploaded to the secure Customer Portal. The Customer Portal is managed by Cloudera Octopai. Cloudera Octopai is triggered by the Customer Portal when new metadata files arrive and they are uploaded to the dedicated Azure environment of the customer. After the upload, the metadata files are deleted from the Customer Portal and the Azure VM Server.

3. **Azure VM Server**
   - Azure Server

     A dedicated Environment is created for the customer in the Cloudera Octopai tenancy on the Azure Cloud Services. The Cloudera Octopai Environment uses the Azure Storage account, the Managed Disk by Microsoft, and includes a dedicated database.
   - Encryption

     All the volumes on Azure are encrypted with the Azure Data Encryption-at-Rest (Encryption Key owned by Microsoft). The SQL Server DB is encrypted through the Transparent Data Encryption (Microsoft TDE). All the disks on Windows are encrypted according to FIPS 140-2.

     Figure 2: Azure Data Encryption

     The Cloudera Octopai application stores and protects the metadata using the Azure Disk Encryption, which uses the Windows BitLocker technology and Linux DM-Crypt to protect both operating system disks and data disks with full volume encryption. Encryption keys and secrets are protected and managed in the Azure Key Vault.
   - Customer's Users

     Secure login using the Azure AD or Okta Connect. Connecting to the customer Active Directory using the B2B collaboration is optional. No segregation exists of the metadata seen by users that share the same customer instance in Cloudera Octopai, however they can be differentiated through separate Cloudera Octopai instances for the customer.
4. **Secure data-in-transit connection**

   Data-in-transit is encrypted with HTTPS signed by the following DigiCert:
   - Standard X.509 certificates
   - Symmetric 256-bit encryption
   - TLS 1.2 Key RSA 4096 bits
   - RSA public-key SHA-2 algorithm

Use case: Applying data lineage for data quality

Learn about the power of comprehensive data lineage with the ability to trace data at column level, within a system, and across systems.

With the help of comprehensive data lineage an organization can quickly identify and resolve data issues, improving overall data quality and trust for effective Data Change Management.

A Data Analyst named Alex in a large financial institution faces a data quality issue.

Alex is tasked with creating a Report to analyze customer transactions using a column named transaction_amount from a central data warehouse.

As he starts the planned change investigation, Alex notices some irregularities in the data. Negative values occur in the transaction_amount column, which is incorrect for this context. Alex begins investigating using the following different layers of data lineage:

1. **End-to-end column lineage** – Alex starts by investigating the end-to-end lineage of the transaction_amount column. He sees various transformations applied to this data point, and where the data point is used in downstream reports. He discovers that the column is derived from the transaction_type (credit or debit) and transaction_value columns in a source system. A transformation is applied to convert debit transaction values to negative.
2. **Inner-system lineage** – Alex then looks at the inner-system lineage within the source system. He notices that the transaction_type column is derived from several fields, including the transaction_code field. A particular transaction code identifies debit transactions, and an error in mapping this code might cause the issue.
3. **Cross-system lineage** – Finally, Alex uses the cross-system lineage to find all systems feeding into the transact ion_type column. He discovers an upstream system where the transaction_code field originates. A recent system update changed the transaction_code values for debit transactions.

Armed with this information, Alex collaborates with the data engineering team to correct the error in mapping the new transaction_codefield and ensures that the transformation logic applied to the transaction_amount column is accurate. As a result, the data quality issue is resolved, and Alex can confidently proceed with his report, trusting the data he is using.

Use case: Applying data lineage for tracing a data propagation error

Learn about how a small error can cause a large-scale impact in complex, interconnected systems. Comprehensive data lineage can act as both a detective and a roadmap for resolving such issues.

Scenario overview

Consider a hypothetical large healthcare organization.

The organization operates a core Electronic Health Records (EHR) system that stores critical patient information. This system feeds multiple downstream platforms, including billing, patient portals, insurance claims, and other operational tools.

During a routine upgrade, a coding error introduces a transformation issue in the EHR system. The patient_ID column, which uniquely identifies patients, is incorrectly mapped to the caregiver_ID column that identifies healthcare providers.

The domino effect

As the EHR system feeds data to downstream systems, the mapping error has the following impacts:

1. In the billing system, patients are billed for services provided to their healthcare providers.
2. The patient portals show the healthcare providers details instead of patient-specific information, leading to privacy breaches.
3. Insurance claims get denied due to incorrect patient information.
4. Patients, healthcare providers, insurance companies, and even the healthcare organization customer service and IT teams are heavily impacted.

Navigating the problem with data lineage

In this situation, data lineage plays a crucial role in both identifying and correcting the issue using the following different layers of data lineage:

- **Cross-System Lineage** – When the first issues that are incorrect billing and privacy breaches arise, the organization's data governance team uses cross-system lineage to trace the patient_ID column data across all systems. They identify that the error originates in the EHR system, which feeds most downstream systems.
- **Inner-System Lineage** – Using inner-system lineage within the EHR system, they realize that the patient_ID column is incorrectly mapped to the caregiver_ID column during a transformation process.

- **End-to-End Column Lineage** – To assess the full impact, the team looks at the end-to-end column lineage of the patient_ID column. They map out all the processes, systems, and reports that use this column. This information is vital for communicating with affected parties and directing corrective measures.

The IT team corrects the erroneous transformation in the EHR system and initiates a massive cleanup operation in all impacted downstream systems.

Prevention using an integrated platform

Having an integrated Cloudera Octopai Data Lineage platform that encompasses data discovery, lineage, and data catalog can significantly enhance the ability to prevent such errors by gaining greater visibility into data flows and transformations. This visibility improves detection and prevention of potential issues, enhancing overall data quality and reliability.

Different roles use the integrated platform to achieve the improvements by performing the following actions:

- **Data Engineers** trace the data lineage during system development and maintenance using the catalog as a reference to check transformation logic and mapping to avoid incorrect data flow. They follow robust testing protocols before and after deploying any changes in data processing or transformation logic.
- **Data Stewards** perform metadata management by ensuring that the platform metadata is up-to-date and accurate. They oversee data quality and leverage the lineage tools to verify proper data flows. They establish and monitor data quality rules aligned with the metadata and lineage information.
- **Data Analysts** understand the origin and transformations of data they use for reporting and analytics. This help them spot potential issues and validate the data they use.
- **IT Security and Compliance** understand where sensitive data is stored, transformed, and used. This enables them to enforce access control and monitoring to prevent unauthorized changes, promptly detect and respond to any suspicious activities, and ensure adherence to regulatory requirements.

This proactive approach is governed by a rigorous change management process, during which any system change requires an impact analysis supported by data lineage. Automated data quality checks are also used to identify out-of-range values or unusual data distributions that can indicate issues.

Comprehensive guide to migrating legacy systems to Snowflake with Cloudera Octopai Data Lineage

Learn about best practices for migrating data from Oracle, SQL Server, and Teradata to Snowflake with Cloudera Octopai Data Lineage.

Introduction

Data migration is a critical part of digital transformation and system upgrades, and it involves transferring data from one system to another. This guide focuses on the migration from Oracle, SQL Server, and Teradata, leading relational database management systems, to Snowflake, a cloud-based data warehousing platform designed for the cloud. It provides a detailed roadmap for a successful data assets migration, including key considerations, challenges, and best practices.

Conducting a comprehensive assessment of the legacy environment, choosing the right migration approach, optimizing data for Snowflake, using the native services of Snowflake, implementing data security and compliance, minimizing downtime and disruption, training the team, and monitoring and optimizing the Snowflake environment after migration are recommended best practices.

Additionally, data and IT teams will need to understand and prepare for differences in how each platform handles things like data types, partitioning, indexing, and cost management to ensure a smooth migration and efficient use of the new system.

Cloudera Octopai Data Lineage plays an integral role in enabling organizations to overcome technical challenges and maximize the value of their data migration from Oracle, SQL Server, and Teradata to Snowflake. Its capabilities are designed to ensure risk mitigation, cost reduction, and efficiency in man hours, thereby contributing to an overall

smoother, more cost-effective, and risk-averse data migration process. Cloudera Octopai Data Discovery tool, for example, automates the process of scanning legacy systems to identify existing data assets, their location, format, and business relevance.

The value proposition of cloud migration

Transitioning from on-premise solutions to cloud-based platforms brings a host of benefits, including cost savings, scalability, increased collaboration, and access to advanced analytics and machine learning capabilities. The move also presents opportunities to adopt new methodologies like DevOps and Agile practices, promoting innovation and reducing time-to-market. However, this transition also has significant implications for Data and IT teams, as they need to manage the shift in technologies, adopt new skill sets, and adjust to different operational practices.

Why Snowflake?

Migration from On-Premise Systems

- Many organizations are moving from traditional on-premise data warehouses to cloud-based solutions like Snowflake. This is due to the scalability, flexibility, and cost-effectiveness of cloud solutions. Common on-premise systems that are often migrated include Oracle, Teradata, and IBM DB2.

Migration from Other Cloud Providers:

- Some organizations are migrating from other cloud data warehouses to Snowflake. This could be due to a variety of reasons, including better performance, cost, or features offered by Snowflake. Common cloud systems that are often migrated include Amazon Redshift, Google BigQuery, and Microsoft Azure SQL Data Warehouse.

Consolidation of Disparate Systems:

- Organizations with data spread across multiple systems (both on-premise and cloud) might choose to migrate to Snowflake to consolidate their data into a single, unified platform. This can simplify data management and improve the ability to derive insights from the data.

Unique aspects of Oracle, SQL Server, Teradata and Snowflake

When discussing the migration from Oracle, SQL Server, and Teradata to Snowflake, you must understand the unique aspects of both platforms that might affect the migration process. The systems have the following key characteristics:

- **Oracle, SQL Server, Teradata**
  - Mature and Robust

    These systems have been around for a long time, and they are known for their robustness and stability. They have a range of features built over time that can handle complex queries and large volumes of data.
  - Scripting

    These systems have their own scripting utilities that allow users to submit SQL commands in batch mode. They are highly versatile and can be used for importing and exporting data, as well as controlling database workflows.
  - Stored Procedures

    These systems support complex stored procedures. This feature might require significant effort to migrate, as the Snowflake scripting and stored procedure capabilities differ.
  - Data Loading and Extraction

    These systems have powerful utilities for data loading and extraction, which might have been used heavily in data and IT teams current setup.

- **Snowflake**
  - Serverless and Fully Managed

    Snowflake is a serverless, highly scalable, and cost-effective cloud data warehouse. It takes away the hassle of managing infrastructure, allowing organizations to focus more on data analysis.
  - Automatic Scaling

    Snowflake automatically scales up and down based on the workload, which means data and IT teams do not need to worry about capacity planning.
  - Real-Time Analysis

    Snowflake allows for real-time analytics on massive streaming data, which is not typically a strength of traditional databases.
  - Integrated with Cloud Platforms

    Snowflake is well-integrated with other cloud services, making it easy to connect with data storage, machine learning, and data processing tools.
  - Data Transfer Service (DTS)

    Snowflake offers DTS for automating data movement from multiple online and offline sources.
  - Security

    Snowflake provides robust security measures, including encryption at rest and in transit, identity and access management, and support for VPC Service Controls.

Migration considerations

Oracle Database

Oracle is a widely used traditional relational database management system. Many businesses are migrating from Oracle to Snowflake due to Snowflake's scalability, flexibility, and cost-effectiveness. The migration process typically involves data extraction from Oracle, data cleaning and transformation, and then loading the data into Snowflake.

Microsoft SQL Server

Another common source for migration is Microsoft SQL Server. This is due to the fact that while SQL Server is powerful and widely used, it can be expensive and lacks the flexibility and scalability that a cloud-native solution like Snowflake offers. The migration process is similar to Oracle, involving extraction, transformation, and loading (ETL).

Teradata

Teradata is a popular data warehouse solution, but many businesses are moving to Snowflake for its superior cloud capabilities, scalability, and cost-effectiveness. The migration process from Teradata to Snowflake involves a similar ETL process, but with additional considerations for the Teradata unique architecture and features.

Migrating from Teradata to Snowflake

Teradata, a popular legacy system, has its own unique challenges when migrating to Snowflake. The process involves understanding the differences between Teradata and Snowflake, such as the handling of NULLs, data types, and stored procedures.

- Data Types: Teradata and Snowflake have different data types. For example, the Teradata TIME data type does not have a direct equivalent in Snowflake. During migration, you must map Teradata data types to their Snowflake equivalents.
- Stored Procedures: The Teradata stored procedures use a different syntax than the Snowflake stored procedures. You must rewrite these procedures in JavaScript, which is the language Snowflake uses for stored procedures.
- NULL Handling: Teradata and Snowflake handle NULLs differently. In Teradata, a comparison with NULL is UNKNOWN, while in Snowflake, it is NULL. You must account for this difference during migration.

Migrating from Oracle and SQL Server to Snowflake

The following considerations apply for migrating from Oracle and SQL Server to Snowflake:

- Data Types: Both Oracle and SQL Server have unique data types that need to be mapped to Snowflake equivalents during migration.
- Stored Procedures: Oracle uses PL/SQL for stored procedures, while SQL Server uses T-SQL. These must be rewritten in JavaScript for Snowflake.
- Indexes: Unlike Oracle and SQL Server, Snowflake does not use indexes. You must rethink any optimization based on indexes for Snowflake.
- Sequences: Oracle and SQL Server handle sequences differently than Snowflake. If your legacy system uses sequences, you must recreate them in Snowflake.

Potential technical and business challenges

In general, migrating from an on-premises solution to a cloud-based one presents a variety of challenges that need to be carefully considered and managed. Migration from Oracle, SQL Server, and Teradata to Snowflake can pose both technical and business challenges, including schema migration, stored procedures conversion, standard and complex views adaptation, handling of scripts, performance considerations and complex risk management.

In some cases, it might be determined that certain tables from the legacy systems are no longer needed, even before the migration, or for example, that specific ETL processes loading data into these tables might be retired. After the migration, it is crucial to review the entire project, looking for opportunities to reduce maintenance, CPU, storage, people resources, and time.

With the right approach and the use of Cloudera Octopai platform that supports Cross Systems, Inner Systems, and End-to-End Column Lineage for both the legacy systems and Snowflake, data and IT teams can ensure a smooth and efficient migration, leading to better data management and utilization in the long run.

The following technical and business challenges can arise:

- **Technical challenges**

  Technical challenges can range from system compatibility issues, data conversion challenges, or issues with connectivity and access rights. They will arise due to differences in data types and SQL dialects, absence of traditional stored procedure support in Snowflake, differences in view handling, managing large files, and distinct performance characteristics of the legacy systems and Snowflake. Consider the following potential technical challenges:

  - Schema Migration: The legacy systems and Snowflake have different data types and SQL dialects. This means that some data transformations might be necessary to convert legacy data types to Snowflake compatible ones. You might need to rewrite SQL queries due to these syntactical differences.
  - Views: In the legacy systems, views are used extensively for data abstraction, simplifying queries, and controlling access. The concept of views exists in Snowflake as well, but the behavior might be different. For instance, the Snowflake views are logical views, not materialized. That means they compute the data when queried, which might affect the performance if not managed properly.
  - Scripting and Stored Procedures: The legacy systems scripts and stored procedures might not be directly compatible with Snowflake, requiring rewriting and transformation. For example, you must consider how to transform the legacy scripts and stored procedures into Snowflake compatible SQL scripts or Cloud Functions, or how to replace the legacy utilities with Snowflake Data Transfer Service.
  - Performance and Data Transfer Considerations: The performance characteristics of the legacy systems and Snowflake can be quite different due to the underlying architecture differences. Snowflake is a serverless, highly scalable, and cost-effective multi-cloud data warehouse designed for business agility, while the legacy systems are popular relational database management systems. Therefore, optimizing for performance might require different strategies in Snowflake as compared to the legacy systems. Moving large amounts of data from one platform to another can be time-consuming and risk-prone. There is a risk of data loss, corruption, or security breaches during the migration process.
  - Utilizing Legacy Utilities: If you are used to the legacy utilities, you will need to figure out how to accomplish these tasks with the Snowflake tools.

- **Business challenges**

  Business challenges can include risk management, cost considerations, training and change management. For instance, there might be costs associated with potential downtime during the migration, training for staff to use the new system, and potential resistance from users who are comfortable with the legacy system. Consider the following potential business challenges:

  - Risk Management: Data migration always carries risk. These can include data loss, corruption, or breaches of security. You must have robust risk management strategies in place to mitigate these risks. This includes thorough testing, backup and recovery plans, and security measures.
  - Cost Considerations: While moving to a cloud-based solution like Snowflake can lead to cost savings in the long run, the migration process itself can be costly. This includes the cost of the migration tools and services, potential downtime during the migration, and the cost of training staff to use the new system.
  - Training and Change Management: Moving to a new system requires users to learn new tools and adapt to new workflows. This can lead to resistance, especially if users were comfortable with the legacy system. You must have a strong change management strategy in place to support users through this transition.
  - Performance Tuning: After migrating, data and IT teams might need to spend time optimizing query performance and cost in Snowflake. This can be a complex task and might require a deep understanding of how Snowflake processes queries.
  - Continuing to Monitor and Optimize: Once you migrated to Snowflake, you must continually monitor your system to ensure that it works effectively and adjust as needed.

Migration process overview and best practices

When migrating to Snowflake, you must follow a structured workflow that includes setting up and configuring data transfer, choosing the right extraction method, and identifying and customizing schema. These are key components of the migration workflow.

The migration process can be broken down into seven distinct phases: data discovery, dependencies and scoping for effort assessment, data cleansing and deprecation to migrate the right data, prioritization according to critical paths, selective migration, data conversion, documentation to capture knowledge, and post migration for onboarding and monitoring.

Assessment - data discovery and scoping

Understanding the existing data landscape is the first step - identifying existing data, its location, format, and business relevance. This phase forms the foundation for all subsequent steps, establishing a comprehensive inventory of data assets and their dependencies, which is critical for determining what data to migrate and how.

Cloudera Octopai Data Discovery will automatically scan your legacy system (Teradata, Oracle, or SQL Server), identify data entities, their relationships, and how they are used. By gathering insights about your data assets, data and IT teams gain a comprehensive view of what is in your system, guiding data teams to build a solid migration plan.

Data cleansing and deprecation

Once the data assets have been scoped, the next step is to cleanse the data in the legacy system. This involves removing redundant, irrelevant, or erroneous data to avoid migrating garbage. This step is also an opportune moment to deprecate unused or unnecessary reports fed by the legacy system. This prevents the migration of garbage, and presents an opportunity to deprecate duplicates, redundant or unused assets.

The Cloudera Octopai cross systems lineage and inner-system lineage tools play a crucial role in this phase. By visualizing the data lineage, Cloudera Octopai pinpoints data assets that are no longer being used and can be deprecated, streamlining the migration process and avoiding the migration of unnecessary data.

Prioritization

Selective migration is the best approach for prioritization of the most critical data assets to be migrated first. This can help minimize downtime and disruptions to key business operations. However, it can also introduce the challenge of

double maintenance - managing both the legacy and new systems concurrently, which can put additional stress on IT and data teams.

For prioritizing and selecting data, the best approach is to migrate the most accessed tables or those critical to business operations first. The prioritization can be based on the data's business impact, data quality, or compliance needs.

The Cloudera Octopai Data Lineage and Catalog capabilities empower the data team to analyze table dependencies, allowing them to collaborate with the business stakeholders and tag critical accounting data assets within the Cloudera Octopai Data Catalog for prioritized migration.

Data conversion

In this phase, the selected data is converted and transformed to suit the schema, stored procedures, and views of the Snowflake system. Data conversion involves converting legacy-specific SQL code into Snowflake standard SQL. For example, the legacy system TIMESTAMP might need to be converted to the Snowflake TIMESTAMP, and similar conversions for other data types. You must consider stored procedures as well since legacy procedures might not work directly in Snowflake due to syntactical differences.

During the data conversion phase, Cloudera Octopai Live lineage functionality proves invaluable in simulating the effects of transformations, allowing for a comprehensive understanding of potential issues and their impact on data lineage. With Cloudera Octopai, the data team can simulate the conversion of legacy-specific SQL code into the Snowflake standard SQL, identifying any syntactical differences or required data type conversions. By testing and verifying the simulated data lineage, the team can confidently address errors, make necessary fixes, and ensure the accuracy and integrity of the data throughout the migration process.

Post-migration challenges and optimization

After the migration, you must review the entire project, looking for opportunities to reduce maintenance, CPU, storage, people resources, and time. Post-migration, regular monitoring is crucial. For example, data teams might set up a daily check of failed SQL jobs in Snowflake. If data and IT teams see a recurring failure, data and IT teams can then dig deeper to understand the issue that can be a data type mismatch, or syntax error, and address the root cause.

Additional considerations exist for adopting new practices and terminologies related to cloud technologies. You must consider the following key points:

- Cloud Terminology: Familiarize yourself with the terminology used in cloud computing and specifically in Snowflake. This includes terms like warehouses, databases, schemas, tables, and more. Understanding these terms will help you navigate and manage your cloud environment effectively.
- Cloud Data Management: Cloud-based data management introduces new practices and tools. You must understand how data is stored, processed, and managed in Snowflake. This includes concepts like data partitioning, clustering, and more. Learn about best practices for optimizing data storage, query performance, and cost management in Snowflake.
- Security and Compliance: Cloud environments have their unique security and compliance considerations. Familiarize yourself with the security features and options provided by Snowflake. Implement proper access controls, encryption, and data governance practices. Ensure compliance with relevant regulations and standards that apply to your data.
- Automation and Orchestration: Cloud platforms offer automation and orchestration capabilities that can streamline data pipelines and processes. Explore tools to automate data transformations, workflows, and data integration tasks. Leverage these tools to optimize your data pipelines in the cloud.
- Scalability and Elasticity: Cloud-based systems provide scalability and elasticity, allowing you to scale your resources up or down based on demand. Understand how to leverage the Snowflake auto-scaling capabilities to handle varying workloads efficiently. Design your data pipelines and infrastructure to take advantage of the cloud scalability benefits.
- Data Cataloguing and Documentation: Ensure that you create a comprehensive data catalog in Snowflake that accurately documents your data assets, including metadata, data lineage, and business descriptions. This will help users find and understand the data in the new system.

- Continuous Learning and Training: Cloud technologies and best practices evolve rapidly. Encourage continuous learning and training for your data and IT teams to stay updated with the latest advancements in Snowflake and cloud computing. Leverage training resources, online documentation, and community forums provided by Snowflake to enhance your knowledge and skills.

Efficient migration to Snowflake with Cloudera Octopai

Cloudera Octopai is a valuable solution that empowers businesses during the migration from legacy systems to Snowflake. It automates data discovery, streamlining the migration process and ensuring a smooth transition. With the Cloudera Octopai lineage and impact analysis capabilities, businesses can enhance data management, validate data integrity, and improve overall reliability. Cloudera Octopai also reduces the time and effort required for identifying and migrating relevant data assets. It optimizes resource allocation by identifying and deprecating unused data, while ensuring data security and compliance through robust lineage tracking. By minimizing downtime and disruptions, Cloudera Octopai facilitates uninterrupted business operations during the migration. Overall, Cloudera Octopai delivers significant benefits, making the migration to Snowflake efficient and successful.

Conclusion

Migrating from Oracle, SQL Server, and Teradata to Snowflake is a significant undertaking that requires careful planning, execution, and monitoring. However, with the right approach and tools, it can be a smooth and efficient process that results in significant benefits for your organization. By following the best practices outlined in this guide and leveraging the capabilities of Cloudera Octopai, you can ensure a successful migration that enhances your organization's data management capabilities and positions you for success in the era of cloud computing.

Comprehensive guide to migrating Teradata to Google BigQuery with Cloudera Octopai Data Lineage

Learn about migrating from Teradata to Google BigQuery, emphasizing best practices, challenges, and solutions.

Introduction

Data migration is a critical part of digital transformation and system upgrades, and it involves transferring data from one system to another. This guide focuses on the migration from Teradata, a leading relational database management system, to Google BigQuery, a serverless, highly scalable, and cost-effective multi-cloud data warehouse. It provides a detailed roadmap for a successful data assets migration, including key considerations, challenges, and best practices.

Conducting a comprehensive assessment of the Teradata environment, choosing the right migration approach, optimizing data for BigQuery, using the BigQuery native services, implementing data security and compliance, minimizing downtime and disruption, training the team, and monitoring and optimizing the BigQuery environment after migration are recommended best practices.

Additionally, data and IT teams will need to understand and prepare for differences in how each platform handles things like data types, partitioning, indexing, and cost management to ensure a smooth migration and efficient use of the new system.

Cloudera Octopai Data Lineage plays an integral role in enabling organizations to overcome technical challenges and maximize the value of their data migration from Teradata to Google BigQuery. Its capabilities are designed to ensure risk mitigation, cost reduction, and efficiency in man hours, thereby contributing to an overall smoother, more cost-effective, and risk-averse data migration process. Cloudera Octopai Data Discovery tool, for example, automates the process of scanning Teradata systems to identify existing data assets, their location, format, and business relevance.

The value proposition of cloud migration

Transitioning from on-premise solutions to cloud-based platforms brings a host of benefits, including cost savings, scalability, increased collaboration, and access to advanced analytics and machine learning capabilities. The move also presents opportunities to adopt new methodologies like DevOps and Agile practices, promoting innovation and

reducing time-to-market. However, this transition also has significant implications for Data and IT teams, as they need to manage the shift in technologies, adopt new skill sets, and adjust to different operational practices.

Teradata compared to Google BigQuery with unique aspects

When discussing the migration from Teradata to Google BigQuery, it is essential to understand the unique aspects of both platforms that might affect the migration process. Teradata and Google BigQuery have the following key characteristics:

- **Teradata**
  - Mature and Robust: Teradata has been around for a long time, and it is known for its robustness and stability. It has a range of features built over time that can handle complex queries and large volumes of data.
  - BTEQ Scripting: BTEQ (Basic Teradata Query) is a utility in Teradata that allows users to submit SQL commands in batch mode. It is highly versatile and can be used for importing and exporting data, as well as controlling database workflows.
  - Stored Procedures: Teradata supports complex stored procedures. This feature might require significant effort to migrate, as BigQuery's scripting and stored procedure capabilities differ.
  - TPump, FastLoad, and MultiLoad: Teradata has powerful utilities for data loading and extraction, which might have been used heavily in data and IT teams current setup.
- **Google BigQuery**
  - Serverless and Fully Managed: BigQuery is a serverless, highly scalable, and cost-effective cloud data warehouse. It takes away the hassle of managing infrastructure, allowing organizations to focus more on data analysis.
  - Automatic Scaling: BigQuery automatically scales up and down based on the workload, which means data and IT teams do not need to worry about capacity planning.
  - Real-Time Analysis: BigQuery allows for real-time analytics on massive streaming data, which is not typically a strength of traditional databases.
  - Integrated with Google Cloud: BigQuery is well-integrated with other Google Cloud services, making it easy to connect with data storage, machine learning, and data processing tools.
  - Data Transfer Service (DTS): BigQuery offers DTS for automating data movement from multiple online and offline sources.
  - Security: BigQuery provides robust security measures, including encryption at rest and in transit, identity and access management, and support for VPC Service Controls.

Potential technical and business challenges

In general, migrating from an on-premises solution to a cloud-based one presents a variety of challenges that need to be carefully considered and managed.

Migration from Teradata to Google Big Query can pose both technical and business challenges, including schema migration, stored procedures conversion, standard and complex views adaptation, handling of BTEQ scripts (Batch Teradata Query), performance considerations and complex risk management.

In some cases, it might be determined that certain tables from Teradata are no longer needed, even before the migration, or for example, that specific ETL processes loading data into these tables might be retired.

After the migration, it is crucial to review the entire project, looking for opportunities to reduce maintenance, CPU, storage, people resources, and time.

With the right approach and the use of Cloudera Octopai platform that supports Cross Systems, Inner Systems, and End-to-End Column Lineage for both Teradata and Google BigQuery, data and IT teams can ensure a smooth and efficient migration, leading to better data management and utilization in the long run.

The following technical and business challenges can arise:

- **Technical challenges**

  Technical challenges can range from system compatibility issues, data conversion challenges, or issues with connectivity and access rights. It will arise due to differences in data types and SQL dialects, absence of traditional stored procedure support in BigQuery, differences in view handling, managing large files, and distinct performance characteristics of Teradata and BigQuery. Consider the following potential technical challenges:

  - Schema Migration: Teradata and Google BigQuery have different data types and SQL dialects. This means that some data transformations might be necessary to convert Teradata data types to BigQuery compatible ones. You might need to rewrite SQL queries due to these syntactical differences.
  - Microviews and Central Views: In Teradata, views (including Microviews and Central views) are used extensively for data abstraction, simplifying queries, and controlling access. The concept of views exists in BigQuery as well, but the behavior might be different. For instance, the BigQuery views are logical views, not materialized. That means they compute the data when queried, which might affect the performance if not managed properly.
  - BTEQ Scripting and Stored Procedures: The Teradata BTEQ scripts and stored procedures might not be directly compatible with BigQuery, requiring rewriting and transformation. For example, you must consider how to transform Teradata BTEQ scripts and stored procedures into BigQuery compatible SQL scripts or Cloud Functions, or how to replace the Teradata utilities with BigQuery Data Transfer Service.
  - Performance and Data Transfer Considerations: The performance characteristics of Teradata and BigQuery can be quite different due to the underlying architecture differences. BigQuery is a serverless, highly scalable, and cost-effective multi-cloud data warehouse designed for business agility, while Teradata is a popular relational database management system. Therefore, optimizing for performance might require different strategies in BigQuery as compared to Teradata. Moving large amounts of data from one platform to another can be time-consuming and risk-prone. There is a risk of data loss, corruption, or security breaches during the migration process.
  - Utilizing Teradata Utilities: If you are used to Teradata utilities, such as TPump, FastLoad, and MultiLoad, you must find a way to accomplish them with the BigQuery tools.
- **Business challenges**

  Business challenges can include risk management, cost considerations, training and change management. For instance, there might be costs associated with potential downtime during the migration, training for staff to use the new system, and potential resistance from users who are comfortable with the legacy system. Consider the following potential business challenges:

  - Risk Management: Data migration always carries risk, including data loss, corruption, or breaches of security. Robust backup and recovery strategies must be in place.
  - Cost: The cost of migration can be high, including the cost of potential downtime, the manpower costs of performing the migration, and the costs of any necessary new software.
  - Training and Change Management: Staff might need to be retrained to use BigQuery, which can require both time and expense. Change management is also a significant concern, as employees need to adapt to the new system.

  Cloudera Octopai is instrumental in addressing data migration challenges, by effectively mitigating risks, cutting costs, and reducing man hours, including wasted time, in transitioning from Teradata to Google BigQuery. Its platform enhances risk management through thorough data mapping, ensuring optimal security while minimizing data loss or corruption. The Cloudera Octopai automation facilities contribute to substantial savings in both costs and time, streamlining the migration process and making staff training more efficient. As such, Cloudera Octopai is fundamental in facilitating a more cost-efficient, safer, and time-saving data migration process.

Migration process overview and best practices

Make sure you are following the migration workflow by setting up and configuring data transfer, choosing the right extraction method, and identifying and customizing schema are key components of the migration workflow.

The suggested migration process consists of the following steps:

- Download the Migration Agent: The process starts with downloading the migration agent which facilitates the data transfer.

- Configure a Transfer in the BigQuery Data Transfer Service: Set up the data transfer based on your requirements.
- Run the Transfer Job: Copy the table schema and data from your data warehouse to BigQuery.
- Monitor Transfer Jobs: You can optionally monitor the transfer jobs using the Google Cloud Console.

During the migration, the BigQuery Data Transfer Service provides two methods for transferring data: the Teradata Parallel Transporter (TPT) tbuild utility and extraction using a JDBC driver with a FastExport connection. The TPT method is faster but if storage space is a constraint, the JDBC driver with FastExport connection is recommended.

The service also provides automatic schema detection and data type mapping during the data transfer. You can alternatively specify a custom schema file, which can be created manually or generated by the migration agent.

The migration process can be broken down into seven distinct phases: data discovery, dependencies and scoping for effort assessment, data cleansing and deprecation to migrate the right data, prioritization according to critical paths, selective migration, data conversion, documentation to capture knowledge, and post migration for onboarding and monitoring.

Assessment - data discovery and scoping

Understanding the existing data landscape is the first step - identifying existing data, its location, format, and business relevance. This phase forms the foundation for all subsequent steps, establishing a comprehensive inventory of data assets and their dependencies, which is critical for determining what data to migrate and how.

The data discovery and scoping phase is fundamental. For instance, in a large company, you might have hundreds of Teradata databases, each with hundreds or thousands of tables. By using Cloudera Octopai Discovery and integrated Data Catalog, Data teams can search, scope and document data entities (tables, views), their owners, data volumes, and how often they are accessed. This information helps to build a solid understanding of what resides in your current databases, guiding your migration plan.

Cloudera Octopai Data Discovery will automatically scan your Teradata system, identify data entities, their relationships, and how they are used. By gathering insights about your data assets, data and IT teams gain a comprehensive view of what is in your system, guiding data teams to build a solid migration plan.

Data cleansing and deprecation

Once the data assets have been scoped, the next step is to cleanse the data in the Teradata system. This involves removing redundant, irrelevant, or erroneous data to avoid migrating garbage. This step is also an opportune moment to deprecate unused or unnecessary reports fed by Teradata. This prevents the migration of garbage and presents an opportunity to deprecate duplicates, redundant or unused assets.

Cloudera Octopai cross-system lineage and inner-system lineage tools play a crucial role in this phase. By visualizing the data lineage, Cloudera Octopai pinpoints data assets that are no longer being used and can be deprecated, streamlining the migration process and avoiding the migration of unnecessary data.

Data teams will operate with Cloudera Octopai Teradata lineage and see which tables and ETL processes are no longer in use. For example, you might have ETL processes or tables created for specific projects that have now ended, and these data assets are no longer used or updated. Identifying and removing these from your migration scope will save considerable time and resources.

Prioritization

Selective migration is the best approach for prioritization of the most critical data assets to be migrated first. This can help minimize downtime and disruptions to key business operations. However, it can also introduce the challenge of double maintenance - managing both the legacy and new systems concurrently, which can put additional stress on IT and data teams. With selective migration, you could start by migrating a small subset of data, perhaps one business area or a specific application. This would allow data and IT teams to identify any issues on a smaller scale before moving on to the rest of the data. It is important to note that this likely requires your team to manage both the old and new systems concurrently, which could increase workload in the short term.

For prioritizing and selecting data, the best approach is to migrate the most accessed tables or those critical to business operations first. Suppose you have a database heavily used by the finance department for monthly reporting.

In that case, this might be prioritized to ensure continued business operation. The prioritization can be based on the data's business impact, data quality, or compliance needs.

After cleansing, data assets are prioritized for migration based on factors like business value, user needs, or other defined criteria. The final selection of data to be migrated is then made by the migration team.

The Cloudera Octopai Data Lineage and Catalog capabilities empower the data team to analyze table dependencies, allowing them to collaborate with the business stakeholders and tag critical accounting data assets within the Cloudera Octopai Data Catalog for prioritized migration. You can ensure a seamless transition, focusing on migrating the essential data assets first and minimizing any potential disruptions to their financial operations during the migration to BigQuery.

Data conversion

In this phase, the selected data is converted and transformed to suit the schema, stored procedures, and views of the BigQuery system.

The migration process involves downloading the migration agent, configuring a transfer in the BigQuery Data Transfer Service, running the transfer job, and optionally monitoring transfer jobs using the Google Cloud console.

Data conversion involves converting Teradata-specific SQL code into BigQuery standard SQL.

For example, the Teradata TIMESTAMP might need to be converted to the BigQuery TIMESTAMP, and similar conversions for other data types. You will need to consider stored procedures as well since Teradata procedures might not work directly in BigQuery due to syntactical differences.

During the data conversion phase, Cloudera Octopai Live lineage functionality proves invaluable in simulating the effects of transformations, allowing for a comprehensive understanding of potential issues and their impact on data lineage. With Cloudera Octopai, the data team can simulate the conversion of Teradata-specific SQL code into the BigQuery standard SQL, identifying any syntactical differences or required data type conversions. By testing and verifying the simulated data lineage, the team can confidently address errors, make necessary fixes, and ensure the accuracy and integrity of the data throughout the migration process.

The Cloudera Octopai Data Lineage Tracking ensures that data is accurately translated during migration. It serves as a validation method, comparing the lineage of data before and after migration.

Special considerations for stored procedures migration

Despite differences in stored procedures between Teradata and Google BigQuery, migration can be achieved by identifying and documenting Teradata stored procedures, learning about BigQuery stored procedures, manually rewriting procedures for BigQuery, testing the new procedures, and iterating on the converted procedures.

When it comes to BTEQ scripts, an important Teradata utility, no direct equivalent exists in BigQuery. This presents a challenge since these scripts often contain business logic. One way to tackle this issue is to convert BTEQ scripts into SQL scripts, then refactor them to comply with BigQuery SQL syntax.

Post-migration challenges

Adopting New Tools and Infrastructure: BigQuery is a part of Google Cloud Platform, and it works well with other Google services. Your team might need to familiarize themselves with a whole new set of tools.

- New Data Management Practices: BigQuery uses a different model of computing resources, and it automatically scales based on the workload. This could change how data and IT teams manage data and resources, including considerations around cost control.
- SQL Syntax Differences: BigQuery uses a slightly different SQL syntax than Teradata. If your team is used to writing SQL for Teradata, they might need to learn some new syntax and conventions.
- Change Management: Adapting to new systems can be difficult. Your team will need to change their workflow and habits, which can lead to resistance.
- Performance Tuning: After migrating, data and IT teams might need to spend time optimizing query performance and cost in BigQuery. This can be a complex task and might require a deep understanding of how BigQuery processes queries.

- Training: Staff might need to be retrained to use BigQuery, which can require both time and expense.
- Continuing to Monitor and Optimize: Once you migrated to BigQuery, you must continually monitor your system to ensure that it works effectively and adjust as needed.

Once the data is successfully migrated, a review and optimization phase is crucial. This phase involves running tests to verify data integrity, analyzing performance, and fine-tuning the system as required. Additionally, feedback from users is collected to further improve and customize the system to better fit their needs. Post-migration, regular monitoring is crucial. For example, data teams might set up a daily check of failed SQL jobs in BigQuery. If data and IT teams see a recurring failure, data and IT teams can then dig deeper to understand the issue - perhaps a data type mismatch, or syntax error, and address the root cause.

Additional considerations exist for adopting new practices and terminologies related to cloud technologies. You must consider the following key points:

- Cloud Terminology: Familiarize yourself with the terminology used in cloud computing and specifically in Google Cloud Platform (GCP) and BigQuery. This includes terms like projects, datasets, tables, buckets, regions, and zones. Understanding these terms will help you navigate and manage your cloud environment effectively.
- Cloud Data Management: Cloud-based data management introduces new practices and tools. You must understand how data is stored, processed, and managed in BigQuery. This includes concepts like data partitioning, clustering, and streaming inserts. Learn about best practices for optimizing data storage, query performance, and cost management in BigQuery.
- Security and Compliance: Cloud environments have their unique security and compliance considerations. Familiarize yourself with the security features and options provided by GCP and BigQuery. Implement proper access controls, encryption, and data governance practices. Ensure compliance with relevant regulations and standards that apply to your data.
- Automation and Orchestration: Cloud platforms offer automation and orchestration capabilities that can streamline data pipelines and processes. Explore tools to automate data transformations, workflows, and data integration tasks. Leverage these tools to optimize your data pipelines in the cloud.
- Scalability and Elasticity: Cloud-based systems provide scalability and elasticity, allowing you to scale your resources up or down based on demand. Understand how to leverage the BigQuery auto-scaling capabilities to handle varying workloads efficiently. Design your data pipelines and infrastructure to take advantage of the cloud scalability benefits.
- Data Cataloging and Documentation: Ensure that you create a comprehensive data catalog in BigQuery that accurately documents your data assets, including metadata, data lineage, and business descriptions. This will help users find and understand the data in the new system.
- Continuous Learning and Training: Cloud technologies and best practices evolve rapidly. Encourage continuous learning and training for your data and IT teams to stay updated with the latest advancements in BigQuery and cloud computing. Leverage training resources, online documentation, and community forums provided by Google Cloud to enhance your knowledge and skills.

Efficient Teradata to BigQuery migration with Cloudera Octopai

Cloudera Octopai is a valuable solution that empowers businesses during the migration from Teradata to Google BigQuery. It automates data discovery, streamlining the migration process and ensuring a smooth transition. With the Cloudera Octopai lineage and impact analysis capabilities, businesses can enhance data management, validate data integrity, and improve overall reliability. Cloudera Octopai also reduces the time and effort required for identifying and migrating relevant data assets. It optimizes resource allocation by identifying and deprecating unused data, while ensuring data security and compliance through robust lineage tracking. By minimizing downtime and disruptions, Cloudera Octopai facilitates uninterrupted business operations during the migration. Overall, Cloudera Octopai delivers significant benefits, making the Teradata to BigQuery migration efficient and successful.

Preparing for using Cloudera Octopai

Cloudera Octopai Data Lineage onboarding process: A guide to successful implementation

Learn about the main stages of Cloudera Octopai Data Lineage onboarding journey to ensure a smooth implementation.

Figure 3: Cloudera Octopai onboarding process

The onboarding process consists of the following main stages, including their importance, and how they contribute to the overall success of implementing Cloudera Octopai Data Lineage within your organization:

Stage 1: Kickoff Meeting

The Kickoff Meeting marks the beginning of the onboarding process and sets the stage for a successful implementation. It is recommended to have the following key individuals attend the meeting: Decision Maker, Project Manager, and a technical focal point. During this meeting, the following topics will be covered:

- Subscription Review

  An overview of the Cloudera Octopai subscription and its key features, ensuring a clear understanding of the platform's capabilities.
- Getting to Know Cloudera Octopai

  Introducing the Cloudera Octopai platform and its benefits, highlighting how it can address your organization's specific data governance and metadata management needs.
- Cloudera Octopai Configuration

  Discussing the configuration process, including integration with existing systems, defining user roles, and establishing security protocols.
- Best Practices

  Sharing industry best practices to maximize the value derived from Cloudera Octopai, providing insights on how to optimize your data governance initiatives.
- Next Steps

  Outlining the action plan for the onboarding process, setting expectations, and establishing timelines for subsequent training sessions.

Stage 2: Admin Training

Admin Training focuses on empowering your administrators with the necessary knowledge and skills to manage Cloudera Octopai effectively. This stage covers the following topics:

- Cloudera Octopai Portal

  A comprehensive overview of the Cloudera Octopai web portal, its navigation, and key functionalities.

- Cloudera Octopai Client

  Exploring the Cloudera Octopai Client application, which provides a user-friendly interface for managing metadata, configurations, and system integrations.
- Cloudera Octopai Platform - Admin Console

  Guiding administrators through the Cloudera Octopai Admin Console, where they can manage users, permissions, and system configurations.

Stage 3: User Training

User Training is designed to familiarize your end-users with the Cloudera Octopai platform and its capabilities. During this stage, participants will learn:

- Cloudera Octopai Platform

  A detailed introduction to the Cloudera Octopai platform, emphasizing its intuitive interface and ease of use.
- Cloudera Octopai Capabilities

  Exploring the various capabilities of Cloudera Octopai such as data discovery, data lineage visualization, impact analysis, and data cataloging.
- Cloudera Octopai Functionalities

  A comprehensive overview of the key functionalities within Cloudera Octopai, including search capabilities, metadata management, and collaboration features.

Use Cases

Presenting real-world use cases to demonstrate how Cloudera Octopai can be leveraged to solve common data governance challenges and optimize data operations.

Q&A Session

Two weeks after the User Training, it is highly recommended to schedule a dedicated Q&A session. This session provides an opportunity for users to ask questions, seek clarification on any aspects of Cloudera Octopai, and further solidify their understanding of the platform.

Conclusion

By following the Cloudera Octopai onboarding process, starting from the Kickoff Meeting, proceeding through Admin Training, User Training, and concluding with a Q&A session, your organization will be well-equipped to successfully implement the Cloudera Octopai data governance and metadata management solution. This comprehensive onboarding process ensures that all stakeholders are empowered with the knowledge and skills required to maximize the value derived from Cloudera Octopai and drive data-driven decision-making across your organization.

General system requirements for Cloudera Octopai Data Lineage Client

Learn about the general system requirements for Cloudera Octopai Data Lineage Client.

Note: Contact Cloudera Support for more details and instructions.

Windows server

- 2019 OS and higher on the organization's domain
- .NET Framework 6 and higher (offline installer)
- .NET Core 8 Runtime (installer)

- 8 Cores (CPU); 16 GB – RAM; 250 GB - Hard Disk
- Reserved ports for local use 5656 & 5657
- Curl is installed. For instructions, see How to Install Curl.
- Continue to Octopai Client - Installation guide.

Linux server

Contact Cloudera Support for more details and instructions.

- Linux distribution - Minimum requirement: RHEL 8 GUI / Oracle 8 GUI
- .NET Core 8 Runtime (link)
- 8 Cores (CPU); 16 GB – RAM; 250 GB - Hard Disk
- Reserved ports for local use 5656 & 5657
- Curl is installed. For instructions, see How to Install Curl.

Browser compatibility - Cloudera Octopai application

- Chrome or MS Edge

Important remarks

- All metadata extractions must be done from the same environment, such as Sandbox or QA
- Run the extractions in admin mode
- If you need assistance, contact Cloudera Support.

Enabling SSO for Cloudera Octopai Data Lineage with Entra ID (former Azure Active Directory)

Learn about enabling SSO for Cloudera Octopai Data Lineage using Entra ID (former Azure Active Directory) by leveraging OAuth 2.0 and OpenID Connect (OIDC) protocols. This integration enhances security, scalability, and compliance, ensuring seamless authentication and secure access to resources.

About this task

OAuth 2.0, serving as an authorization framework, allows third-party services to exchange web resources on behalf of users, utilizing access tokens over HTTP. When paired with OIDC, it introduces an authentication layer, verifying user identity before any authorization or data exchange, enhancing the security model.

This authorization has the following key benefits:

- Delegated Access – The synergy of OAuth 2.0 and OIDC is ideal for scenarios requiring the Cloudera Octopai systems to access metadata on behalf of users securely, without directly handling user credentials.
- Scalability and Flexibility – This combination supports a wide array of applications, meeting diverse client requirements from desktop and mobile apps to server-side and client-side applications.
- Enhanced Security and Azure AD Compatibility – Integration with Azure AD ensures a reliable and secure ecosystem, crucial for enterprise environments.

Implementing OAuth 2.0 with OIDC enriches our SSO capabilities, streamlining authentication across services with a single set of credentials. This demonstrates the reliability and security of the protocols, endorsed by critical sectors like banking, manufacturing, and healthcare.

For those upholding the highest data security and governance standards, our alignment with OAuth 2.0 and OIDC, facilitated by Azure Active Directory, signifies a proactive, security-first approach. It ensures our platform interactions adhere to stringent security protocols, evolving with digital threat landscapes and regulatory demands.

The Cloudera Octopai adoption of OAuth 2.0, enhanced by OIDC for authentication, underscores our commitment to security and meeting the complex needs of our diverse clientele. Cloudera Octopai is dedicated to continuous

improvement, aligning with industry-best practices to deliver the most secure, efficient, and compliant data lineage solutions available.

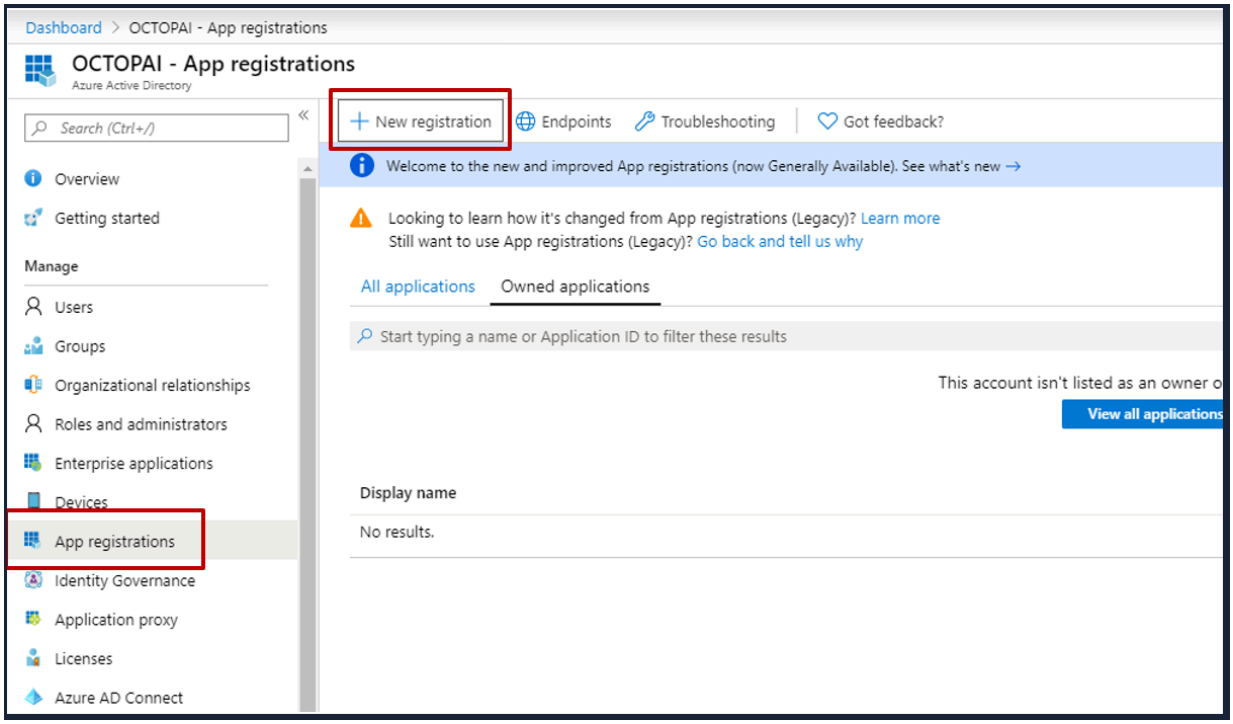
Note:

The Supported Protocol is JWT over OAuth 2.0.

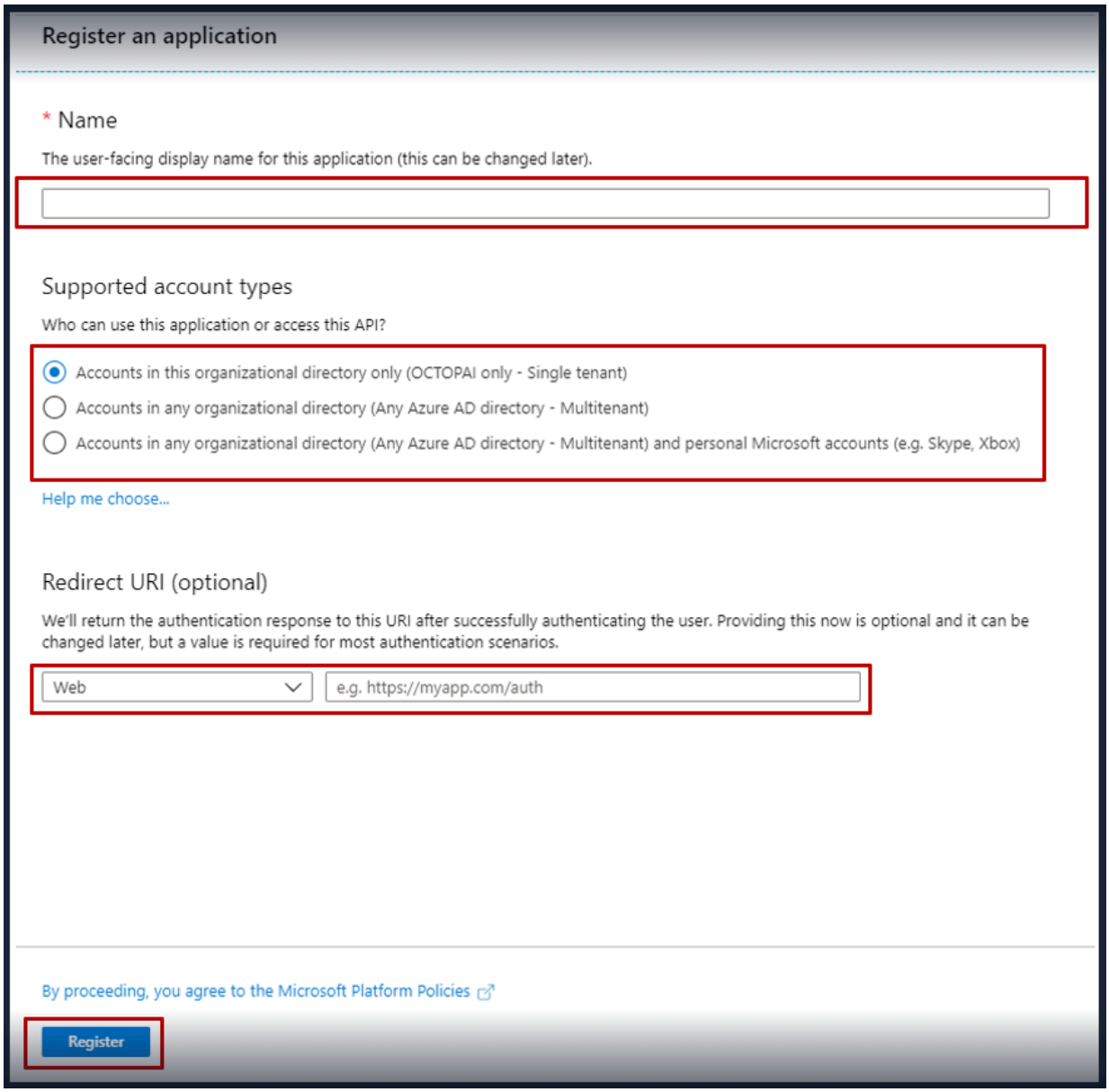
To set up SSO authentication for Cloudera Octopai using Entra ID (former Azure Active Directory), perform the following steps:

Procedure

1. Sign in to the Azure portal.
2. If you have more than one tenant, choose your Azure AD tenant by selecting your account in the top-right corner, clicking Switch Directory, and selecting the appropriate tenant.
3. Select Azure Active Directory in the left-hand navigation pane.
4. Choose App registrations and then select New application registration.



5. Fill in the redirect URI using your Cloudera Octopai application URL.



6. For web applications, provide the sign-on URL that is the base URL where users can sign in, https://app.octopai.cloud/customer_Login/Home/SignIn.vav
7. Set the redirect URL to https://app.octopai.cloud/customer_login/Register/index.

Results

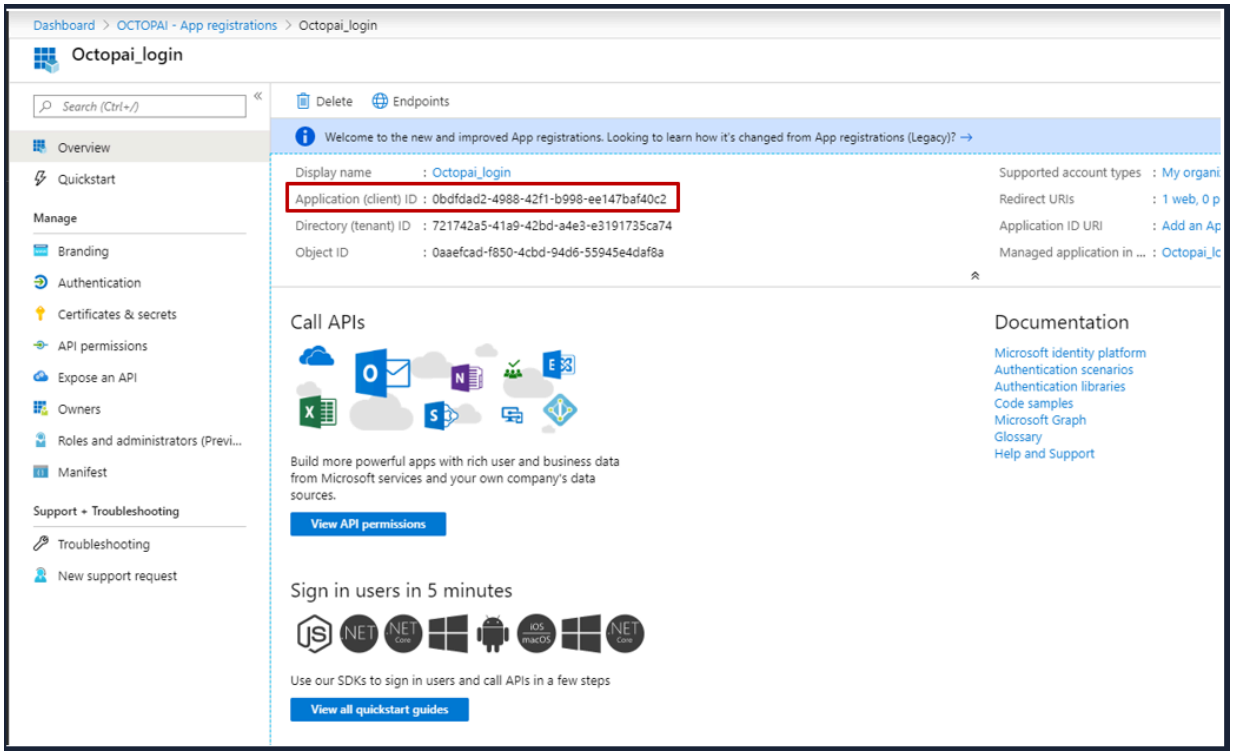
Once you completed registration, Azure AD will assign your application a unique client identifier, the Application (Client) ID. Copy this ID from the application page and send it to Cloudera Octopai.

Warning: After you submit your settings information, the configuration will be finalized by the Cloudera Support. Only once this is done, the SSO will be ready to use.

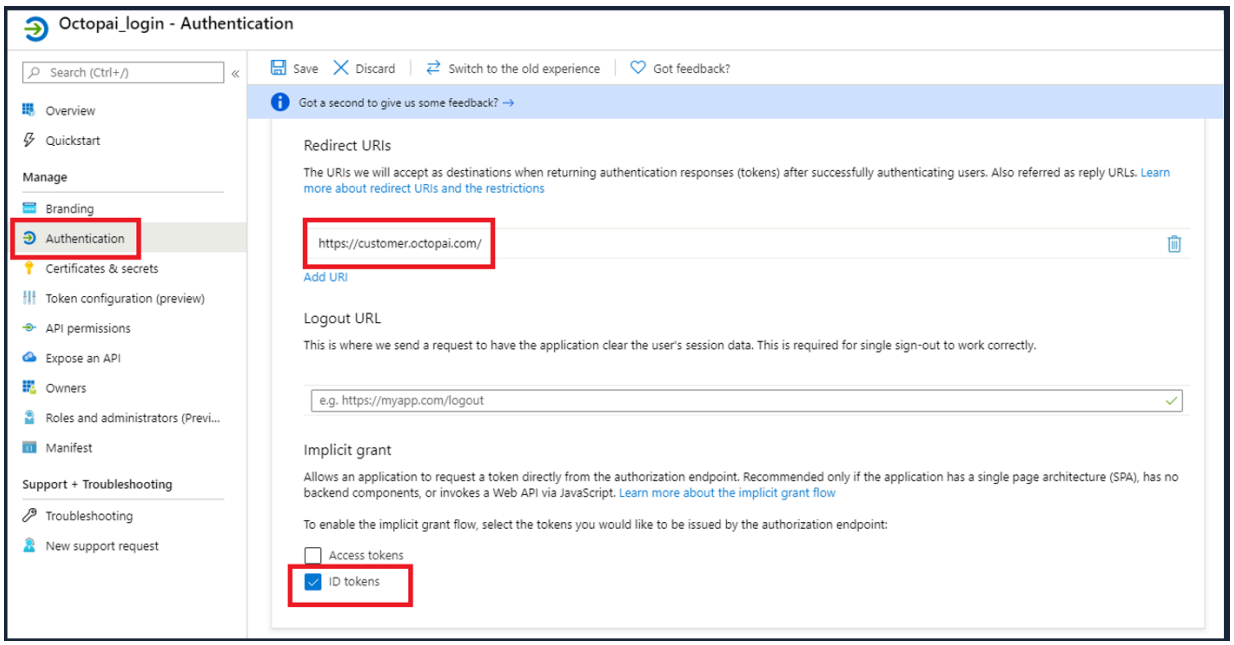
Once you complete the configuration setup, provide the Cloudera Support team with the following details:

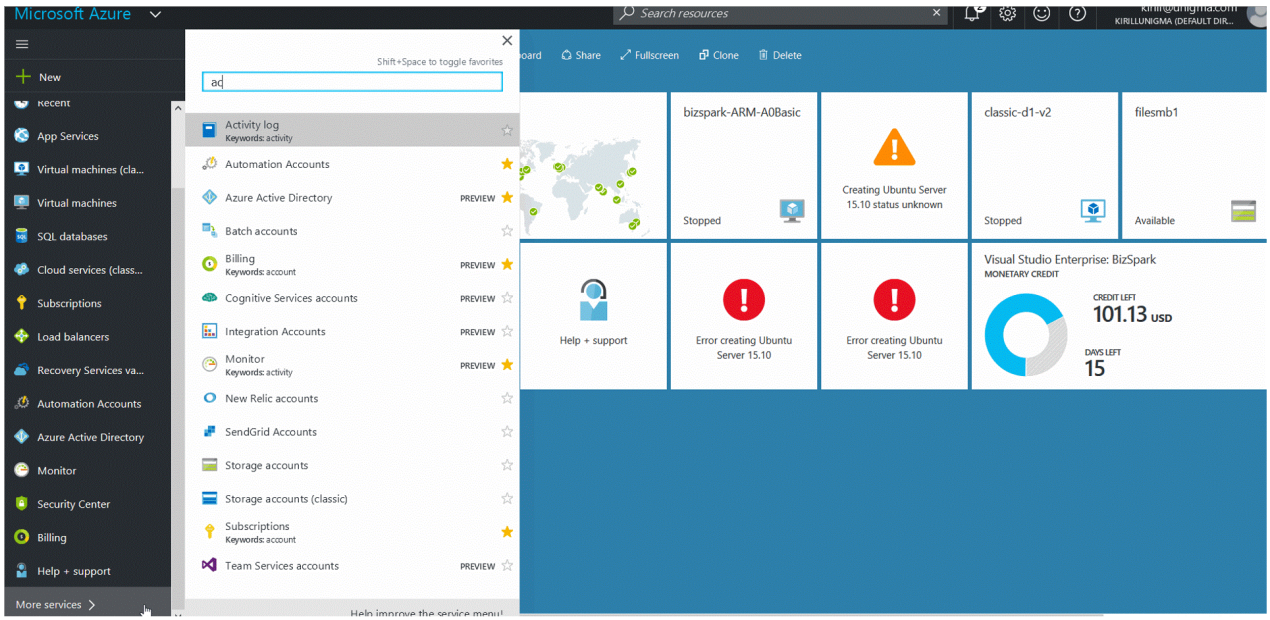
- Your tenant name or ID.

- Your application (client) ID.



- The redirect URI you set, for example https://*YOURNAME*.octopai.com/.





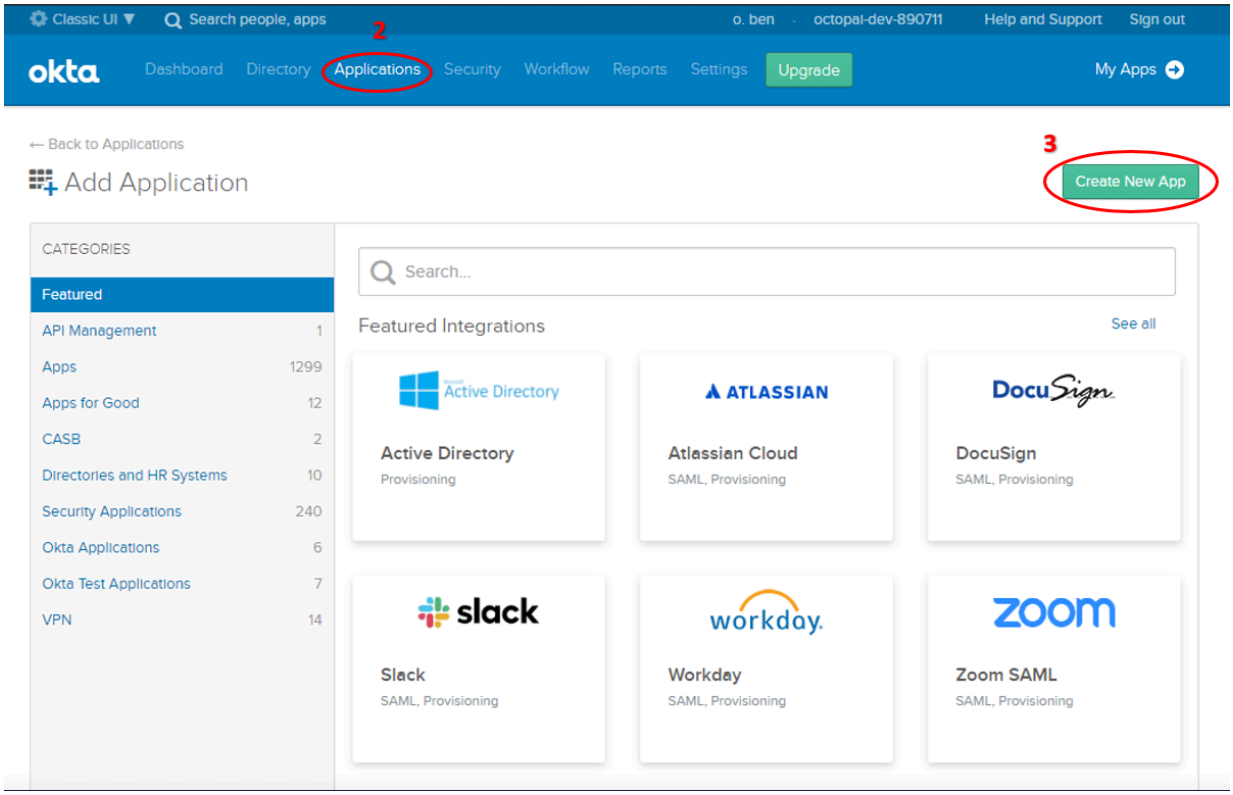
Setting up SSO authentication for Cloudera Octopai Data Lineage using Azure Active Directory (Azure AD) or Microsoft Entra ID

Learn about setting up SSO authentication for Cloudera Octopai Data Lineage using OKTA.

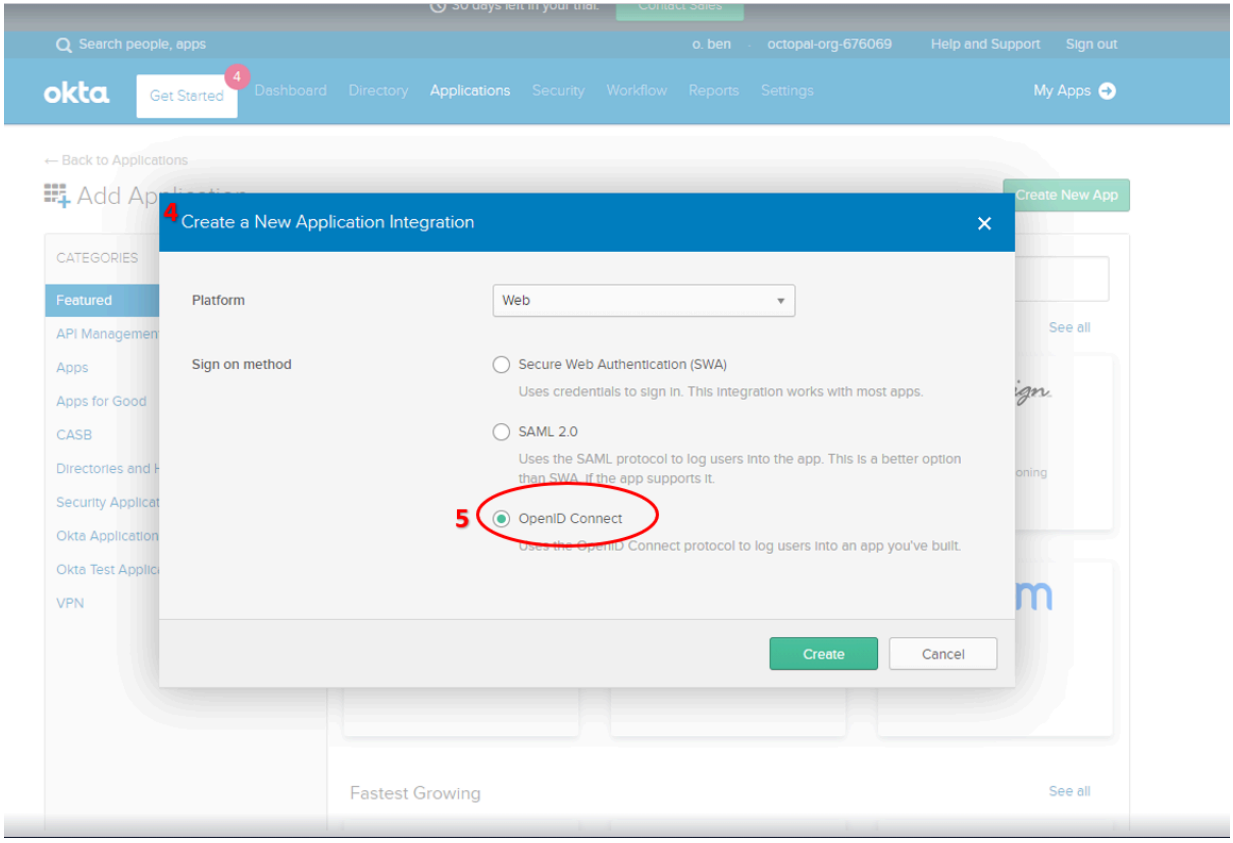
Procedure

1. Login to OKTA and access Applications

   - Login to OKTA with the Admin User of your OKTA instance.
   - Go to the Applications tab in the top menu.
   - Click Create New App.



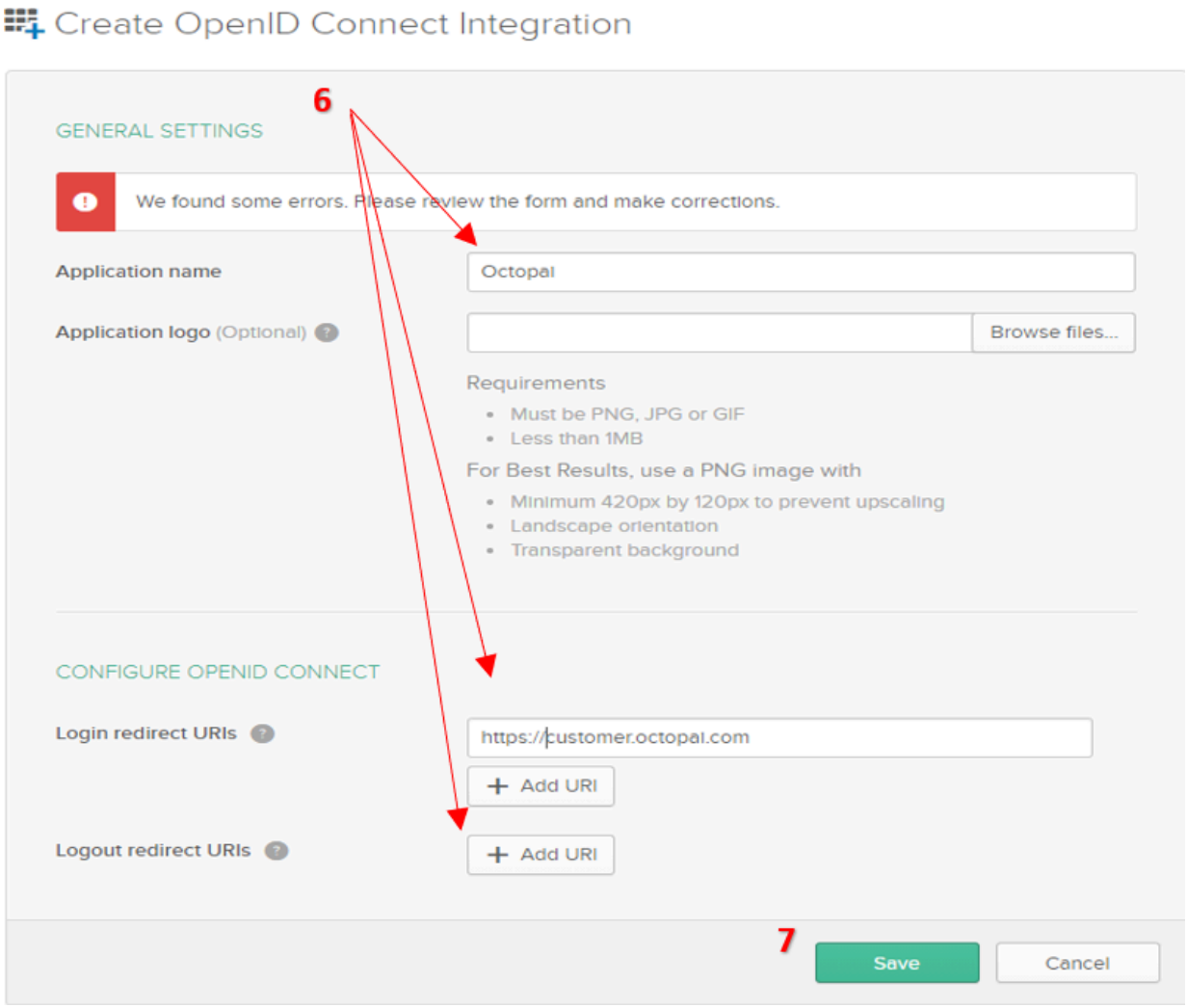
2. Configure the Application Integration.
   a) The **Create New Application Integration** window opens up.
   b) Select the OpenID Connect sign-on method.



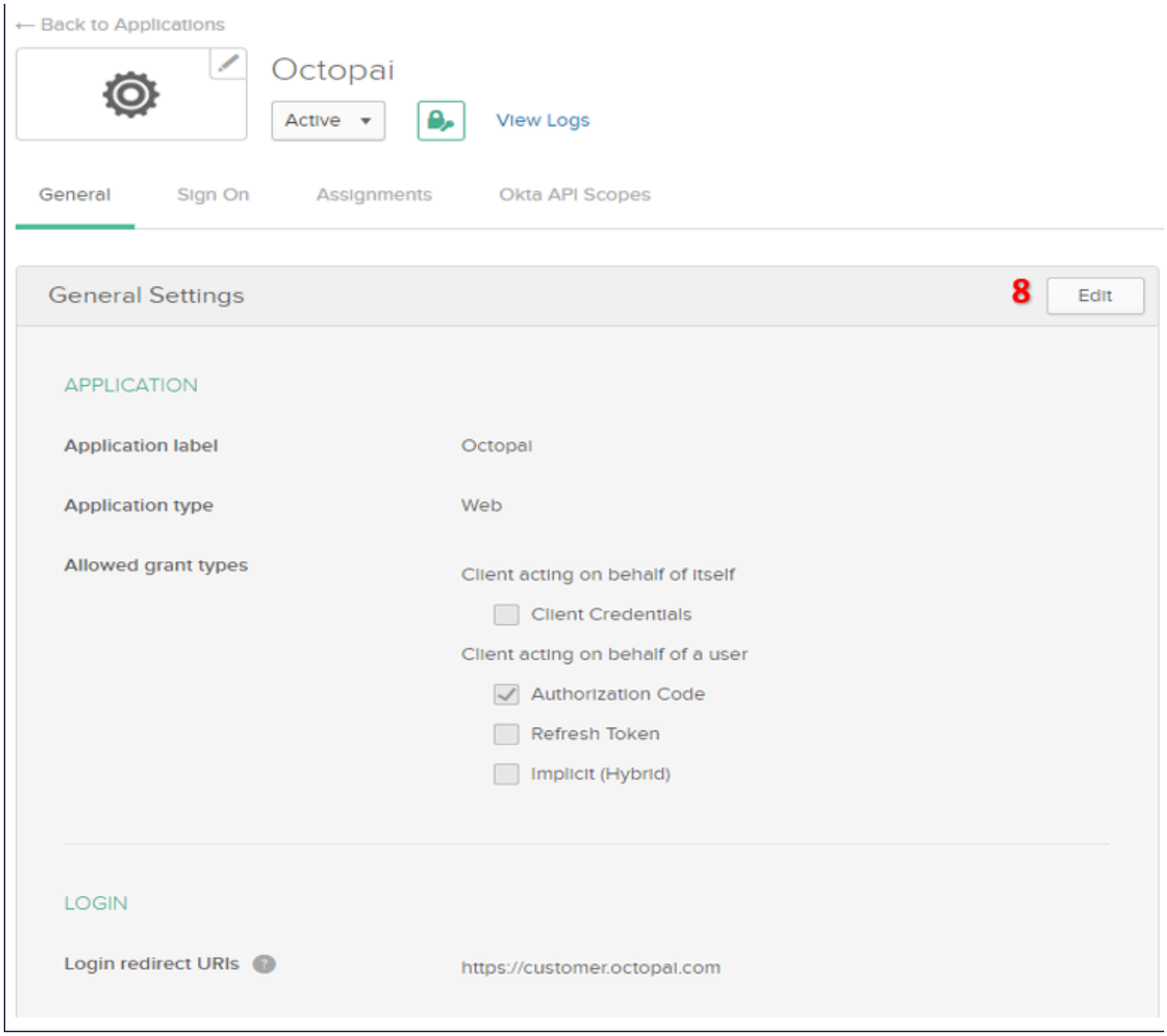
3. Fill in the Application parameters.
   a) Fill all fields in the **Create OpenID Connect Integration** window.

   Note: The values of the Login redirect URIs and Logout redirect URIs fields must be the same as your Cloudera Octopai application URL.

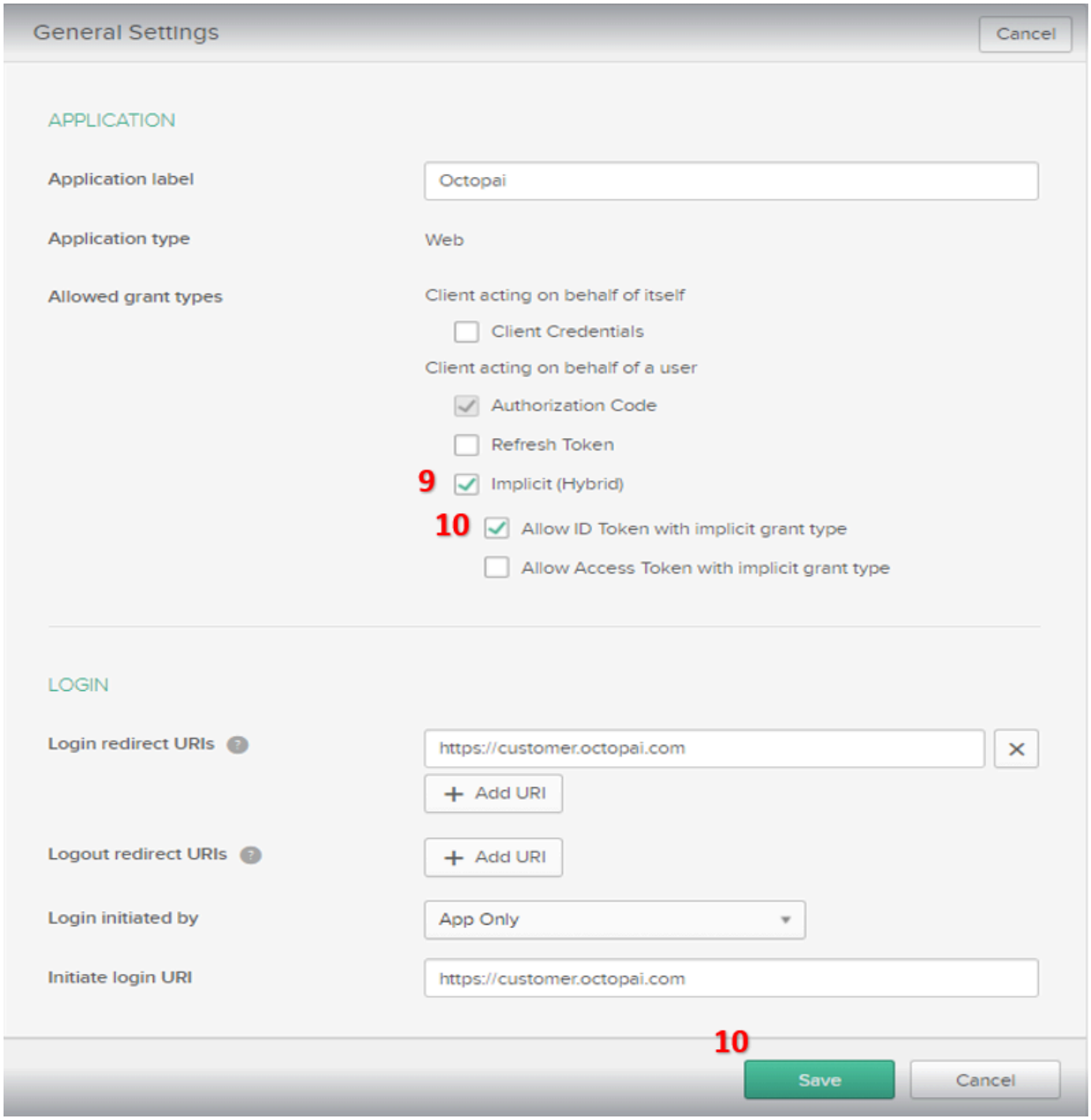
   b) Click Save to complete the creation of the new application connection.



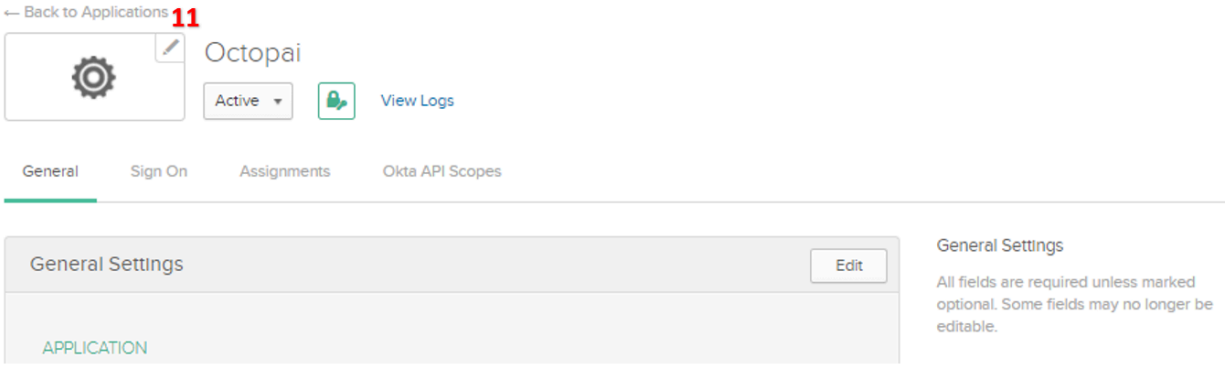
4. Click Edit in the **General Settings** to edit the application settings.



5. Configure the Grant Type.
   a) Select the Implicit (Hybrid) checkbox for the **Allowed grant type**.
   b) Select the Allow ID Token with implicit grant type checkbox.
   c) Click Save.



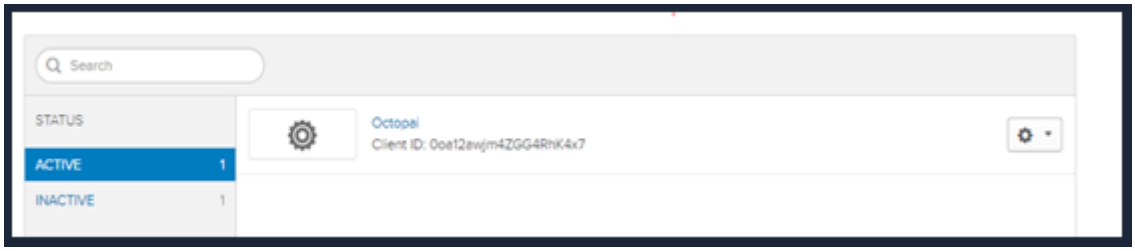
6. Click # Back to Applications to return to the Applications.



What to do next

As a final step, send your Cloudera Octopai representative the following details to add OKTA SSO to your Cloudera Octopai application authentication:

- Client ID. For more information, see Authorization | Okta



- OKTA Redirect URI. For more information, see Authorization | Okta
- OKTA Issuer. For more information, see Issuer
- SSO well-known (Metadata URI). This can be found in Security API Authorization server Edit Settings Metadata URI .

Note: After you submit your settings information, the configuration will be finalized by the Cloudera Support. Only once this is done, the SSO will be ready to use.

Automating Cloudera Octopai Data Lineage metadata extractions with Microsoft Task Scheduler

Learn about automating Cloudera Octopai Data Lineage metadata extractions using Microsoft Task Scheduler on Windows, including task creation, configuration, and scheduling to streamline metadata updates efficiently.

Before you begin

Automating metadata extractions in Cloudera Octopai Data Lineage can greatly streamline data management processes. By utilizing Microsoft Task Scheduler, you can schedule the execution of the Cloudera Octopai BAT files, which are located by default in the C:\Program Files (x86)\Octopai\Service\BAT directory. Scheduling the execution of BAT files simplifies and streamlines the process, ensuring timely and accurate metadata updates. Leverage the power of Task Scheduler to enhance your Cloudera Octopai workflow and improve overall data management efficiency.

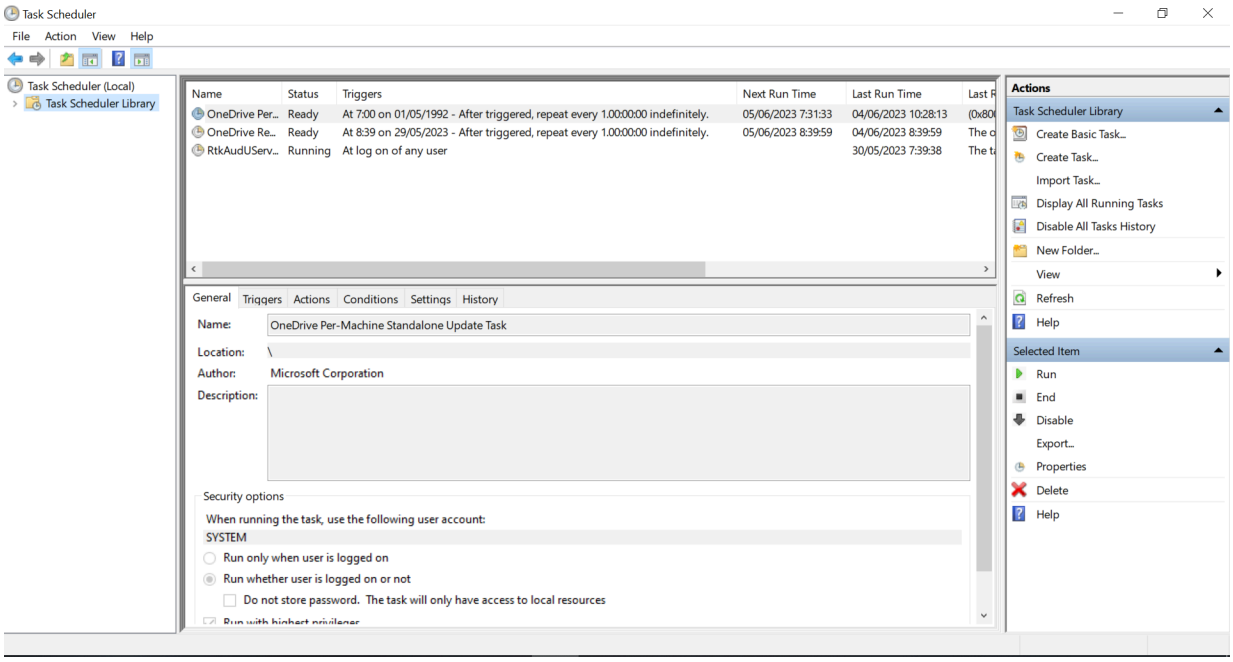
To successfully run the extractions you must have rights to log on as a batch job. For more information, see Log on as a batch job.

Procedure

1. Launch the Microsoft Task Scheduler application on your Windows computer.

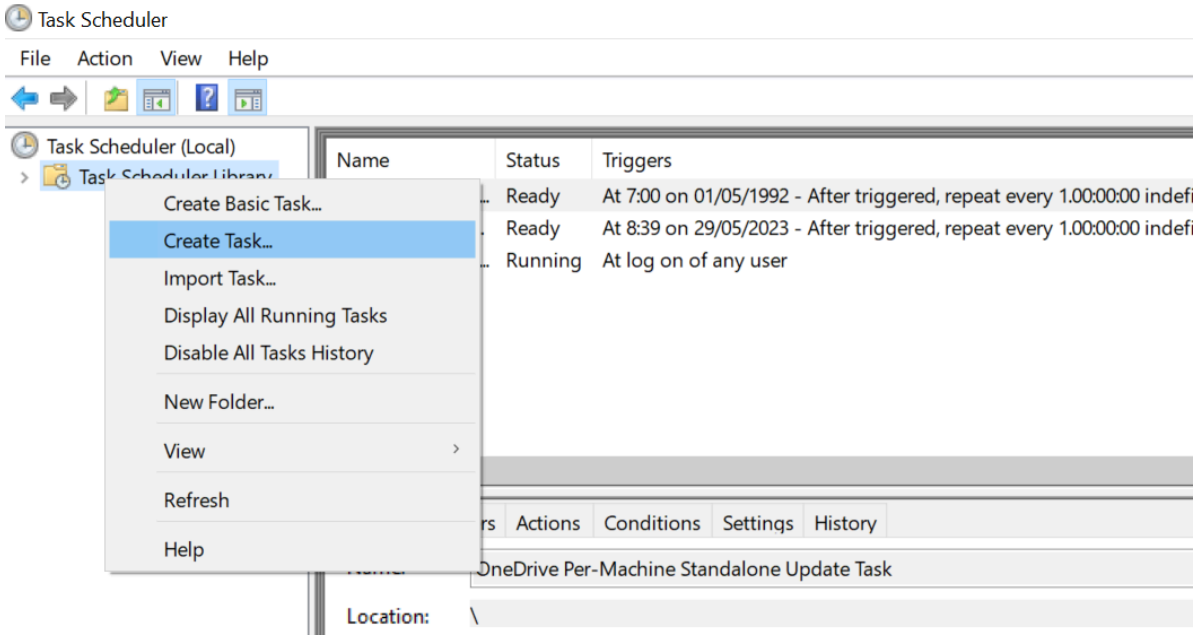
   It can typically be found by searching for Task Scheduler in the **Start** menu.

Figure 4: Microsoft Task Scheduler application



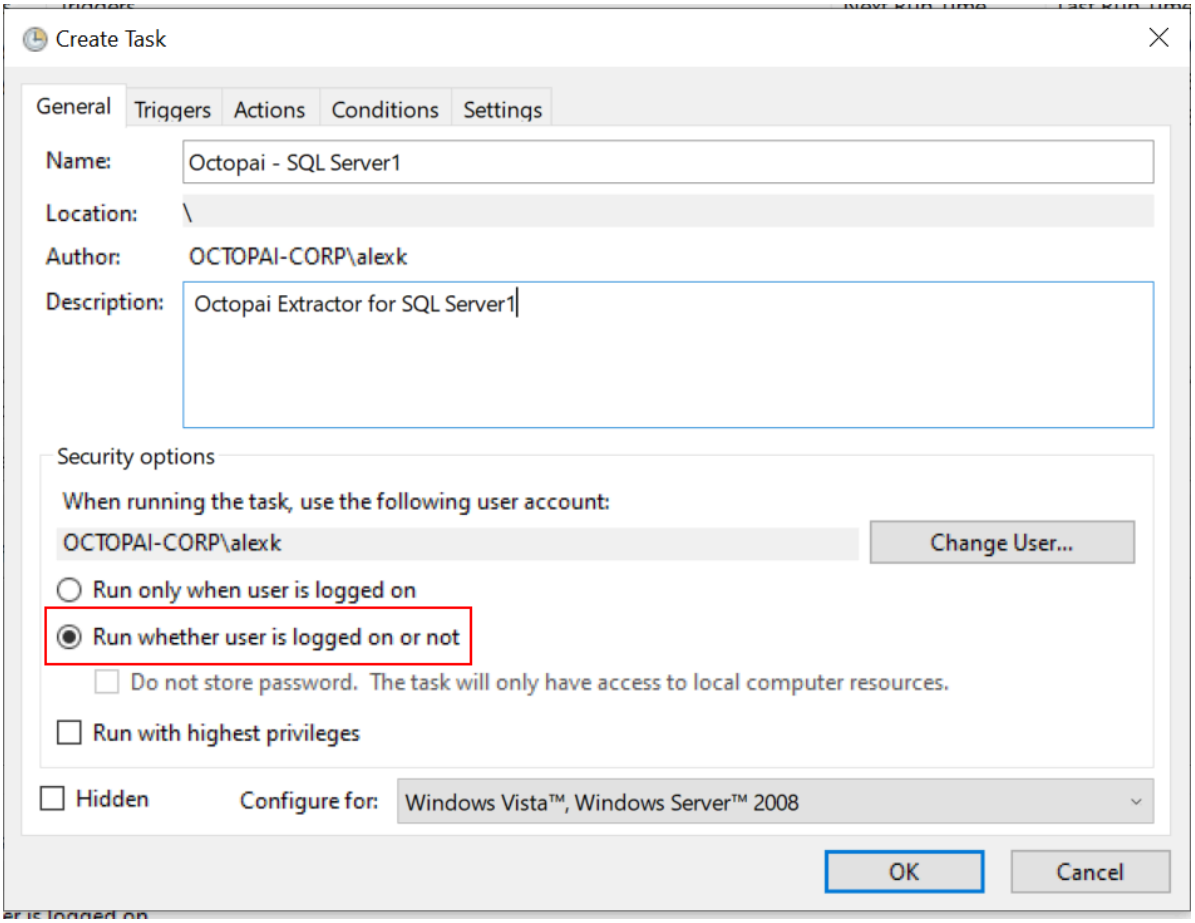
2. Create a new task.
   a) Right-clicking on the **Task Scheduler Library** folder or the desired folder where you want to save the Cloudera Octopai schedule task and selecting the Create Task... action.

Figure 5: Create a task



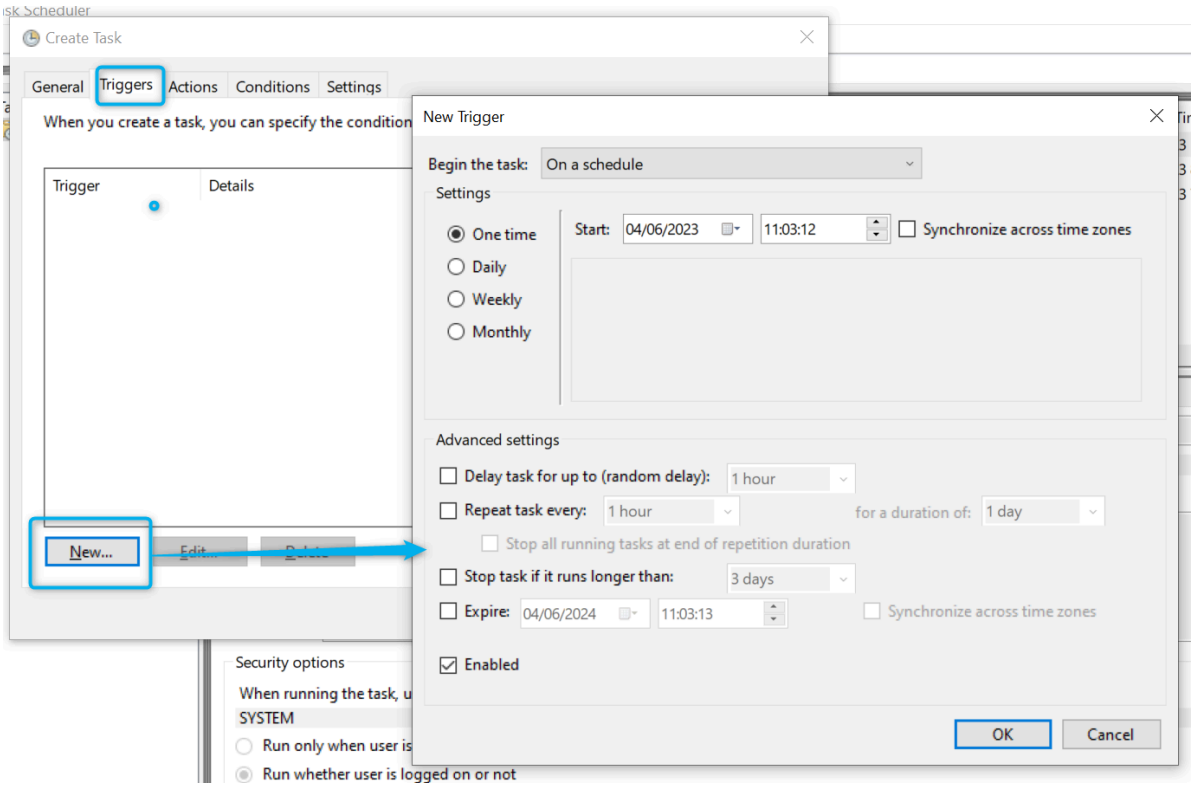
   b) In the **General** tab of the **Create Task** window, provide the necessary details.

Figure 6: Create a task General tab



c) In the **Triggers** tab of the **Create Task** window, click New to open the **New Trigger** window and define the schedule recurrence. Cloudera Octopai recommends setting it to once a week.

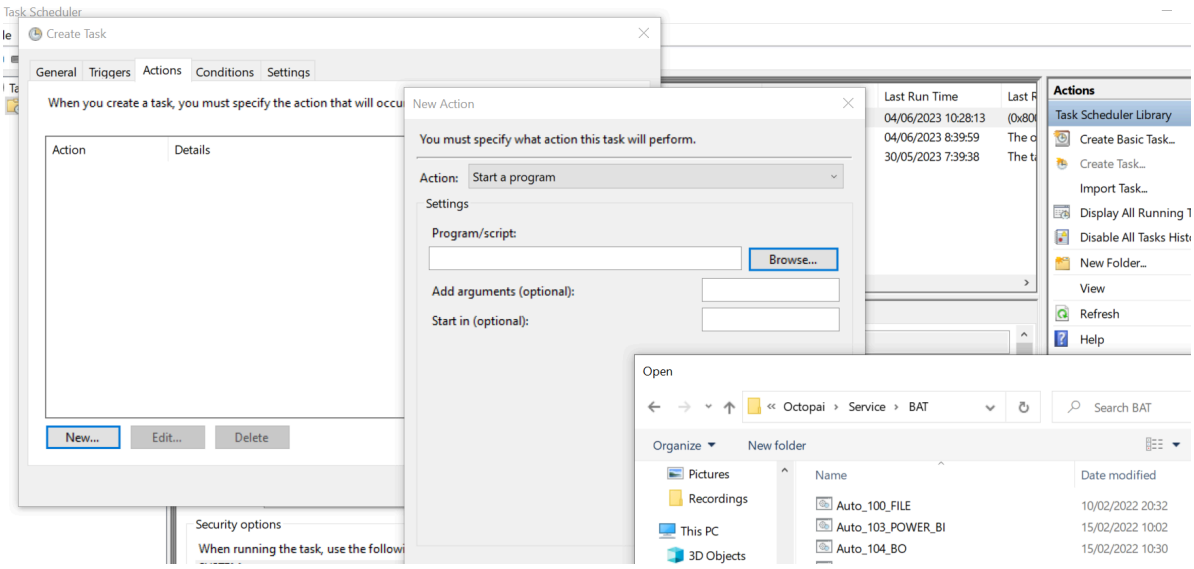
Figure 7: Create a task Triggers tab



d) Configure actions.

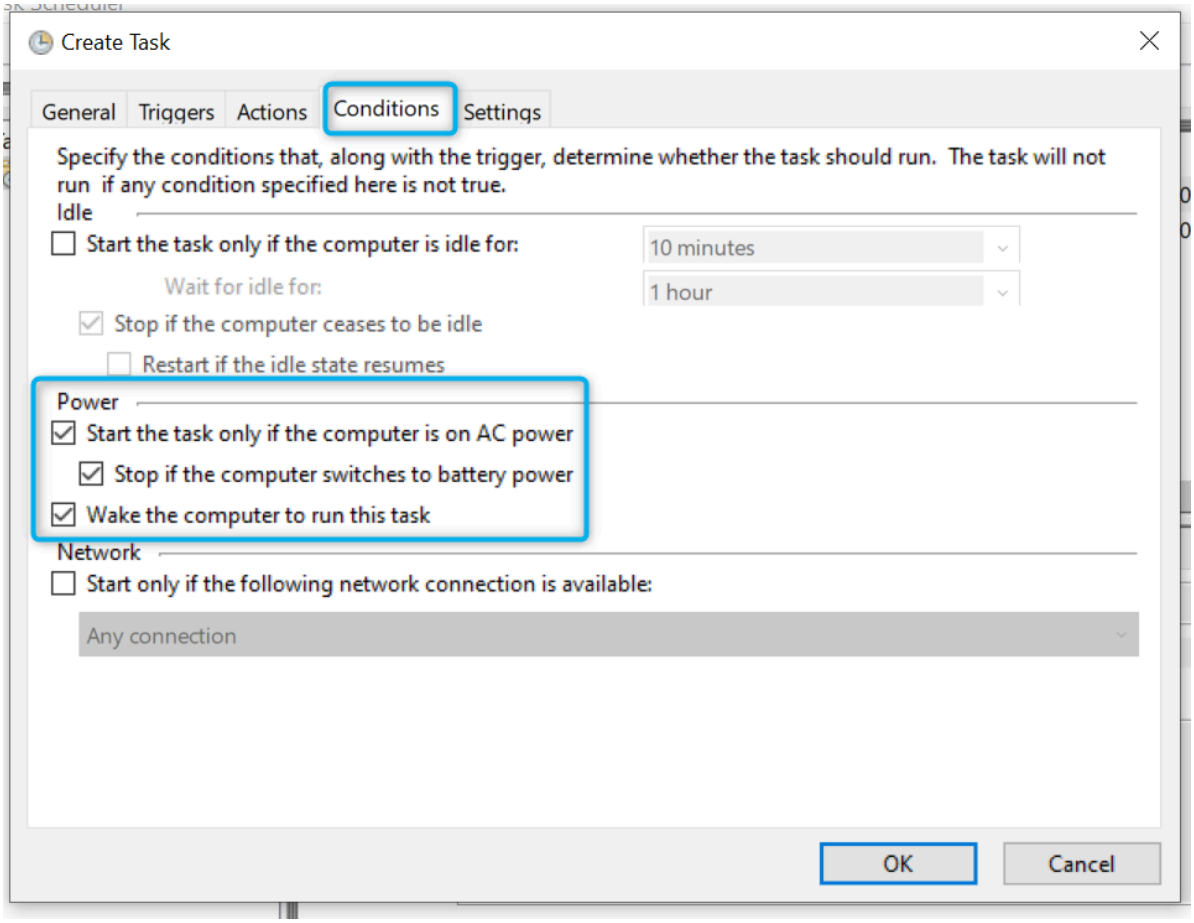
1. In the **Actions** tab of the **Create Task** window, click New to open the **New Action** window.
2. Browse for the relevant BAT file by clicking Browse or manually entering the file path. The default location for BAT files is the folder where Cloudera Octopai Client is installed that is C:\Program Files (x86)\Octopai\Service\BAT by default.
3. Select the file and click OK to save.

Figure 8: Create a task Actions tab



e) In the **Conditions** tab of the **Create Task** window, select all the checkboxes in the **Power** section and go to the **Settings** tab without clicking OK.

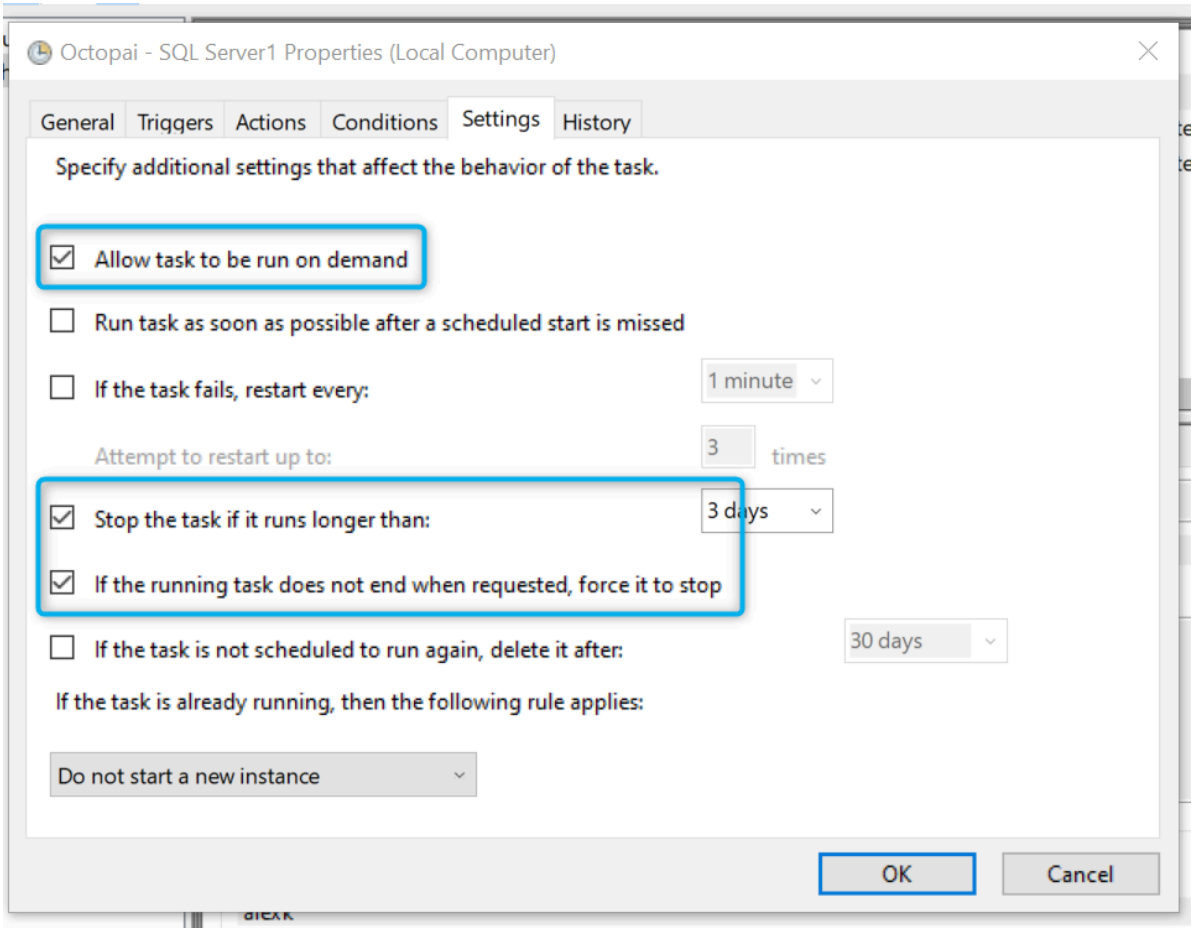
Figure 9: Create a task Conditions tab



f) In the **Settings** tab of the **Create Task** window, select the following checkboxes:

- Allow task to be run on demand
- Stop the task if runs longer than:
- If the running task does not end when requested, force it to stop

Figure 10: Create a task Settings tab



g) Click OK to save the task and the configurations.

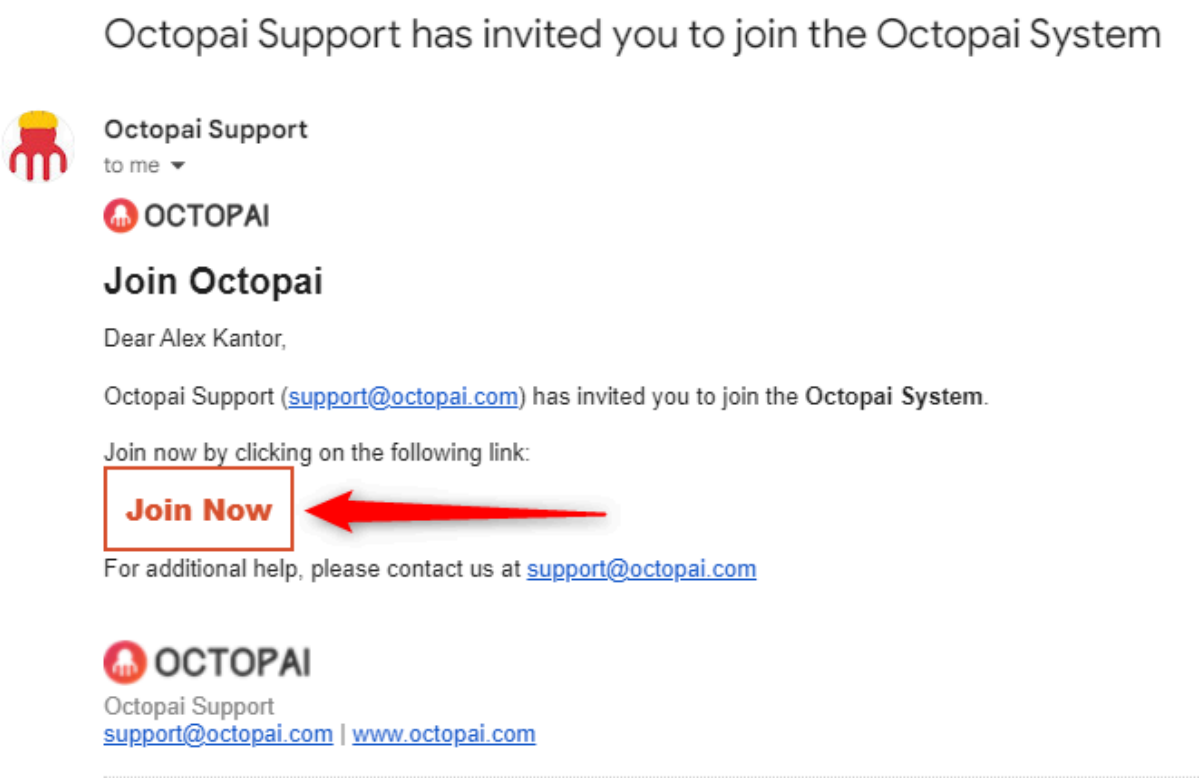
Product guides

Signing up without SSO for the first time

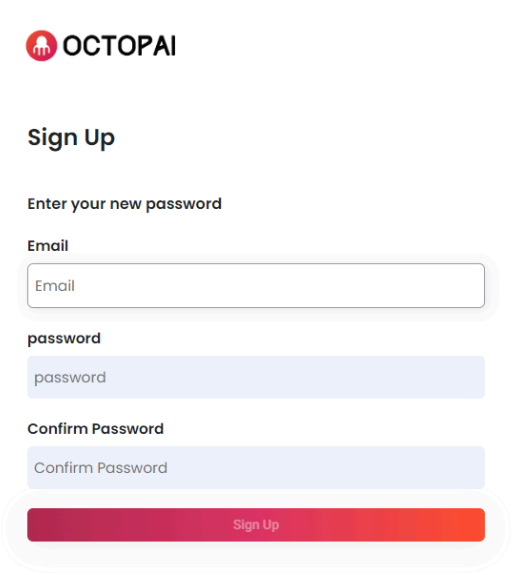
Learn about signing in for the first time to Cloudera Octopai Data Lineage to access the comprehensive data management capabilities of the platform.

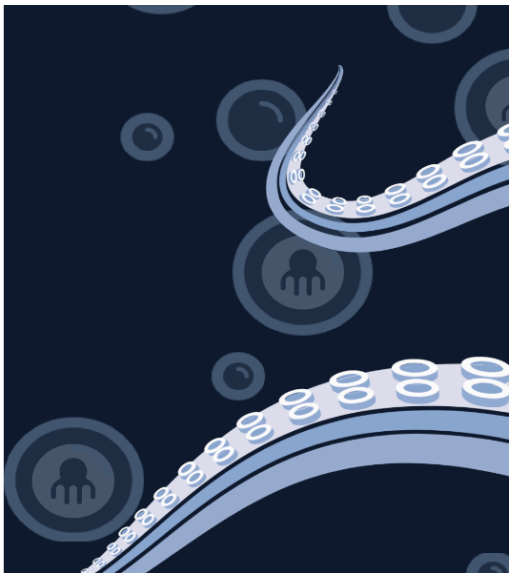
Procedure

1. Click Join Now in the Cloudera Octopai invite mail you received.



2. The **Sign Up** window displays in your default browser. Chrome or Edge is supported.





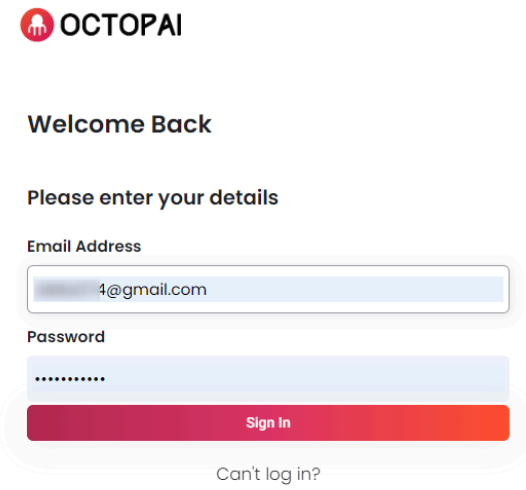
   a. Create your own Password.

      A password must have the following characteristics:

      - Have at least 8 characters
      - Have at least one letter
      - Have at least one number
      - Include both uppercase and lowercase characters
      - The password must NOT contain 4 consecutive characters, for example 11111, 12345, abcde, or qwert.

   b. Click Sign Up to get the **Login** window.

   c. Fill in your data and click Sign In.



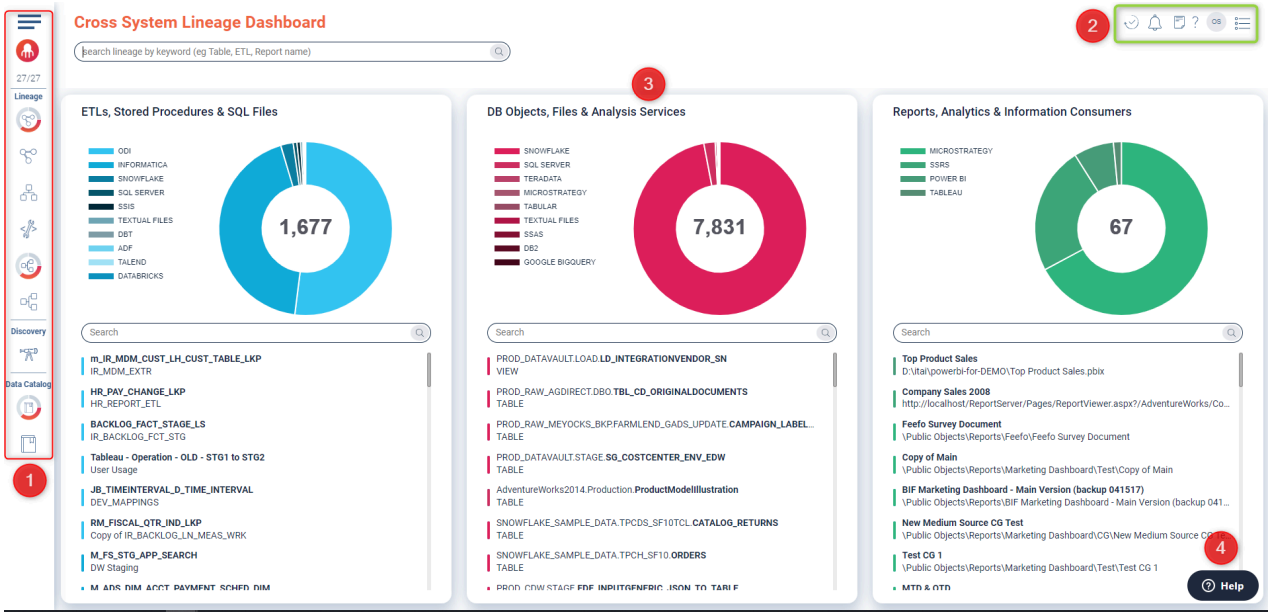
Cloudera Octopai Data Lineage User Guide

Tour the Cloudera Octopai Data Lineage web platform interface, navigation areas, and key tools available to administrators and analysts.

Cloudera Octopai web platform

The Cloudera Octopai web platform provides central access to lineage, discovery, and knowledge features for data teams.

Figure 11: Cloudera Octopai web platform home page



Section 1 - Spaces menu

Select the menu icon to expand the Spaces menu and view detailed navigation options.

The left menu is divided into the following spaces:

1. **Lineage Space**
   - Cross System Lineage
   - Inner System Lineage
   - Live Lineage
   - E2E Column Lineage
2. **Discovery Space**

   Octopai Automated Discovery Space
3. **Knowledge Hub Space**

   Knowledge Hub User Guide

Section 2 - General options

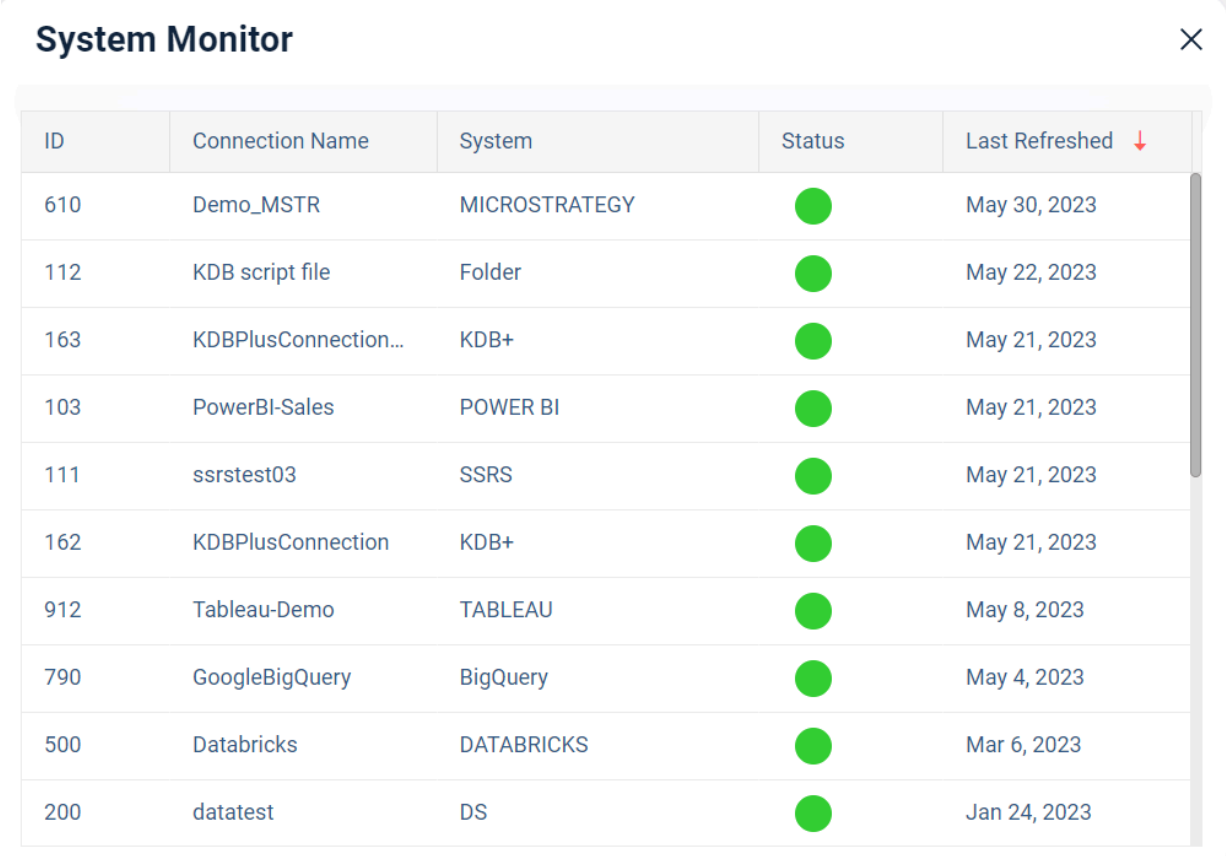
The toolbar across the top of the interface provides quick access to monitoring, notifications, and account tools.

You can choose the following options in the **General options** section:

- **System monitor**

  The **System Monitor** pop-up window lists connectors by number, name, system, status, and last refreshed date.

  Figure 12: Connector status summary



  Review connector status details to confirm ingestion health before refreshing metadata.

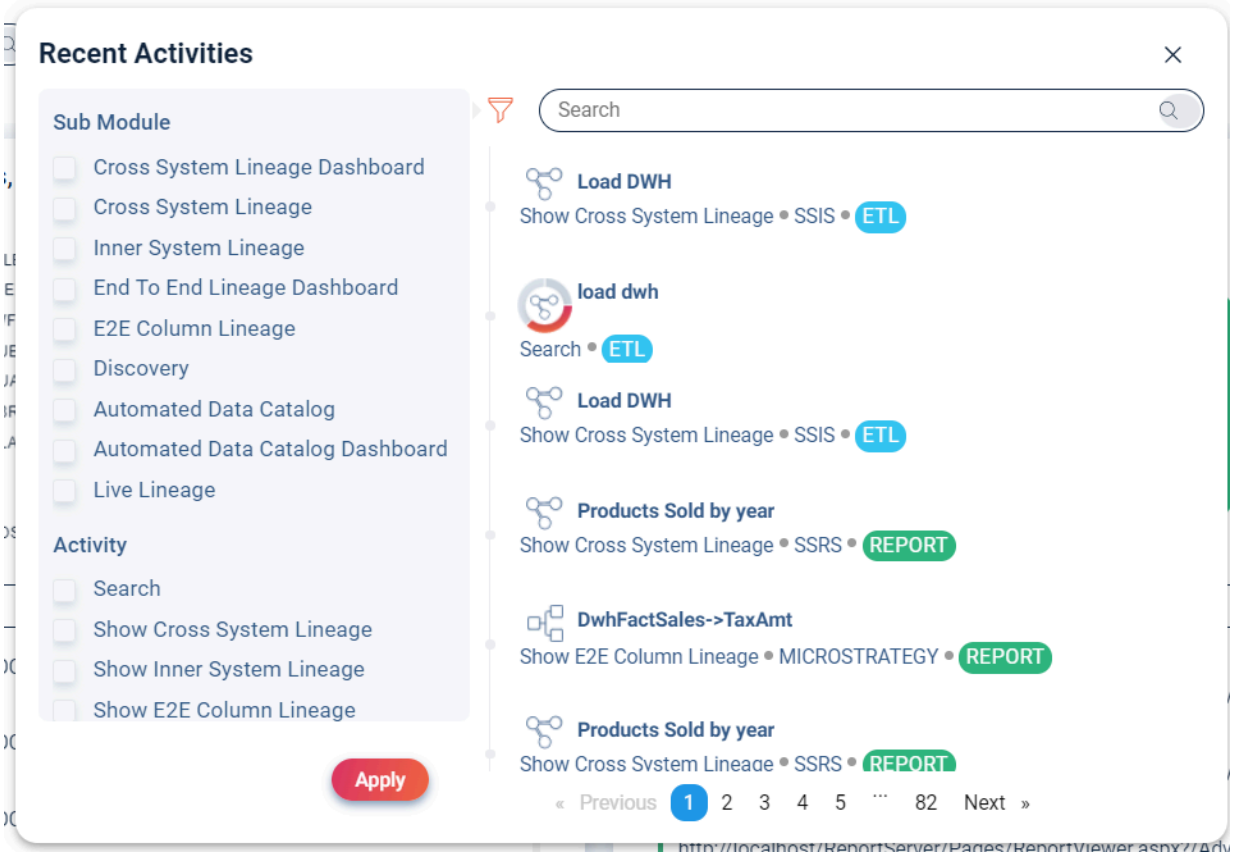
- **User notifications**

  User notifications display license information and other account-wide alerts.

- **Recent activities**

  The **Recent Activities** list identifies the most recent searches performed in the environment. Use the red funnel icon to refine the list by submodule, activity, tool name, or tool type.

  Figure 13: Recent activity filter options



- **Knowledge Center access**

  Select Need Help? to open the Cloudera Octopai Knowledge Center.

- **User avatar options**

  Figure 14: User avatar menu

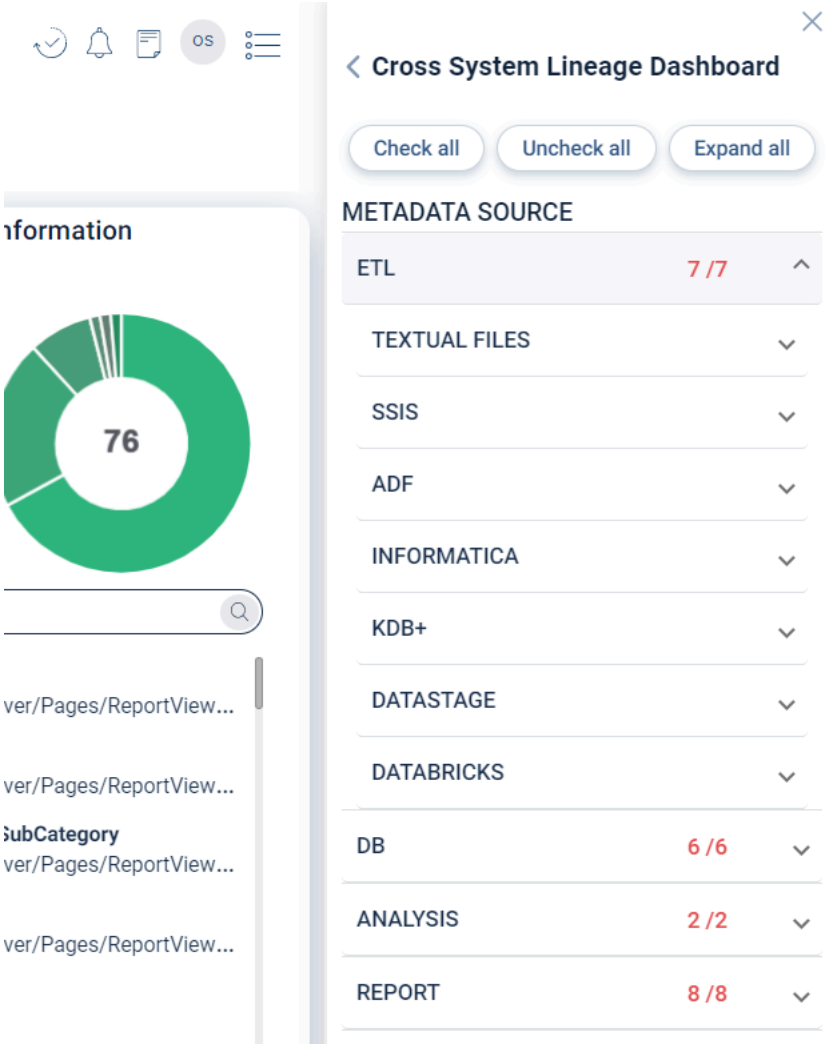
  The user avatar menu provides shortcuts to profile settings. Administrators can open the **Admin Console**, and all users can sign out from the logout option.

- **Metadata connections**

Figure 15: Metadata connections panel

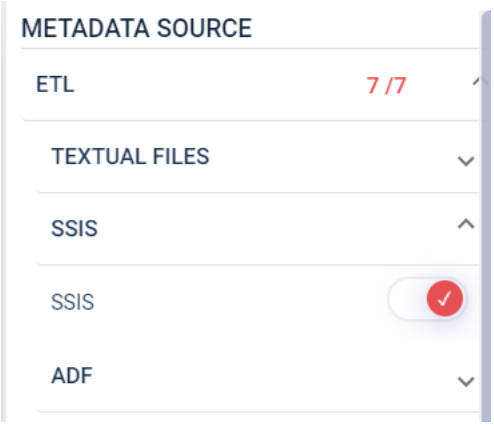
The Metadata Connections panel lists every metadata source extracted by tool type.

Figure 16: Metadata source controls



Toggle each metadata source on or off to control which systems appear in lineage views.

Figure 17: Cross System Lineage toggle



Note:

The Metadata Connections control expands only when you are on the Cross System Lineage dashboard.

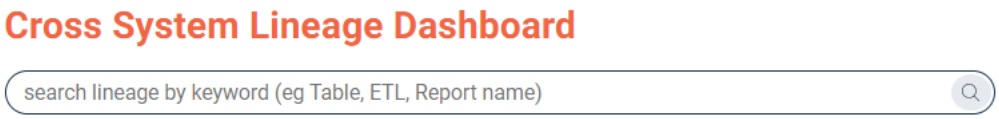
Section 3 - Main pane

The main pane shows the current dashboard with search, lineage objects, and color-coded visual cues in the following sections:

- **Search box**

  The Search box supports fuzzy searches across report and process names (columns are not included).

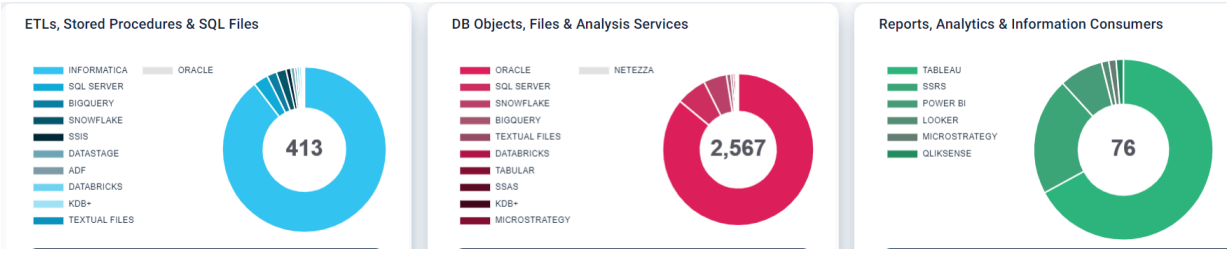
  Figure 18: Search box controls



- **Data objects list by type**

  The **Data Objects List by Type & Color** option groups assets by the area they belong to.

  Figure 19: Data object legend



  The color legend refers to the following elements:

  - Blue area: ETL, stored procedures, and SQL files.
  - Red area: Database objects, Analysis Services items, and physical objects.
  - Green area: Reports.

Section 4 - Help button

Select the help icon to contact Cloudera Support.

The help resources open in a separate browser tab so you can continue working in the platform.

Admin User - Creating users for Cloudera Octopai Portal

Learn about creating your user for the Cloudera Octopai Portal using CSM.

About this task

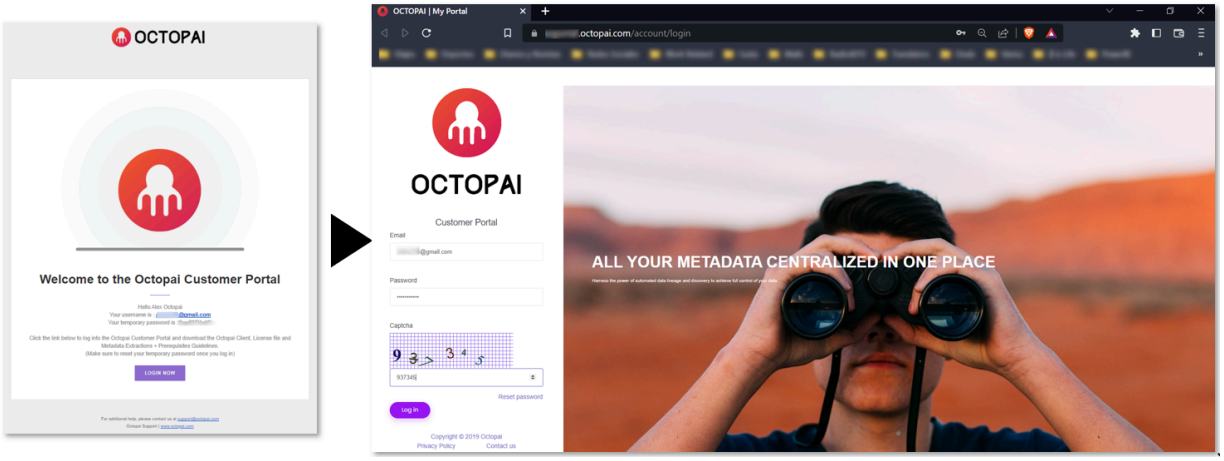
Note: In Cloudera Octopai Portal, you will find the Cloudera Octopai Client, licenses, and files that have been last extracted. The Portal can also be used to exchange large files, always in a zip format.

For the CSM to create your user for the Cloudera Octopai Portal, perform the following steps:

Procedure

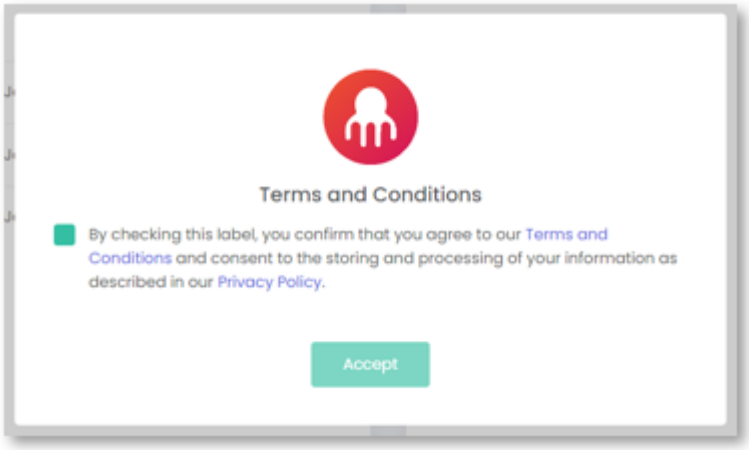
1. Click the Login Now link in the email you received from Cloudera Octopai Data Lineage.
2. Log in to the Cloudera Octopai Portal.

Figure 20: Cloudera Octopai Portal login page



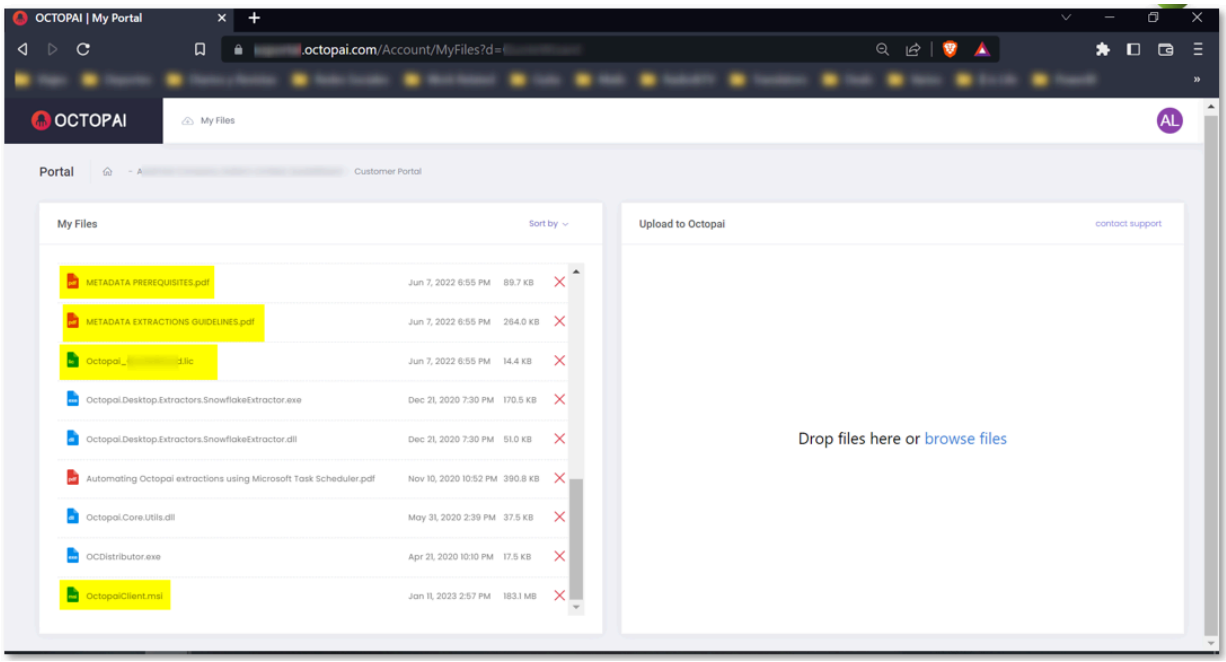
3. Accept the Terms and Conditions.

Figure 21: Terms and conditions modal window



4. Access you local files in the **My Files** section. The following files are your main files:
   - Metadata Prerequisites
   - Metadata Extractions Guidelines
   - Cloudera Octopai License - OC (*.lic)
   - Cloudera Octopai Client (*.msi)

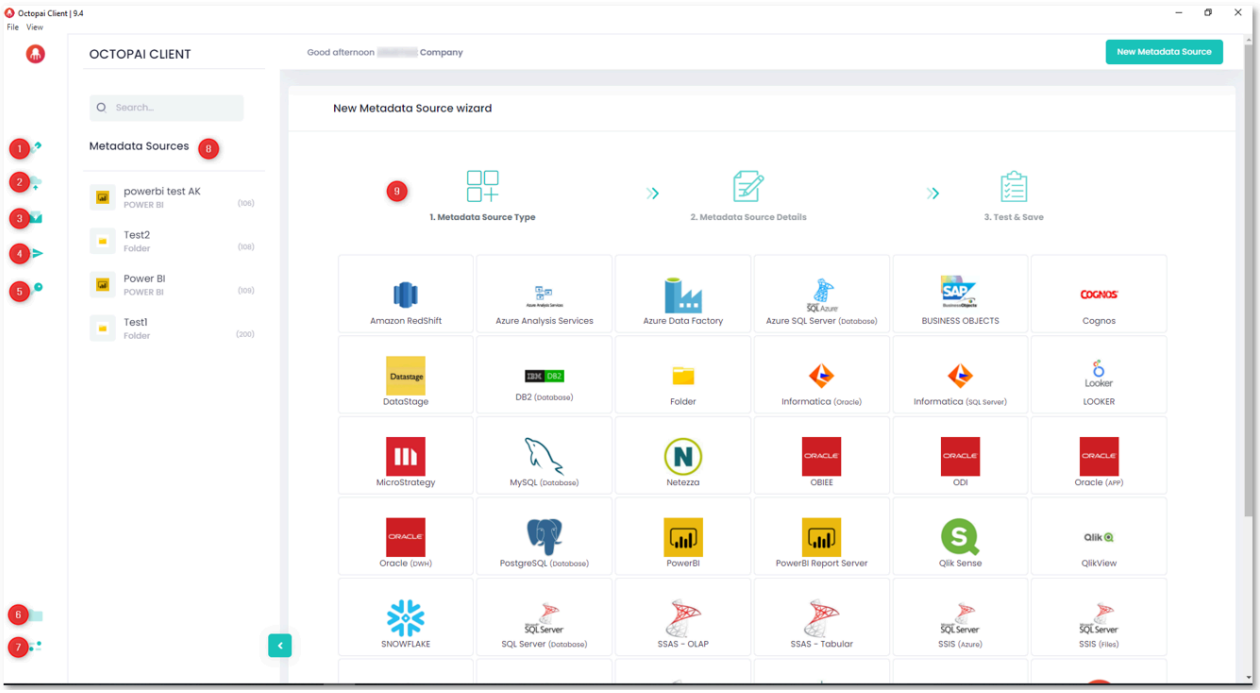
Figure 22: Cloudera Octopai Portal My Files section



5. Upload your files to Cloudera Octopai. Only files in zip format are accepted.

Admin User - Cloudera Octopai Client

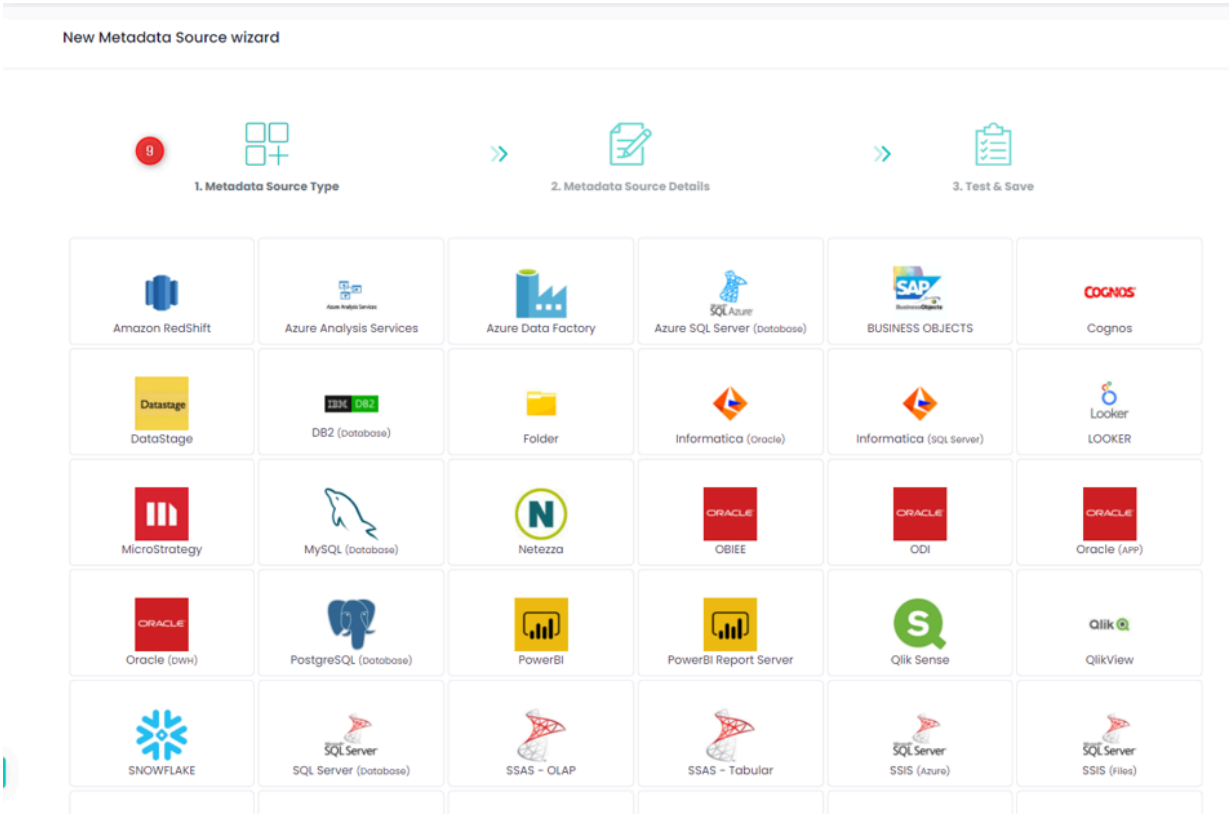
With the Cloudera Octopai Client, you are able to extract metadata from your tools and upload them to the Cloudera Octopai Cloud.



The Cloudera Octopai Client application consists of the following sections:

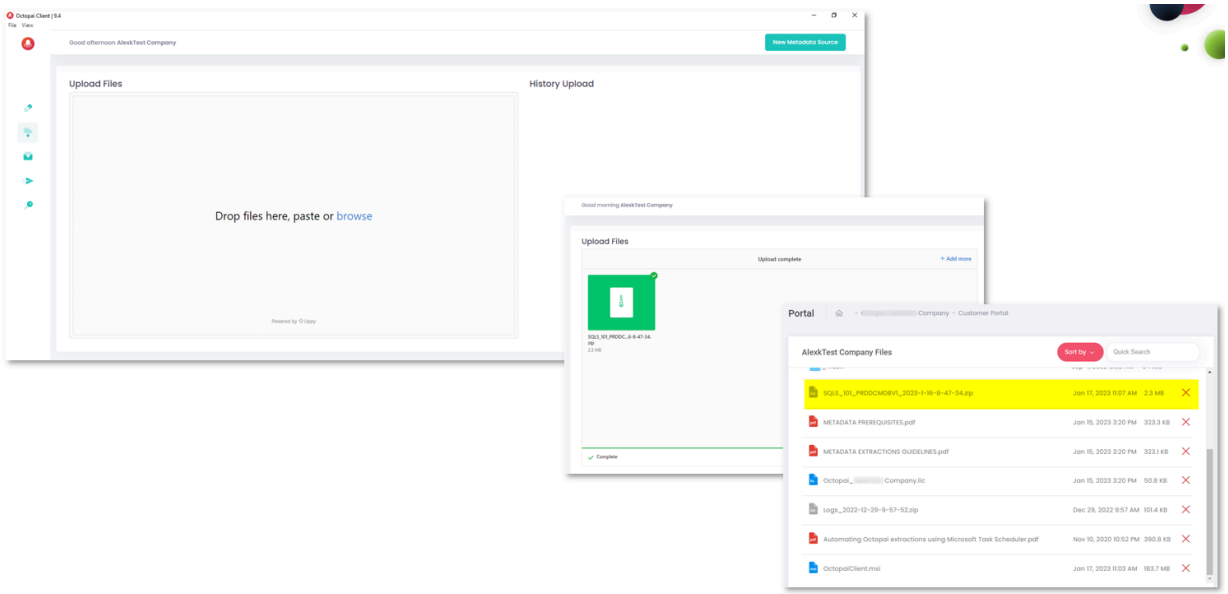
1. **Metadata Sources**

   The main panel contains the out-of-the-box connections in the purchased license.



2. **Upload Files to Cloudera Octopai**

   Upload files to Cloudera Octopai. Only files in zip format are accepted. When the upload is completed, the files are displayed on your Portal.

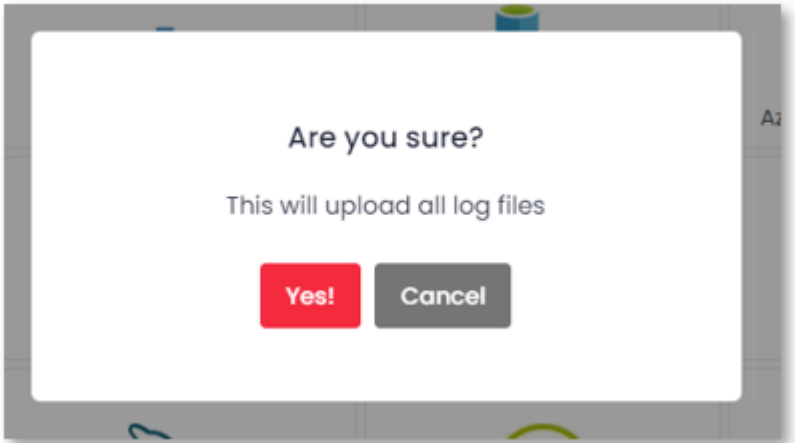


3. **Contact Support**

   If you need assistance, contact Cloudera Support.

4. **Send Logs to Cloudera**

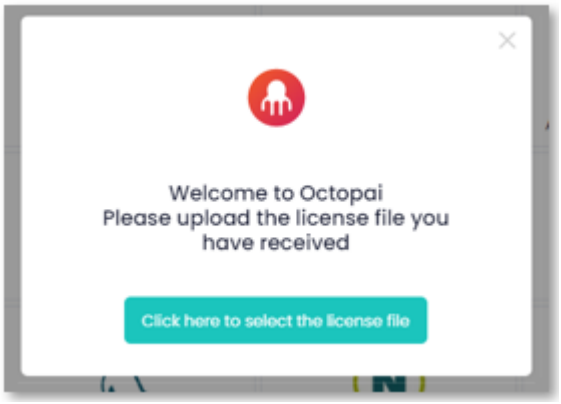
   Logs are helpful to solve issues. Click the icon to send the last logs to Cloudera to be analyzed by the Support team.



5. **Update License**

   After opening the Cloudera Octopai Client for the first time, or after a Version Upgrade, a window to install the license pops up.

   Click Click to select the license file to browse for the .lic file and click Open to install the license.

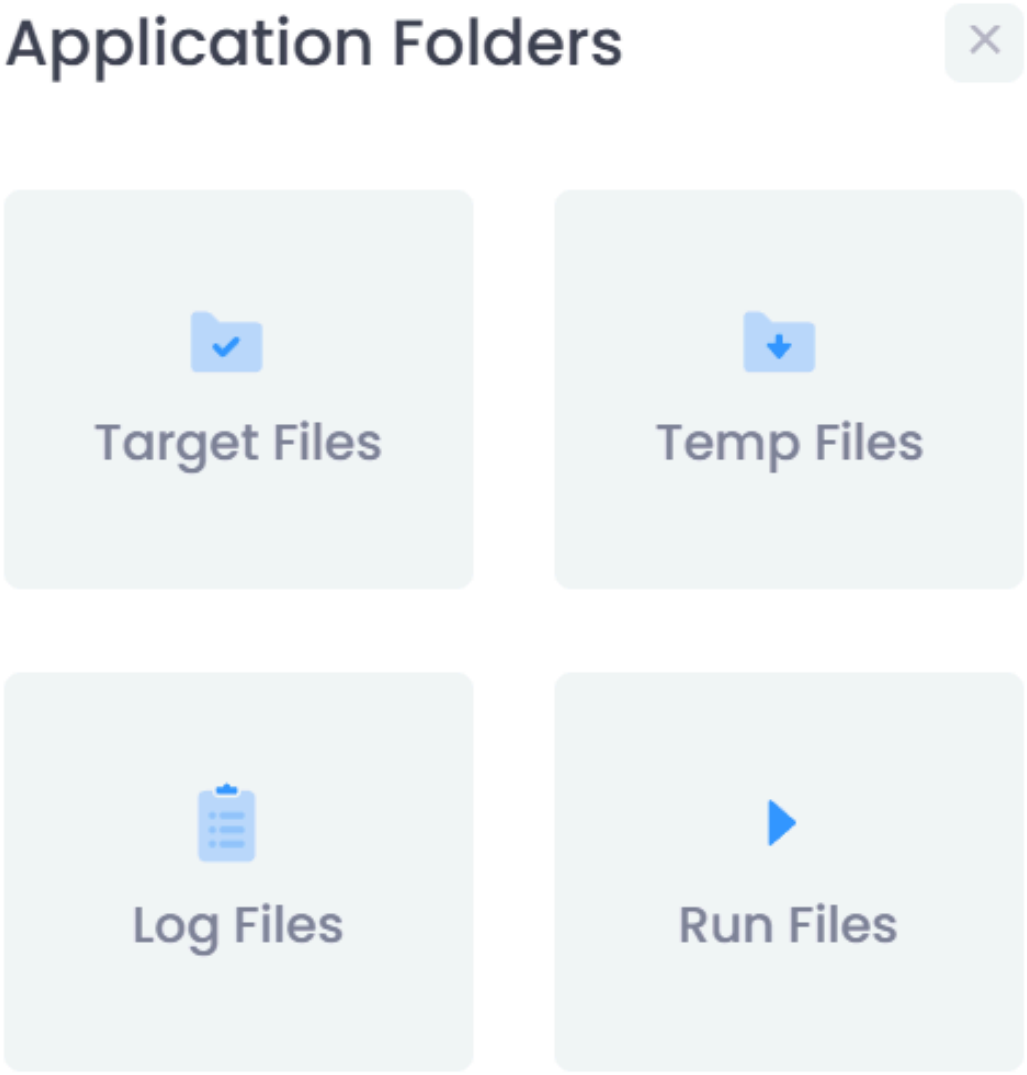


6. **Application Folders**

   The Cloudera Octopai application folders play a crucial role in managing metadata extraction processes and ensuring the quality and integrity of metadata. By providing a structured storage environment, these folders enable

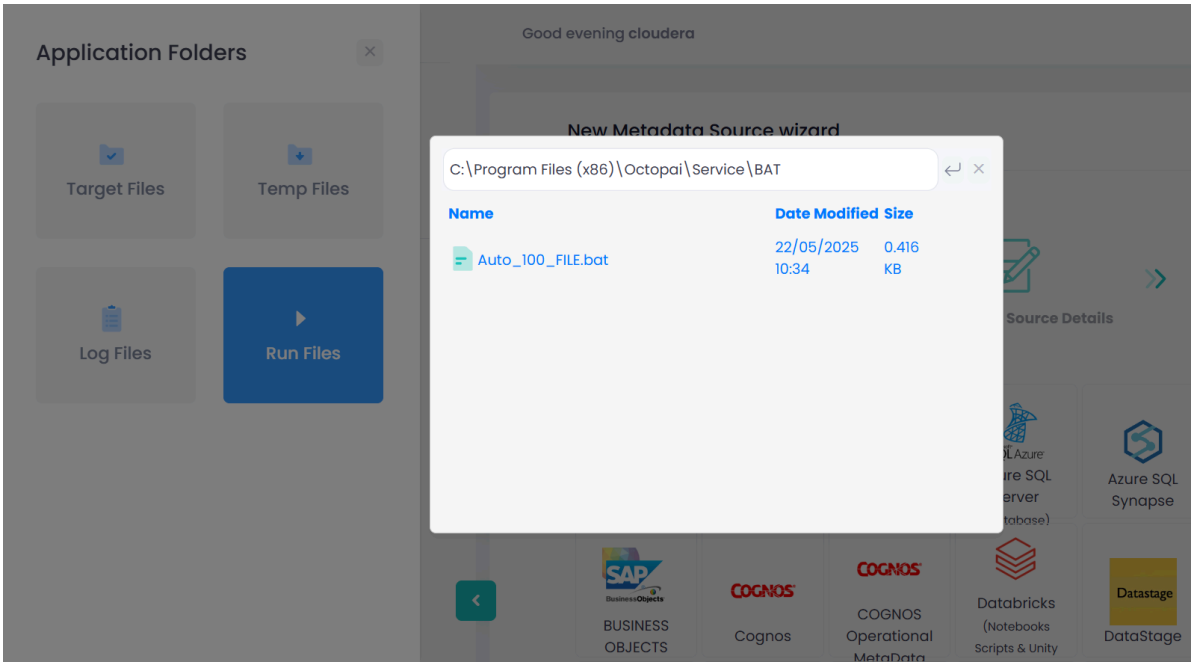
users to verify metadata accuracy and address any potential issues. You can find the following folders in Cloudera Octopai Client:

- Target Files – The Target Files folder serves as a repository for the zip files uploaded to Cloudera Octopai.
- Temp Files – During the metadata extraction process, Cloudera Octopai creates temporary files to facilitate data processing and transformation.
- Log Files – Cloudera Octopai generates log files during the metadata extraction process to capture valuable information about the extraction activities.
- Run Files – To automate the metadata extraction process, Cloudera Octopai provides Run (bat/sh) files that can be used in conjunction with Microsoft Task Scheduler. For instructions on using Microsoft Task Scheduler, see Automating Octopai metadata extractions with Microsoft Task Scheduler.



a. Click Run Files.

b. Select the respective Batch or shell script file, depending on your operating system, to download to your computer for automation activity. Point your scheduler of choice.


Application Folders
Target Files
Temp Files
Log Files
Run Files
Good evening cloudera
New Metadata Source wizard
C:\Program Files (x86)\Octopai\Service\BAT
Name
Date Modified
Size
Auto_100_FILE.bat
22/05/2025 10:34
0.416 KB
Source Details
Azure SQL Synapse
BUSINESS OBJECTS
Cognos
COGNOS Operational MetaData
Databricks (Notebooks Scripts & Unity
Datastage
DataStage

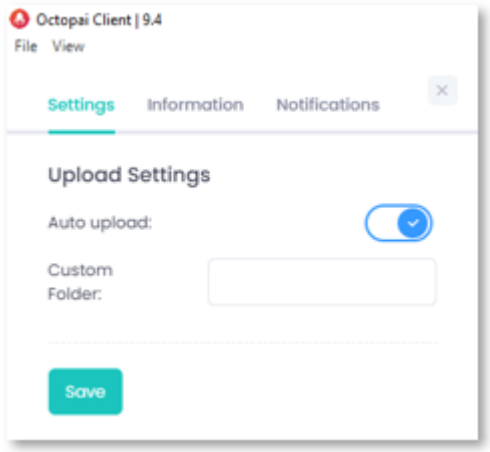
7. **Information and Settings**

   You can see the following tabs in the Information and Settings section:

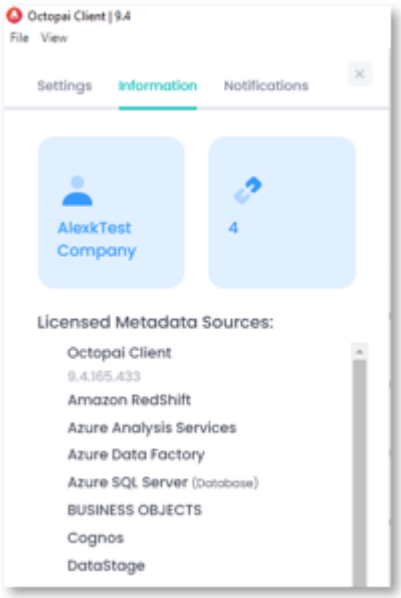
   - The **Settings** tab defines if the extractions are automatically uploaded to the Cloudera Octopai Cloud or are done manually.

     Set the **Auto Upload** to ON to enable the extractions to be automatically transferred to the Cloudera Octopai Cloud.

     Set the **Auto Upload** to OFF to review the files before they are being transferred to the Cloudera Octopai Cloud.

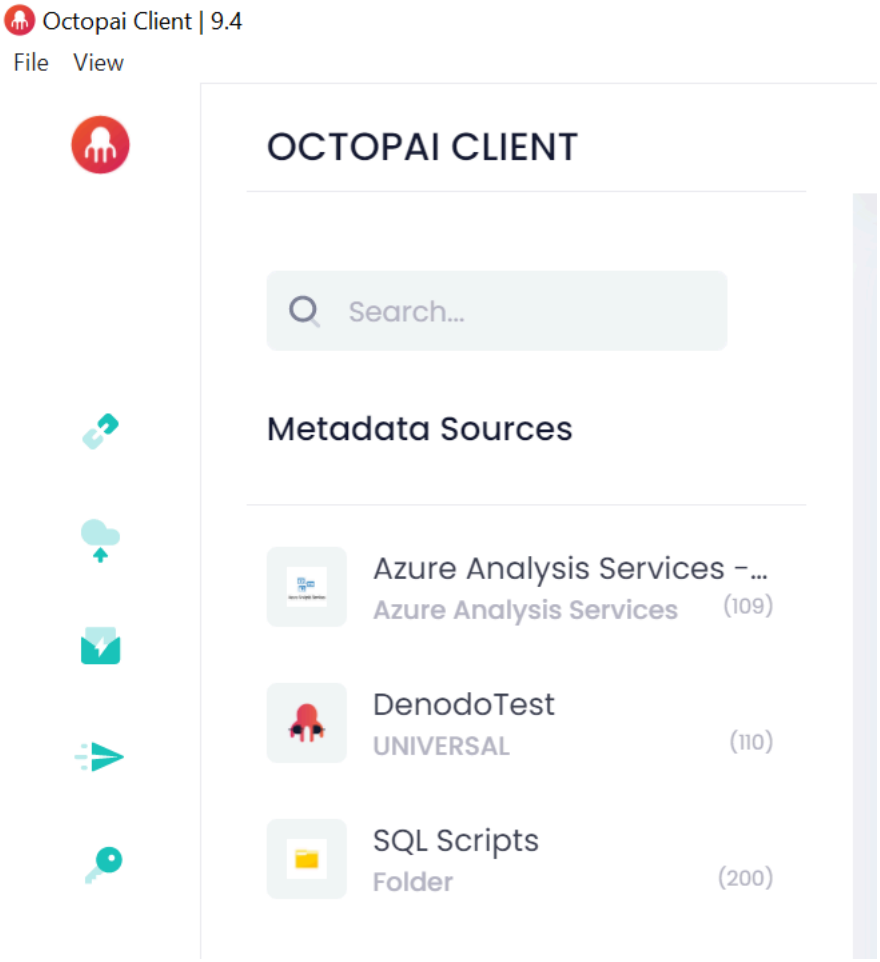


   - The **Information** tab describes the Cloudera Octopai client version installed and the list of tools in the purchased license.



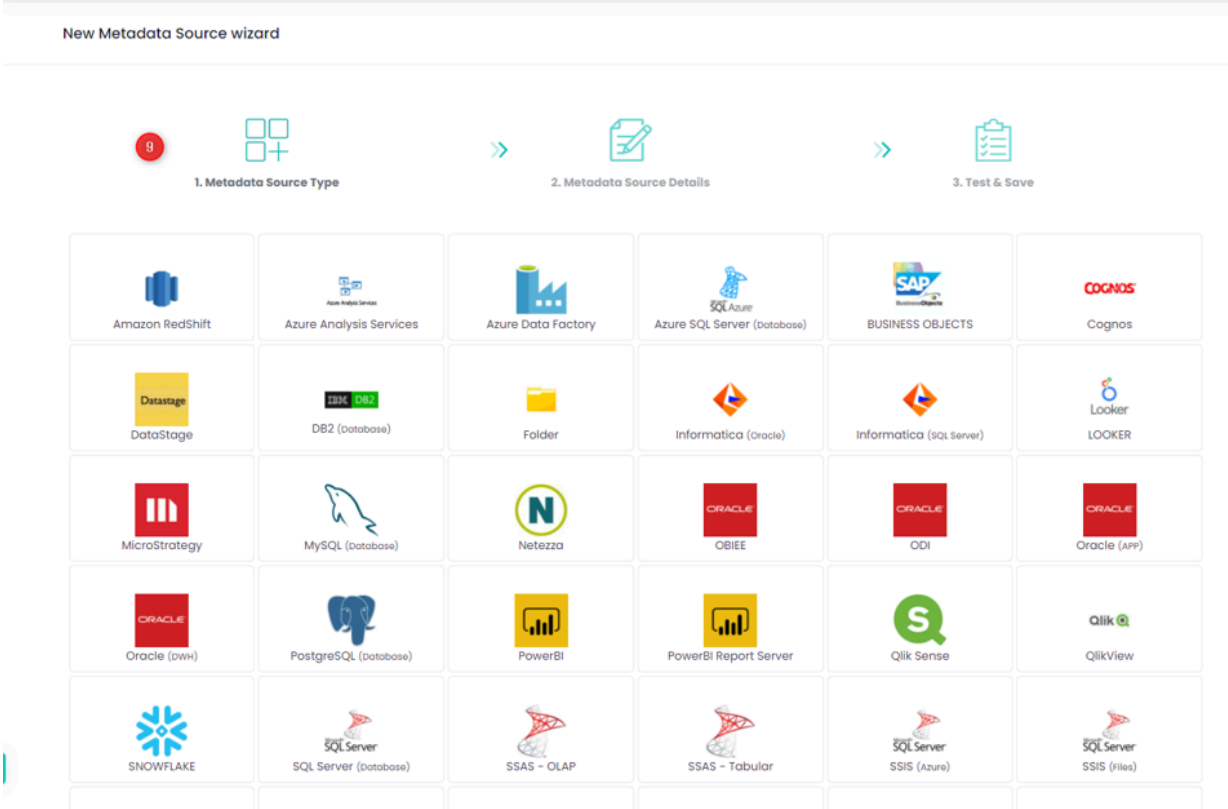
8. **List of Connectors**

If you create a new connector, it is shown on the **Metadata Sources** list or List of Connectors. The new source set gets a Connector Number, a common denomination between you and Cloudera Octopai Support.



9. **Tools List**

   All the tools purchased to create connectors to extract metadata are displayed in the main pane.



Admin User - Cloudera Octopai Admin Console

Admin Console is the central location where administrators can configure settings and manage the platform.

You can access the Admin Console by clicking on the avatar icon. You can leverage Admin Console features to streamline user management, configure system settings, manage metadata, and gain valuable insights.

Figure 23: Access Admin Console

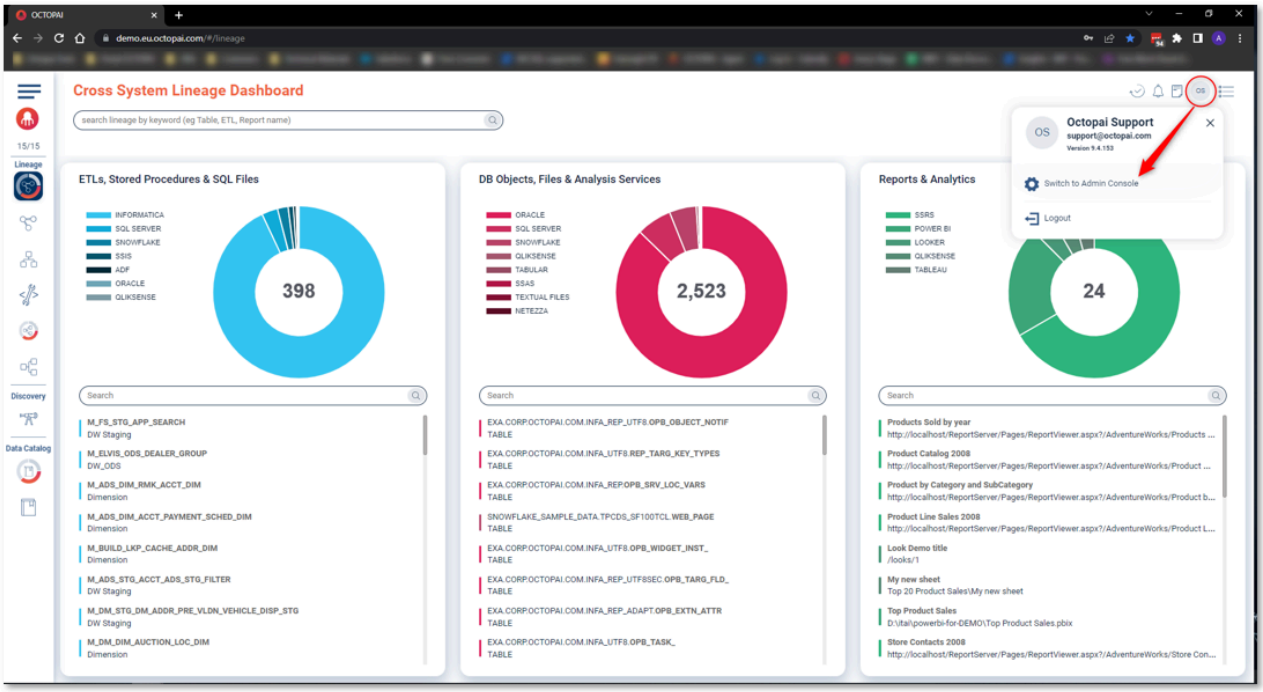
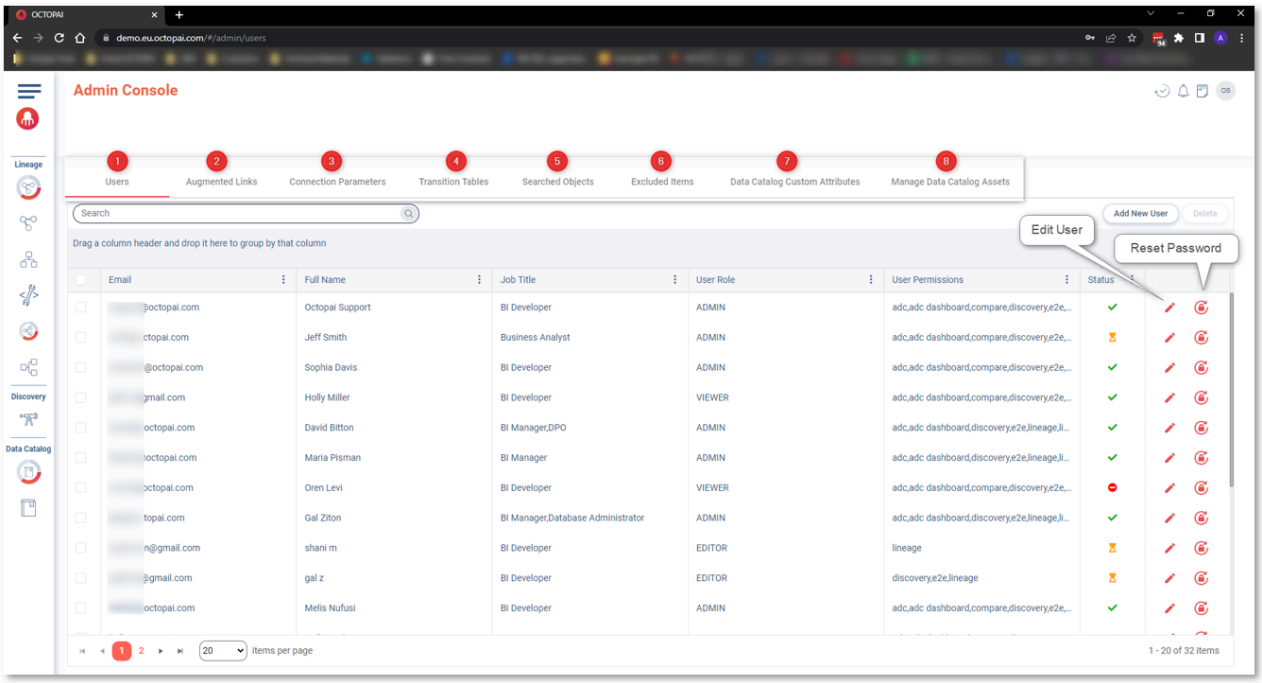


Figure 24: Admin Console sections



Admin Console consist of the following sections:

1. **User Management**

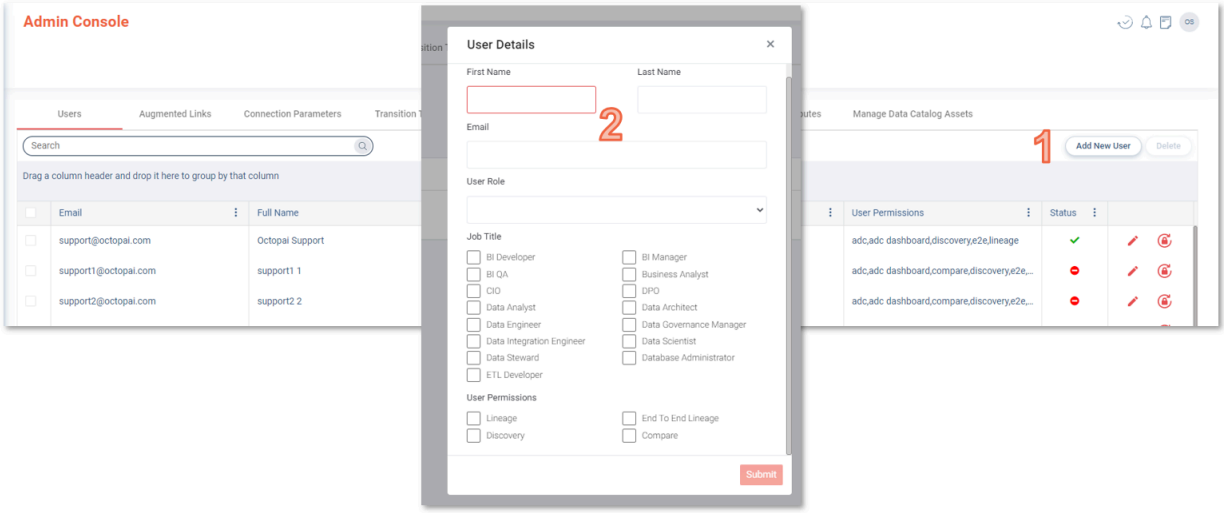
   You can create users manually by performing the following steps:

   a. Click Add a New User.

   b. Fill in the fields in the **User Details** model window. The following fields are mandatory:

   - **First Name**
   - **Last Name**
   - **Email**
   - **User Role** – Can be Admin, Editor, or Viewer.
   - **Job Title** – You can choose only one.

   c. Click Submit.



Users can have the following statuses:

User Deactivated

User Active

Waiting for first Login

The new user will receive an automatic notification email.

2. **Augmented Links**

   Add Database metadata from your systems into Cloudera Octopai Data Lineage to get the full picture, completing the cross system lineage.

Figure 25: Augmented Links tab in Admin Console

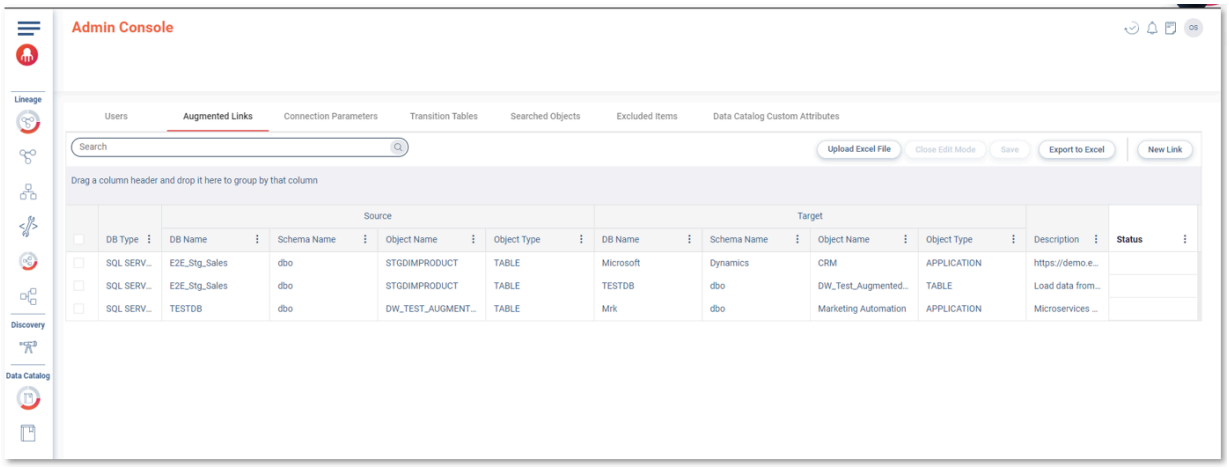
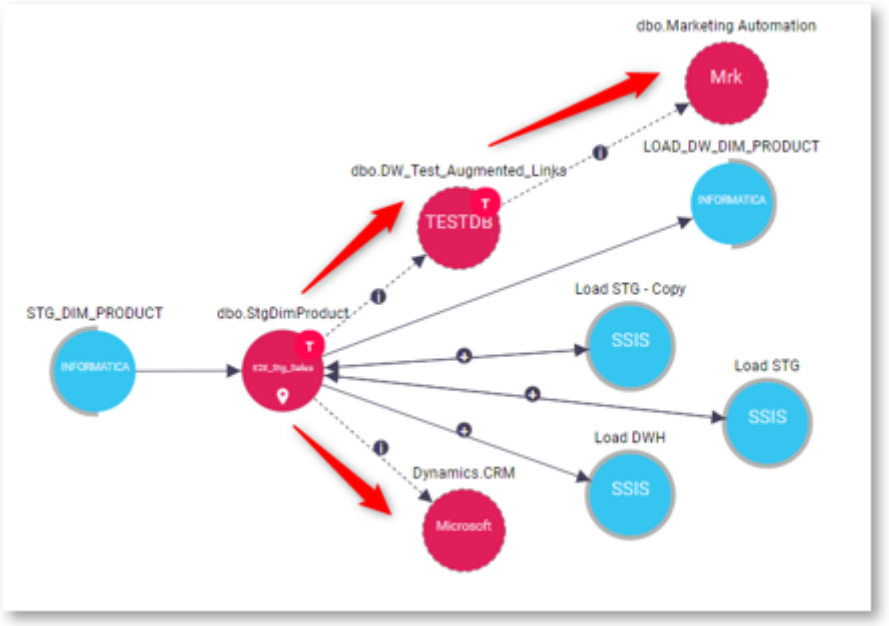


Figure 26: Cross system lineage example



For more information about this capability, see Augmented Links.

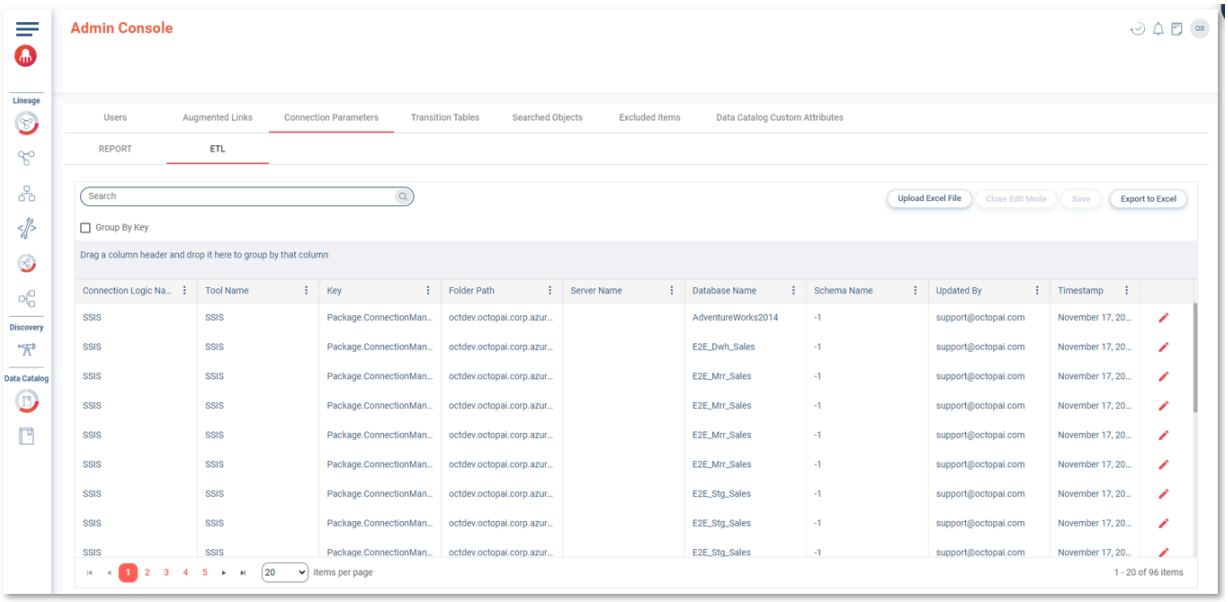
3. **Connection Parameters**

   Cloudera Octopai helps address broken lineages by providing tools and functionalities to analyze and repair them. By leveraging metadata relationships and lineage information, Cloudera Octopai identifies and resolves gaps in

data lineages. It assists users in tracing the missing connections, identifying the causes of lineage breaks, and restoring the lineage flow.

Through a combination of automated processes, manual interventions, and intelligent algorithms, Cloudera Octopai helps reconstruct accurate data lineages. This enables data professionals to regain visibility into data flow, understand dependencies, and ensure data integrity throughout their systems.

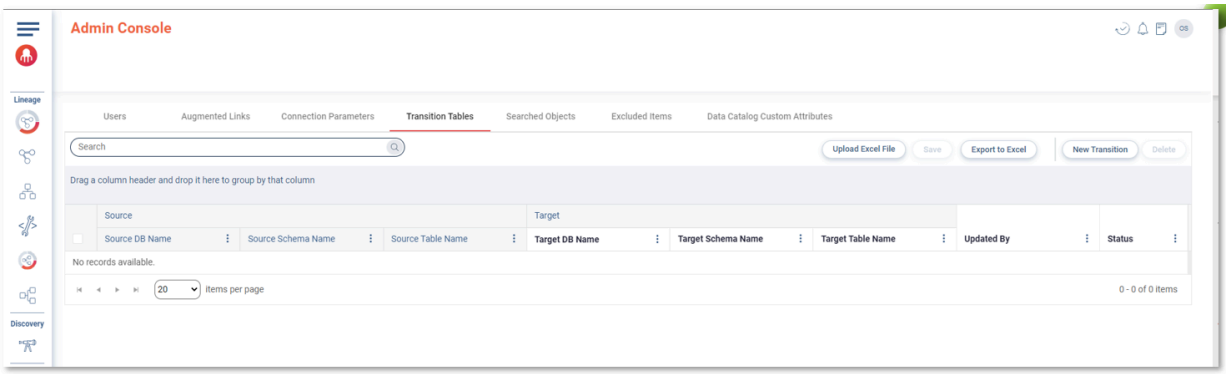
Figure 27: Connection Parameters tab in Admin Console



4. **Transition Tables**

Cloudera Octopai facilitates the use of the `EXCHANGE PARTITION` command to swap tables and implement bulk loads for partitioned tables. It also establishes connections between two tables, recognizing ETL processes that write to the source table and reflecting the same connection in the **Transition Table** tab.

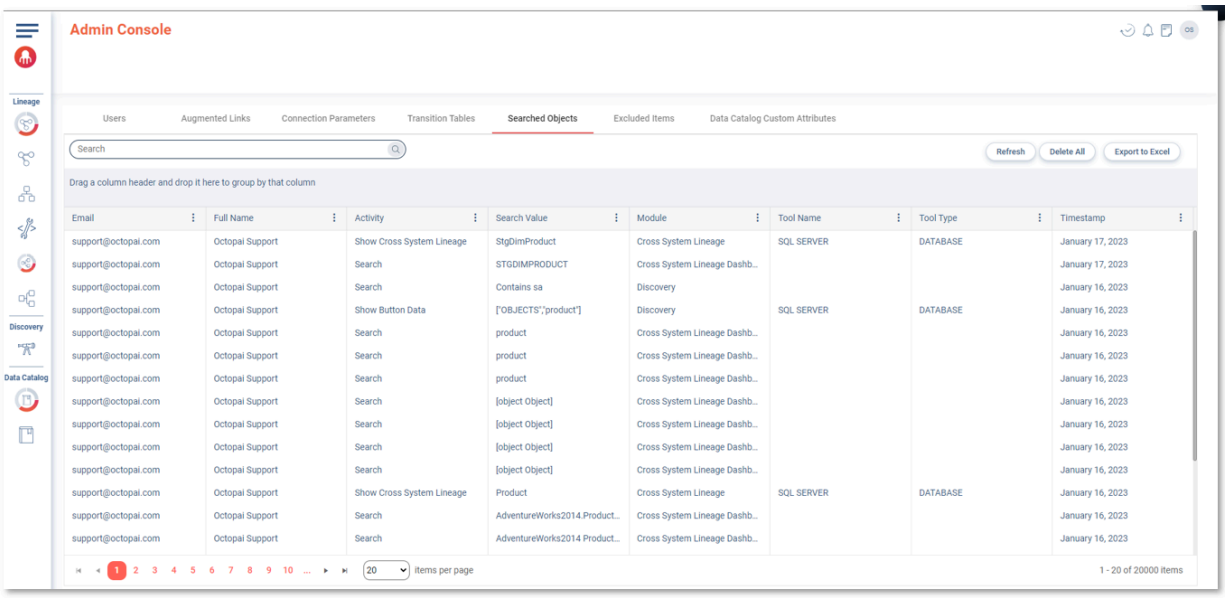
Figure 28: Transition Tables tab in Admin Console



5. **Searched Objects**

Cloudera Octopai collects user search statistics for analysis and insights.

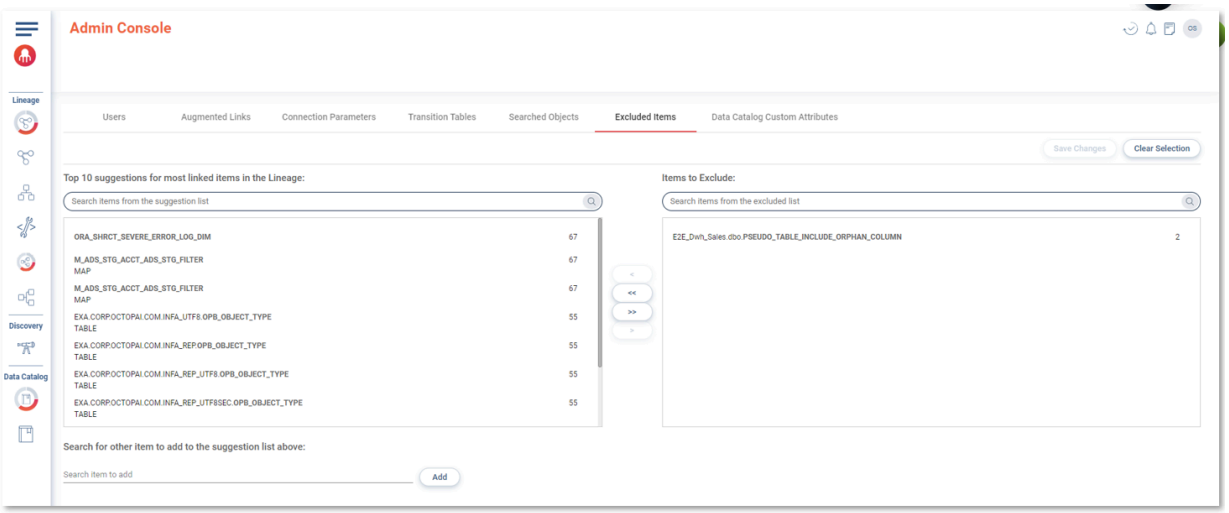
Figure 29: Searched Objects tab in Admin Console



6. **Excluded Items**

Cloudera Octopai allows for excluding specific data objects from lineages. Objects such as logs and synonyms can be excluded from lineage display, simplifying the visualization and providing a cleaner representation. The aim is to streamline the lineage and enhance its visual clarity.

Figure 30: Excluded Items tab in Admin Console



7. **Knowledge Hub Custom Attributes**

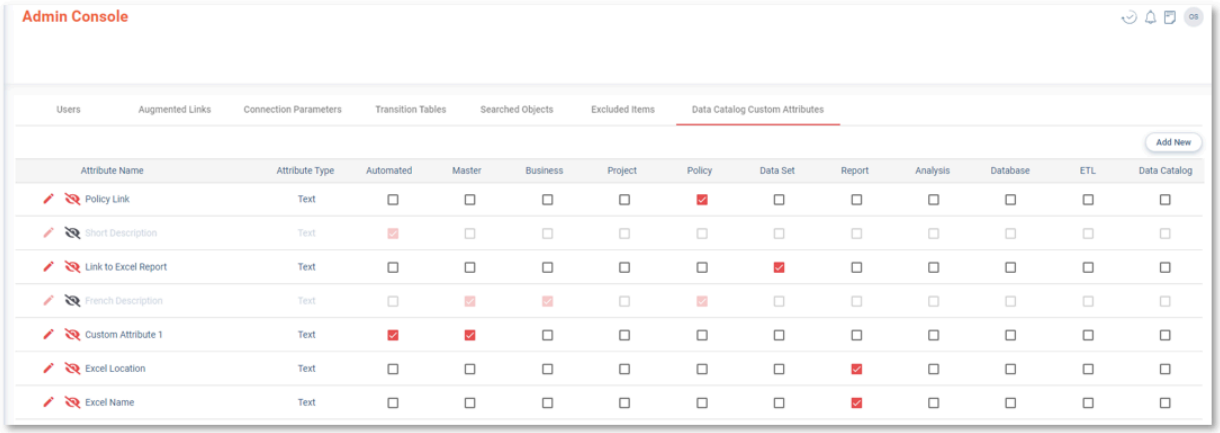
   Administrators can create and manage additional attributes in the Knowledge Hub Custom Attributes tab. Administrators can perform the following actions:

   a. Create a new attribute.
   b. Select the applicable asset types by checking the relevant checkboxes. Assets generated through automation are grouped in the Automated column.
   c. Edit the attribute name.
   d. Sort the attributes using a simple drag-and-drop interface.
   e. Hide attributes.

   Note: Data already entered for hidden attributes will be preserved, but the attribute will no longer be displayed for any asset type.

   Users can access and view the details of the additional attributes in the **Overview** tab of the **Catalog**, within the **Properties** section.

Figure 31: Data Catalog Custom Attributes tab in Admin Console



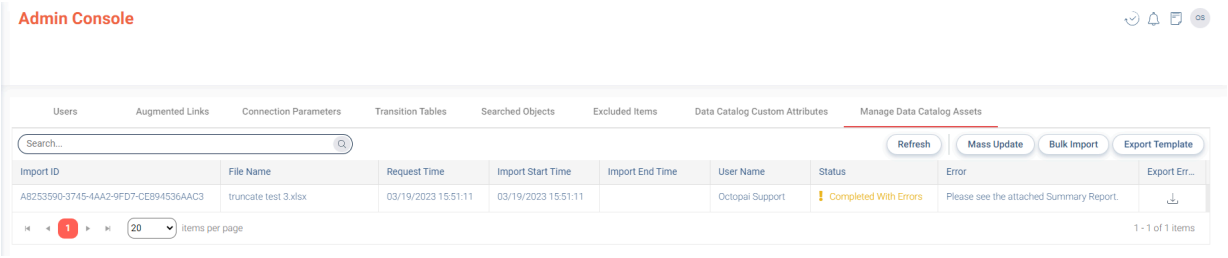
Important: Custom attributes are limited to 40 per Asset. Each custom attribute definition can contain a maximum of 500 characters.

8. **Manage Knowledge Hub Assets**

   Administrators can update assets in bulk by importing the same template spreadsheet that was exported from the **Total Assets** section. This functionality allows for efficient and streamlined updates to multiple assets simultaneously.

   Figure 32: Total Assets section

   Figure 33: Manage Data Catalog Assets tab in Admin Console



Exporting items from Lineage

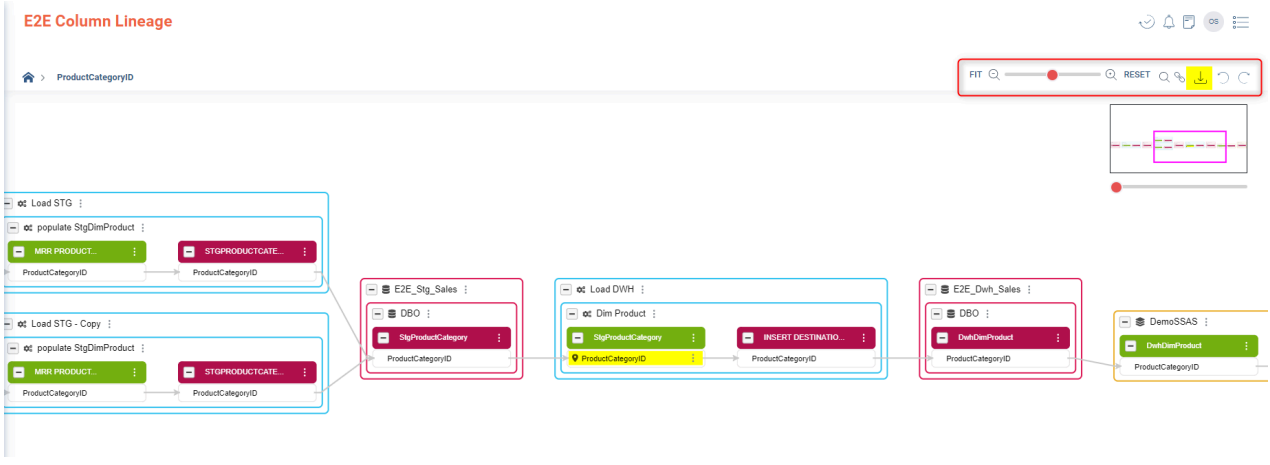
Learn about exporting items from Lineage to streamline data management and analysis.

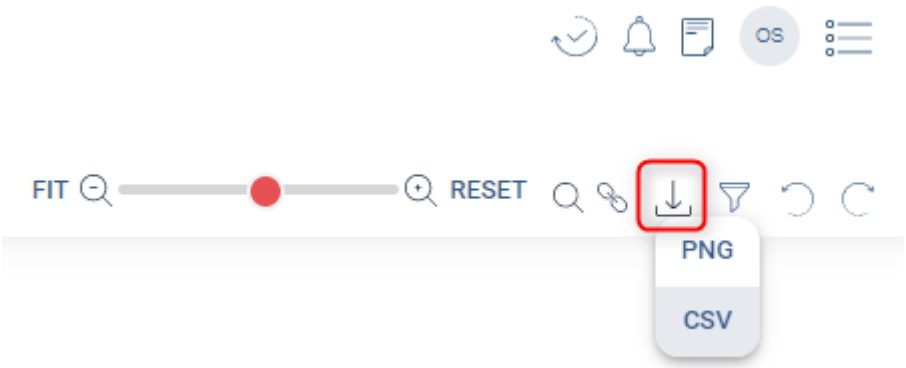
Procedure

Export all Lineage levels by clicking the Download icon located in the upper-right corner of the main screen.

Cloudera Octopai Data Lineage supports downloading items either as an image in PNG format or as a spreadsheet in CSV format.

Figure 34: Export items





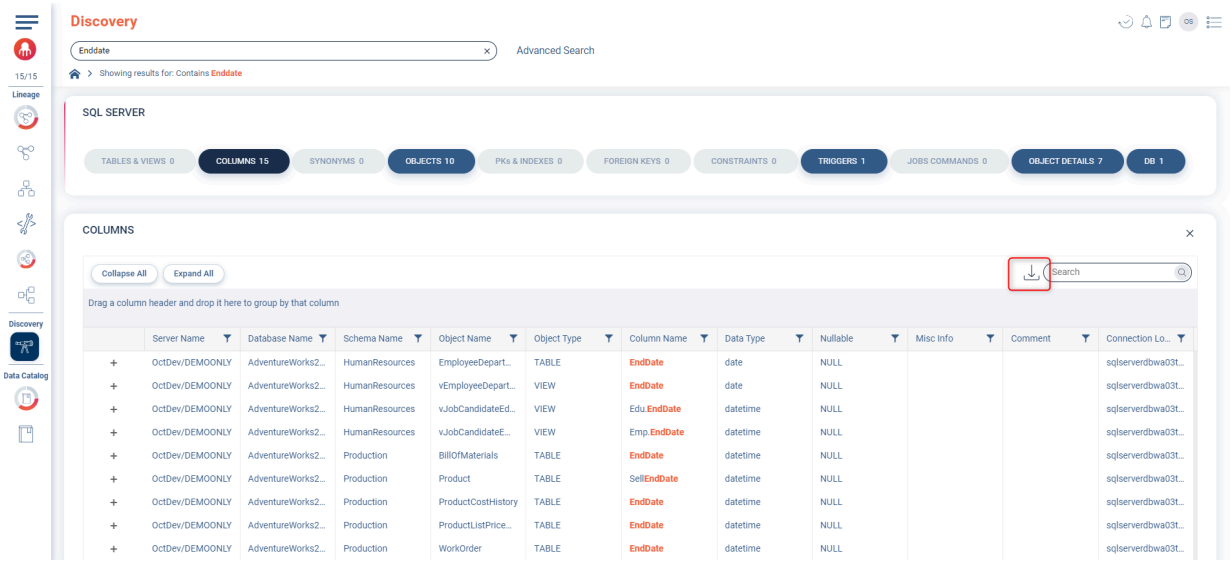
Export items from Discovery

Learn about exporting the Discovery Module results from Cloudera Octopai Data Lineage.

Procedure

1. Search for your item in the **Discovery** module.
2. Click Download to export the list to an Excel file in XLSX format.

Figure 35: Export items from Discovery



Change management - Best practices for using Cloudera Octopai for CI/CD DataOps

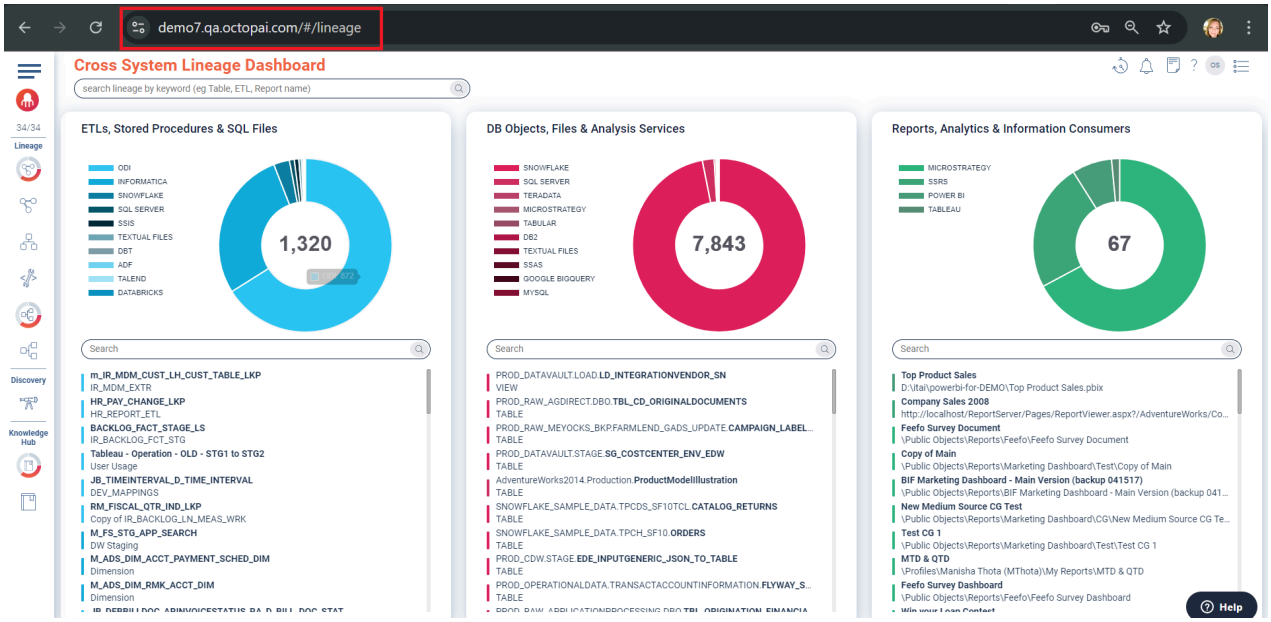
As data developers and data governance teams, learn about how to effectively use the Cloudera Octopai Data Lineage automated data lineage solution in managing changes to data flows within a CI/CD DataOps process through best practices for environment setup, impact analysis, automated refreshes, pre-production testing, and troubleshooting to ensure data integrity and governance.

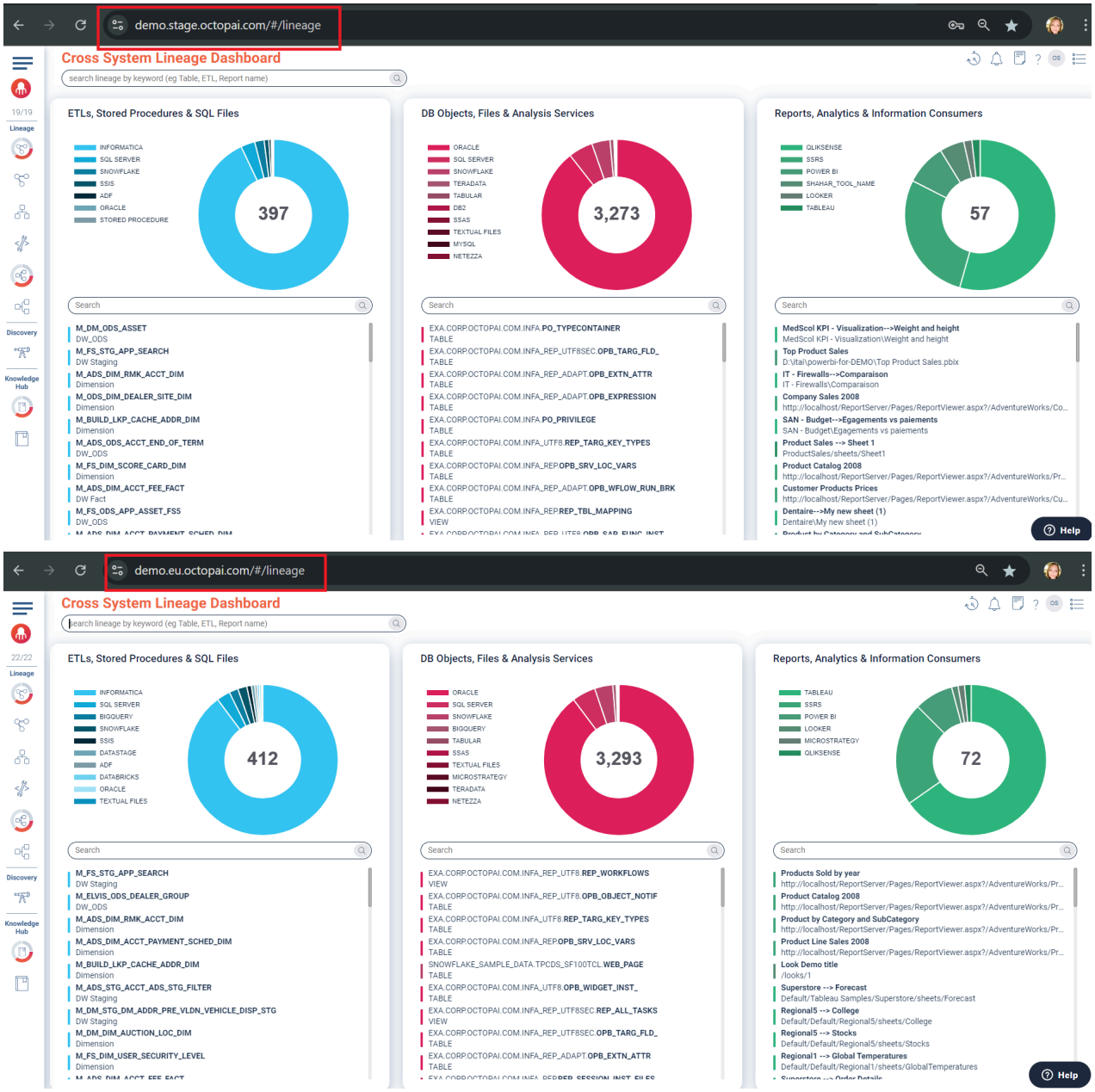
Your environments

For data developers, establishing multiple environments is crucial to manage changes efficiently. This setup allows you to rigorously test changes before they reach production, while also aligning with data governance protocols. The best practice is to effectively use the following environments within Cloudera Octopai:

- **Development Environment** – In this environment, data developers introduce and iterate on changes, ensuring that the changes meet initial requirements.
- **QA Environment** – In this environment, data governance teams can validate changes against governance policies and standards by executing test plans to ensure compliance and data integrity.
- **Staging Environment** – In this environment, data developers and governance teams can mimic production and perform final validation.
- **Production Environment** – In this live environment, end-users and business processes actively use data with strict governance oversight.

Important: You must ensure that your organization's contract includes licensing for multiple environments. Coordinated this setup with the Cloudera Octopai Support team as part of your licensing agreement. Having this in place allows you to fully leverage Cloudera Octopai capabilities across all necessary environments, ensuring a smooth and compliant change management process.





Applying impact analysis and risk assessment

Understanding the potential impact of changes on data flows is critical for both data developers and governance teams by performing the following actions:

1. Trigger impact analysis.
   - Use Cloudera Octopai to identify upstream and downstream dependencies that might be affected by the change.
   - Ensure that the impact analysis aligns with governance policies, documenting any risks or compliance issues.
2. Conduct risk assessment.
   - Evaluate the technical risks associated with the change, such as potential disruptions to dependent systems.
   - Assess the risks from a compliance perspective, ensuring that all regulatory requirements are met.

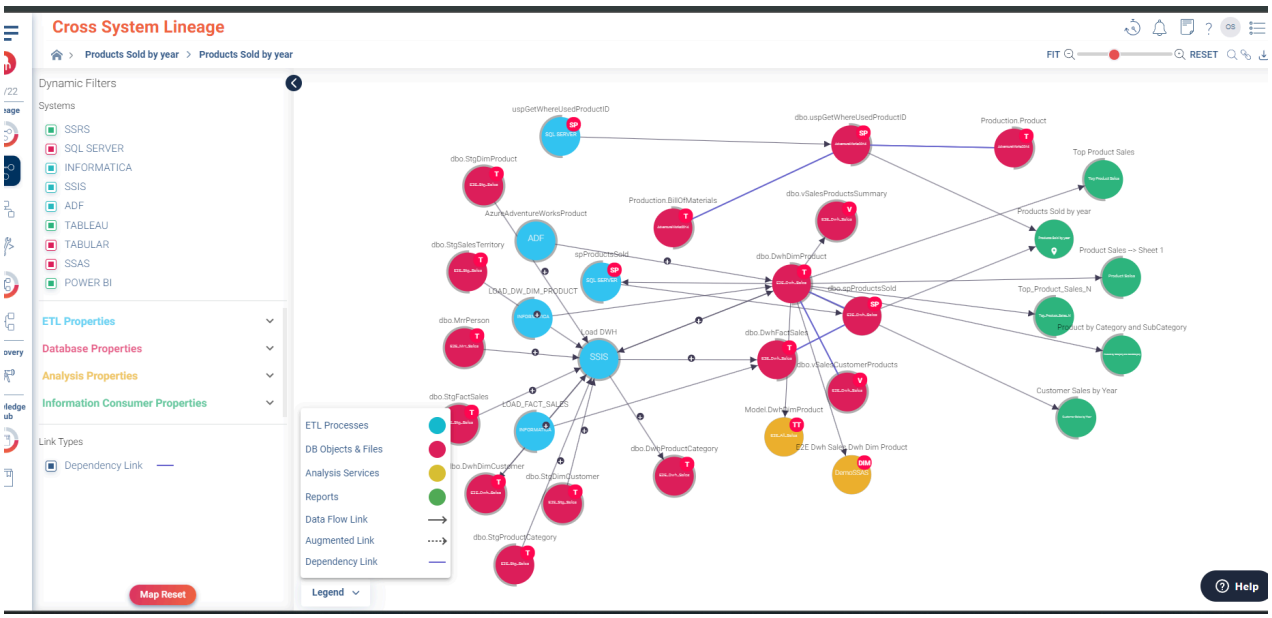


Figure 36: Upstream impact analysis

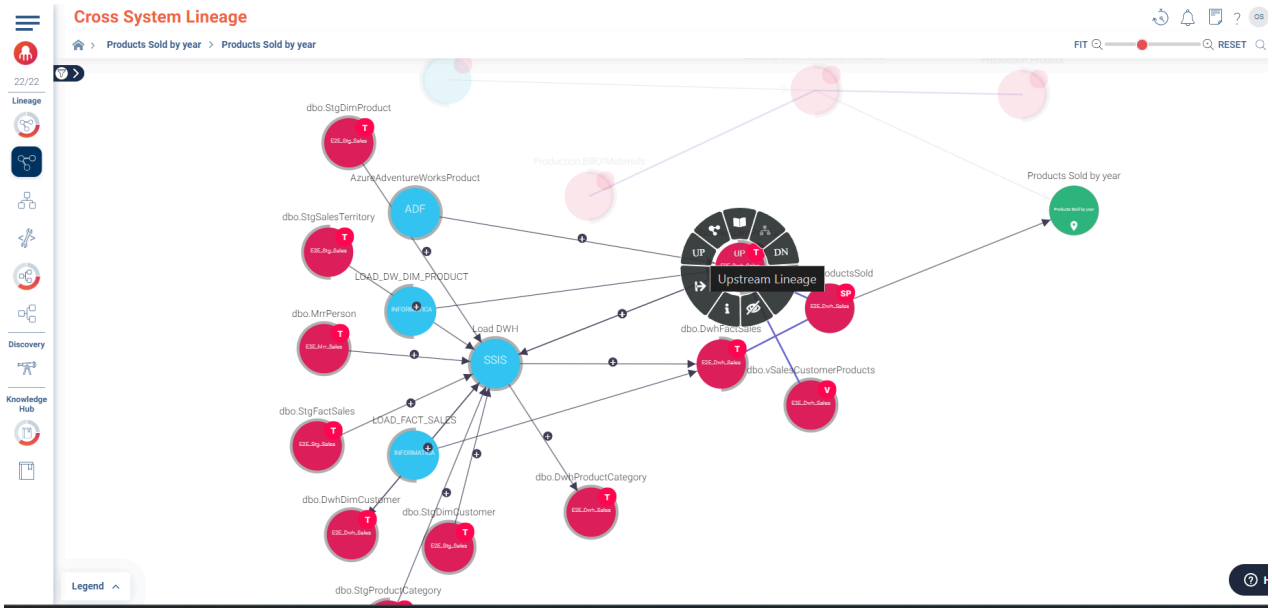
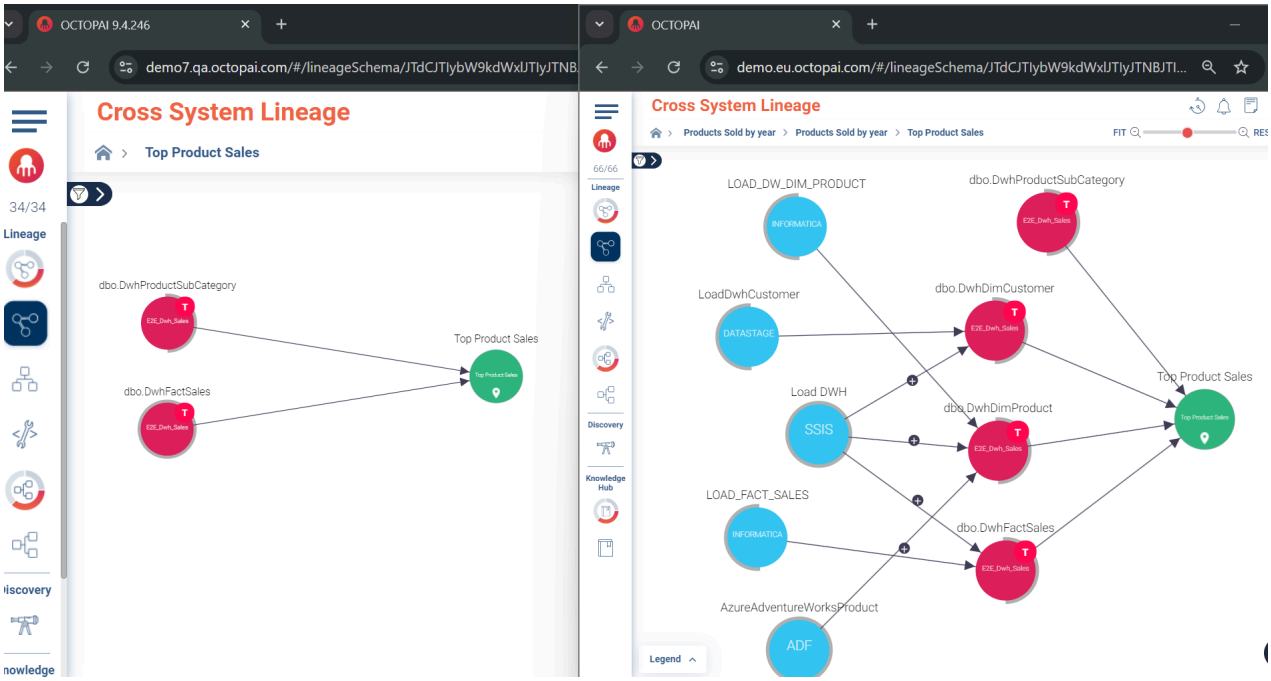


Figure 37: Comparison upstream impact analysis between QA and production



Automating data refresh for testing

Automating the data refresh process in Cloudera Octopai ensures that all environments reflect the most recent changes, which is essential for both effective development and governance. To automate data refresh, consider the following recommendations:

- Use Jenkins. For teams using Jenkins in their CI/CD pipeline, Cloudera Octopai can be integrated to automate the data refresh process. This allows data developers to work with the latest data lineage and governance teams to ensure continuous compliance.
- Use Cloudera Octopai Client built-in feature for scheduling automatic data refreshes. This ensures that the lineage is always up-to-date, facilitating both development and governance efforts.
- Apply Jenkins in environments with established CI/CD pipelines where rigorous testing and governance are required. If not, the Cloudera Octopai scheduling features can serve as a simpler alternative.
- Apply the Cloudera Octopai automatic refresh solution on a connection level that is metadata source level.

For more information, see https://docs.cloudera.com/octopai/latest/getting-started/topics/oct-admin-user-octopai-client.html.

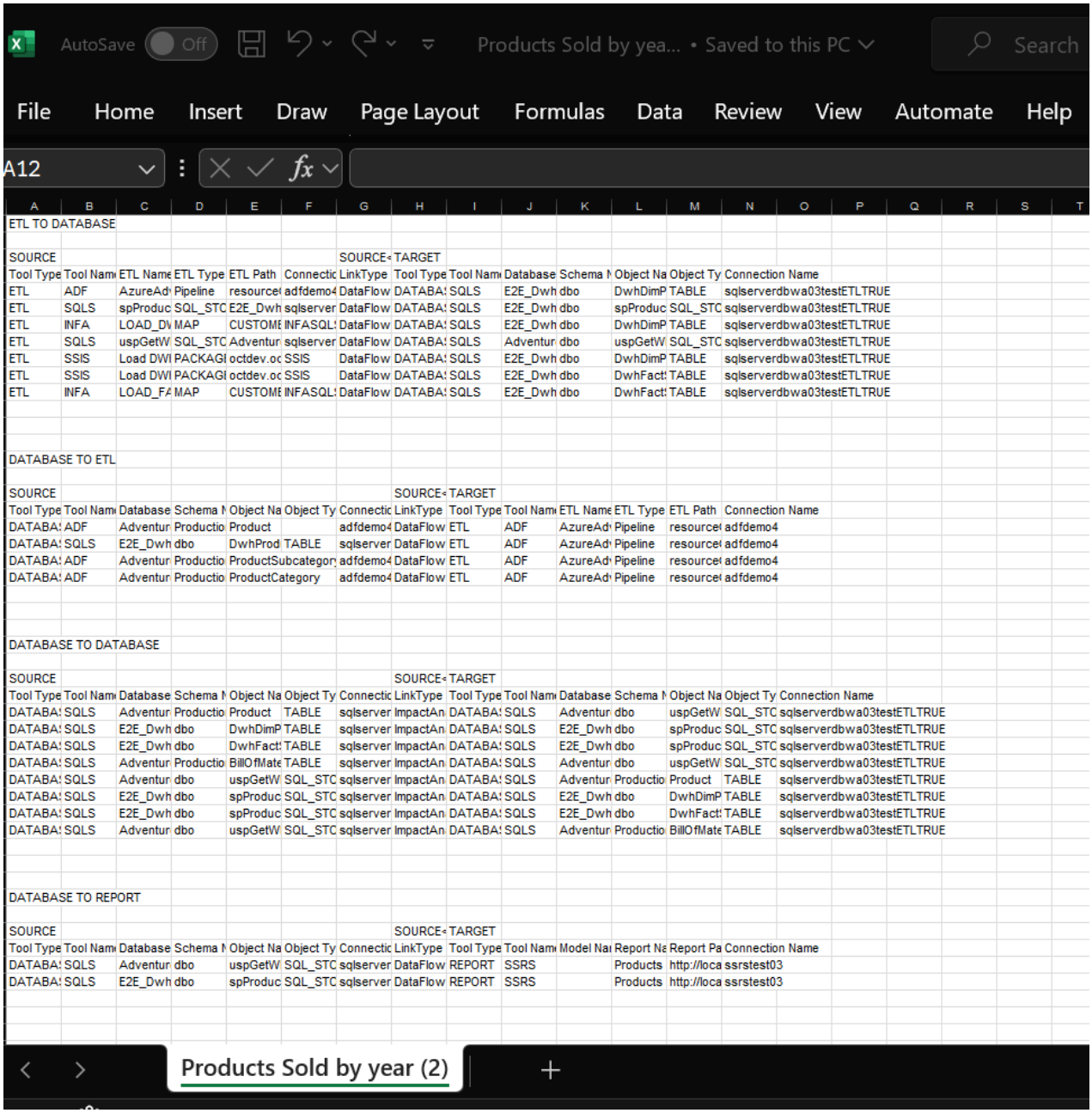
Simulating changes in pre-production

Simulating changes in a pre-production environment is essential to avoid unintended consequences in production, particularly when governance standards must be met. To simulate changes, perform the following actions:

- Simulate impact.
  - Use Cloudera Octopai to simulate the change and understand its technical implications across environments.
  - Ensure that the simulated impact is analysed for compliance risks, validating that the change adheres to governance policies.
- Create a regression test plan.
  - Develop a test plan that covers all critical data flows impacted by the change.
  - Validate that the test plan includes checks for compliance and data integrity.

Use the Export of Cloudera Octopai E2E Column Lineage capability as a foundation for your test plan.

Figure 38: Export of Cloudera Octopai E2E Column Lineage



ETL TO DATABASE

SOURCE						SOURCE<	TARGET										
Tool Type	Tool Nam	ETL Name	ETL Type	ETL Path	Connectic	LinkType	Tool Type	Tool Nam	Database	Schema N	Object Na	Object Ty	Connection Name				
ETL	ADF	AzureAd	Pipeline	resource	adfdemo4	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	DwhDimP	TABLE	sqlserverdbwa03testETLTRUE				
ETL	SQLS	spProduc	SQL_STC	E2E_Dwh	sqlserver	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserverdbwa03testETLTRUE				
ETL	INFA	LOAD_DV	MAP	CUSTOME	INFASQL	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	DwhDimP	TABLE	sqlserverdbwa03testETLTRUE				
ETL	SQLS	uspGetW	SQL_STC	Adventur	sqlserver	DataFlow	DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserverdbwa03testETLTRUE				
ETL	SSIS	Load DWI	PACKAGI	octdev.oc	SSIS	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	DwhDimP	TABLE	sqlserverdbwa03testETLTRUE				
ETL	SSIS	Load DWI	PACKAGI	octdev.oc	SSIS	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	DwhFact	TABLE	sqlserverdbwa03testETLTRUE				
ETL	INFA	LOAD_FA	MAP	CUSTOME	INFASQL	DataFlow	DATABA	SQLS	E2E_Dwh	dbo	DwhFact	TABLE	sqlserverdbwa03testETLTRUE				

DATABASE TO ETL

SOURCE							SOURCE<	TARGET					
Tool Type	Tool Nam	Database	Schema N	Object Na	Object Ty	Connectic	LinkType	Tool Type	Tool Nam	ETL Name	ETL Type	ETL Path	Connection Name
DATABA	ADF	Adventur	Productio	Product		adfdemo4	DataFlow	ETL	ADF	AzureAd	Pipeline	resource	adfdemo4
DATABA	SQLS	E2E_Dwh	dbo	DwhProd	TABLE	sqlserver	DataFlow	ETL	ADF	AzureAd	Pipeline	resource	adfdemo4
DATABA	ADF	Adventur	Productio	ProductSubcategor		adfdemo4	DataFlow	ETL	ADF	AzureAd	Pipeline	resource	adfdemo4
DATABA	ADF	Adventur	Productio	ProductCategory		adfdemo4	DataFlow	ETL	ADF	AzureAd	Pipeline	resource	adfdemo4

DATABASE TO DATABASE

SOURCE								SOURCE<	TARGET						
Tool Type	Tool Nam	Database	Schema N	Object Na	Object Ty	Connectic	LinkType	Tool Type	Tool Nam	Database	Schema N	Object Na	Object Ty	Connection Name	
DATABA	SQLS	Adventur	Productio	Product	TABLE	sqlserver	ImpactAn	DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	E2E_Dwh	dbo	DwhDimP	TABLE	sqlserver	ImpactAn	DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	E2E_Dwh	dbo	DwhFact	TABLE	sqlserver	ImpactAn	DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	Adventur	Productio	BillOfMate	TABLE	sqlserver	ImpactAn	DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserver	ImpactAn	DATABA	SQLS	Adventur	Productio	Product	TABLE	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserver	ImpactAn	DATABA	SQLS	E2E_Dwh	dbo	DwhDimP	TABLE	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserver	ImpactAn	DATABA	SQLS	E2E_Dwh	dbo	DwhFact	TABLE	sqlserverdbwa03testETLTRUE	
DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserver	ImpactAn	DATABA	SQLS	Adventur	Productio	BillOfMate	TABLE	sqlserverdbwa03testETLTRUE	

DATABASE TO REPORT

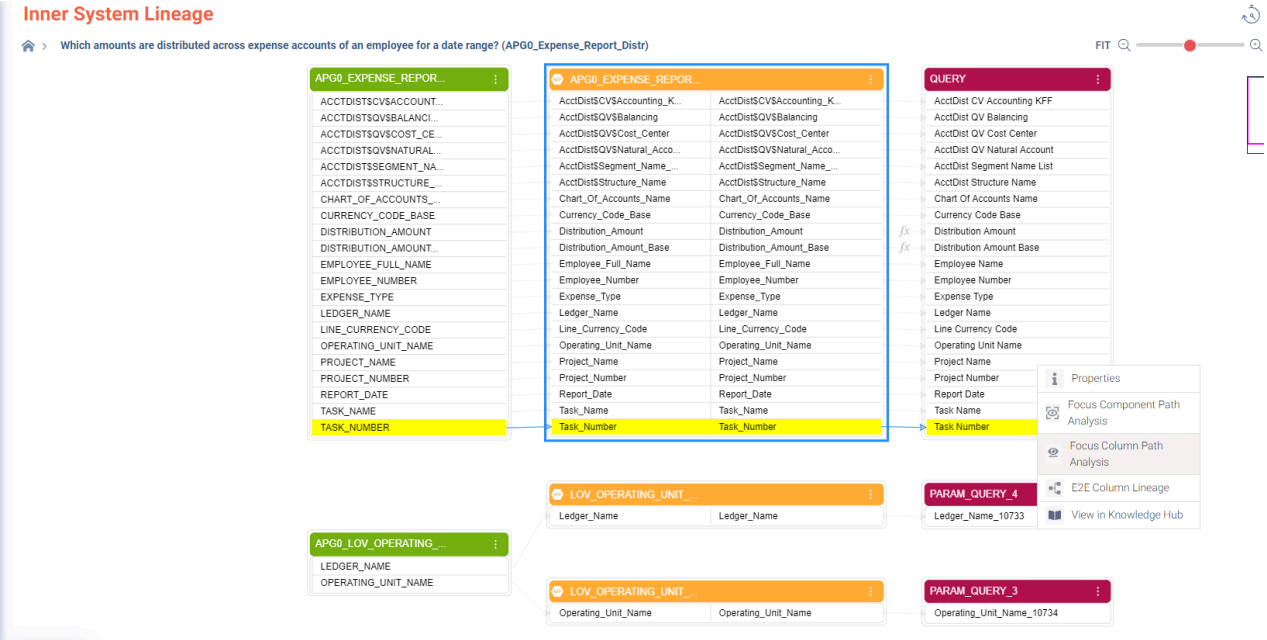
SOURCE							SOURCE<	TARGET					
Tool Type	Tool Nam	Database	Schema N	Object Na	Object Ty	Connectic	LinkType	Tool Type	Tool Nam	Model Na	Report Na	Report Pa	Connection Name
DATABA	SQLS	Adventur	dbo	uspGetW	SQL_STC	sqlserver	DataFlow	REPORT	SSRS		Products	http://loca	ssrstest03
DATABA	SQLS	E2E_Dwh	dbo	spProduc	SQL_STC	sqlserver	DataFlow	REPORT	SSRS		Products	http://loca	ssrstest03

Troubleshooting in production

When a change leads to an issue in production, Cloudera Octopai helps both data developers and governance teams quickly identify and resolve the problem.

Use lineage information for troubleshooting by performing the following actions:

- Trace the issue back to its source within the data flow, identifying the root cause.
- Ensure that the resolution process aligns with governance standards, updating documentation as needed.



Document the solution by performing the following actions:

- Document the technical steps taken to resolve the issue.
- Update governance documentation to reflect the resolution and any changes to compliance processes.

Schema management - Best practices for handling changes and impact analysis

Learn about best practices for managing schema changes and conducting impact analysis using Cloudera Octopai Data Lineage.

When comparing schemas across different environments, several types of changes are commonly encountered. These changes can have varying degrees of impact on your data flows and applications. By focusing on the strengths of Cloudera Octopai, such as its Dynamic Filters, inner system maps, Discovery Space, and detailed impact analysis, you can effectively manage and compare schema changes across environments, ensuring consistency, compliance, and operational efficiency.

You might need to address the following typical changes:

- **Table structure modifications**
  - Addition or removal of columns – A common change when new columns are added to or existing columns are removed from a table. This can affect how data is transformed and used downstream.
  - Data type changes – Modifying the data type of a column, for example changing an INT to a VARCHAR. Such changes can lead to data mismatches or errors if not handled consistently across environments.
  - Column renaming – Renaming a column can break dependencies in data flows if the new name is not updated throughout the pipeline.
- **Creating or dropping indexes** – Indexes are often added or removed to optimize performance. However, these changes can impact query performance differently across environments, potentially leading to inconsistent results.
- Adding or dropping constraints – Constraints like primary keys, foreign keys, and unique constraints ensure data integrity. Changes to these can lead to different behaviour in data validation and integrity checks.
- **Stored procedures and triggers**
  - Modifications to business logic – Changes to stored procedures or triggers, which encapsulate business logic, can have cascading effects on data operations and need to be carefully managed and tested across environments.

- **Partitioning and clustering changes**
  - Adjusting partition schemes – Modifying how tables are partitioned can impact query performance and storage efficiency, requiring careful comparison to ensure consistency in data processing across environments.
- **New or altered views** – Views are often used to simplify complex queries or present data differently. Any changes to views should be examined for their impact on dependent reports or applications.

Recommended best practices

Conduct impact analysis with Cloudera Octopai

With the Cloudera Octopai robust data lineage capabilities, you can perform detailed impact analysis before implementing schema changes across environments. This helps you foresee how modifications will affect data flows and downstream systems in different environments, ensuring smooth transitions and preventing disruptions.

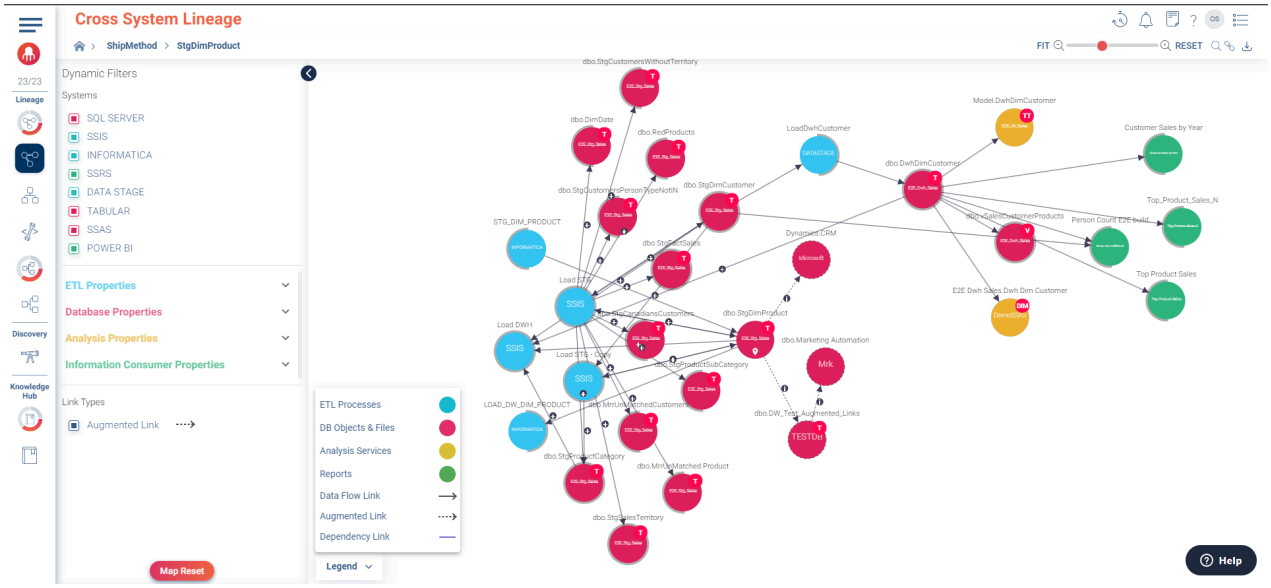
Cross-environment schema management

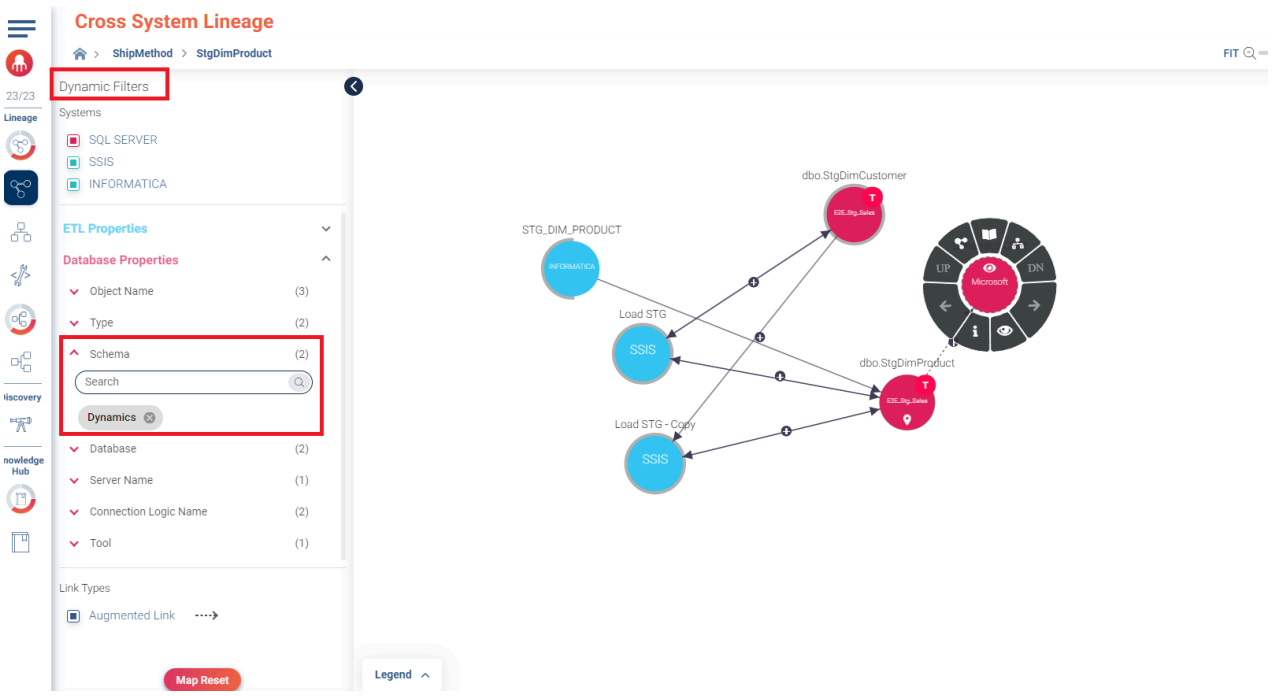
Use the Cloudera Octopai metadata management features to maintain consistent schemas across different environments, such as development, QA, and production. By carefully monitoring and comparing metadata, you can ensure that all changes are implemented consistently, reducing the risk of discrepancies and data integrity issues across environments.

Leverage dynamic filters for focused analysis

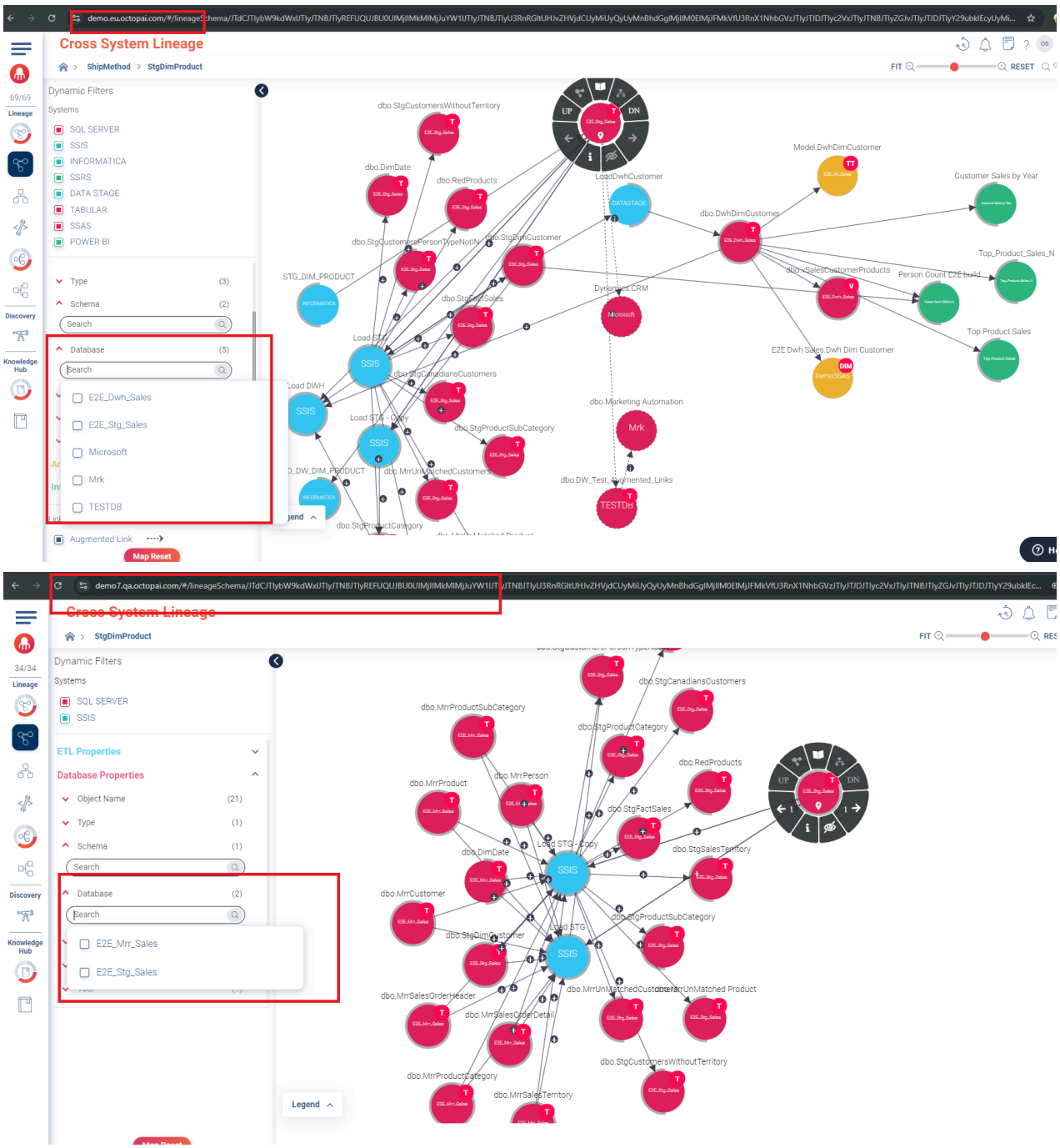
Cloudera Octopai Dynamic Filters feature is a powerful tool for managing and analysing schema changes across different environments. Dynamic Filters allow you to narrow down the scope of your analysis by focusing on specific environments, systems, or metadata properties. You can utilize operational metadata properties of an asset to concentrate on specific object characteristics, such as schema, type, and database. This capability helps ensure that you are analysing the most relevant aspects of your data lineage, making it easier to compare environments and identify discrepancies.

Figure 39: Dynamic filters




Cross System Lineage
ShipMethod > StgDimProduct
FIT
Dynamic Filters
Systems
SQL SERVER
SSIS
INFORMATICA
ETL Properties
Database Properties
Object Name (3)
Type (2)
Schema (2)
Search
Dynamics
Database (2)
Server Name (1)
Connection Logic Name (2)
Tool (1)
Link Types
Augmented Link
Map Reset
Legend
dbo.StgDimCustomer
STG_DIM_PRODUCT
Load STG
SSIS
dbo.StgDimProduct
Load STG - Copy
SSIS
UP
DN
Microsoft

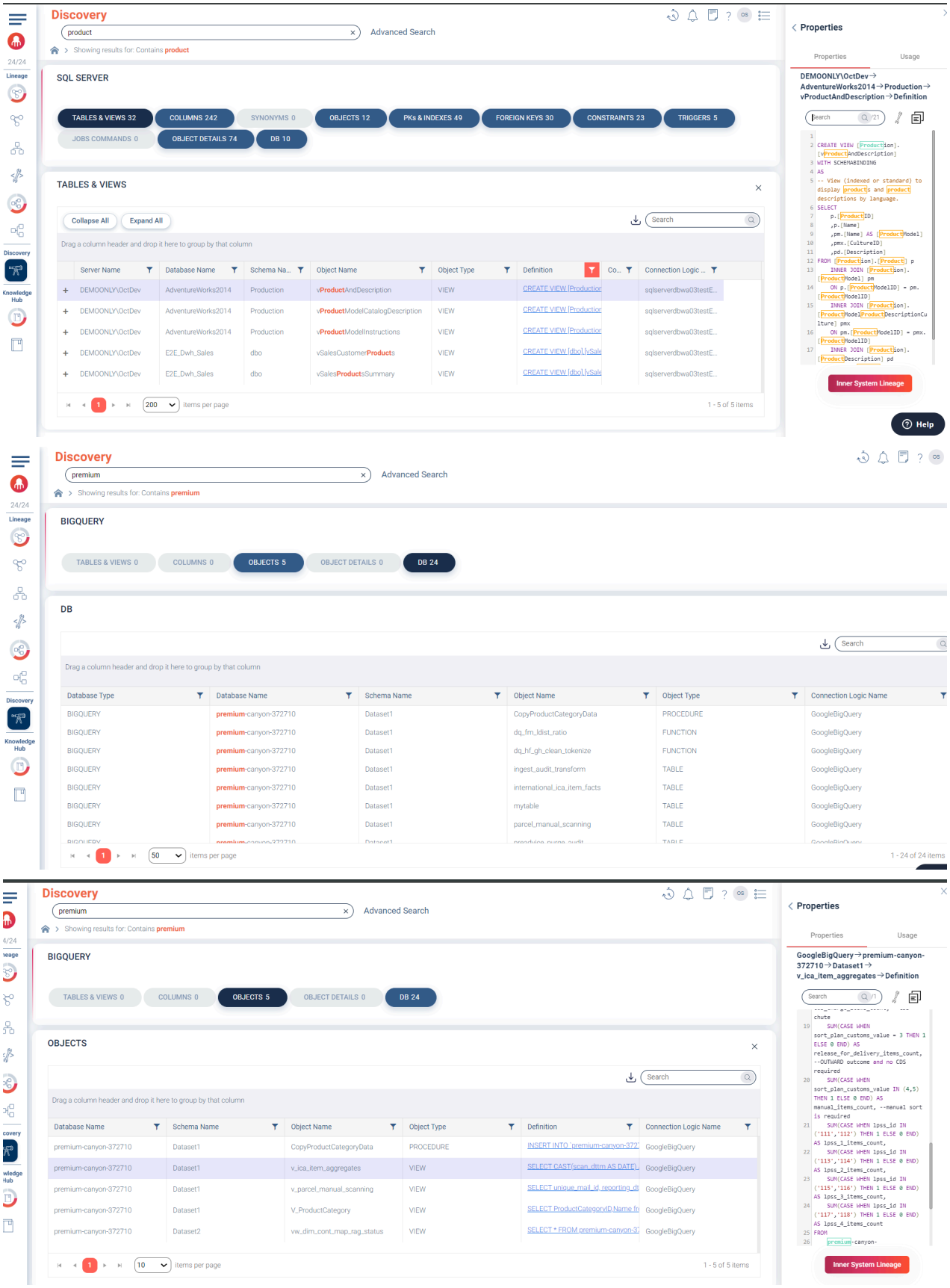

Cross System Lineage
ShipMethod > StgDimProduct
Dynamic Filters
Systems
SQL SERVER
SSIS
INFORMATICA
ETL Properties
Database Properties
Object Name (3)
Type (2)
Search
APPLICATION
TABLE
Schema (2)
Search
dbo
Dynamics
Database (2)
Search
E2E_Stg_Sales
Server Name (1)
Connection Logic Name (2)
Tool (1)
Link Types
Augmented Link
Legend
dbo.StgDimCustomer
STG_DIM_PRODUCT
Load STG
SSIS
dbo.StgDimProduct
Load STG - Copy
SSIS
Dynamics CRM



Utilize the discovery space for in-depth environment comparison

The Discovery Space in Cloudera Octopai is an essential feature for comparing environments and gaining deeper insights into your data landscape. It allows you to visually explore and compare data assets, their relationships, and dependencies across different systems and environments. This space is particularly useful for identifying potential issues, understanding the broader context of specific data elements, and conducting in-depth analysis of how different environments align or diverge. You can drill down into specific assets within the Discovery Space, explore their lineage, and uncover hidden connections, making it a powerful tool for comprehensive environment comparison.

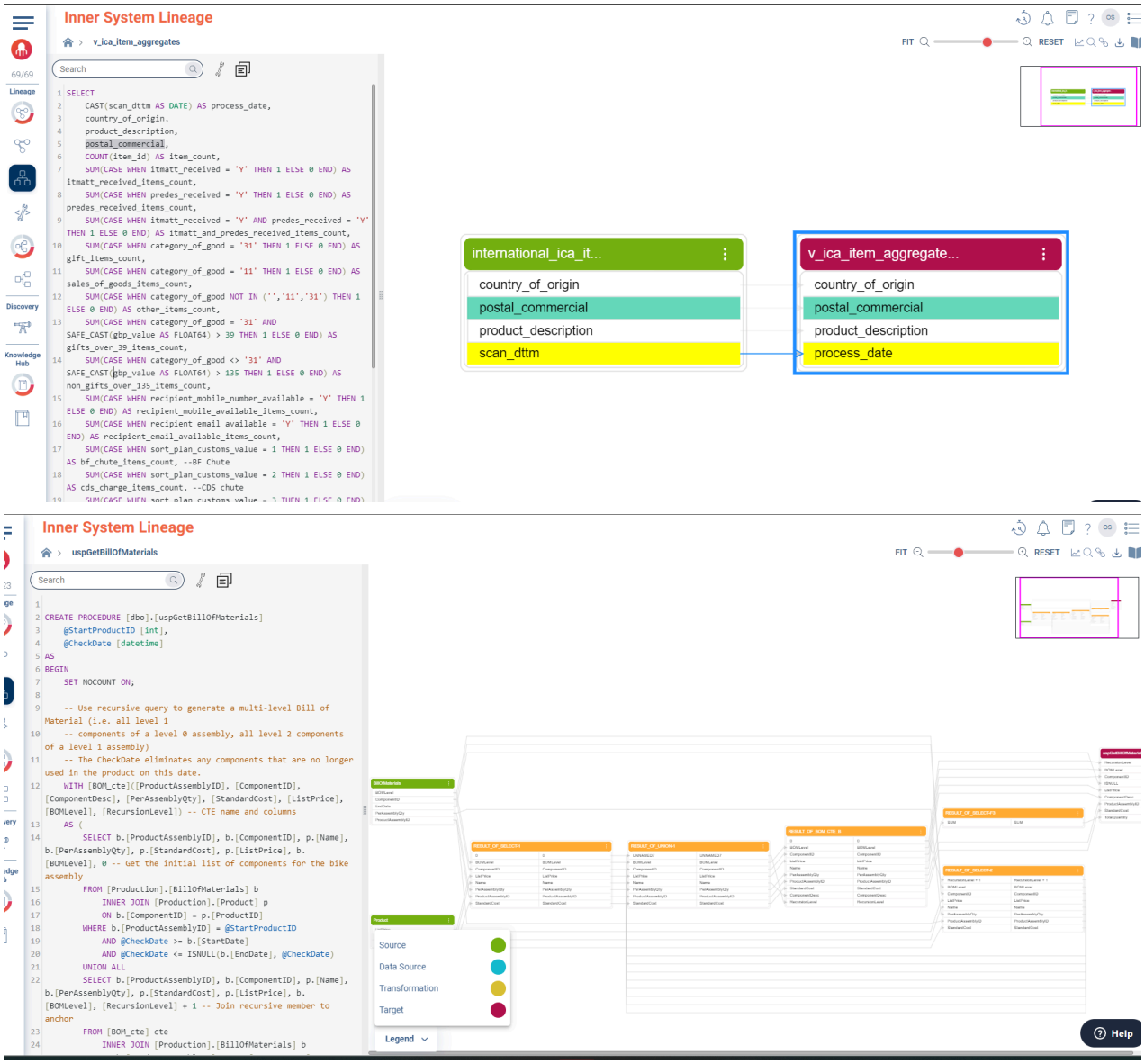
Figure 40: Schema discovery

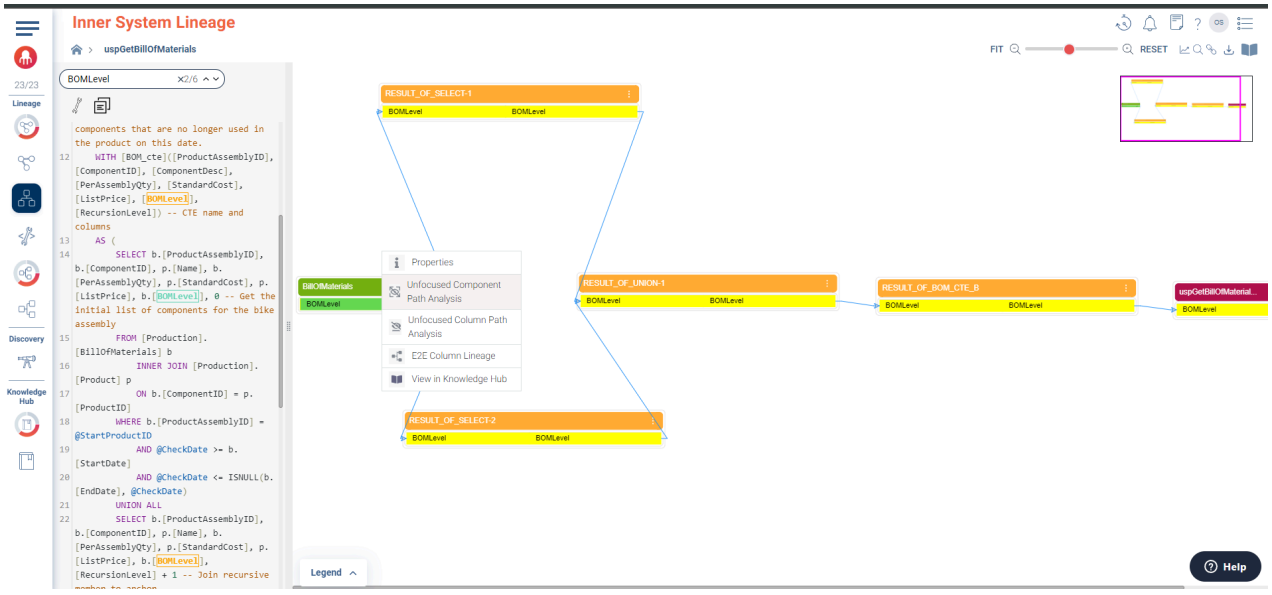


Use inner system maps for detailed comparison

Within Cloudera Octopai, you can also utilize inner system maps to perform focused component and asset analysis. This feature enables you to compare specific elements between different environments with precision. By honing in on specific components, you can identify and address differences that might impact data consistency or performance across your environments, ensuring that all environments are aligned and functioning optimally.

Figure 41: Inner system lineage





Documentation and compliance

Cloudera Octopai excels in creating comprehensive documentation of your data environments. This documentation is essential for compliance, as it provides a clear trail of how data moves and changes across your systems. Regularly updated lineage documentation supports audits and helps maintain a clear understanding of your data landscape across all environments.

Post-deployment verification

After deploying changes, the Cloudera Octopai tools can be used to verify that your environments remain consistent and functional. This step is crucial for ensuring that your production environment continues to operate smoothly and efficiently.

Databricks Lineage in Cloudera Octopai Data Lineage

Cloudera Octopai Data Lineage provides automated data lineage and metadata visibility for Databricks environments, supporting Unity Catalog, Hive Metastore, and hybrid configurations. It extracts lineage from system metadata and notebook parsing, offering insights into data flows across pipelines, notebooks, and downstream systems.

Overview

Cloudera Octopai Data Lineage automates data lineage and metadata visibility for Databricks environments. Support for Databricks in Cloudera Octopai accommodates various catalog architectures, including Unity Catalog, Hive Metastore, and hybrid deployments that combine Unity Catalog and Hive Metastore.

Cloudera Octopai extracts lineage directly from Databricks system metadata and enhances it with advanced parsing capabilities for notebook-based workloads. This enables organizations to gain deeper insights into data flows across Databricks pipelines, notebooks, and downstream systems.

Supported Databricks catalog configurations

Cloudera Octopai supports lineage extraction from the following Databricks catalog setups:

Unity Catalog

Unity Catalog is Databricks' centralized governance layer. When Unity Catalog is enabled, Cloudera Octopai extracts lineage using Databricks system lineage tables and catalog metadata, as well as other parsing capabilities.

Unity Catalog is most suitable for customers using it as their primary metastore.

Hive Metastore

For Databricks workspaces still using the legacy Hive Metastore, Cloudera Octopai supports metadata and lineage extraction without relying on Unity Catalog system tables.

Hive Metastore is most suitable for customers not yet migrated to Unity Catalog.

Hybrid: Unity Catalog and Hive Metastore

Some environments operate with both Unity Catalog and Hive Metastore simultaneously. Cloudera Octopai supports these hybrid deployments and can provide lineage across both catalogs, offering broader visibility compared to native Databricks capabilities.

Hybrid configurations are most suitable for customers transitioning from Hive Metastore to Unity Catalog.

Lineage coverage and supported workloads

Cloudera Octopai supports lineage extraction from Databricks environments using either Unity Catalog system metadata, notebook parsing, or a combination of both. Lineage behavior is determined by the catalog configuration of the customer's workspace.

Supported notebook languages

Cloudera Octopai supports SQL, Python, and PySpark notebook lineage consistently across Databricks environments. Lineages for Scala and R are available only when Unity Catalog system metadata provides lineage records.

This table provides an overview of the lineage capabilities available in different Databricks catalog configurations, including Unity Catalog, Hive Metastore, and hybrid (Unity Catalog and Hive Metastore) environments.

Catalog Configuration	SQL	Python	PySpark	Scala	R
Unity Catalog	Supported	Supported	Supported	Supported when Databricks provides lineage	Supported when Databricks provides lineage
Hive Metastore	Supported	Supported	Supported	Not supported	Not supported
Unity Catalog and Hive Metastore (hybrid)	Supported	Supported	Supported	Unity Catalog-only when Databricks provides lineage	Unity Catalog-only when Databricks provides lineage

Lineage behavior by catalog type

Cloudera Octopai Databricks lineage extraction varies based on whether the workspace uses Unity Catalog, Hive Metastore, or both.

Unity Catalog lineage (including hybrid Unity Catalog and Hive Metastore deployments)

In Unity Catalog environments, Cloudera Octopai integrates system lineage metadata with notebook parsing to deliver comprehensive coverage. It uses two complementary sources:

- Databricks Unity Catalog system lineage tables
- Notebook-level parsing for supported scripts, including SQL, Python, and PySpark

This feature provides the following benefits:

- Authoritative lineage recorded natively by Databricks
- Additional lineage relationships derived from notebook code analysis

Native Unity Catalog lineage:

Unity Catalog system lineage captures persistent, governed operations, including reads and writes to managed tables.

Operations that do not produce persistent table or storage writes, such as intermediate DataFrame transformations, in-memory processing, or pandas-based manipulations, may not appear in Unity Catalog lineage metadata.

Notebook parsing enhancement:

Cloudera Octopai also parses notebook scripts to enrich lineage coverage, including cases where native system lineage may be incomplete.

Temporary views or non-persistent transformations may not appear in Unity Catalog system metadata, but may still be partially reflected through parsing where possible.

Hybrid (Unity Catalog and Hive Metastore) environments:

In hybrid environments, Cloudera Octopai uses the same Unity Catalog extraction approach while also including Hive Metastore assets. This approach offers broader visibility across both catalog types compared to the Databricks native lineage UI.

Hive Metastore lineage support (Hive Metastore only):

In Hive Metastore-only environments, Cloudera Octopai derives lineage primarily through notebook script parsing.

Cloudera Octopai supports lineage extraction from the following notebook types:

- Python notebooks
- PySpark notebooks
- SQL notebooks

Lineage is determined based on the notebook code.

Script-based parsing limitations:

In Hive Metastore environments, lineage is inferred from notebook scripts, resulting in the following limitations:

- Highly dynamic transformations (such as code-generated queries, function-driven logic, user-defined functions (UDFs), loops that write files programmatically, or indirect write operations) can limit lineage resolution or prevent full identification of sources and targets..
- Lineage resolution requires explicit table and column references in the code.
- If a table is referenced without a fully qualified database or schema name, Cloudera Octopai might not be able to resolve the database context. For example, the query SELECT * FROM sales_table might appear in the lineage without its associated database.
- Fully qualified references, such as db.schema.sales_table, provide the most complete lineage results.

In Hive Metastore environments, Cloudera Octopai does not currently support lineage for the following cases:

- Scala notebooks
- R notebooks
- Databricks pipelines or jobs

Column-level lineage support (Unity Catalog and Hive Metastore):

Column-level lineage is supported when column mappings are available in the metadata source:

- In Unity Catalog environments, column lineage is captured primarily from Databricks system lineage tables. It may also be supplemented by notebook parsing if column references are explicitly defined.
- In Hive Metastore environments, columns are captured only when explicitly referenced in notebook scripts (for example, within SQL queries).

Shared platform limitations

The following limitations apply across both Unity Catalog and Hive Metastore environments:

Pipelines and jobs not represented as entities

Databricks jobs and pipelines are not currently modeled as standalone entities in Cloudera Octopai lineage. In Unity Catalog environments, lineage generated by jobs or pipelines is captured. However, the lineage is reported through the underlying notebook activity instead of appearing as a separate job or pipeline asset.

Cloudera Octopai captures lineage at the notebook execution level and focuses on the underlying data transformations, rather than modeling orchestration constructs (such as jobs or pipelines) as standalone lineage entities.

Unity Catalog environments provide the most complete lineage coverage. Hive Metastore environments rely entirely on notebook parsing and require explicit table and column references.