Cloudera Octopai Data Lineage 1.0.0

Resources

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

https://docs.cloudera.com/

Legal Notice

Contents

Knowledge Hub

The Knowledge Hub serves as a comprehensive resource center, offering user guides, implementation best practices, and dashboards tailored for various personas. It provides essential tools and insights to optimize the use of Cloudera Octopai across diverse roles and workflows.

Knowledge Hub User Guide

How to navigate the Knowledge Hub to streamline data discovery and management.

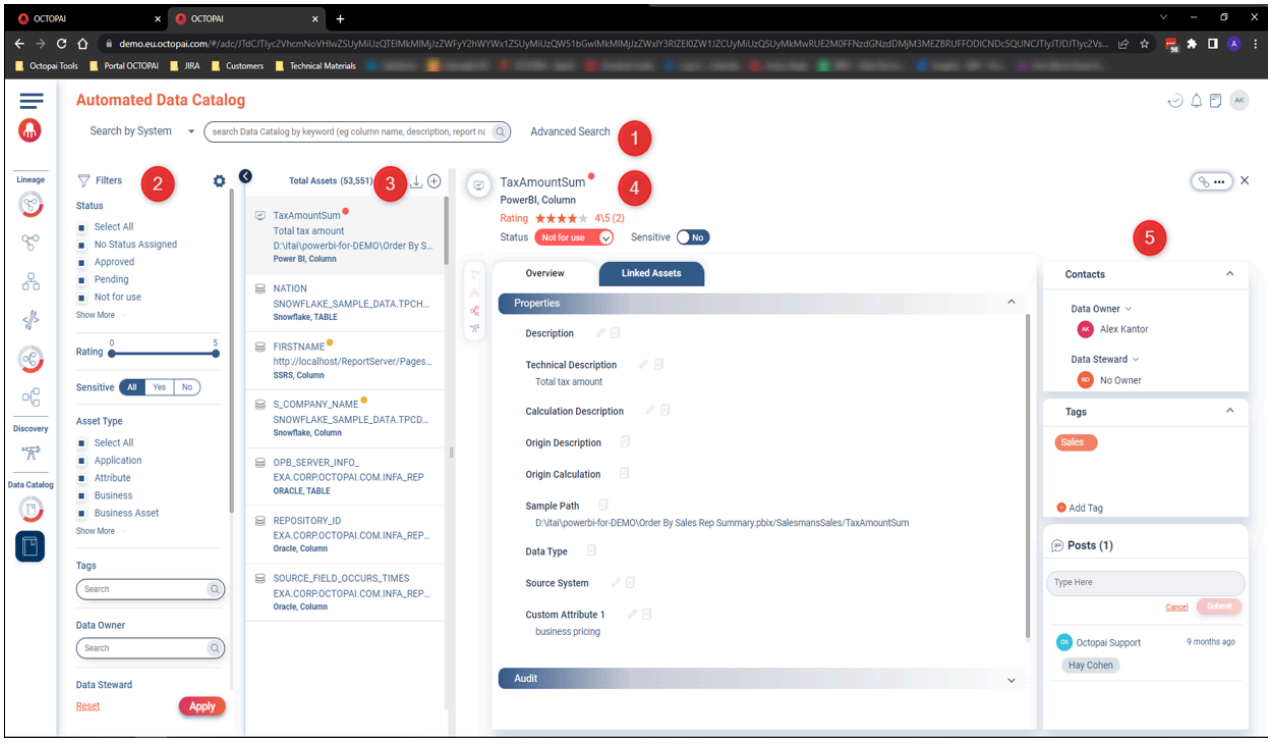
About this task

This guide helps you effectively navigate and utilize the Knowledge Hub to simplify data discovery and management. By centralizing metadata and offering advanced search, collaboration, and integration features, the Knowledge Hub enhances productivity and facilitates informed decision-making.

Key features and benefits:

- Metadata Management: Capture and organize metadata, enabling users to understand data structure and quality.
- Search and Discovery: Find specific datasets, tables, or files quickly using advanced search and navigation options.
- Collaboration and Data Governance: Share, comment, and collaborate on data assets while ensuring security and compliance.
- Integration with Data Ecosystem: Seamlessly integrate with data sources, management tools, and analytics platforms.

Target Audience: Data analysts, scientists, business intelligence professionals, and data engineers can benefit from this guide.



- Access the Knowledge Hub interface

- Navigate to the Knowledge Hub main interface to begin your data discovery journey.

Search

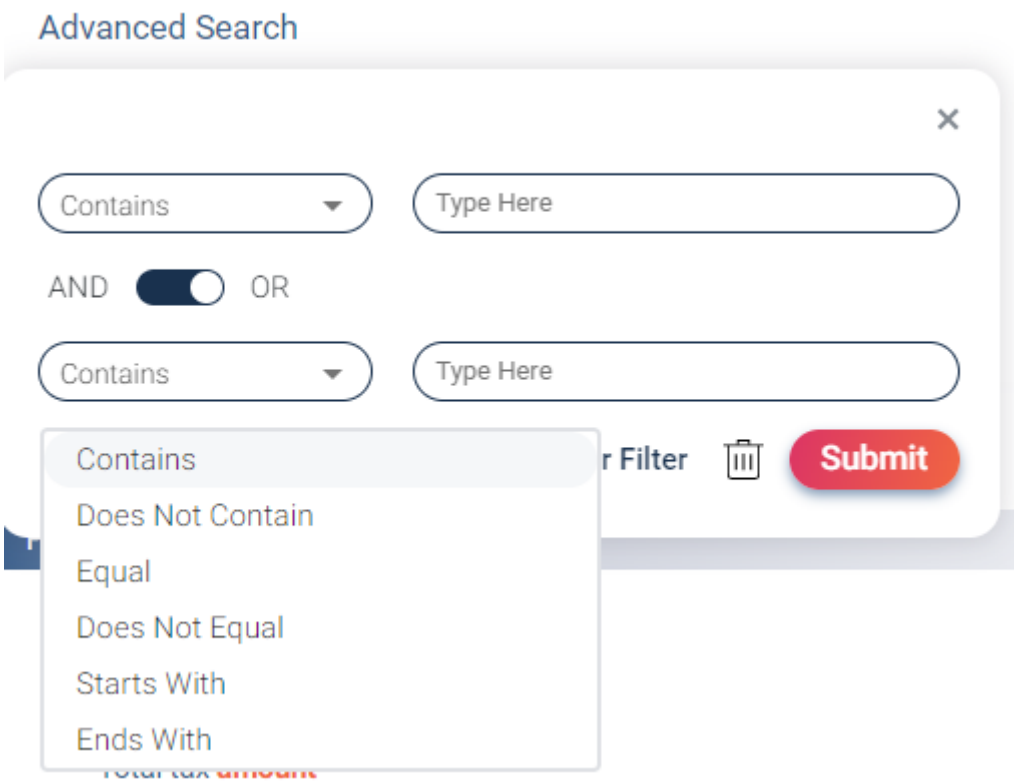
Search by system

Improve your search by filtering on system.

Search by term (fuzzy search)

Enter a term in the search box and Cloudera Octopai finds it within the asset definitions.

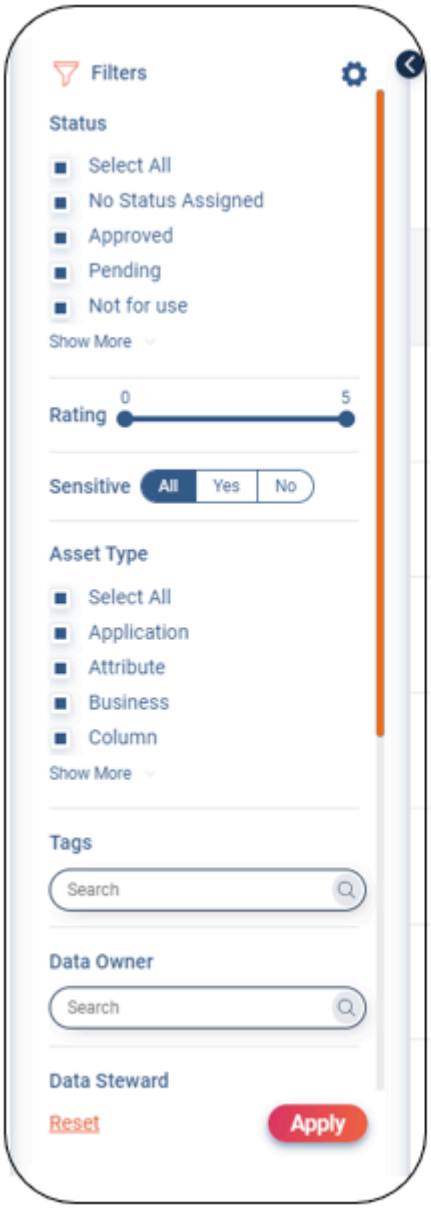
Advanced search

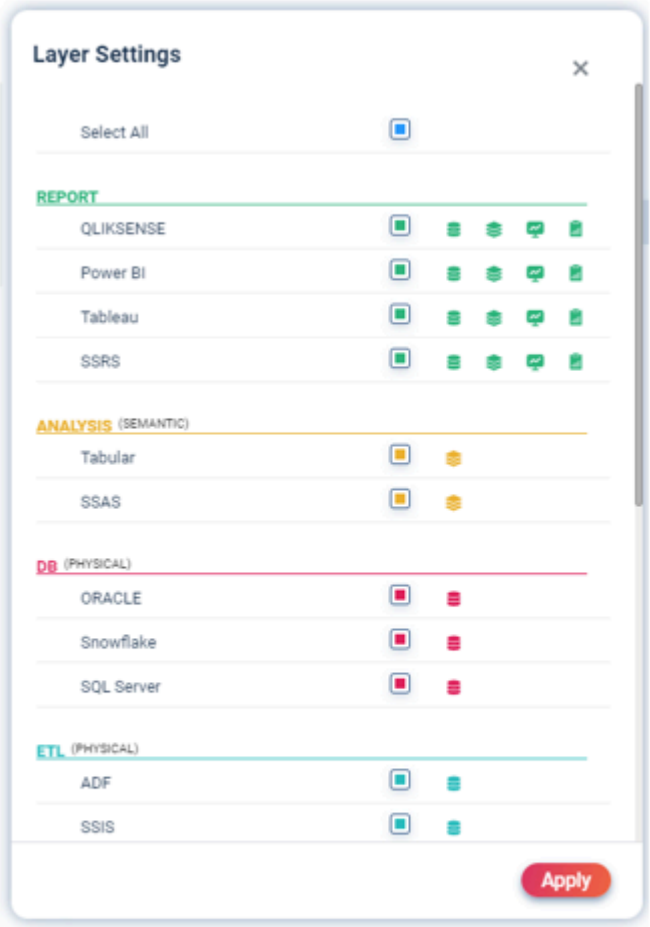


With advanced search, you can refine your query by combining up to three options using AND or OR.

Filter Section

The search filters in the Cloudera Octopai platform include various attributes that help users quickly find the desired data assets, offering a seamless shopping-like experience. Once all the attribute filters are set, simply click on "Apply" to filter the results.





- Collapse or expand the filter section.

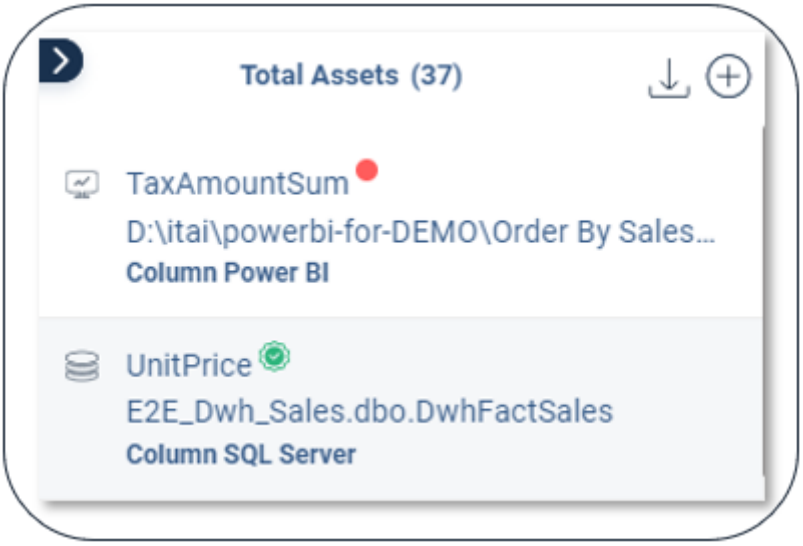
- Layer settings let you customize asset visibility by selecting specific layers using the Layer Setting selector. The selected layer setting is saved for the user and remains unchanged between sessions.

Layer Type	Layer Name	Comments
Reports	Report Name	
Presentation	Assets of Identical name are aggregated per tool.	
Semantic		

Physical	Assets of Identical name are aggregated per tool.	
Analysis	Semantic	
Database	Physical	
ETL	Physical	
Custom Asset Types	Custom Asset Types	Assets that are created within the Data Catalog

Assets Results

The section displays all assets based on the search and filter criteria, including the Asset Layer, Name, Path, Type, and Tool. Users have the option to export the list to an Excel Spreadsheet and also create new Augmented Assets.

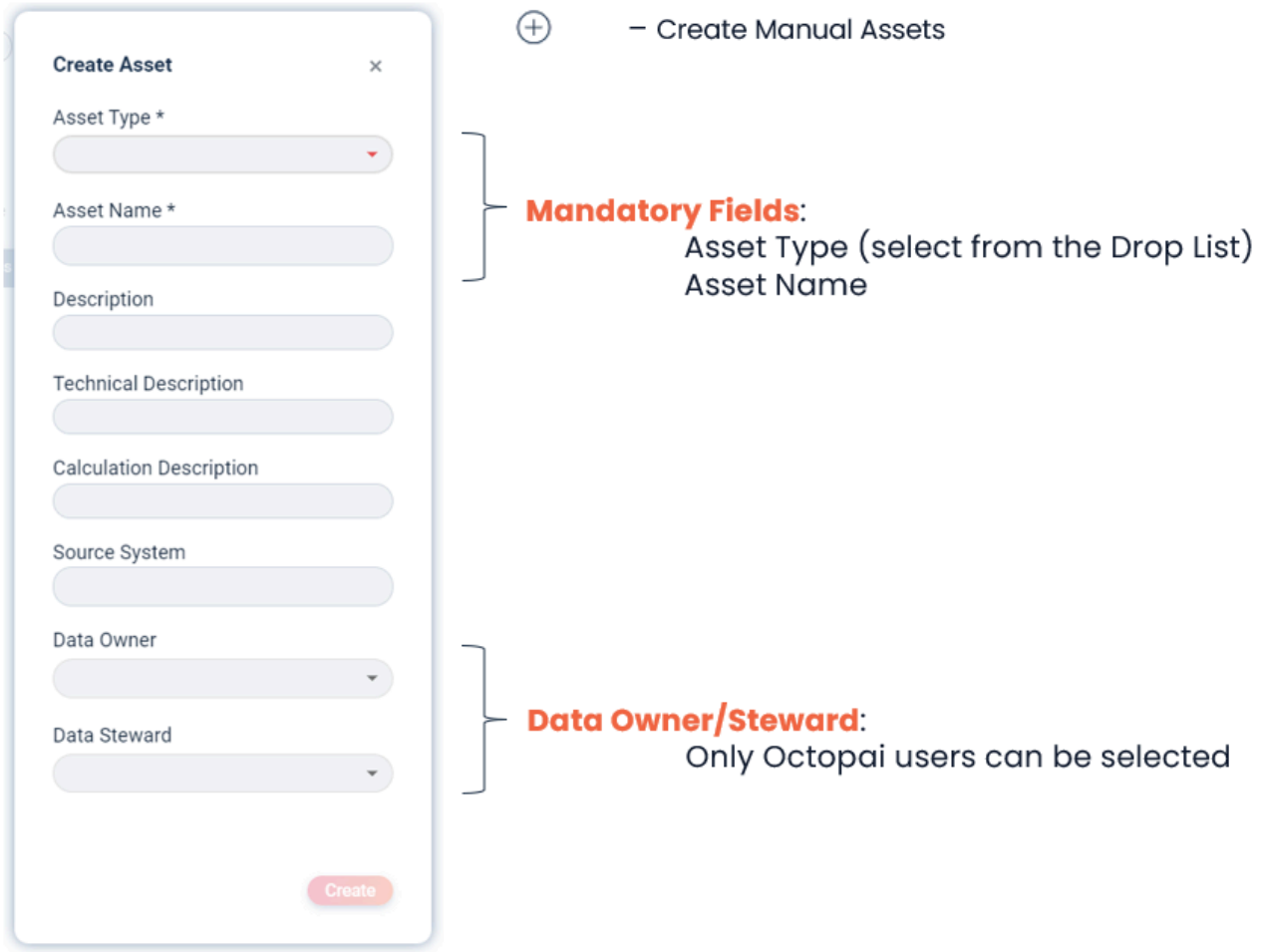


Total assets: After searching and filtering, this indicates the number of assets found.

Each asset row shows:

- Layer icon (hover to see the description).
- Asset name.
- Asset path.
- Asset type and asset tool.
- Export the total assets to an Excel spreadsheet (limited to 50,000 values).

- Create augmented assets.

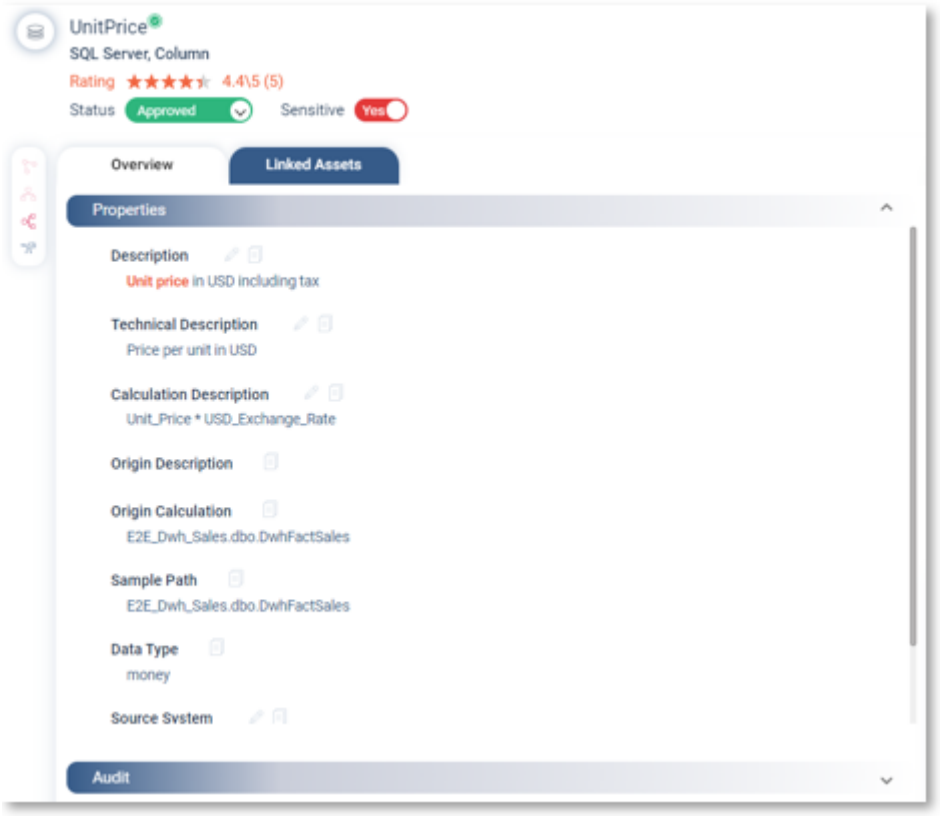


Asset Type	Purpose
Master	To create a single asset to represent multiple assets that would be linked to it.
Business	To hold the business terminology / to create a single asset to represent multiple assets that would be linked to it.
Project	An asset that represents a Project, its scope, who is responsible for it, etc. and can be linked to other assets relevant to the project.
Policy	An asset that represents a Policy, its description, what it applies to, who is responsible for it, etc. and can be linked to other assets associated to the policy.
Report	To Represent Report objects that were not automatically harvested (for example, from an unsupported system).
Analysis	To Represent Analysis objects that were not automatically harvested (for example, from an unsupported system).
Database	To Represent Database objects that were not automatically harvested (for example, from an unsupported system)
ETL	To Represent ETL objects that were not automatically harvested (for example, from an unsupported system)
Data Catalog	General for any other type of asset. (formerly 'ADC Asset')

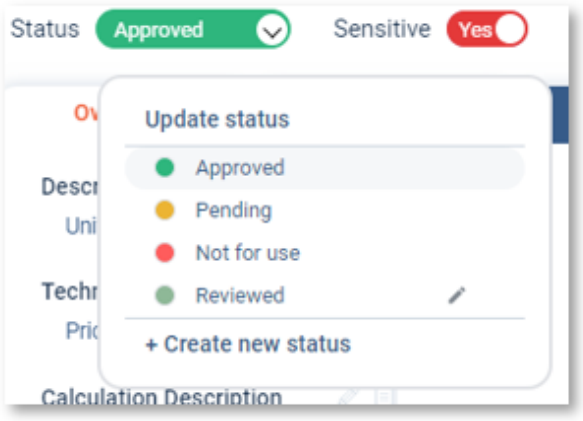
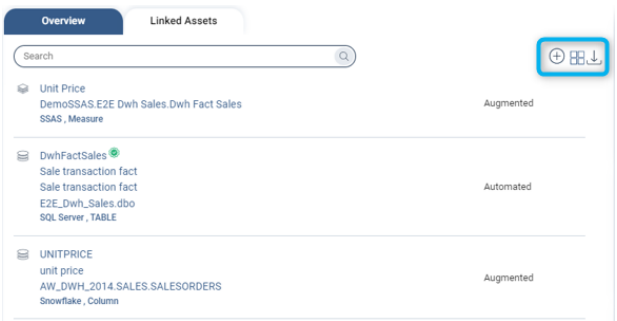
Additional resources

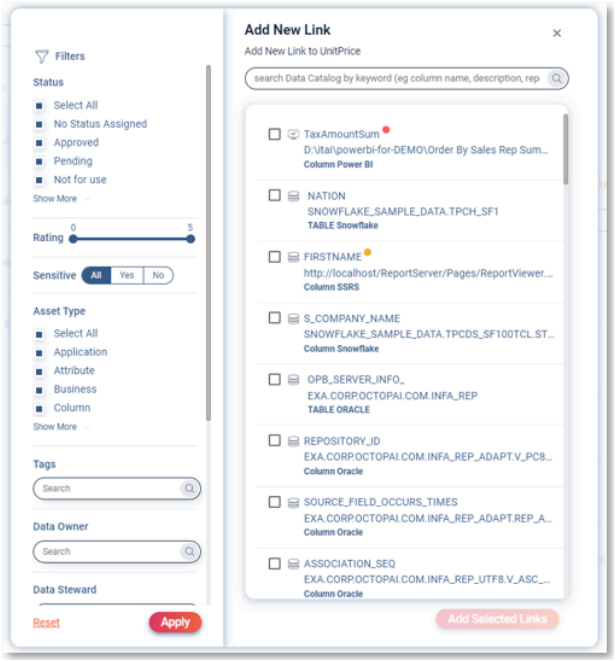
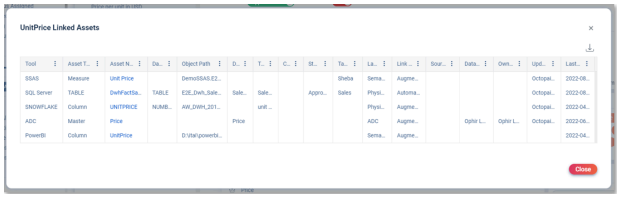
Use the in-product help links and tooltips for contextual guidance while working in the Knowledge Hub.

Asset Details Pane

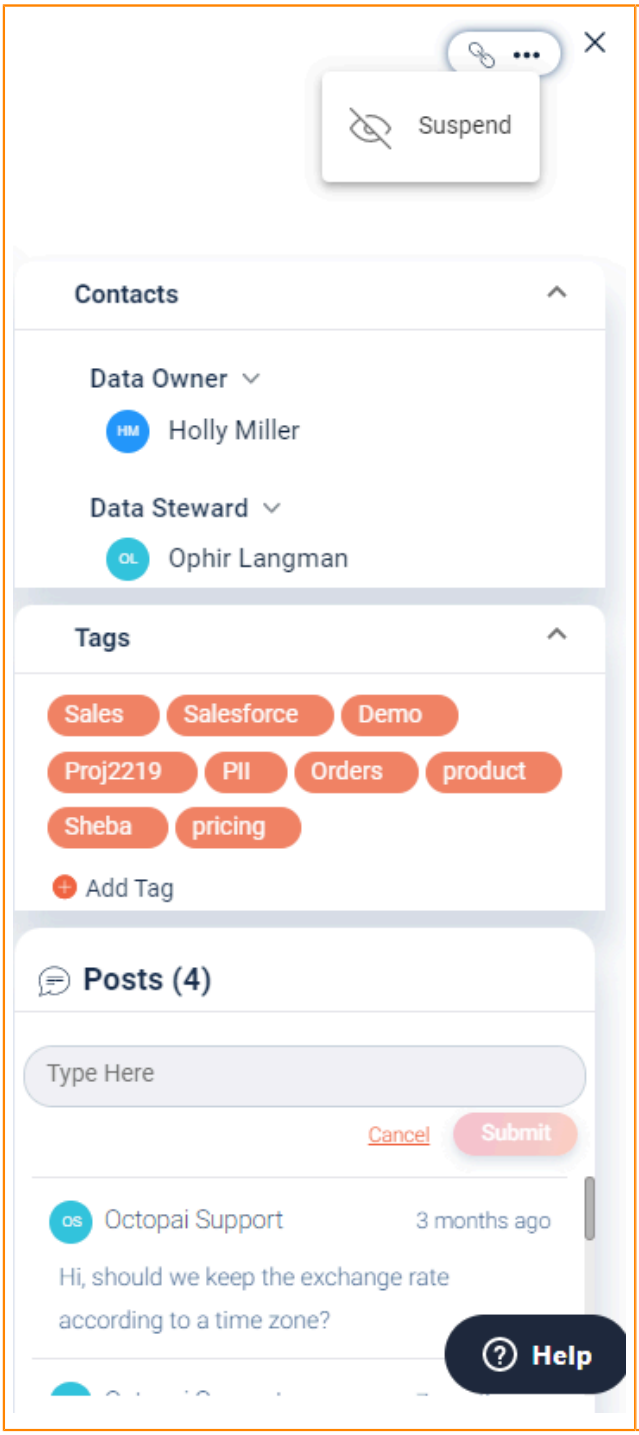


Icon	Description
	**Layer -** (Physical / Semantic / Presentation)
	Asset Name + Status Badge Asset Tool | Asset Type
	**Rating -** Average rating is intended to imply of the quality of the data asset as perceived by the users, the detail pane will display the avg rating as well as amount of ratings. Clicking on the rating will give: a. option to rate b. list of all users that rated the asset and their rating. Range 1-5

Icon	Description
	**Status** - indicates whether the Assets can be used/trusted. Each status is assigned a color to easily be identified. Default statuses ‘Approved', ‘Pending', ’Not for use’ Approved assets add a badge to their asset in the result pane. Admins can add/edit new statuses including assigning a relevant color. **Sensitive** - Assign sensitivity of asset to indicate how the asset can be used. (Yes/No)
	Presentation/Physical Columns integrate with End to End Column Lineage Reports, Views, Procedures, Processes and Functions integrate with Inner System and Cross System Lineage Tables integrate with Cross System Lineage Discovery - Searches for the Asset by Name in the Discovery module
	**Overview** Properties Description Technical Description Calculation Description Origin Description Origin Calculation Sample Path Data Type s Source system (List of Custom Attributes) Audit (Refers to Cloudera Octopai Only) Updated By Entry Date Last Update
	**Linked Assets** **Automated** - non-editable links created as a result of analysis (Example, Report Presentation Assets linked to Report, View columns linked to View). **Augmented** - links that can be added/removed by user.

Icon	Description
 	Create Manual (Augmented) Linked Assets Select from the proposed list the existing asset to create an augmented link.
 	**Grid -** Pop-up window detailing the Linked Assets Properties The list can be downloaded to Excel.
	**Export** - The lists can be downloaded to Excel.

Contacts / Tags / Posts



Suspend (...) - Suspend an asset to disable modifications and gray out the details.

Chain icon – Share the asset location to increase collaboration (copy the unique URL).

X – Close the detail pane.

Contacts – Data owner (technical owner) and data steward (administrative owner).

Tags - Select from existing tags or create new ones.

Posts - Create posts for collaborative discussions.

Within the Knowledge Hub, foster collaboration by creating posts and mentioning other users using the "@" sign and selecting their name from the dropdown list.

When you mention users in a post, they receive email notifications that include the asset name, post details, and a direct link to the relevant asset, ensuring efficient communication and easy access to the discussed information.

All posts are saved and publicly visible, allowing for transparency and enabling users to benefit from shared knowledge and insights.

Knowledge Hub - General Terms

Definitions of key terms related to Cloudera Octopai Knowledge Hub.

Term	Description	Who can edit?
Asset	Any item in the Knowledge Hub is an asset, assets can be automatically created from metadata or users can create new assets within Cloudera Octopai Knowledge Hub.	Changes per attribute

Term	Description	Who can edit?
Knowledge Hub Asset	Assets that are created by users within Cloudera Octopai Knowledge Hub.	·Admin ·Editor
Search	Optional search using free text.	·Admin ·Editor ·Viewer
Advanced Search	Search for free text with advanced conditions.	·Admin ·Editor ·Viewer
Rating	AVG rating is intended to imply the quality of the data asset, the detail pane will display the average rating as well as the number of ratings. Clicking on the rating will give a. option to rate b. list of all users that rated the asset and their rating. Range 1-5	·Admin ·Editor ·Viewer
Status	Status indicates whether the Assets can be used/trusted. Each status is assigned a color to easily be identified. Default statuses ‘Approved', ‘Pending', ’Not for use’ Approved assets add a badge to their asset in the result pane. Admins can add/edit new statuses including assigning a relevant color.	·Admin ·Editor
Sensitive	Assign sensitivity of asset to indicate how the asset can be used. Yes/No	·Admin ·Editor
Technical Description	Short Description - Free text.	·Admin ·Editor
Description	Full Description - Free text.	·Admin ·Editor
Calculation Description	Calculation or Calculation description - Free text.	·Admin ·Editor
Origin Description	The original description as analyzed from metadata.	Non-editable
Origin Calculation	The original Calculation as analyzed from metadata (where applicable).	Non-editable
Asset Type	Asset type e.g. Knowledge Hub, Asset, Table, Column, Report, Process, etc.	Non-editable
Data Type	The original Data Type as analyzed from metadata (where applicable).	Non-editable

Term	Description	Who can edit?
Sample Path	A sample path for where the asset can be found for Aggregated assets such as ‘Report Presentation Layer’ assets and 'Report Physical Layer'. All other Asset types show the direct path.	Non-editable
Source System	Source System - Free text.	·Admin ·Editor
Data Owner	Single selection from drop down list of all system users.	·Admin ·Editor
Data Steward	Single selection from drop down list of all system users.	·Admin ·Editor
Updated By	Last user that updated the asset.	Non-editable
Last Update	Date of last update to the asset.	Non-editable
Entry Date	First appearance of the asset	Non-editable
Tags	Selection from existing tags or create new.	·Admin ·Editor
Linked Assets	Automated - non-editable links created as a result of analysis (Example, Report Presentation Assets linked to Report, View columns linked to View). Augmented - links that can be added/removed by user.	·Admin ·Editor
Suspend	Suspend of an asset will disable modification of any detail and gray them out.	·Admin ·Editor
Search Filters	Search filters contain different attributes related to the assets to enable quickly locating desired data assets with a shopping-like experience, once all attributes are set click on **Apply** to filter results.	·Admin ·Editor ·Viewer
Reset Search Filter	Quickly reset all search filters by clicking on the **Reset** button and then on **apply** .	·Admin ·Editor ·Viewer
Sort results	Sorts search results.	·Admin ·Editor ·Viewer
Post	Collaborate with other users by creating a post and mention other users by typing '@' sign and selecting the user from the dropdown list. Users mentioned in a post will receive an email notification including the asset name, post details and link to the relevant asset. All posts are saved and publicly available.	·Admin ·Editor ·Viewer

Term	Description	Who can edit?
Integration with Lineage	Presentation/Physical Columns integrate with End to End Column Lineage Reports, Views, Procedures, Processes and Functions integrate with Inner System and Cross System Lineage. Tables integrate with Cross System Lineage	Non-editable Useable for any user who has permissions to the integrated module
Search in Discovery	Searches for the Asset by Name in the Discovery module	Non-editable Useable for any user who has permissions to the Discovery Module

Knowledge Hub - Dashboard Terms

Definitions of key terms related to the Knowledge Hub dashboard.

Term	Description	Who can edit?
Search	Optional search using free text.	Admin Editor Viewer
Active	Active - All assets that have been edited in the selected date range.	Non-editable
Missing Description	All assets that have no description in any of the description fields.	Non-editable
Suspended	All suspended assets within the selected date range.	Non-editable
Dropped	All dropped assets within the selected date range.	Non-editable
Top 10 tags	Top 10 most used tags.	Non-editable
Date Selector	Select Date range for Active, Dropped & suspended charts.	Admin Editor Viewer
Sample Result Grid	Sample 1k results per chart with quick actions.	Admin Editor Viewer
Search Filters	Search filters contain different attributes related to the assets to enable quickly locating desired data assets with a shopping-like experience, once all attributes are set click on **Apply** to filter results.	Admin Editor Viewer
Reset Search Filters	Quickly reset all search filters by clicking on the **Reset** button and then on **Apply**.	Admin Editor Viewer

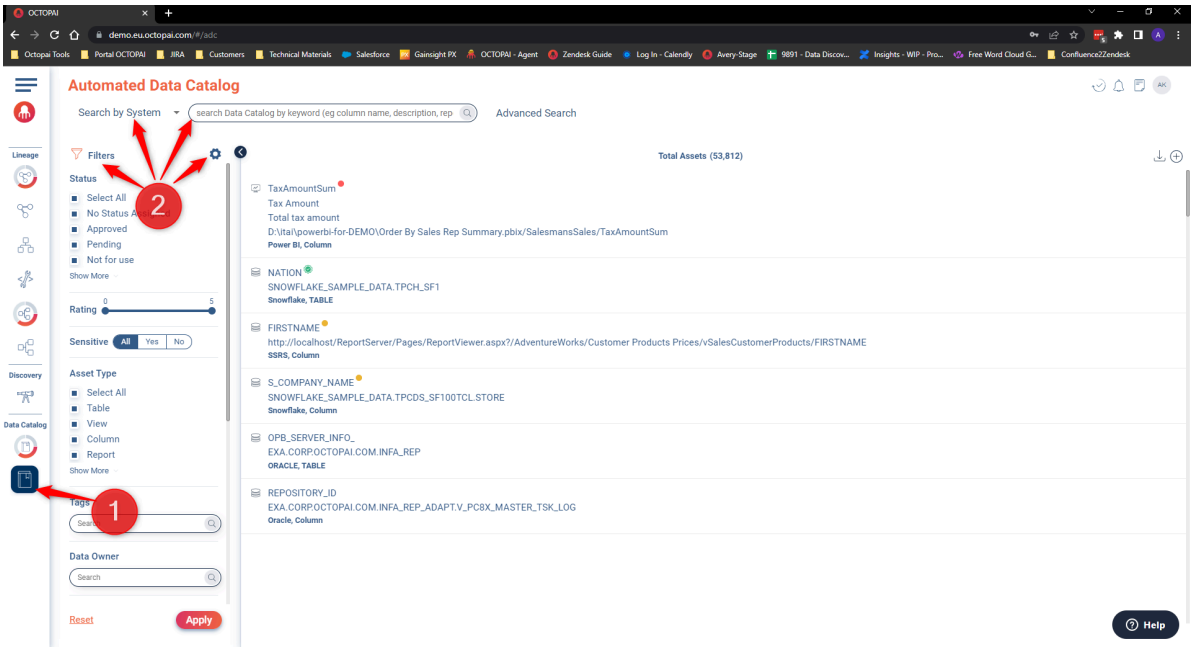
Knowledge Hub - Mass Update

Learn about simplifying Cloudera Octopai Knowledge Hub curation by downloading assets to an Excel spreadsheet to enable effortless information definition and organization and dividing the workload among team members for efficient collaboration.

Procedure

1. Download assets to an Excel spreadsheet.
   a) Select a group of assets.

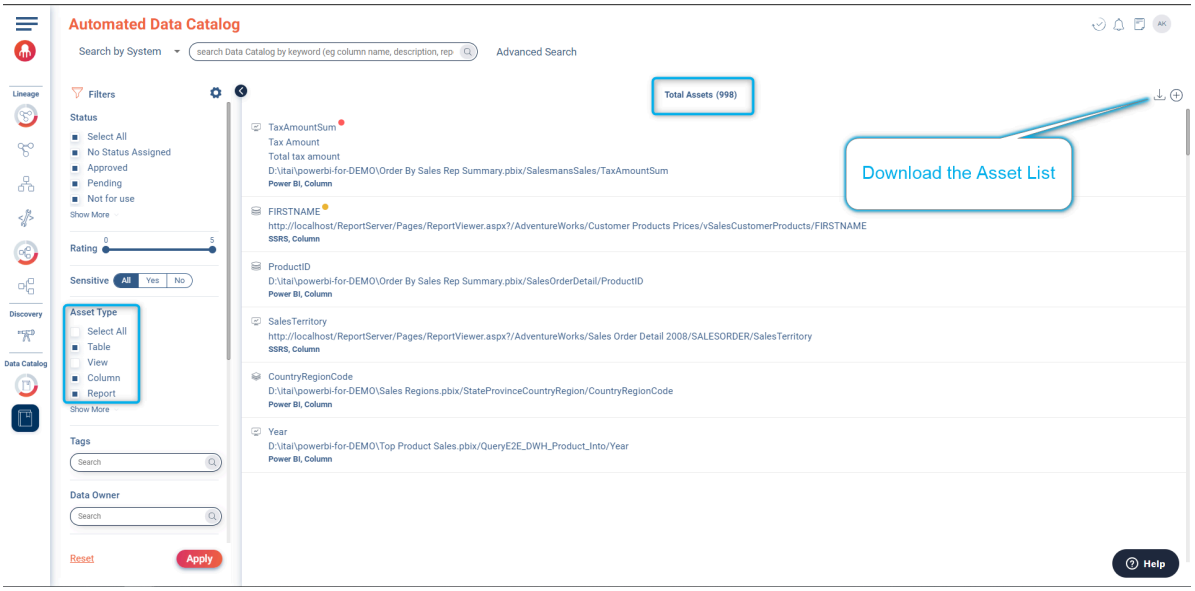
Figure 1: Select a group of assets



1. Go to the Knowledge Hub.
2. Combine the Search by System, Filters, Search box / Advanced Search, and Layer Settings Knowledge Hub functionalities to better filter your assets.

b) Click the icon to download the Assets List.

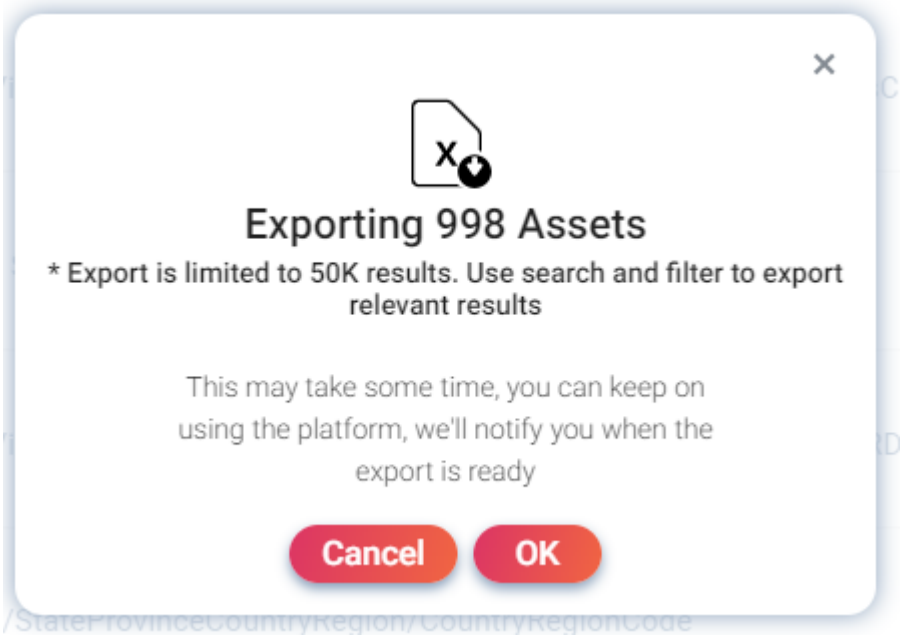
Figure 2: Asset List download



c) Click OK in the pop-up window.

Note: The export is limited to 50,000 results.

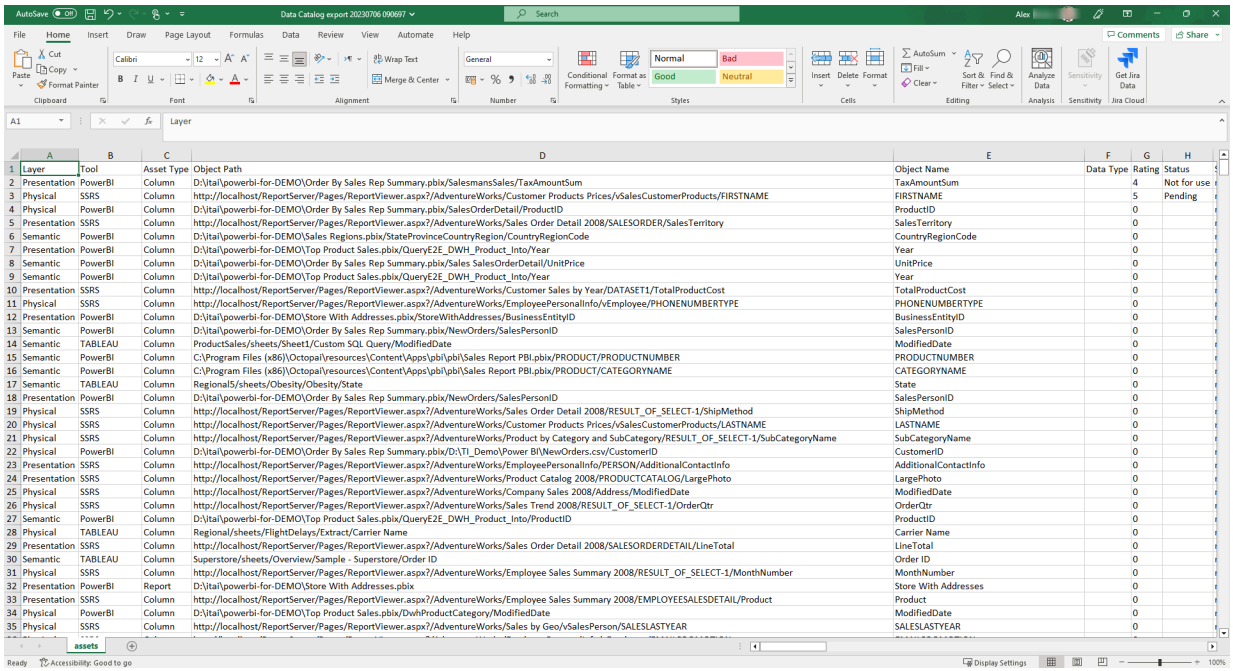
Figure 3: Export pop-up window



2. Curate the Knowledge Hub.

Warning: Do not change column sorting, since the Cloudera Octopai Knowledge Hub will be updated in bulk using this template.

Figure 4: Knowledge Hub curation



Layer	Tool	Asset Type	Object Path	Object Name	Data Type	Rating	Status
Presentation	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/SalesmansSales/TaxAmountSum	TaxAmountSum		4	Not for use
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Customer Products Prices/vSalesCustomerProducts/FIRSTNAME	FIRSTNAME		5	Pending
Physical	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/SalesOrderDetail/ProductID	ProductID		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Sales Order Detail 2008/SALESORDER/SalesTerritory	SalesTerritory		0	
Semantic	PowerBI	Column	D:\tal\powerbi-for-DEMO\Sales Regions.pbix/StateProvinceCountryRegion/CountryRegionCode	CountryRegionCode		0	
Presentation	PowerBI	Column	D:\tal\powerbi-for-DEMO\Top Product Sales.pbix/QueryE2E_DWH_Product_Into/Year	Year		0	
Semantic	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/Sales SalesOrderDetail/UnitPrice	UnitPrice		0	
Semantic	PowerBI	Column	D:\tal\powerbi-for-DEMO\Top Product Sales.pbix/QueryE2E_DWH_Product_Into/Year	Year		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Customer Sales by Year/DATASET1/TotalProductCost	TotalProductCost		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/EmployeePersonalInfo/vEmployee/PHONENUMBERTYPE	PHONENUMBERTYPE		0	
Presentation	PowerBI	Column	D:\tal\powerbi-for-DEMO\Store With Addresses.pbix/StoreWithAddresses/BusinessEntityID	BusinessEntityID		0	
Semantic	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/NewOrders/SalesPersonID	SalesPersonID		0	
Semantic	TABLEAU	Column	ProductSales/sheets/Sheet1/Custom SQL Query/ModifiedDate	ModifiedDate		0	
Semantic	PowerBI	Column	C:\Program Files (x86)\Octopai\resources\Content\Apps\pbi\pbi\Sales Report PBI.pbix/PRODUCT/PRODUCTNUMBER	PRODUCTNUMBER		0	
Semantic	PowerBI	Column	C:\Program Files (x86)\Octopai\resources\Content\Apps\pbi\pbi\Sales Report PBI.pbix/PRODUCT/CATEGORYNAME	CATEGORYNAME		0	
Semantic	TABLEAU	Column	Regional5/sheets/Obesity/Obesity/State	State		0	
Presentation	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/NewOrders/SalesPersonID	SalesPersonID		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Sales Order Detail 2008/RESULT_OF_SELECT-1/ShipMethod	ShipMethod		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Customer Products Prices/vSalesCustomerProducts/LASTNAME	LASTNAME		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Product by Category and SubCategory/RESULT_OF_SELECT-1/SubCategoryName	SubCategoryName		0	
Physical	PowerBI	Column	D:\tal\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/D:\TI_Demo\Power BI\NewOrders.csv/CustomerID	CustomerID		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/EmployeePersonalInfo/PERSON/AdditionalContactInfo	AdditionalContactInfo		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Product Catalog 2008/PRODUCTCATALOG/LargePhoto	LargePhoto		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Company Sales 2008/Address/ModifiedDate	ModifiedDate		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Sales Trend 2008/RESULT_OF_SELECT-1/OrderQtr	OrderQtr		0	
Semantic	PowerBI	Column	D:\tal\powerbi-for-DEMO\Top Product Sales.pbix/QueryE2E_DWH_Product_Into/ProductID	ProductID		0	
Physical	TABLEAU	Column	Regional/sheets/FlightDelays/Extract/Carrier Name	Carrier Name		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Sales Order Detail 2008/SALESORDERDETAIL/LineTotal	LineTotal		0	
Semantic	TABLEAU	Column	Superstore/sheets/Overview/Sample - Superstore/Order ID	Order ID		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Employee Sales Summary 2008/RESULT_OF_SELECT-1/MonthNumber	MonthNumber		0	
Presentation	PowerBI	Report	D:\tal\powerbi-for-DEMO\Store With Addresses.pbix	Store With Addresses		0	
Presentation	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Employee Sales Summary 2008/EMPLOYEESALESDETAIL/Product	Product		0	
Physical	PowerBI	Column	D:\tal\powerbi-for-DEMO\Top Product Sales.pbix/DwhProductCategory/ModifiedDate	ModifiedDate		0	
Physical	SSRS	Column	http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Sales by Geo/vSalesPerson/SALESLASTYEAR	SALESLASTYEAR		0	

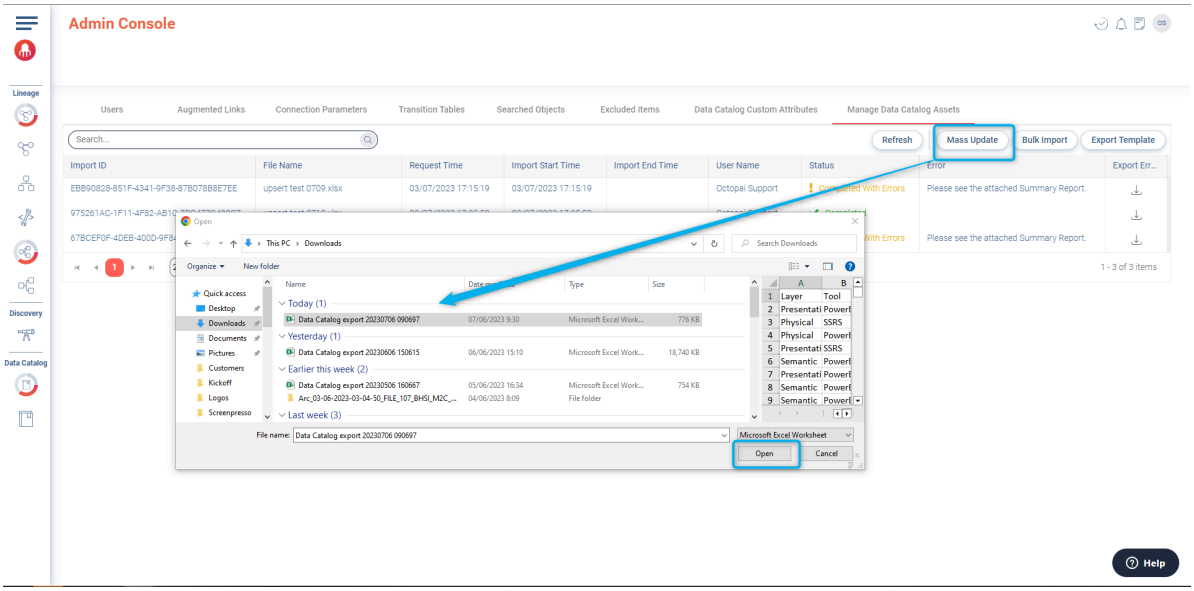
Customize data distribution based on your company requirements, allowing different team members to contribute the necessary information. Consolidate all curated definitions into the original downloaded file.

Tip: Provide the following information based on the requirements:

- Description: Full Description
- Technical Description: Short Description
- Calculation Description: Calculation or Calculation description
- Source System: Free Text
- Data Owner: Fill in the email of the owner that must be an existing Cloudera Octopai user
- Data Steward: Fill in the email of the owner that must be an existing Cloudera Octopai user
- Tags: When tagging an asset with multiple values, use a comma (,) as a separator within the same cell.

3. Perform mass update.
   a. Go to Admin Console.
   b. Go to the **Manage Knowledge Hub Assets** tab.
   c. Click Mass Update and browse the curated excel file.

Figure 5: Mass update



   d. Click Open.

   During the upload, assets will be updated. The completion time of the process will vary based on the number of assets and available machine resources.

Results

Check the process status.

Status	Description
✔ Completed	All the rows were updated successfully.
! Completed With Errors	Potential errors occurred during the update. To identify them, download the summary report by clicking the icon.

Knowledge Hub Implementation Best Practices

Best practices for implementing Knowledge Hub.

Step	Description
Define a clear purpose and scope	Clearly define the purpose and scope of the Knowledge Hub, including: • The types of data that will be included. • The intended audience. • The business goals that the Knowledge Hub is intended to support.
Identify and involve stakeholders	Identify key stakeholders in the data team and in the business teams and involve them in the design and implementation process to ensure that the Knowledge Hub meets their needs and requirements.

Step	Description
Establish data governance policies	Establish data governance policies and workflows within the organization. This will ensure that the Knowledge Hub is accurate, up-to-date, and secure. This includes defining data standards, access controls, and data quality measures.
Use Knowledge Hub metadata standards	Knowledge Hub metadata standards and data models such as (the same header, a description must-have, etc.) to ensure that the Knowledge Hub is consistent and interoperable with other systems and data sources.
Automate metadata capture	Use Cloudera Octopai to capture metadata from various sources.
Define milestones	Defining milestones is an important part of the process of populating your Knowledge Hub. Here are some steps you can take to define milestones for populating your Knowledge Hub: • Identify the data assets to be documented: prioritize (according to the next step) the data assets that will be populated with descriptions in the Knowledge Hub. This could include databases item only, Reporting assets, or assets by important projects. • Define metadata requirements: Define the metadata requirements for each data asset, including the level of detail required and additional asset information not captured by Cloudera Octopai. • Create a timeline: Create a timeline that identifies key milestones for populating the Knowledge Hub. This timeline should include the start and end dates for the project, as well as specific milestones for each phase of the project. • Define phases of the project: Define the phases of the project, such as Knowledge Hub population by different projects, business groups, critical assets, and other asset types as appear in the next step. • Assign responsibilities: Assign responsibilities for each phase of the project to ensure that all tasks are completed on time and to the required quality standards. • Establish quality control measures: Establish quality control measures to ensure that the metadata captured is accurate, complete, and consistent with established standards. • Monitor progress: Monitor progress against the timeline and adjust the plan as necessary to ensure that the project stays on track and meets its milestones. By following these steps, you can create a comprehensive plan for populating your Knowledge Hub and define clear milestones that will help you track progress and ensure that the project is completed on time and to the required quality standards.

Step	Description
Data Assets Prioritization	When populating your Knowledge Hub, it's important to prioritize the data assets that are most critical to the organization's operations and that are likely to have the greatest impact on business outcomes. Here are some guidelines for what to prioritize when populating your Knowledge Hub: • Business-critical data: Start by populating your Knowledge Hub with descriptions for the data assets that are most critical to the organization's operations, such as financial data, customer data, and product data. • High-value data: Prioritize data assets that are high in value, either because they are frequently used, asked about, or because they have a high impact on business outcomes. • Frequently used data: Prioritize data assets that are frequently used by the organization, as these are likely to have the greatest impact on productivity and efficiency. • Data that is difficult to find: Identify data assets that are difficult to find or that are scattered across multiple systems or applications. Populating your Knowledge Hub with metadata for these assets can help improve accessibility and reduce the time and effort required to locate and use the data. • New data assets: When new data assets are added to the organization, prioritize adding metadata for these assets to the Knowledge Hub as soon as possible to ensure that they are discoverable and accessible to users.
Populate the Knowledge Hub	Collaborate with data owners: Work with the data owners or subject matter experts to obtain information about the data assets they own. This can include metadata such as the data source, data lineage, data quality, and data usage information. You can use this information to populate your Knowledge Hub manually.
Provide user-friendly search and discovery capabilities	Train users on how to search and discover in Cloudera Octopai, which enables users to quickly find and access the data they need. This includes providing filters, tags, owners, and other Cloudera Octopai Knowledge Hub search capabilities.
Monitor usage and adoption	Monitor usage and adoption of the Knowledge Hub to ensure that it is meeting the needs of the organization and that users are leveraging its capabilities effectively.
Provide ongoing maintenance and support	Provide ongoing maintenance and support for the Knowledge Hub, including regular updates and enhancements to ensure that it remains relevant and useful over time.

Introduction to Knowledge Hub Implementation for Point Person

Cloudera Octopai Knowledge Hub automates metadata harvesting from ETL, DB, and Reporting systems to populate data assets, reducing manual effort and enhancing efficiency.

Take maximum advantage of the Knowledge Hub automation

If you have additional documentation, such as excel spreadsheet with descriptions, Cloudera Octopai can upload this to the Knowledge Hub as well. Everything you give it will enable it to produce fuller, more comprehensive data asset entries, necessitating less human involvement and saving your users time and effort.

Knowledge Hub entries still need human review and enrichment to be of maximum benefit. To execute this in the most effective way possible, consider the following recommendations:

- **Concentrate initial efforts on enriching the Knowledge Hub to the most used data assets.**

  As almost everything else, your company’s data assets subscribe to the 20-80 rule. About 20% of your data assets are responsible for 80% of your company’s data use and value you derive from data. Identify those assets and assign responsible parties to actively go in and enrich their Knowledge Hub entries.

- **Emphasize to participants how this Knowledge Hub will make life and work easier.**

  The friendly introductions Cloudera Octopai provided for you have that focus. Use that model when you communicate personally as well. For example, when assigning a technical user the responsibility to enrich a Knowledge Hub entry that means more work for him, Cloudera Octopai recommends couching it the following way:

  So, Joe, I see you created this great data asset that so many people in our organization could gain from. But very few people know about it, and of those who do, almost all of them have the same questions that you have to answer over and over again. I'm going to put you down in the Knowledge Hub as the data asset owner so that you can add anything users should know in order to use your data effectively. You'll also be able to answer any questions - once and only once - in the notes on the asset entry. Many more people will be able to leverage the asset you created, but fewer will disturb you unnecessarily.
- Add updating the Knowledge Hub to pre-existing data management routines.
- **Integrate human checks and updates to Knowledge Hub entries into your processes for handling change requests, user inquiries, new data assets and other pre-existing data management responsibilities, including the following actions:**

  - Make it a standard part of a new project to document its data assets in the Knowledge Hub
  - Make it a step in closing a change request or a deployment to update the Knowledge Hub for the specific asset that the change request addressed

  Just adding these checkboxes to your existing checklists and routines can be a powerful Knowledge Hub management tool. Slowly and painlessly, your Knowledge Hub will become more comprehensive and even more effective.

Cloudera Octopai aims to help you leverage your Data Catalog. If you have any questions or issues, contact Cloudera Support.

Introduction to Knowledge Hub Implementation for Data Engineering Teams

Learn about how the Knowledge Hub empowers data teams to focus on impactful projects while ensuring consistent and transparent access to data assets across the organization.

A story of data team freedom

The data team at a large enterprise spent an inordinate amount of their time putting out fires by answering urgent questions or requests, such as Why is this metric not showing up in this report?, We need it for an audit - tomorrow!, and fielding self-service BI user questions, such as I need a dataset that shows export sales per product category for last quarter, but I found three datasets that seem to cover that and they all have slightly different numbers. What's the difference between them?.

The situation was most extreme for senior members of the team who had amassed a large amount of tribal knowledge. They consistently received more than their fair share of questions, both from colleagues within the data team and from the rest of the organization.

The data team's passion was developing new data projects, but they had much less time to spend on that aspect of their work than they wanted. The constant minor emergencies and customer support requests interrupted the flow of and affected the quality of their work.

Then the enterprise invested in Cloudera Octopai Knowledge Hub. The initial construction of the Knowledge Hub was completely automated. Cloudera Octopai harvested and ingested their data systems' metadata and then integrated all the external documentation uploaded by the data team.

Already the data team sensed that they were getting fewer support questions. And this was only the beginning.

Then the time came to actively enhance and enrich the Knowledge Hub assets. The data team identified the data assets they most commonly got questions about and they added descriptions and roles to those asset entries. Any time a team member received a question, they directed the questioner to the Knowledge Hub built-in collaboration tools,

requesting that they ask the question there. When the relevant data engineer wrote out the answer and added it to the asset entry, they knew they were answering the question once and for all.

Over the next four months, the data engineering team saw the number of questions within the Knowledge Hub increase and the amount of communication outside the Knowledge Hub decrease significantly. In addition, the Knowledge Hub became the primary, and eventually the only, source of documentation on data assets, putting every department and every employee in the company on the same page. Finally, there was one source of truth.

With the increased consistency, accuracy and transparency around the company's data assets, the enterprise data team found themselves called upon less to do troubleshooting and emergency management, leaving them with time and peace of mind to do the data work that they truly enjoyed.

When the Knowledge Hub can save the day

The following cases can help you understand the benefits of the Knowledge Hub:

- This is the fifth time this week a business user has asked me 'how is customer life time value calculated?' Every time it's an interruption from the work I'm trying to get done.

  Your Knowledge Hub empowers your business users to find the answers on their own, with definitions, descriptions and more for every data asset in your systems. If a user has a question that calls for more expert or technical knowledge, built-in communication tools let them ask and you answer right there in the Knowledge Hub entry. So you only have to answer any question once - and then the answer is forever recorded and easily accessible.
- Whenever I receive a change request for a data asset, it takes me longer to find documentation on what's been done on that data asset already than it does to actually make the change.

  Your Knowledge Hub functions as a one-stop-shop for all data information and documentation. This centralized, always up-to-date Knowledge Hub significantly shortens the amount of time you need to put into preparing for a change to a data asset or a business process. Communication about data assets is always about the most updated version: no more misunderstandings or miscommunications. Every additional asset you organize or documentation detail you add will pay exponential returns in future time and resources saved.
- Regulations require us to identify, mask and limit access to PII. But it feels like such a waste of time to have to check all our assets manually every time we have an audit coming up.

  Knowledge Hub entries provide a place for noting the sensitivity level or access permissions of each data asset. The search function of your Knowledge Hub enables you to easily locate all data assets with any given sensitivity level. Audit preparation - or any other task involving identification of PII - becomes much quicker and easier.

How you can make the Knowledge Hub a success

Knowledge Hub entries still need human review and enrichment to be of maximum benefit. To execute this in the most effective way possible, consider the following recommendations:

- **Concentrate initial efforts on enriching the Knowledge Hub to the most used data assets.**

  As almost everything else, your company's data assets subscribe to the 20-80 rule. About 20% of your data assets are responsible for 80% of your company's data use and value you derive from data. Identify those assets and assign responsible parties to actively go in and enrich their assets.
- Add updating the Knowledge Hub to pre-existing data management routines.
- **Integrate human checks and updates to Knowledge Hub entries into your processes for handling change requests, user inquiries, new data assets and other pre-existing data management responsibilities, including the following actions:**

  - Make it a standard part of a new project to document its data assets in the Knowledge Hub
  - Make it a step in closing a change request or a deployment to update the Knowledge Hub for the specific asset that the change request addressed

  Just adding these checkboxes to your existing checklists and routines can be a powerful Knowledge Hub management tool. Slowly and painlessly, your Knowledge Hub will become more comprehensive and even more effective.

Introduction for Knowledge Hub for Business Users

Learn about how Cloudera Octopai Knowledge Hub empowers business users to find, trust, and collaborate around critical data assets.

Data is supposed to help us make more informed, accurate business decisions.

But no matter how much data your organization has, if it's hard to find, complicated to evaluate and tricky to use, it might as well not be there.

A data catalog makes finding, understanding and using business data as intuitive as choosing and purchasing a new shirt from an online marketplace.

When data powers sales and profits

The marketing team at a medium-sized consumer brand was tasked with coming up with ideas for campaigns promoting the company's new product line. In the past, gathering data upon which to base the marketing campaigns had been a frustrating experience strewn with landmines:

- It was hard to locate all the relevant data on previous marketing campaigns and audience targeting.
- Each manager thought that his/her report is the most accurate report (i.e. no one source of truth).
- When numbers looked odd, the marketing team had no way of knowing for sure if the data was accurate or not.
- Inquiries to other departments to find the right person to ask about the inaccuracies and the versions rarely succeeded in turning up helpful information. Often they did eventually track someone down, but sometimes the party in question only had partial answers, and sometimes it was too late to make a difference for the campaign they had to deliver the next day.

The marketing team rarely felt fully comfortable with the data they were using to guide their campaign creation. Either they reluctantly used what data they had, having nothing better, or they just put the data to the side, not trusting it enough to make using it worth it. The marketing team often felt that their campaigns could be stronger and more successful if only they had the right data resources, but what to do? They didn't.

This time, however, was different. The company had invested in a data catalog a few months earlier and the business team had been involved in the implementation, tagged the relevant marketing data, and validated its relevancy. So, the marketing team's first stop in searching for relevant data from which to create the new campaign was the data catalog's search and tagging functionality. Within seconds they had a comprehensive list of all marketing-, campaign- and audience-related datasets.

Now it was time to choose the most relevant datasets. The marketing team narrowed down the list by tags and data owners, preferring data assets that had been used in more than 100 reports and were recently maintained.

Of the remaining data assets, two of them looked like they were measuring audience engagement with the previous season's campaigns, but the numbers were significantly different. One of the marketing managers posted a question about the difference in the communication section of each data asset's catalog entry. Within an hour, they had an expert answer. The marketing team now understood what the difference was and which asset was more relevant for their purposes, and this exchange was saved in the catalog for future reference.

This season's product marketing campaigns were compiled faster, and much more agile, targeted and effective than any the marketing team had designed before. These data catalog-powered campaigns took them less time to research and plan than usual. At the same time, they experienced less stress and felt less dependent on the company's data team.

More sales and more profit, in less time and with less stress. Knowledge Hub win for business.

When the Knowledge Hub can save the day

The following scenarios can help you understand the benefits of the Knowledge Hub:

- I spent a long time creating a report (or a dataset)… and after I finished I found out that it existed already.

  It is a frustrating waste of time and energy to duplicate what already exists. It is almost inevitable when data assets are not organized.

  Your Knowledge Hub platform serves as a single, searchable repository for all data assets your company possesses.
- It takes me SO long to find the right data for my project.

  Without a Knowledge Hub, getting the data you need for a project is reminiscent of a poor blind dating experience. Based on whatever documentation exists, and some talking to colleagues, you are set up with a data asset you think has promise but are sometimes severely disappointed.

  Your Knowledge Hub removes the blindness from your data matchmaking experience. You can search for what you think would make a good match, and then see your potential dates in the illuminating context of objective information about them, for example what accomplishments have they had, where do they usually go on dates, how many people have gone out with them, and subjective information, for example what prior relationships say about them. You can even ask questions to people who know them. The chances of meeting Mr. or Ms. Right Data go way, way up.
- Whenever I have a question about a data asset, it takes forever to get in touch with someone who can give me an answer.

  Who is the data owner? Who is the data steward? Is there a subject matter expert in the house or did she jump ship a month ago without a forwarding address?

  Your data catalog can help in the following ways:

  - Clearly identifies the important roles for every data asset.
  - Provides communication channels from within the catalog itself for getting answers and clarifications from those responsible for the data asset.
  - Records these questions and conversations within the data asset entry, making this valuable tribal knowledge available to every future user who looks at the entry.

How you can make the Knowledge Hub a success

To help Knowledge Hub be a success, consider the following recommendations:

- **Add to it.**

  The main bulk of the Knowledge Hub is created using automated processes. Once that is done, the data asset entries can be enriched with information that is not usually found in documentation, such as calculations, topic tags, descriptions, or ratings.

  If you have created a data asset, making you the owner of that asset, check the automation created entry of that asset and fill in anything that is missing. If you are asked questions in the Knowledge Hub's communication and collaboration tools, answer them. The more complete a data asset documentation entry is, the more people in your company can find, use and benefit from the data, while fewer will disturb you unnecessarily with questions that you have already answered.

  If you use a data asset and find it helpful (or unhelpful), leave a brief user review in the asset documentation entry so that your colleagues can benefit from your experience and advice.

- **Use it.**

  With a Knowledge Hub, you can see what data you have and find what data you need. You can rapidly locate decision-critical data and put it to use, speeding up time to insight and increasing business agility.

  If you are looking over a colleague's report, you can easily check up on which data assets were used in the report, what details about the data the report includes, and understand the scope of the data and any constraints on the dataset or on the report itself.

  Additionally, the more you use the Knowledge Hub, the more it will be enriched with tribal knowledge, and the more helpful it will get in its search results and recommendations.

  When you want to do research, create a report or understand someone else's report, make the Knowledge Hub your first stop.

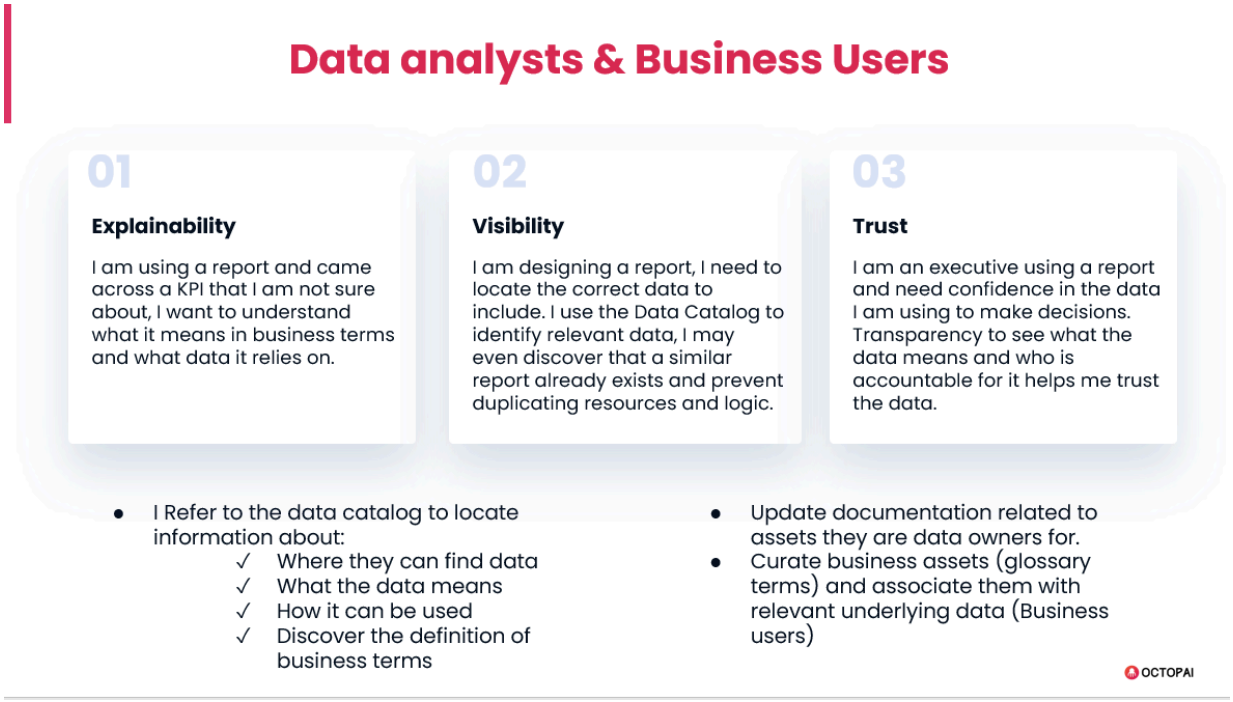
Knowledge Hub Personas

Overview of the primary personas who collaborate inside the Knowledge Hub.

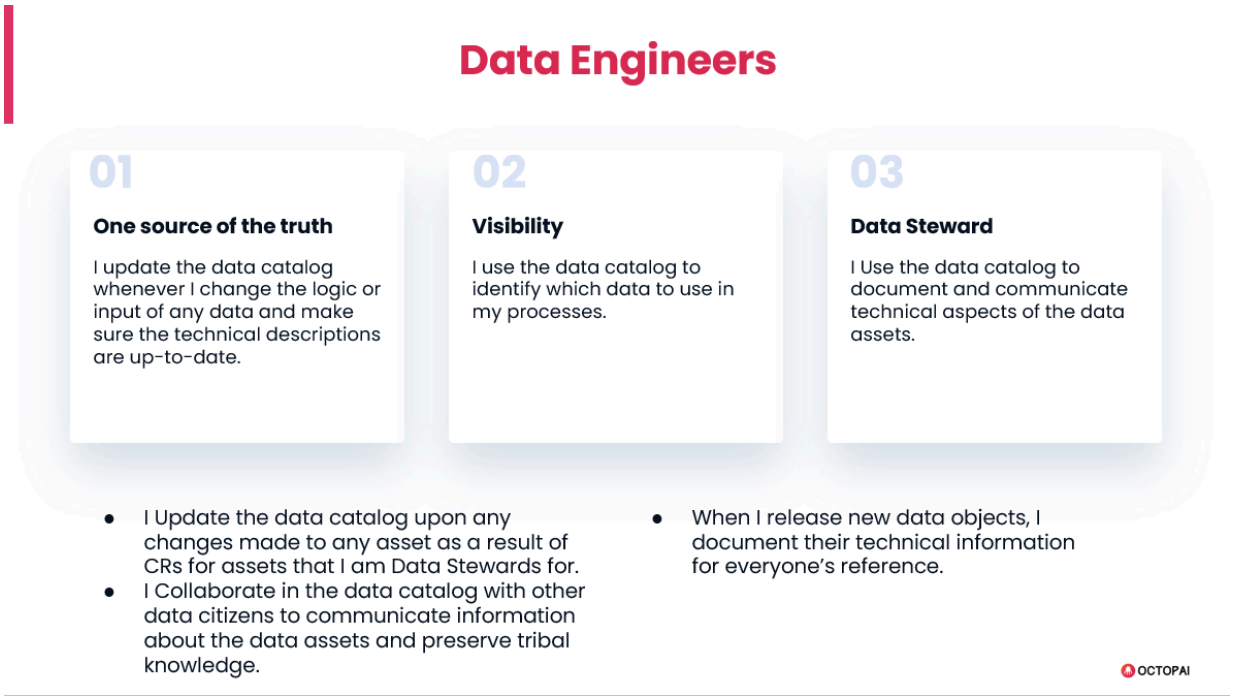
The following personas collaborate withing the Knowledge Hub:

- Collaboration

- Data Analysts and Business Users



- Data Engineers



- Data Governance Managers

Cloudera Octopai Knowledge Hub and Insight Dashboard

This guide explores the key product capabilities of Cloudera Octopai Knowledge Hub, along with tips to maximize its usage.

Introduction

Cloudera Octopai Knowledge Hub is designed to empower data teams in managing and ensuring the integrity of their data assets. By seamlessly integrating data assets and data lineage, providing flexible search options, and offering a pre-defined dashboard for monitoring, collaboration, and auditing, Cloudera Octopai enables organizations to effectively manage their data, maintain compliance, and make informed decisions.

Asset Management and Data Lineage Integration

Cloudera Octopai Knowledge Hub seamlessly integrates data assets and data lineage, allowing users to navigate between them effortlessly. Key capabilities include:

1. Data Lineage Integration: Cloudera Octopai Knowledge Hub provides full integration with the Cloudera Octopai platform, enabling users to easily navigate from a data asset to its data lineage and vice versa. This integration ensures a comprehensive view of data, making it easier to trace the origin, transformations, and dependencies of data assets.
2. Utilization for Auditing: The Knowledge Hub's integration with data lineage allows for comprehensive auditing capabilities. Users can track the lineage of data assets, ensuring compliance with regulatory requirements and providing transparency for auditing purposes.
3. Indication of Sensitive Data Assets: Cloudera Octopai Knowledge Hub allows for the identification and indication of sensitive data assets. With customizable attributes and tags, users can label and flag sensitive data, ensuring proper handling, access controls, and compliance with data privacy regulations.
4. Fast Implementation: Cloudera Octopai Knowledge Hub is designed for rapid implementation, allowing data teams to quickly set up and start leveraging its capabilities. This enables organizations to accelerate their data management initiatives and realize value in a short timeframe.

Flexible Search and Custom Filters

Cloudera Octopai Knowledge Hub offers a flexible search functionality and custom filters, allowing users to quickly locate and narrow down specific data assets. Additional capabilities include:

1. Flexible Search: The Knowledge Hub provides a powerful search feature, enabling users to search for data assets using free text. This flexibility simplifies the process of finding relevant data assets based on specific keywords or phrases.
2. Custom Filters: Users can create custom filters to refine search results further. This feature allows data teams to tailor their searches to specific criteria, such as asset types, data owners, tags, or sensitive data indicators, streamlining the discovery of relevant data assets.

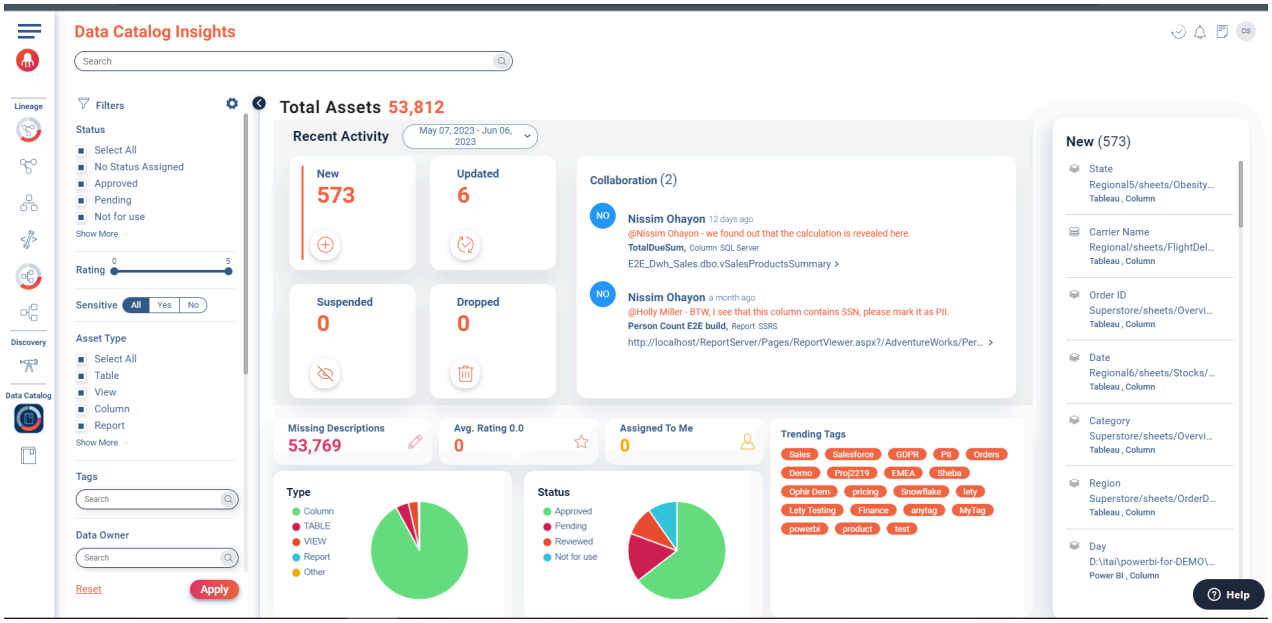
Knowledge Hub - Insight Dashboard and Collaboration

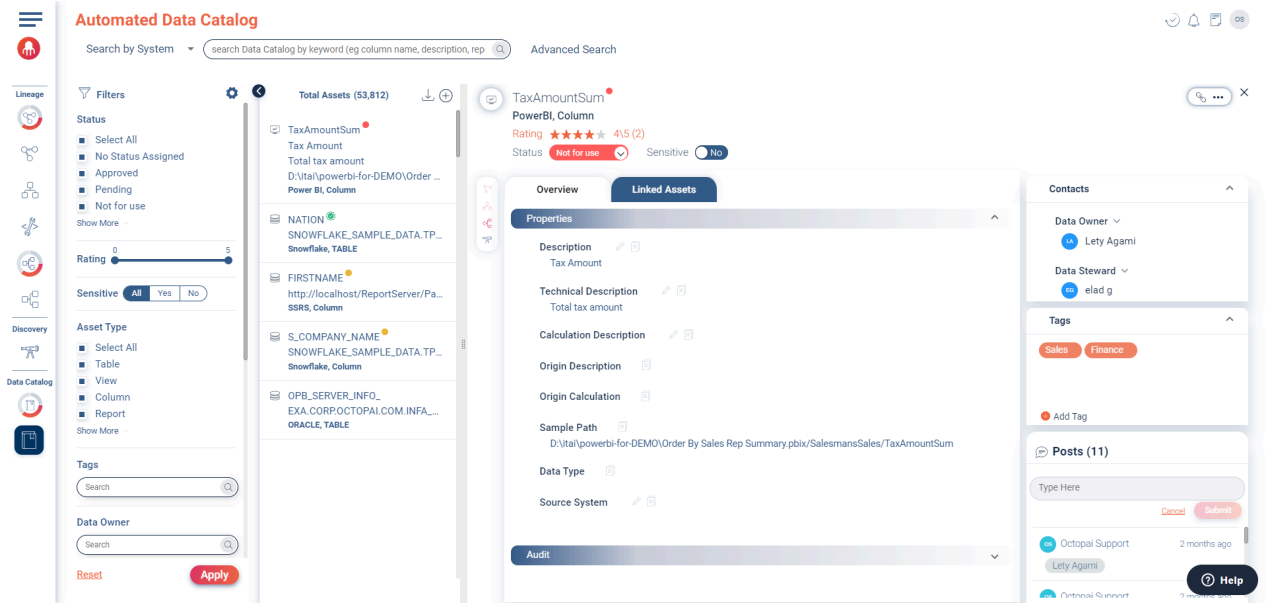
Cloudera Octopai Knowledge Hub offers a pre-defined dashboard that provides valuable insights into data assets and promotes collaboration among users. Key capabilities of the Knowledge Insight dashboard include:

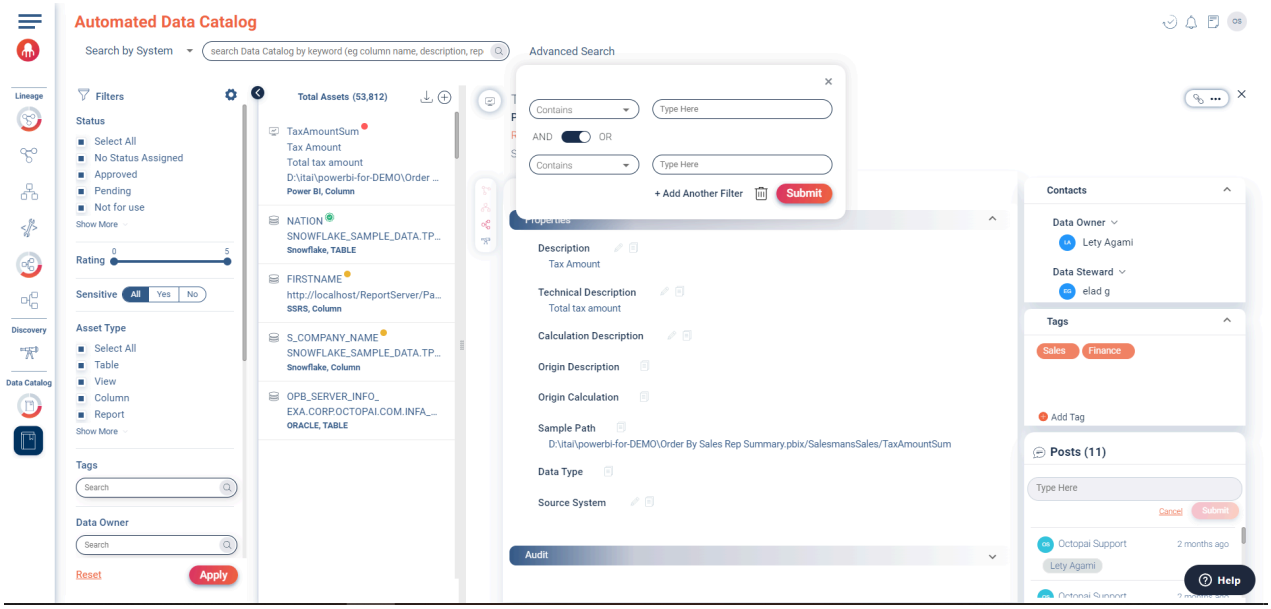
1. Monitoring and Reporting: The dashboard allows users to monitor the latest activities within the Knowledge Hub, providing real-time updates on asset modifications, user interactions, and overall data asset health. This helps users stay informed and make data-driven decisions based on current information.
2. Collaboration and Governance: The Knowledge Hub facilitates collaboration among users, particularly between asset owners and data stewards. This collaboration ensures that data assets are managed, maintained, and governed by the responsible individuals, promoting data integrity, accuracy, and compliance.

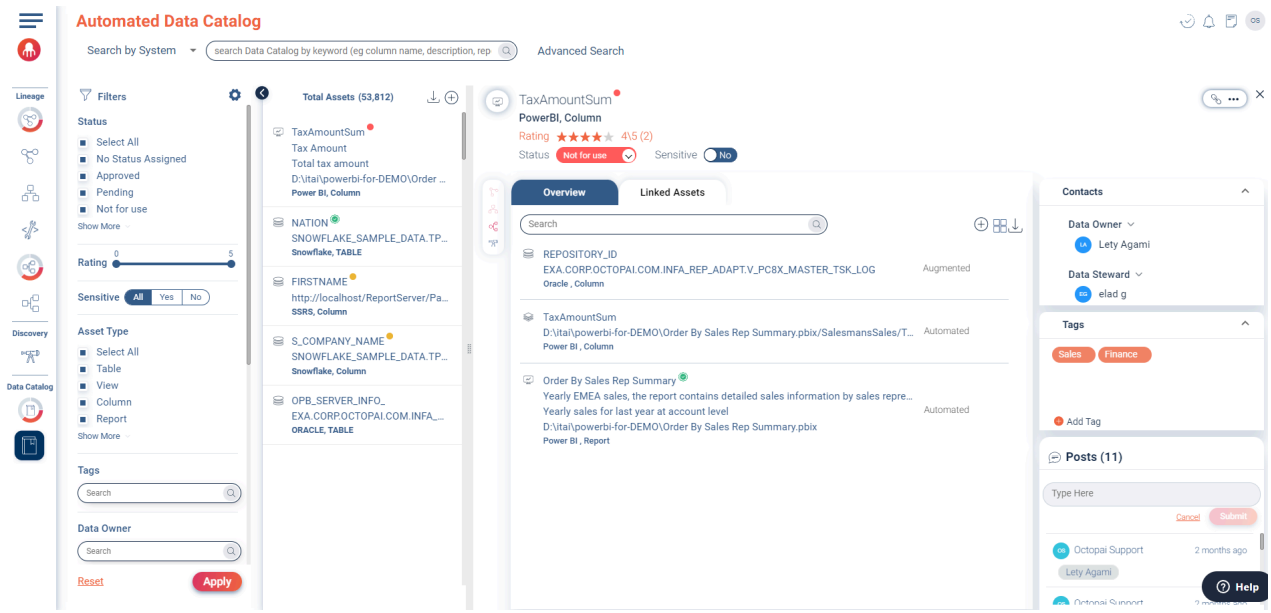
Benefits

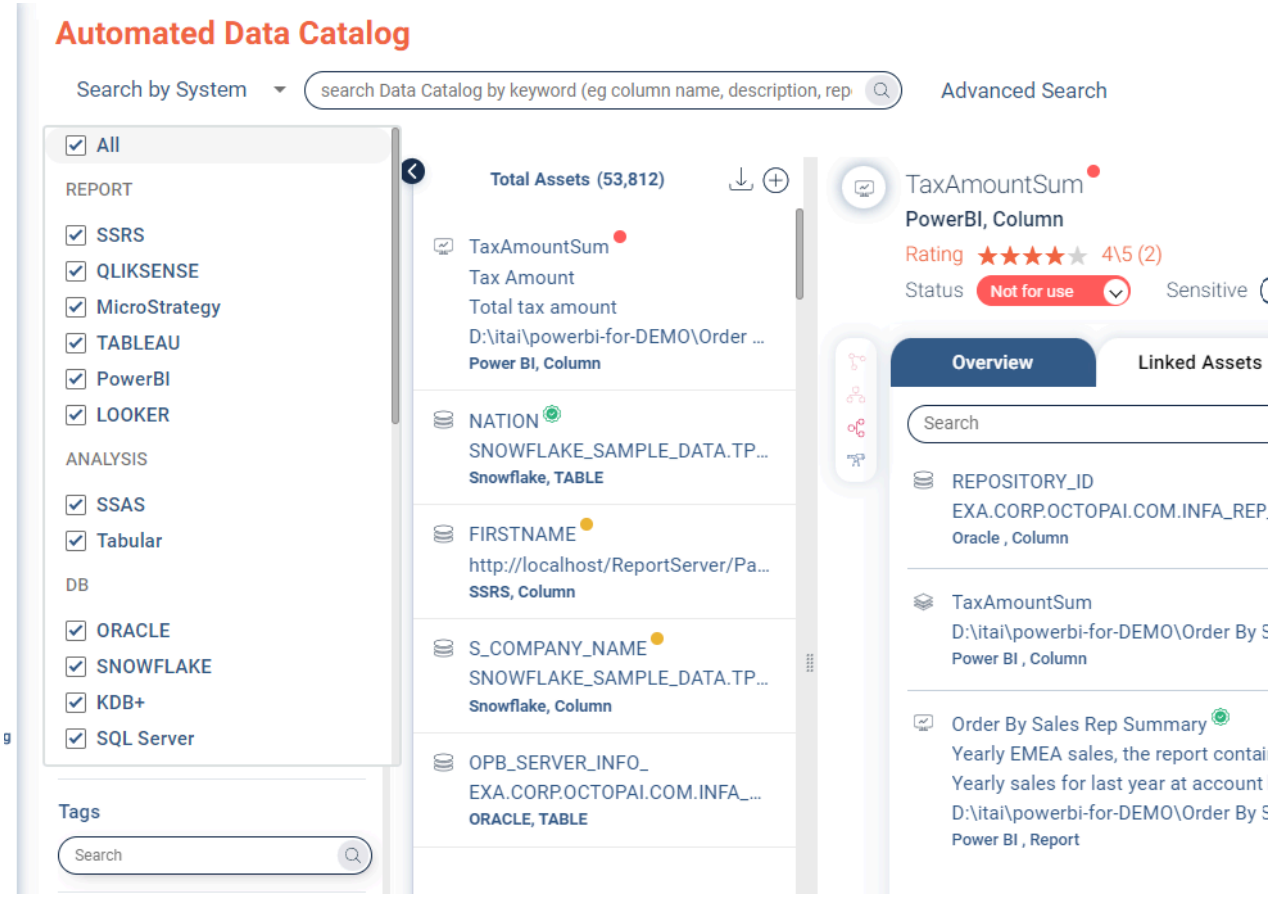
1. Improved Data Discoverability: quickly search for specific data assets using keywords, tags, or other attributes. This saves time and effort by eliminating the need to manually browse through multiple systems or databases.
2. Enhanced Data Governance: establish a consistent and standardized approach to data management. It provides a centralized location for defining and documenting data definitions, metadata, and data lineage. This ensures data governance practices are followed and promotes data quality and integrity.
3. Increased Collaboration and Knowledge Sharing: collaborate within the Knowledge Hub by creating posts and mentioning other users. This promotes knowledge sharing, allows for discussions about data assets, and facilitates better decision-making based on shared insights.
4. Improved Data Confidence: By providing comprehensive information about data assets, such as descriptions, ratings, and usage statistics, a Knowledge Hub enhances users' confidence in the data they are working with. It helps users understand the context and quality of the data, leading to more accurate analysis and decision-making.
5. Time and Cost Savings: quickly locate and access the data they need, reducing the time spent searching for information. It also eliminates the need for manual documentation and maintenance of data assets, saving both time and resources.
6. Regulatory Compliance: compliance with data protection and privacy regulations by providing visibility into sensitive data assets. It allows organizations to track data lineage and ensure proper data access controls are in place.

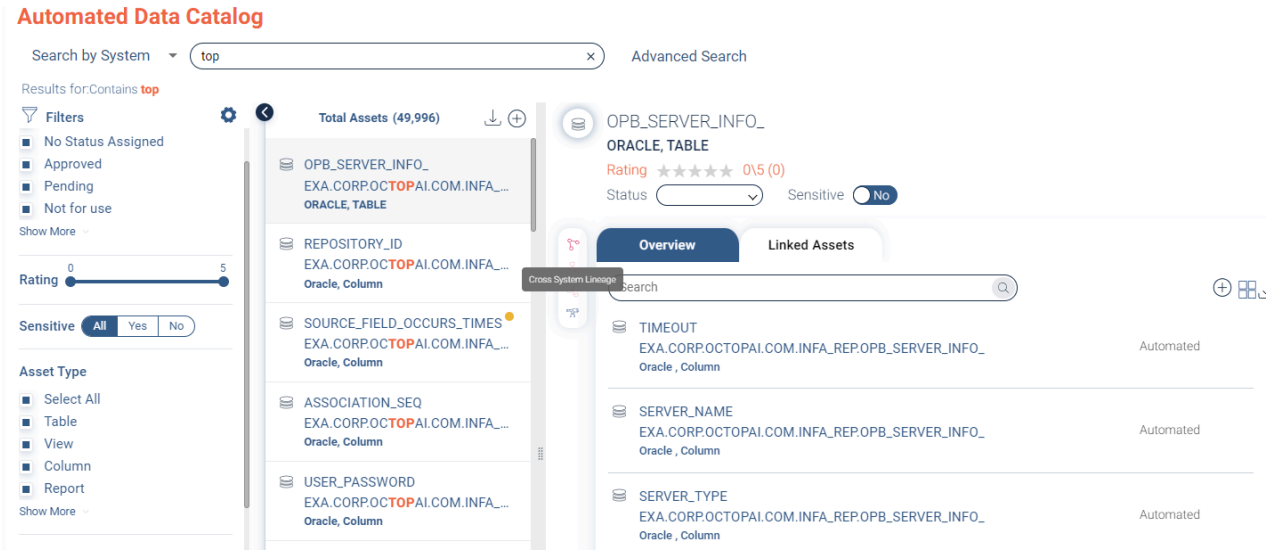

Data Catalog Insights
Search
Filters
Status
Select All
No Status Assigned
Approved
Pending
Not for use
Show More
Rating
Sensitive
All
Yes
No
Asset Type
Select All
Table
View
Column
Report
Show More
Tags
Data Owner
Reset
Apply
Total Assets 53,812
Recent Activity
May 07, 2023 - Jun 06, 2023
New
573
Updated
6
Suspended
0
Dropped
0
Collaboration (2)
Nissim Ohayon 12 days ago
@Nissim Ohayon - we found out that the calculation is revealed here.
TotalDueSum, Column SQL Server
E2E_Dwh_Sales.dbo.vSalesProductsSummary >
Nissim Ohayon a month ago
@Holly Miller - BTW, I see that this column contains SSN, please mark it as PII.
Person Count E2E build, Report SSRS
http://localhost/ReportServer/Pages/ReportViewer.aspx?/AdventureWorks/Per... >
Missing Descriptions
53,769
Avg. Rating 0.0
0
Assigned To Me
0
Trending Tags
Sales
Salesforce
GDPR
PII
Orders
Demo
Proj2219
EMEA
Sheba
Ophir Dem
pricing
Snowflake
lety
Lety Testing
Finance
anytag
MyTag
powerbi
product
test
Type
Column
TABLE
VIEW
Report
Other
Status
Approved
Pending
Reviewed
Not for use
New (573)
State
Regional5/sheets/Obesity...
Tableau , Column
Carrier Name
Regional/sheets/FlightDel...
Tableau , Column
Order ID
Superstore/sheets/Overvi...
Tableau , Column
Date
Regional6/sheets/Stocks/...
Tableau , Column
Category
Superstore/sheets/Overvi...
Tableau , Column
Region
Superstore/sheets/OrderD...
Tableau , Column
Day
D:\itai\powerbi-for-DEMO\...
Power BI , Column
Help


Automated Data Catalog
Search by System
search Data Catalog by keyword (eg column name, description, rep
Advanced Search
Filters
Total Assets (53,812)
TaxAmountSum
Tax Amount
Total tax amount
D:\itai\powerbi-for-DEMO\Order ...
Power BI, Column
NATION
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, TABLE
FIRSTNAME
http://localhost/ReportServer/Pa...
SSRS, Column
S_COMPANY_NAME
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, Column
OPB_SERVER_INFO_
EXA.CORP.OCTOPAI.COM.INFA_...
ORACLE, TABLE
TaxAmountSum
PowerBI, Column
Rating
4\5 (2)
Status
Not for use
Sensitive
No
Overview
Linked Assets
Properties
Description
Tax Amount
Technical Description
Total tax amount
Calculation Description
Origin Description
Origin Calculation
Sample Path
D:\itai\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/SalesmansSales/TaxAmountSum
Data Type
Source System
Audit
Contacts
Data Owner
Lety Agami
Data Steward
elad g
Tags
Sales
Finance
Add Tag
Posts (11)
Type Here
Cancel
Submit
Octopai Support
2 months ago
Lety Agami
Help


Automated Data Catalog
Search by System
search Data Catalog by keyword (eg column name, description, rep
Advanced Search
Filters
Status
Select All
No Status Assigned
Approved
Pending
Not for use
Show More
Rating
Sensitive
All
Yes
No
Asset Type
Select All
Table
View
Column
Report
Show More
Tags
Search
Data Owner
Search
Reset
Apply
Total Assets (53,812)
TaxAmountSum
Tax Amount
Total tax amount
D:\itai\powerbi-for-DEMO\Order ...
Power BI, Column
NATION
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, TABLE
FIRSTNAME
http://localhost/ReportServer/Pa...
SSRS, Column
S_COMPANY_NAME
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, Column
OPB_SERVER_INFO_
EXA.CORP.OCTOPAI.COM.INFA_...
ORACLE, TABLE
Contains
Type Here
AND
OR
+ Add Another Filter
Submit
Description
Tax Amount
Technical Description
Total tax amount
Calculation Description
Origin Description
Origin Calculation
Sample Path
D:\itai\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/SalesmansSales/TaxAmountSum
Data Type
Source System
Audit
Contacts
Data Owner
Lety Agami
Data Steward
elad g
Tags
Sales
Finance
Add Tag
Posts (11)
Type Here
Cancel
Submit
Octopai Support
2 months ago
Lety Agami
Help


Automated Data Catalog
Search by System
search Data Catalog by keyword (eg column name, description, rep
Advanced Search
Filters
Status
Select All
No Status Assigned
Approved
Pending
Not for use
Show More
Rating
Sensitive
All
Yes
No
Asset Type
Select All
Table
View
Column
Report
Show More
Tags
Search
Data Owner
Search
Reset
Apply
Total Assets (53,812)
TaxAmountSum
Tax Amount
Total tax amount
D:\itai\powerbi-for-DEMO\Order ...
Power BI, Column
NATION
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, TABLE
FIRSTNAME
http://localhost/ReportServer/Pa...
SSRS, Column
S_COMPANY_NAME
SNOWFLAKE_SAMPLE_DATA.TP...
Snowflake, Column
OPB_SERVER_INFO_
EXA.CORP.OCTOPAI.COM.INFA_...
ORACLE, TABLE
TaxAmountSum
PowerBI, Column
Rating
4\5 (2)
Status
Not for use
Sensitive
No
Overview
Linked Assets
Search
REPOSITORY_ID
EXA.CORP.OCTOPAI.COM.INFA_REP_ADAPT.V_PC8X_MASTER_TSK_LOG
Oracle , Column
Augmented
TaxAmountSum
D:\itai\powerbi-for-DEMO\Order By Sales Rep Summary.pbix/SalesmansSales/T...
Power BI , Column
Automated
Order By Sales Rep Summary
Yearly EMEA sales, the report contains detailed sales information by sales repre...
Yearly sales for last year at account level
D:\itai\powerbi-for-DEMO\Order By Sales Rep Summary.pbix
Power BI , Report
Automated
Contacts
Data Owner
Lety Agami
Data Steward
elad g
Tags
Sales
Finance
Add Tag
Posts (11)
Type Here
Cancel
Submit
Octopai Support
2 months ago
Lety Agami
Help





Main capabilities of the Knowledge Hub and how to utilize them

This guide will walk you through the key capabilities of the Knowledge Hub and how to utilize them.

1. Bulk Import

- The bulk import capability allows users to import assets into the Cloudera Octopai Knowledge Hub in large quantities.
- Users can provide existing documentation or data in bulk and have it added to the Knowledge Hub.
- Instead of manually entering each asset, users can provide a complete XLS file with the relevant asset details.

- The provided XLS file should follow a specific format or template provided by Cloudera Octopai.
- The bulk import process enables users to quickly populate the Knowledge Hub with a large number of assets.
- This capability provides customers with independence in updating and adding assets using external files.

2. Asset Management

- Any item in the Knowledge Hub is considered an asset.
- Assets can be automatically created from metadata or manually created by users.
- Examples of assets include tables, columns, reports, processes, and more.

3. Editing Permissions

- Editing capabilities are based on user roles:
  - Admin: Full editing privileges.
  - Viewer: Read-only access.

4. Search Functionality

- The Knowledge Hub provides a search feature to find assets using free text.
- Admins, editors, and viewers can use the search functionality.

5. Rating

- Users can rate assets to indicate their quality.
- Average ratings are displayed in the asset's detail pane.
- Admins, editors, and viewers can rate assets.

6. Status

- Assets can be assigned different statuses, such as "Approved," "Pending," or "Not for use."
- Statuses are color-coded for easy identification.
- Admins and editors can manage asset statuses.

7. Sensitivity

- Assets can be assigned sensitivity levels to indicate how they can be used.
- Sensitivity options include "Yes" or "No."
- Admins and editors can manage asset sensitivity.

8. Descriptions

- Assets have short and full descriptions.
- Short descriptions provide a brief overview.
- Full descriptions offer detailed information.
- Admins and editors can manage asset descriptions.

9. Origin Description and Calculation

- Origin descriptions and calculations are analyzed from metadata.
- These attributes are non-editable.

10. Asset Type and Data Type

- Asset type represents the category of the asset (e.g., Knowledge Hub, Asset, Table, Column, Report).
- Data type represents the original data type as analyzed from metadata.
- These attributes are non-editable.

11. Sample Path

- Sample path shows the location of the asset.
- Aggregated assets have sample paths for different layers.
- Direct paths are shown for other asset types.
- This attribute is non-editable.

12. Ownership and Stewardship

- Data owner and data steward attributes help assign responsibility for assets.
- Admins and editors can select owners and stewards from a drop-down list.

13. Updates and Entry Date

- The last update attribute shows the date of the last modification to the asset.
- The entry date represents the initial appearance of the asset.
- These attributes are non-editable.

14. Tags

- Tags can be assigned to assets for better categorization and organization.
- Admins and editors can select existing tags or create new ones.

15. Linked Assets

- Linked assets can be automated or augmented.
- Automated links are created through analysis.
- Augmented links can be added or removed by users with appropriate permissions.

16. Suspension

- Assets can be suspended to prevent modification of their details.
- This feature is available to admins and editors.

17. Search Filters and Sorting

- Search filters allow users to locate specific data assets.
- Filters can be applied based on various attributes.
- Sorting options help organize search results.

18. Collaboration with Posts

- Users can collaborate by creating posts and mentioning other users.
- Mentioned users receive email notifications with relevant asset information.
- Posts are publicly available.

19. Lineage Integration

- Presentation and physical columns integrate with end-to-end column lineage.
- Reports, views, procedures, processes, and functions integrate with inner system and cross-system lineage.
- Tables integrate with cross-system lineage.

20. Search in Discovery

- Users can search for assets by name in the Discovery module.

21. Dashboard

- The Knowledge Hub provides a dashboard for insights and quick actions.
- Users can monitor the health of the Knowledge Hub and perform tasks efficiently.

22. Export Assets

- Users can export, update, and create assets using external files.
- Export functionality allows for future updates.
- Update and create capabilities are available in the Admin Console.
- Error handling and user-friendly input are supported.

23. Knowledge Hub with Technical Descriptions for Linked Assets

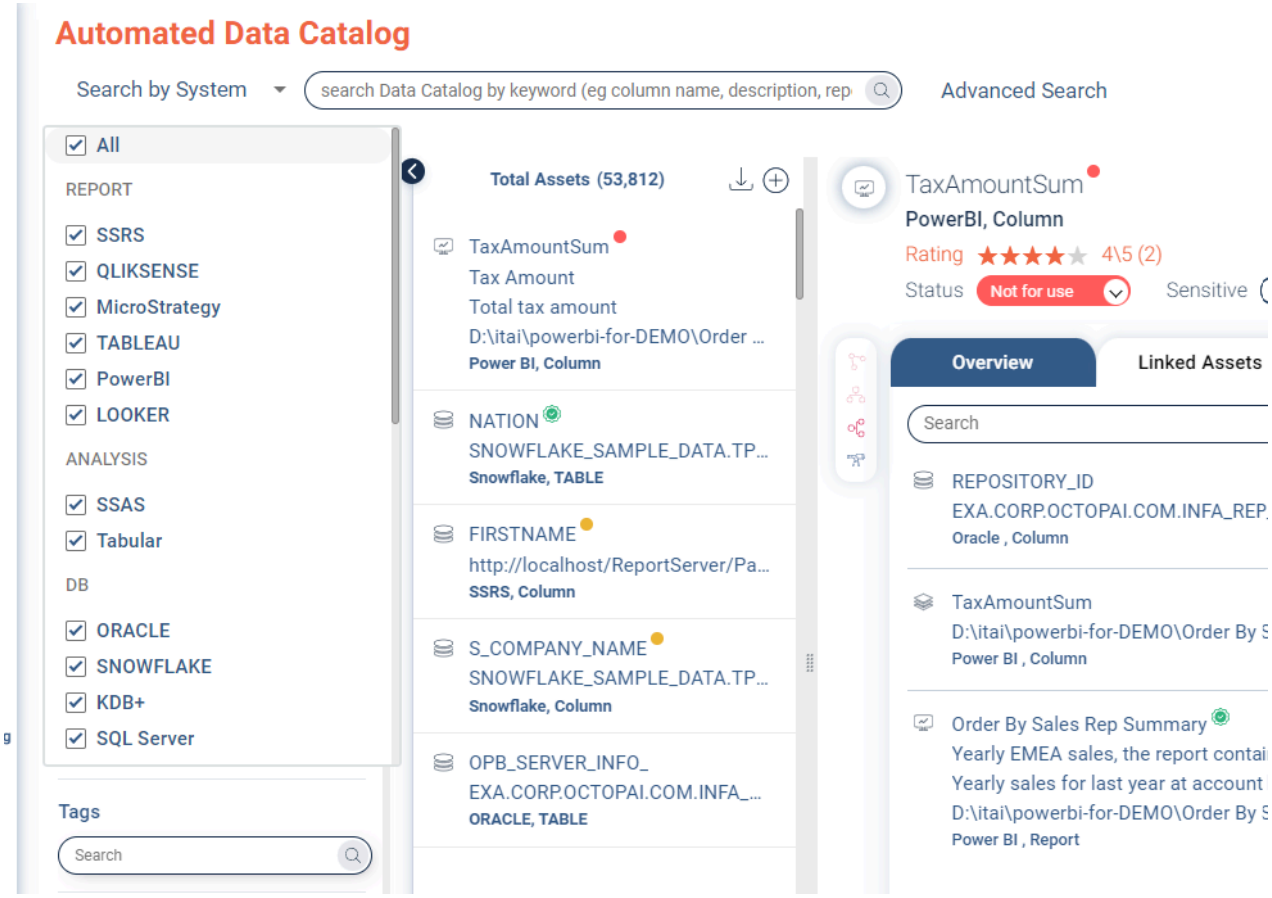
- Technical descriptions are now included in the Knowledge Hub for linked assets.
- The technical description can be viewed alongside the regular description.
- Excel download files include technical descriptions.

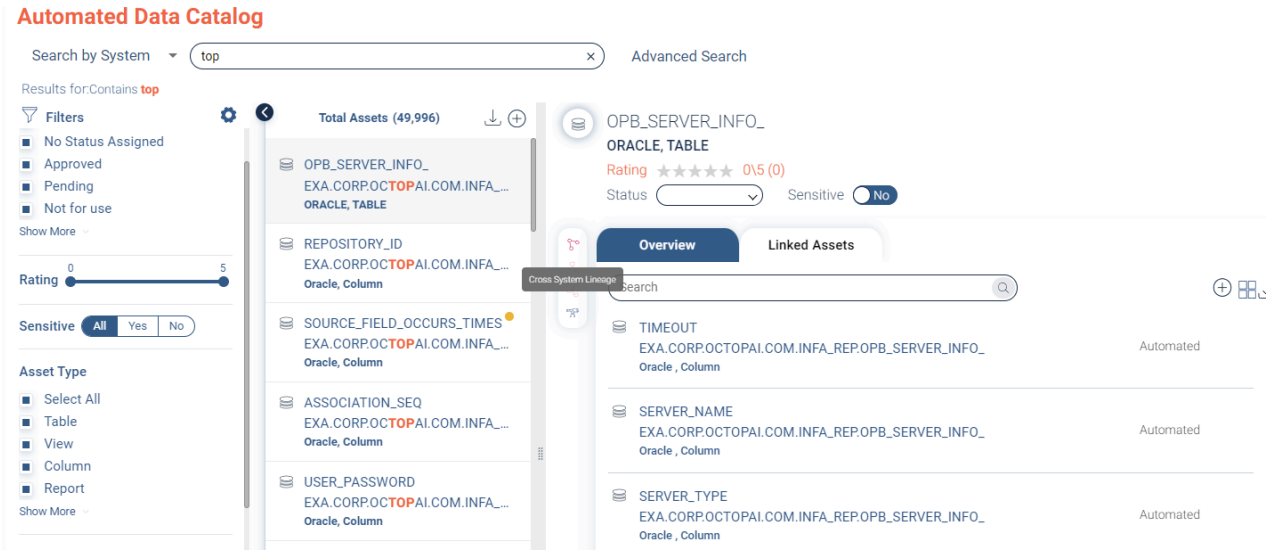
24. Grid Mode

- Grid mode include path and description columns.
- All linked assets columns are available.
- Filtering columns facilitate data exploration.

25. Mass Update

- The Data Catalog provides a mass update capability to efficiently update attributes of multiple assets at once.
- Users can utilize an Excel template to make bulk updates to attributes such as descriptions and tags.
- The template includes all relevant attributes, including custom attributes and technical descriptions.
- Only filled cells in the template will be used for updates, preventing unintended modifications.
- Tags can be entered as a comma-separated list or in separate columns.
- Custom attributes can be included as columns in the CSV file.
- Usernames are captured correctly for assets uploaded through the Excel template.
- Users receive notifications indicating the number of assets successfully updated and any errors encountered.
- An error output file is provided for assets that could not be updated, along with the reasons for failure.





Now that you have an overview of the capabilities of the Data Catalog, you can effectively manage and work with your data assets in Cloudera Octopai. Utilize the various features to search, organize, collaborate, and gain insights from your data catalog.

Cloudera Octopai FAQ

The Cloudera Octopai FAQ addresses common questions and challenges, including system refreshes, password resets, and managing unknown data objects. It provides clear solutions to enhance user experience and streamline troubleshooting for Cloudera Octopai.

Data Owner or Data Steward - What is the difference?

Cloudera Octopai Knowledge Hub provides a centralized repository for storing and managing data metadata, which helps to ensure that both data owners and data stewards have the information they need to do their jobs effectively.

In Cloudera Octopai Knowledge Hub, both data owners and data stewards are important roles in the data governance process.

Additionally, Cloudera Octopai Knowledge Hub provides a number of features that help to improve data quality, such as data lineage and data profiling.

These features can help data owners and data stewards to identify and resolve data quality issues, which ultimately helps to improve the overall quality of the data.

Note: Both Data Owner and Data Steward must be existing users in Cloudera Octopai.

The following table summarizes the key differences between data owners and data stewards:

Data Owner	Data Steward
Accountable for the data's quality, accuracy, and compliance.	Responsible for the day-to-day management of the data.
Has a deep understanding of the business context of the data.	Works with data owners to ensure that the data is properly classified, stored, and protected.
Resolves data quality issues.	Promotes the use of the data across the organization.
Provides the strategic direction for data management.	Ensures that the data is managed in a way that meets the needs of the organization.

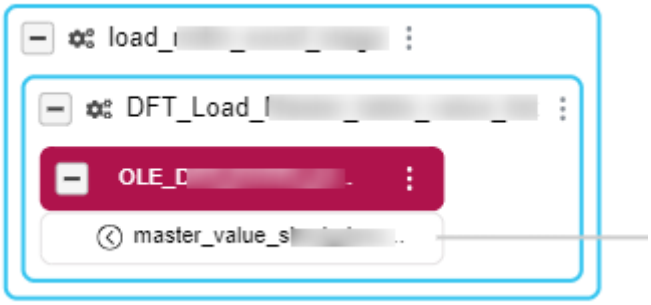
Non-Expansion Issues in E2E Column Lineage

The platform assesses the availability of more physical layer elements for expansion. If these elements are found, the expansion process smoothly carries on, ensuring a comprehensive E2E column lineage.

Note: The E2E Column Lineage feature exclusively displays connections between physical data objects that have been extracted and uploaded to Cloudera Octopai.

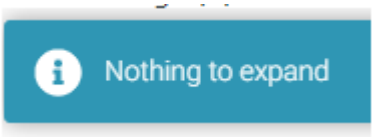
In the context of end-to-end (E2E) column lineage, Cloudera Octopai is programmed to expand the scope of analysis by incorporating additional physical layer elements along with the transformation or semantic layer elements.

Your action initiates the expansion discovery process. Upon an Expand action trigger,  , the platform assesses the availability of more physical layer elements for expansion. If these elements are found, the expansion process smoothly carries on, ensuring a comprehensive E2E column lineage.



This indicator implies that there could be additional sources or targets associated with this column. By clicking on the icon, Cloudera Octopai will attempt to expand the lineage online and conduct a more comprehensive analysis.

In the event the platform cannot identify additional physical layer elements, it will alert you with a warning message and will not expand further. This warning is not a cause for concern. It simply means the expansion process is paused due to the current unavailability of additional physical layer elements. As a result, your E2E column lineage remains accurate and complete with the available data.



If the expansion icon is not displayed on the left side of the column, it indicates that no other sources or targets are linked to this particular column.

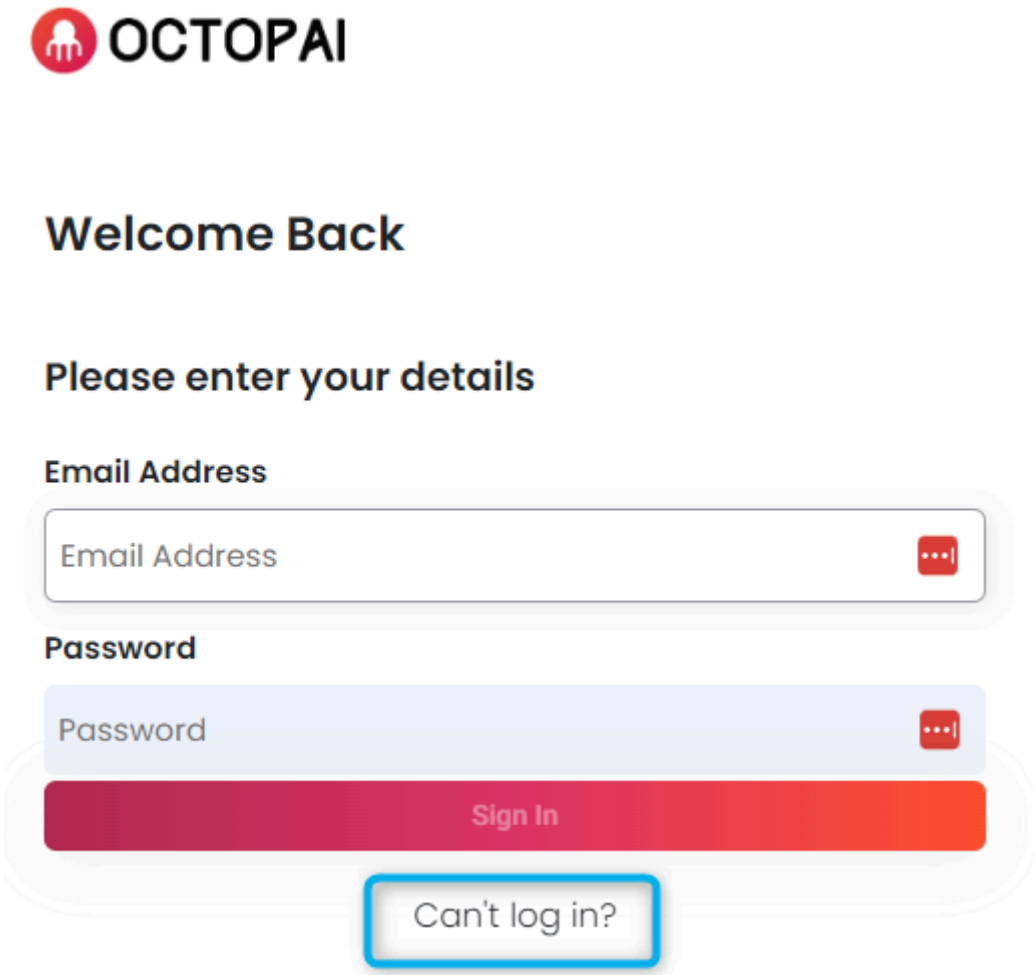
How to reset your password

Learn about resetting your password.

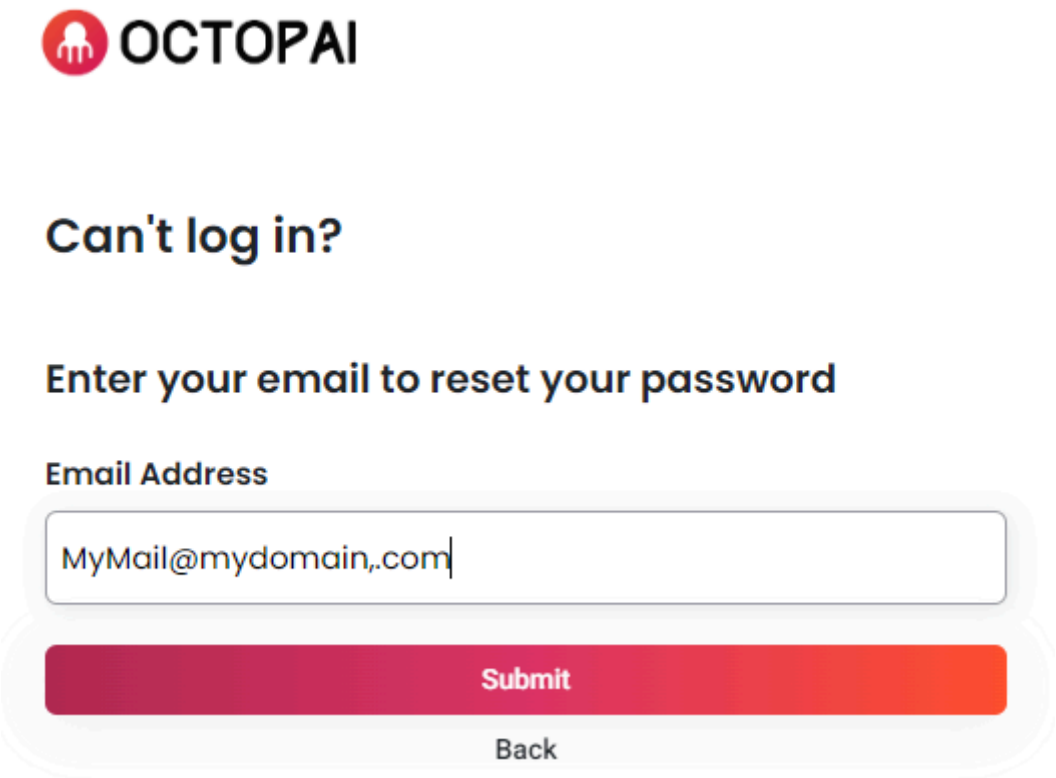
Procedure

1. Access your platform URL.

2. Click Can't Login on the login page.

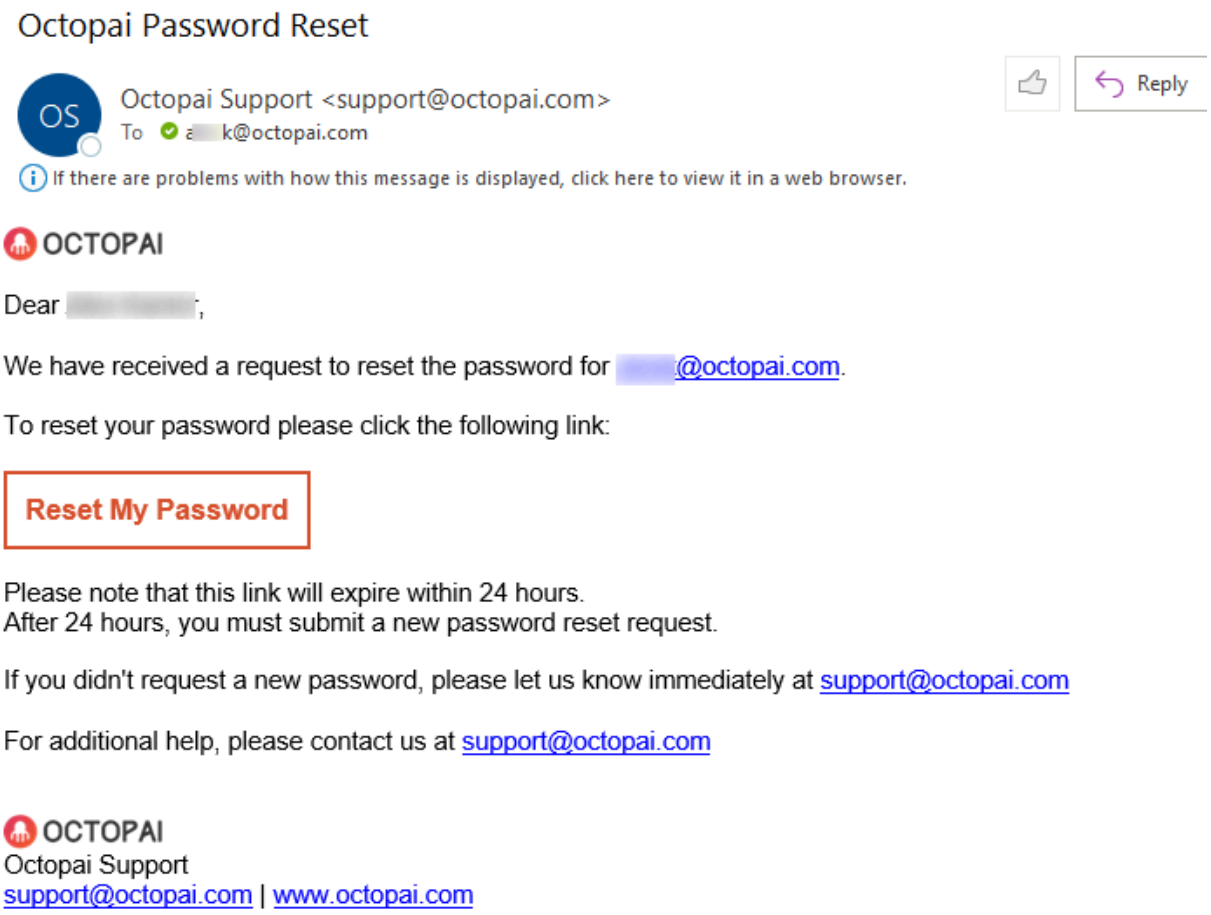


3. Enter you password and click Submit to submit your password reset request.



4. Check your mailbox for the password reset email.

Figure 6: Example password reset email



Octopai Password Reset

OS Octopai Support <support@octopai.com>
To a[redacted]k@octopai.com

Reply

If there are problems with how this message is displayed, click here to view it in a web browser.

OCTOPAI

Dear [redacted],

We have received a request to reset the password for [redacted]@octopai.com.

To reset your password please click the following link:

Reset My Password

Please note that this link will expire within 24 hours.
After 24 hours, you must submit a new password reset request.

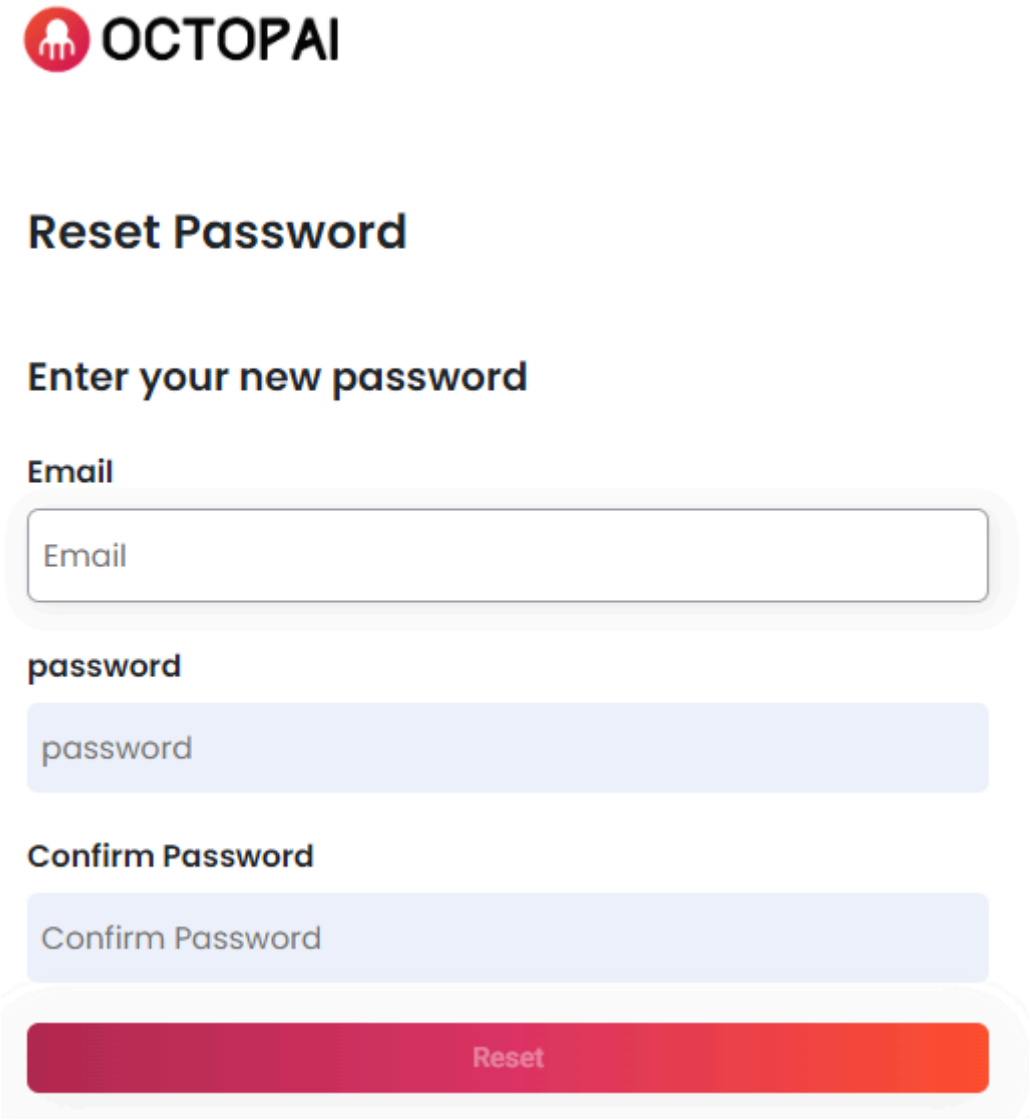
If you didn't request a new password, please let us know immediately at support@octopai.com

For additional help, please contact us at support@octopai.com

OCTOPAI
Octopai Support
support@octopai.com | www.octopai.com

5. Click Reset My Password in the email.

6. Complete the form with your email and new password, and click Reset.



Results

You are redirected to the login page where you can use your new password.

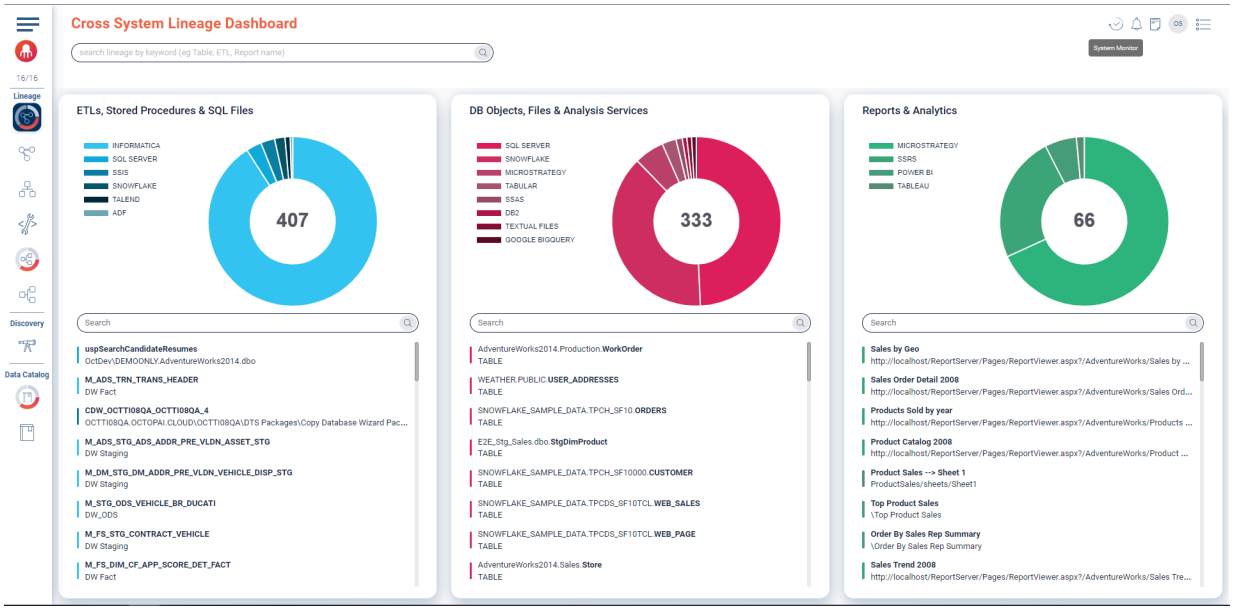
How to check when the system was last refreshed

Learn about how to check when the Cloudera Octopai Data Lineage system was last refreshed.

Procedure

1. Go to the Cloudera Octopai platform.

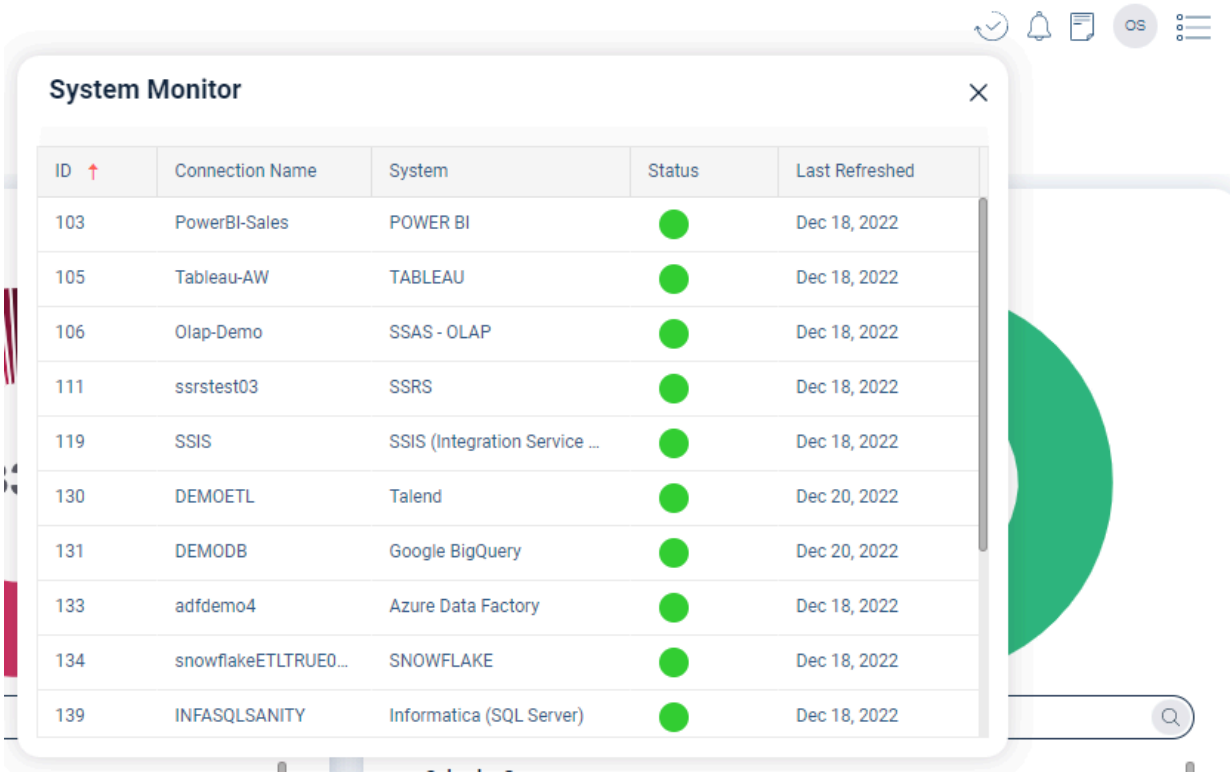
2. Click the System Monitor icon in the upper-right corner.



The System Monitor window opens and shows the following information:

- **ID** that is the connection ID number
- **Connection Name**
- **System** that is the system name
- **Status**
- **Last Refreshed** that is the date when the analysis was published to the site

Figure 7: System Monitor



How to manage Unknown Data Objects (UNK)

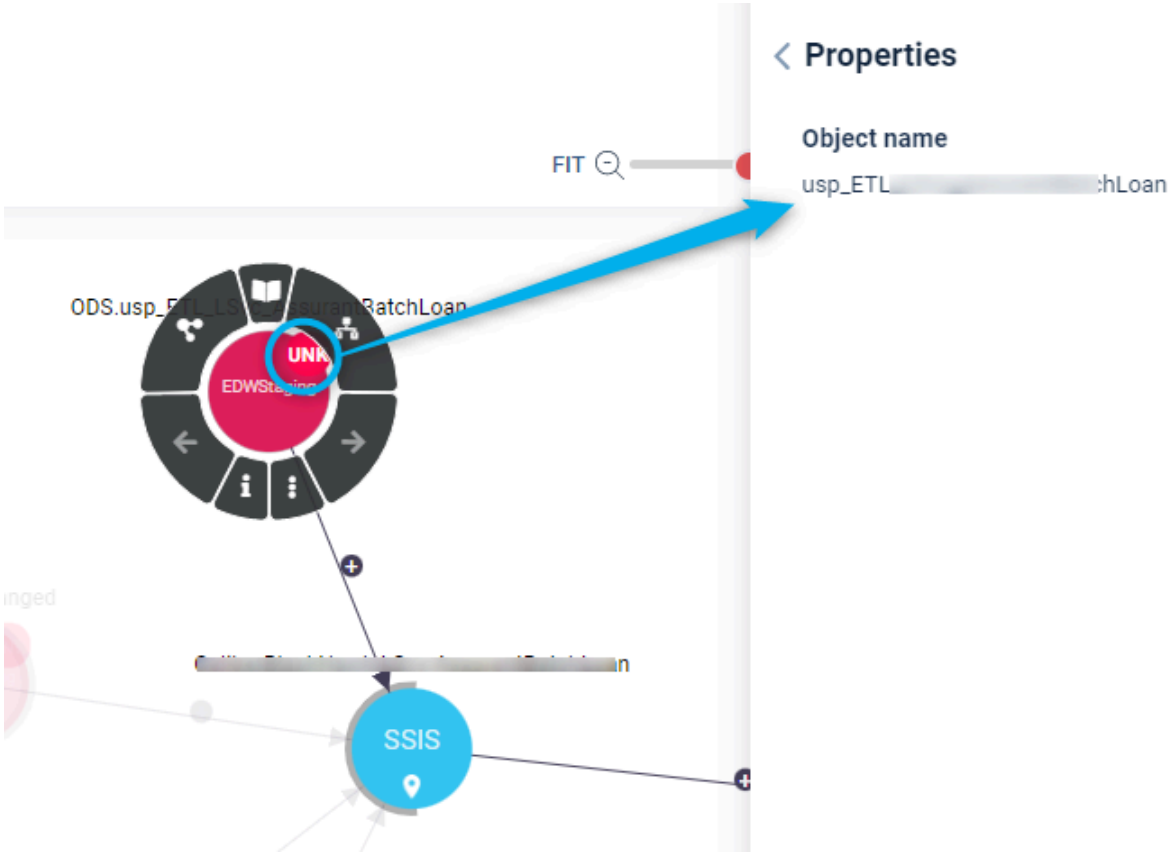
Encountering Unknown Data Types (UNK) for certain Data Objects in Cloudera Octopai is a common issue that can occur when an object has not been extracted and uploaded to Cloudera Octopai, but appears in the lineage due to its association with a script.

About this task

When viewing the lineage in Cloudera Octopai, you might come across Data Objects that are displayed as Unknown Data Type (UNK). This happens when Cloudera Octopai lacks sufficient information to determine the type of the Data Object, even though it is directly related to a Report or ETL. Cloudera Octopai fails to recognize the associated Database Name and Schema through their keys, resulting in the UNK classification.

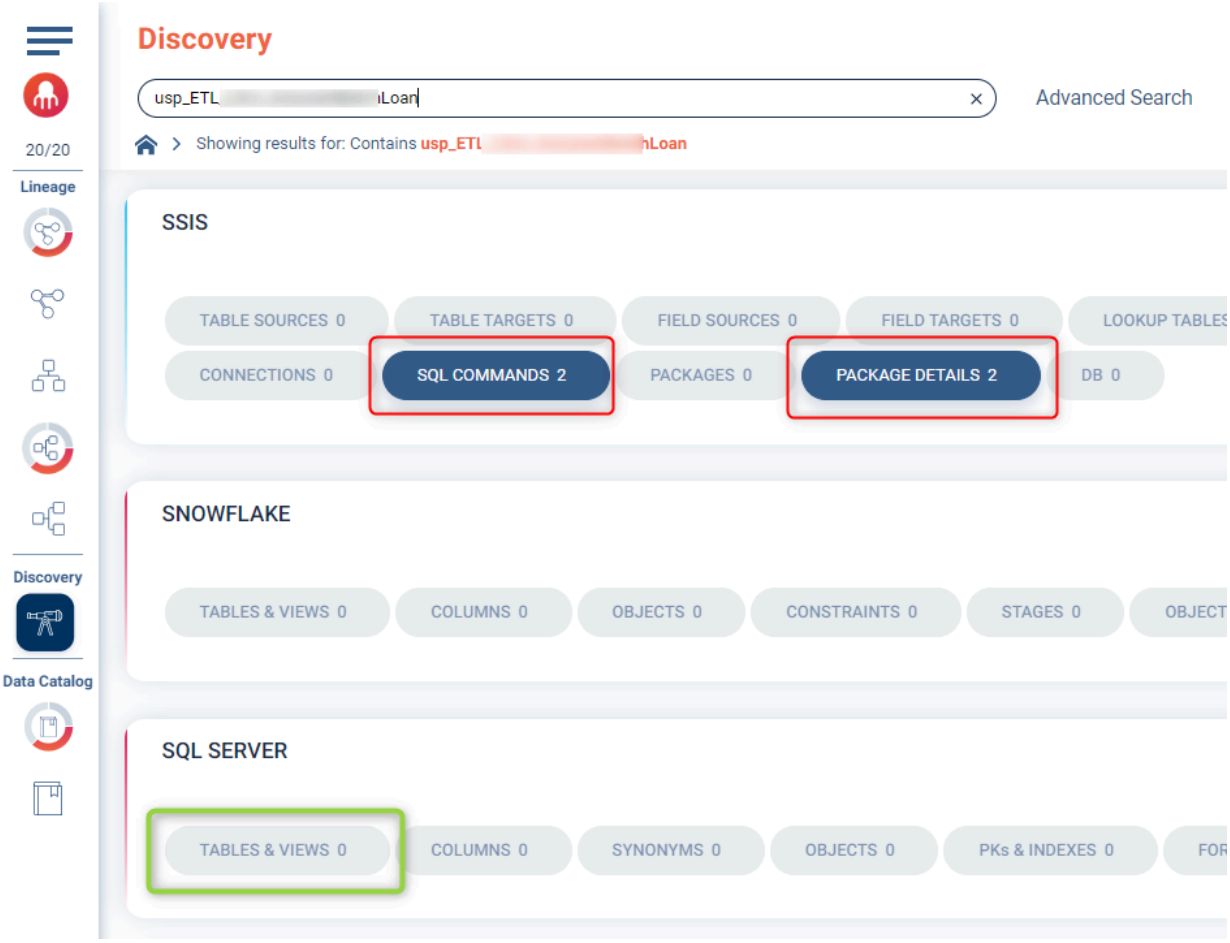
Procedure

1. Click the Information i button in the object radial button, and copy the object name from the **Properties** section.



2. Paste the object name in the Cloudera Octopai **Discovery** space.

Figure 8: Discovery space



You can identify that the Unknown Data Object is part of SSIS scripts but was not physically uploaded to Cloudera Octopai during the Metadata extraction process.

To resolve this issue, perform the following steps:

a. Verify the permissions on the Database associated with the object, based on the tool prerequisites.

b. Take note of the path of the Report or ETL shown in the Data Object **Properties** section and paste it to the Admin Console Connection Parameters tab. The example is valid for ETL.

Figure 9: Path in the Properties section

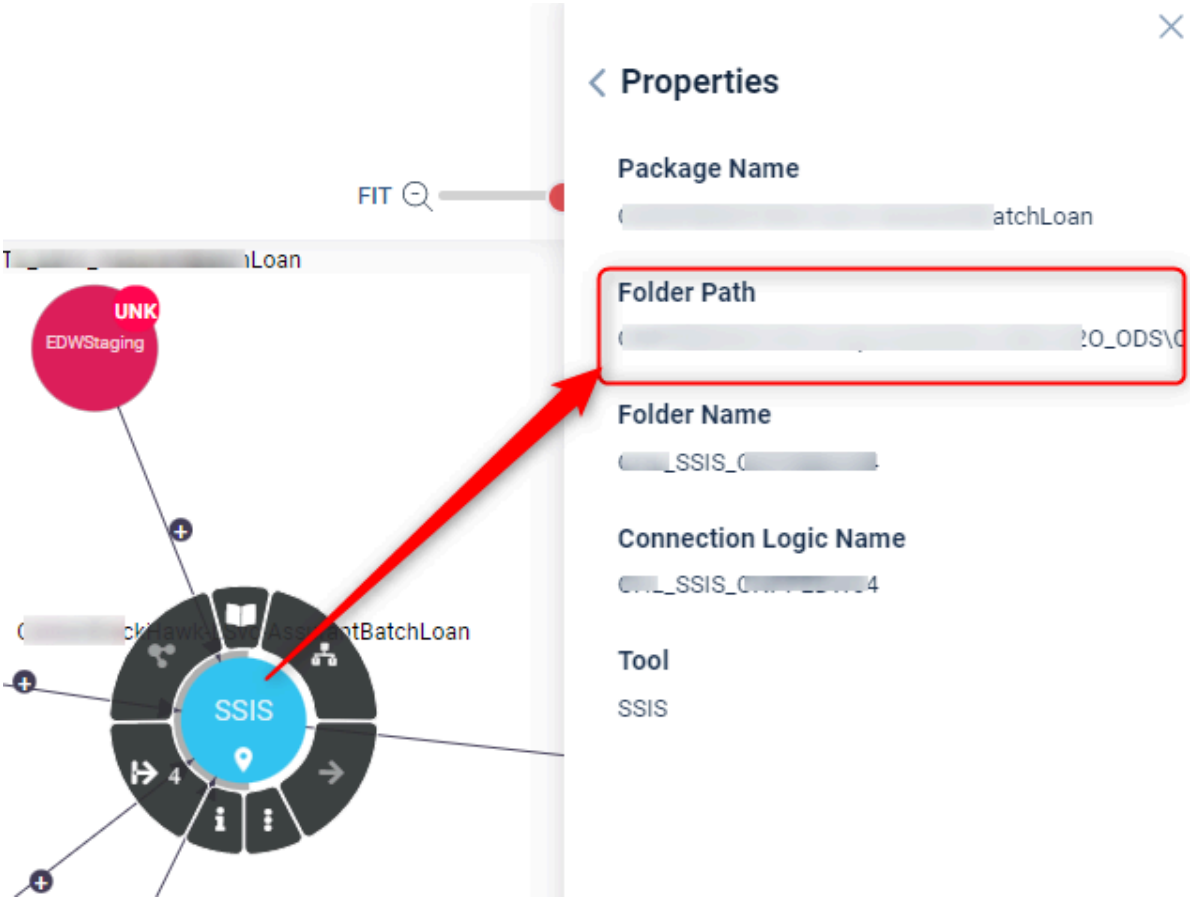


Figure 10: Switch to Admin Console

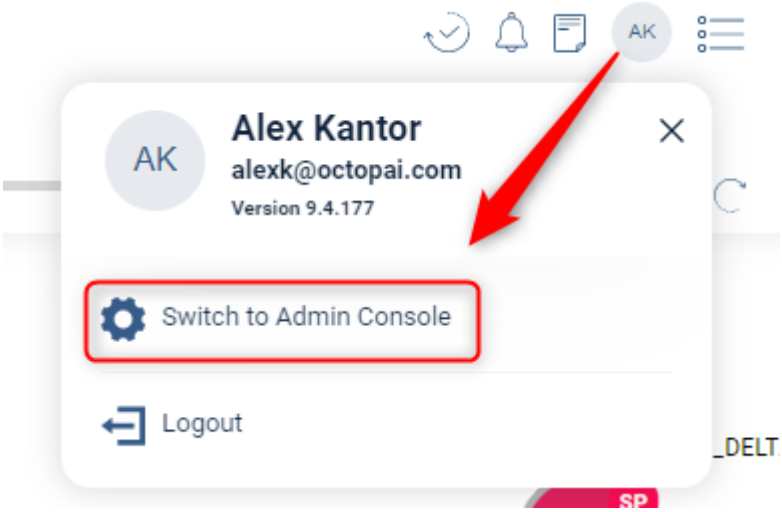


Figure 11: Admin Console

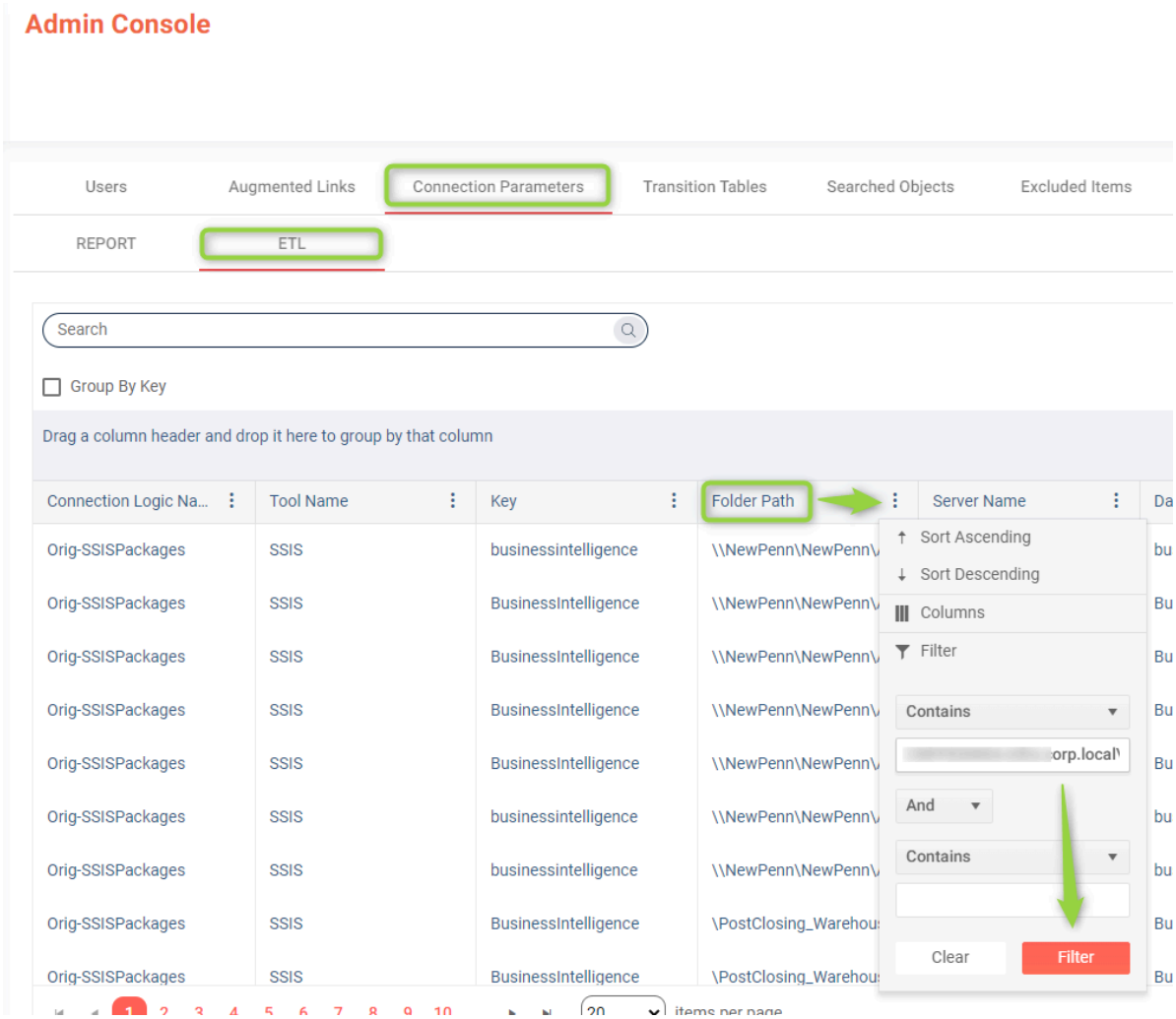
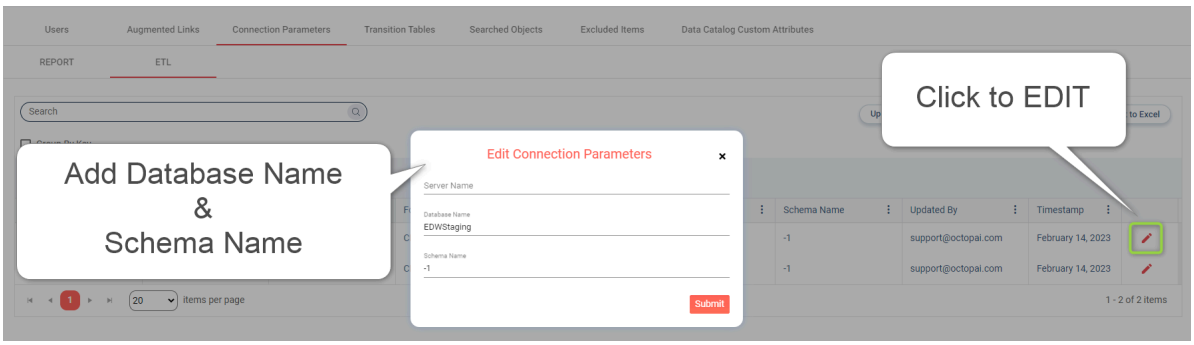
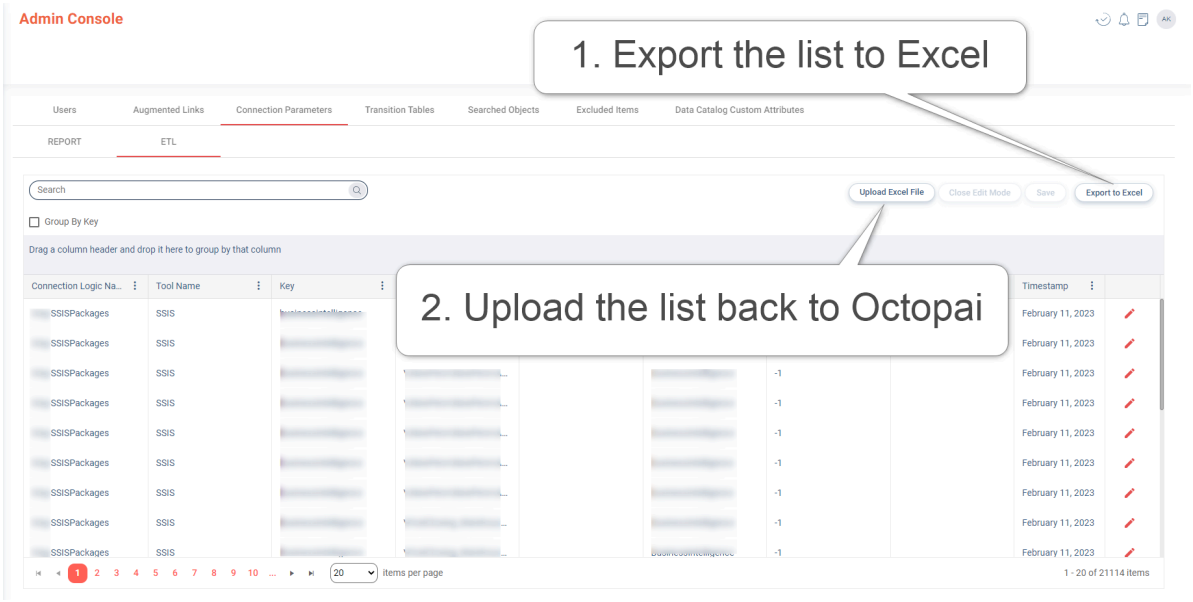


Figure 12: Add the connection parameters



c. An Cloudera Octopai night job will handle the rest.

3. Edit bulk connection parameters.
   a) Export the list to Excel.



   b) Complete the Reports - Database and Schemaand ETL - Database missing data in the Excel file.
   c) Upload the modified file to Cloudera Octopai.
   d) An Cloudera Octopai night job will handle the rest.

What to do next

If you still encounter UNK objects, contact Cloudera Support.

How to clear your site data (clear cache)

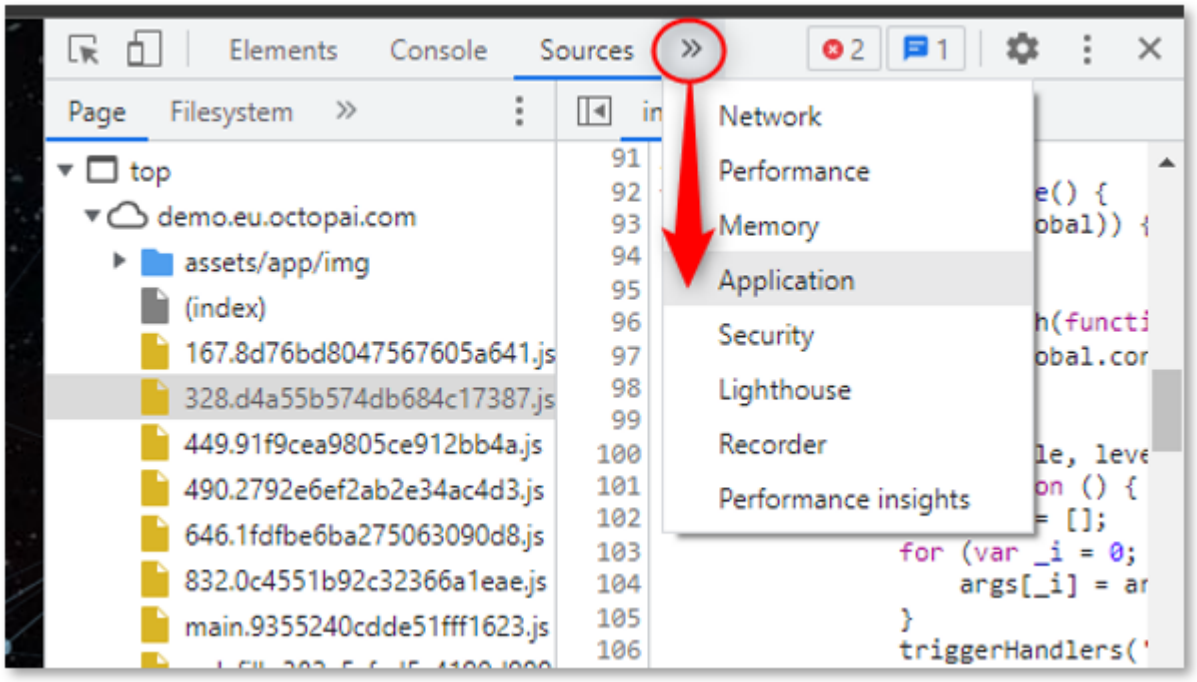
Learn about clearing your site cache to solve some basic issues, such as the performance of Cloudera Octopai or applying a hotfix.

Procedure

1. Press F12 on your browser.
   The Dev Tools opens.

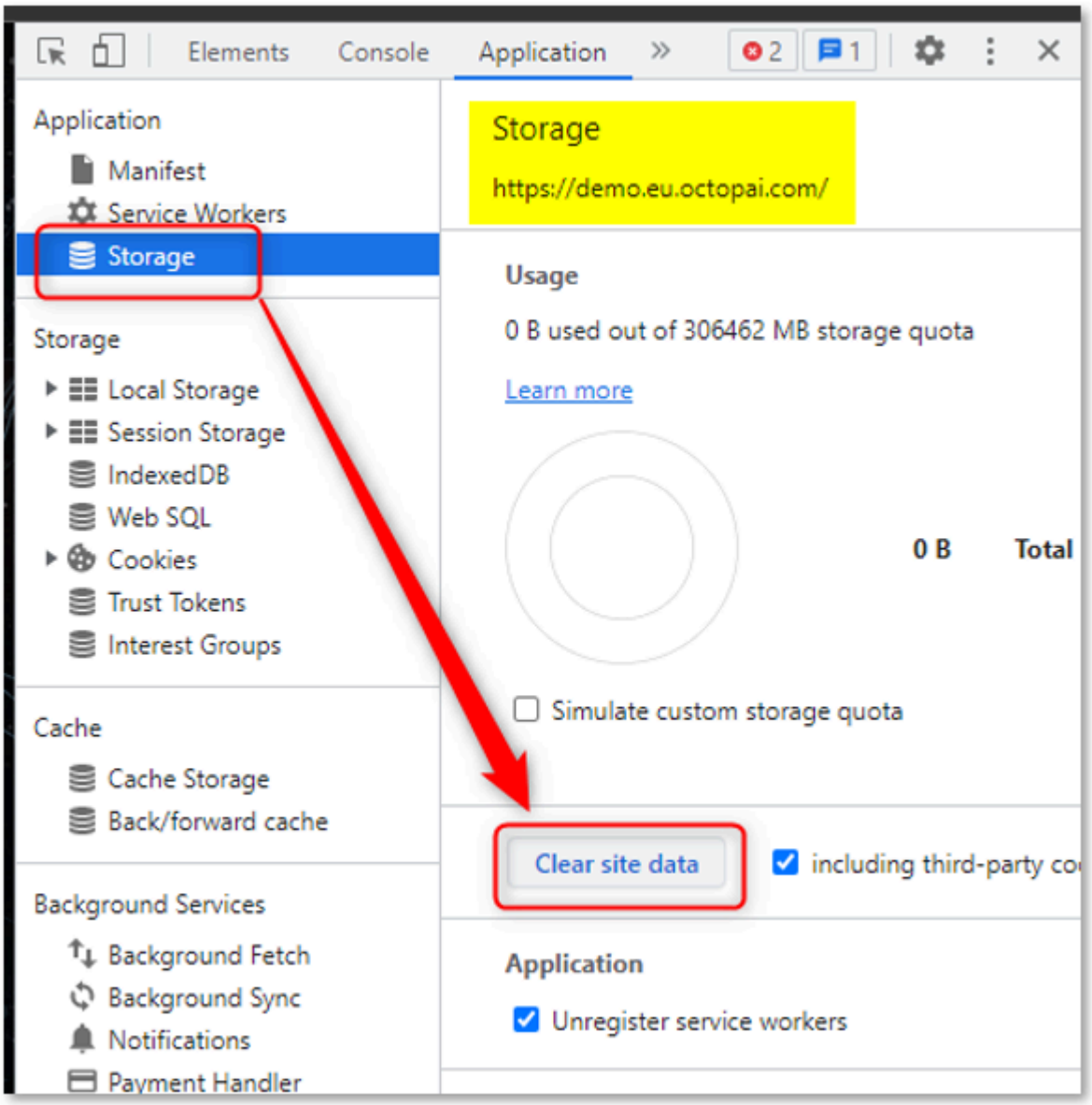
2. Go to **Application**.

Figure 13: Application option



3. Go to **Storage** and click Clear site data.

Figure 14: Clear cache



Results

This operation clears the stored data for your Cloudera Octopai Platform Site only.

Your browser data, including cookies from other sites, will not be affected.

Cloudera Octopai API

The Cloudera Octopai API enables developers to integrate Cloudera Octopai seamlessly into their workflows, enabling efficient and secure data operations.

Authentication

The Cloudera Octopai API ensures secure communication through OAuth 2.0 and HTTPS protocols.

The Cloudera Octopai API supports the following security protocols:

- OAuth 2.0 for token-based authentication
- HTTPS enforced for all endpoints to ensure secure communication

Use the following API request to get access_token:

- **Endpoint**

```
POST https://<your URL name>.octopai.com/api/UserAccount/Login
```

- **Request body**

```
{"Username": "<YOUR USER EMAIL>","Password": "<YOUR PASSWORD>"}
```

The request body has the following parameters:

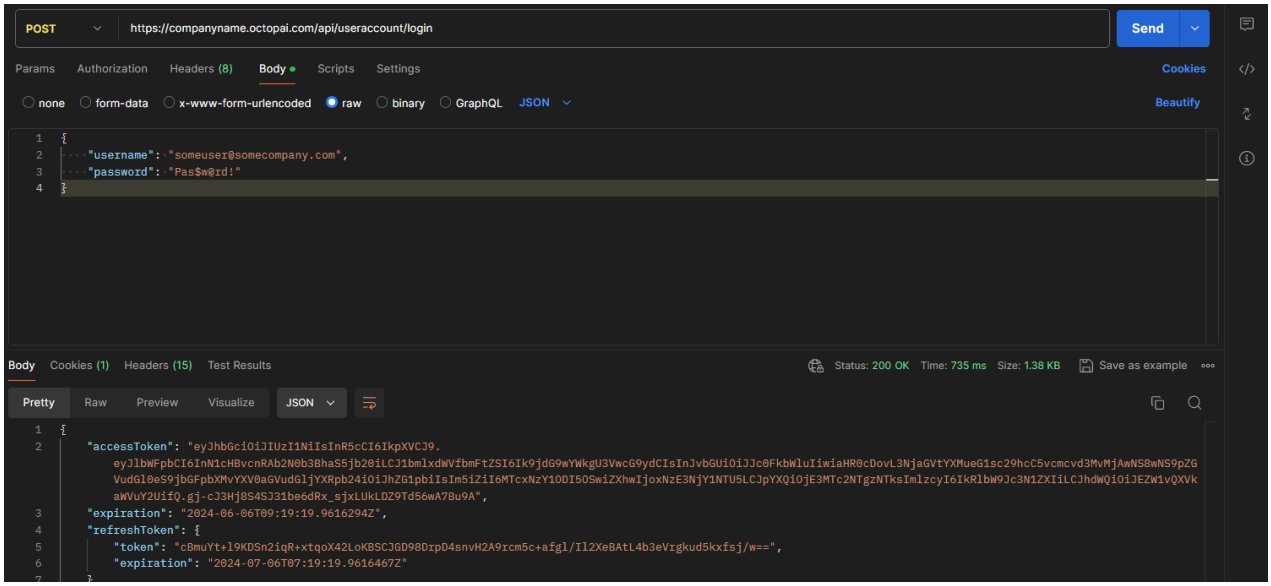
Parameter	Type	Description	Example
userName	string	Octopai user name	someuser@somecompany.com
password	string	Octopai user password	Pas$w@rd

- **Sample response**

```
{    "accessToken": "eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJlbWFpbCI6In
N1cHBvcnRAb2N0b3BhaS5jb20iLCJ1bmlxdWVfbmFtZSI6Ik9jdG9wYWkgU3VwcG9ydCIsIn
JvbGUiOiJJc0FkbWluIiwiaHR0cDovL3NjaGVtYXMueG1sc29hcC5vcmcvd3MvMjAwNS8wNS
9pZGVudGl0eS9jbGFpbXMvYXV0aGVudGljYXRpb24iOiJhZG1pbiIsIm5iZiI6MTcxNzY1OD
I5OSwiZXhwIjoxNzE3NjY1NTU5LCJpYXQiOjE3MTc2NTgzNTksImlzcyI6IkRlbW9Jc3N1ZX
IiLCJhdWQiOiJEZW1vQXVkaWVuY2UifQ.gj-cJ3Hj8S4SJ31be6dRx_sjxLUkLDZ9Td56wA7
Bu9A",    "expiration": "2024-06-06T09:19:19.9616294Z",    "refreshToken":
 {        "token": "cBmuYt+l9KDSn2iqR+xtqoX42LoKBSCJGD98DrpD4snvH2A9rcm5c
+afgl/Il2XeBAtL4b3eVrgkud5kxfsj/w==",        "expiration": "2024-07-06T07
:19:19.9616467Z"    },    "userName": "someuser@somecompany.com",    "di
splayName": "Some User",    "userEmail": null,    "isAdmin": true,    "e
rror": null,    "groupMember": [],    "authLevel": "ADMIN",    "displayM
odules": "lineage;e2e;discovery;compare;adc;adc dashboard;livelineage"}
```

Use access_token for all following API requests.

Postman example



Cloudera Octopai Extraction APIs

Learn about the API endpoints that are designed to facilitate the extraction of lineage data from the Cloudera Octopai system.

For optimal results, start with a call to /assets/query to obtain a comprehensive list of assets. Then call /lineage to retrieve detailed lineage information for each asset identified in the initial query. This two-step approach delivers a structured and efficient data retrieval flow and provides clear insight into asset lineage within the Cloudera Octopai ecosystem.

Assets API

This API retrieves data about assets such as columns and tables so you can review their properties.

Use this API to locate the asset _key value that is required when you call /lineage.

Key function

Search assets.

Endpoint

GET /api/v2.0/assets/query

Request body

```
{
  "ConnectionIds": [],
  "AssetNames": [],
  "ToolNames": [],
  "ToolTypes": [],
  "DatabaseName": "",
  "SchemaName": "",
  "LayerName": "",
  "TableName": "",
  "assetType": 0,
  "batchSize": 0,
  "nextID": ""
}
```

Request body parameters

Parameter	Required	Type	Description
ConnectionIds	No	list	Search by connection identifiers. Example: ["101","102","103"]
AssetNames	No	list	Match assets that contain one of the supplied names.
ToolNames	No	list	Filter by tool names.
ToolTypes	No	list	Filter by tool type. Set assetType to 1 for cross-system lineage (DATABASE / ETL / REPORT). Set assetType to 2 for column-level lineage (DB / ETL / REPORT).
DatabaseName	No	string	Filter by database name.
SchemaName	No	string	Filter by schema name.
LayerName	No	string	Filter by layer name.
TableName	No	string	Filter by table name.
assetType	Yes	number	1 – Cross-system lineage. 2 – Column-level lineage.
batchSize	Yes	number	Maximum results returned per page up to 10,000.
nextID	Yes (from the second call)	string	Use the supplied value to retrieve the next result page.

Request example

```
{
  "assetNames": ["storeid"],
  "schemaName": "SALES",
  "tableName": "Customer",
  "databaseName": "AdventureWorks2014",
  "toolTypes": ["DB"],
  "assetType": 2,
  "limit": 1000
}
```

Response structure

The response body contains the main columns array along with the following fields to control pagination:

- hasMore (boolean) – Indicates whether more result pages are available.
- id (string) – Cursor identifier.
- columns (array) – List of asset objects with metadata such as _key, connection details, object type, and timestamps.

Pagination

To retrieve subsequent batches of results, call GET /api/v2.0/assets/query/scroll/{nextID} with the nextID value from the previous response. Continue this process until the hasMore field in a response is false.

Response example

```
{
  "items": [
    {
```

```
      "_key": "8D1A993B6739DAB65392B54D45F7052E",
      "connectionId": "251",
      "connLogicName": "sqlserverdbwa03testETLTRUE",
      "toolName": "SQLS",
      "toolType": "DB",
      "displayConnectionId": "251",
      "containerObjectName": "sp_built_MrrPerson_MrrCustomer",
      "containerObjectPath": "AdventureWorks2014.dbo",
      "containerObjectType": "SQL_STORED_PROCEDURE",
      "controlFlowPath": "sp_built_MrrPerson_MrrCustomer",
      "controlFlowName": "sp_built_MrrPerson_MrrCustomer",
      "objectID": "ADVENTUREWORKS2014SALESCUSTOMERTABLESTOREID",
      "objectGUID": "8D1A993B6739DAB65392B54D45F7052E",
      "objectType": "TABLE",
      "assetName": "StoreID",
      "dataType": "",
      "precision": "",
      "scale": "",
      "layerName": "Customer",
      "schemaName": "SALES",
      "databaseName": "AdventureWorks2014",
      "tableName": "Customer",
      "isObjectData": true,
      "isMap": false,
      "isVisible": true,
      "updatedDate": "2024-05-26T10:21:51.224Z",
      "createDate": "0001-01-01T00:00:00Z",
      "serverName": "",
      "isSrcColumnOrphan": null
    }
  ],
  "hasMore": false,
  "cursorId": null
}
```


Lineage API

Use the Lineage API to retrieve lineage data that describes the relationships and dependencies between assets.

Key function

Return lineage details for a specific asset.

Endpoint

GET /api/v2.0/lineage

Request body

```
{
  "assetKey": "",
  "depth": 0,
  "direction": 3,
  "limit": 0,
  "assetType": 0
}
```

Request body parameters

Parameter	Required	Type	Description
assetKey	Yes	string	Asset key that defines the lineage starting point
depth	Yes	number	Total number of hops to traverse from the start asset

Parameter	Required	Type	Description
direction	No	number	The default value is 3. 1 – Input relations only 2 – Output relations only 3 – Both directions
limit	Yes	number	Total number of assets returned in the lineage graph
assetType	Yes	number	1 – Cross-system lineage 2 – Column-level lineage

Request example

```
{
  "assetKey": "8D1A993B6739DAB65392B54D45F7052E",
  "depth": 6,
  "limit": 1000,
  "assetType": 2,
  "direction": null
}
```

Response structure

- nodes (array): Asset objects that participate in the lineage graph.
- edges (array): Links that connect the nodes.
- startNode (object): The originating asset for the requested lineage.
- depth and direction: Metadata describing the traversal depth and direction.

Response example

```
{
  "nodes": [
    {
      "_key": "8D1A993B6739DAB65392B54D45F7052E",
      "toolName": "SQLS",
      "toolType": "DB",
      "objectType": "TABLE",
      "assetName": "StoreID",
      "databaseName": "AdventureWorks2014",
      "tableName": "Customer"
    }
  ],
  "edges": [
    {
      "_from": "05E69AF36C41C5FA59208CD75C937AA7",
      "_to": "F6DC4A4D0BA1C88011F843A31BB6BFD4",
      "isCompressed": true
    }
  ],
  "startNode": {
    "_key": "8D1A993B6739DAB65392B54D45F7052E",
    "toolName": "SQLS",
    "assetName": "StoreID",
    "databaseName": "AdventureWorks2014",
    "tableName": "Customer"
  },
  "depth": 12,
  "direction": "Both"
}
```

Cloudera Octopai API: UserEvents documentation for user audit trails

Learn about how to interact with the Cloudera Octopai API, covering the `/api/UserAccount/Login` and `/api/UserAccount/UserEvents` endpoints. The utilization of these endpoints can serve various business use cases including User Behavior Analysis, Security and Fraud Detection, Compliance and Auditing, System Monitoring and Performance, Customer Support, and Product Development.

About this task

An audit trail is a security-relevant chronological record that provides documentary evidence of the sequence of activities that have affected a specific operation, procedure, or event. In the context of this API, it records the sequence of user activities or events.

Audit trails are a crucial aspect of security and compliance for many organizations. They are used to detect security incidents, performance issues, and to aid in the recovery from incidents. Additionally, they support the investigation and forensic analysis of how an incident occurred.

The UserEvents functionality contributes to a user audit trail the following ways:

- User authentication: The API logs events related to user authentication, such as successful and failed login attempts. This can help detect potential security risks, like repeated failed login attempts that might indicate a brute-force attack.
- User activity: The API records various user activities like page loads. By tracking these events, administrators can establish a pattern of normal behavior per user, making it easier to identify anomalous actions that could signify a breach.
- Timestamps: Every event logged by the API includes a timestamp. This allows administrators to reconstruct the sequence of events leading up to a particular incident, which is vital in forensic investigations.
- Data source: The IP addresses from which events originate are also recorded. This can be used to identify suspicious activity from unfamiliar sources.

By extracting and analyzing data from the `UserEvents` API, organizations can maintain a comprehensive audit trail that helps uphold security, facilitate incident response, and ensure regulatory compliance.

Before you begin

- You must be familiar with HTTP methods, specifically POST.
- You must be able to use command-line tools like curl.
- You must have valid Cloudera Octopai user credentials that is email and password.

Procedure

1. User Login

   To authenticate a user, send a POST request to the following endpoint:

   ```
   https://[***YOUR URL NAME***].octopai.com/api/UserAccount/Login
   ```

   Construct and send this request using curl.

   ```
   curl --location 'https://[***YOUR URL NAME***].octopai.com/api/UserAccount/Login' \
   --header 'Content-Type: application/json' \
   --data-raw '{
   "Username": "[***YOUR USER EMAIL***]",
   "Password": "[***YOUR PASSWORD***]"
   }'
   ```

Replace *[***YOUR URL NAME***]*, *[***YOUR USER EMAIL***]*, and *[***YOUR PASSWORD***]* with your actual values.

2. Extract Access Token

   After a successful login, the API returns a JSON response containing an accessToken, which is needed for subsequent authenticated requests.

   Successful response example

```
{
"accessToken": "[***YOUR ACCESS TOKEN***]",
"expiration": "2023-07-23T12:36:37.0355877Z",
"refreshToken": {
"token": "[***YOUR REFRESH TOKEN***]",
"expiration": "2023-08-22T10:36:37.035615Z"
},
"userName": "[***YOUR USER NAME***]",
"displayName": "[***YOUR DISPLAY NAME***]",
"userEmail": null,
"isAdmin": true,
"error": null,
"groupMember": [],
"authLevel": "ADMIN",
"displayModules": "..."
}
```

   Extract the accessToken from this response for the next step.

3. User Events Request

   To retrieve user events, send a POST request to the /api/UserAccount/UserEvents endpoint.

   This request requires the accessToken obtained from the previous step and the user must have an Admin role.

   Construct and send the request.

```
curl --location 'https://[***YOUR URL NAME***].octopai.com/api/Use
rAccount/UserEvents' \
--header 'Content-Type: application/json' \
--header 'Authorization: Bearer [***ACCESS TOKEN***]' \
--data '{
"PageNumber":"1",
"RowsInPage":"50",
"fromDate": "2024-01-10 00:00:01",
"toDate": "2024-01-10 23:59:59"
}'
```

   Replace *[***YOUR NAME***]* and *[***ACCESS TOKEN***]* with your actual values.

4. Understanding the Response

   The response from the /api/UserAccount/UserEvents endpoint is an array of objects, each representing a user event.

   Successful response example

```
[
{
```

```
"type": "LOGIN SUCCESS",
"userName": null,
"source": "[***IP***]",
"time": "2023-07-23T10:39:33.84"
},
{
"type": "LOGIN FAIL",
"userName": null,
"source": "[***IP***]",
"time": "2023-07-18T13:11:17.6"
},
{
"type": "PAGE LOAD",
"userName": null,
"source": "[***IP***]",
"time": "2023-07-18T13:10:58.2"
}
]
```


Each object contains the following properties:

- ‘type’: The type of the event, for example, "LOGIN SUCCESS", "LOGIN FAIL".
- ‘userName’: The username associated with the event. This might be null or "UNKNOWN USER".
- ‘source’: The IP address from which the event originated.
- ‘time’: The timestamp of when the event occurred.

What to do next

Store your accessToken securely and refresh it as needed to prevent it from being compromised. Always use HTTPS to make your requests to ensure data security during transit.

Contact Cloudera Support for more details and instructions.