Dell Data Lakehouse and Diskover: Creating AI Datasets from Unstructured Data

Transforming unstructured data into AI-Driven insights

February 2025

H04428

White Paper

Abstract

This white paper highlights how Dell Data Lakehouse and Diskover integrate to convert unstructured data into contextual, actionable datasets. By combining Diskover's metadata inventory capabilities with Dell Data Lakehouse's analytics power, the solution streamlines data indexing and contextualization.

DELL Technologies

Copyright

Contents

Executive summary

Overview

Unstructured data dominates enterprise environments but often remains untapped due to its complexity. The integration of Dell Data Lakehouse (DDLH) and Diskover revolutionizes how organizations transform this raw data into meaningful datasets for AI and generative AI (GenAI) workloads. Diskover's metadata inventory capabilities efficiently index and organize unstructured data, while DDLH's analytics unlock insights and contextual value. This seamless combination streamlines workflows, reduces redundancies, and enhances data-driven decision making. It enables businesses to aggregate and contextualize metadata, creating rich datasets ideal for machine learning (ML) and generative AI applications.

Scalable processing ensures the solution can handle vast datasets as businesses grow. GenAI applications benefit significantly from the contextualized data, improving training and deployment outcomes. Automation reduces manual effort, driving productivity while fostering innovation. Designed to adapt to both current and future data management needs, this solution positions organizations to innovate rapidly and respond to new challenges. By leveraging this integration, enterprises unlock the full potential of their unstructured data, fueling smarter operations and innovative advancements in AI and GenAI.

Audience

This document is intended for enterprises with data lakes or a data lake strategy interested in empowering their organizations to act more quickly, effectively, and efficiently on their data. Audience roles include:

- Data and application administrators
- Data engineers
- Data scientists
- IT decision makers

A data lakehouse can assist more traditional analytics customers looking to modernize their data collection. It can also help analytics systems to get more value from their data or standardize their data for modern analytics workloads.

Revisions

Date	Part number/ revision	Description
February 2025	H04428	Initial release

We value your feedback.

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Authors: Kirankumar Bhusanurmath, DA/AI Specialist | Dell Technologies

Chris Park, Brandon Langley | Diskover

Note: For links to other documentation for this topic, see the Data Analytics Info Hub.

Solution Components

Dell Data Lakehouse

DDLH is a turnkey solution, boasting the Dell Data Analytics Engine, a potent federated data lake query engine powered by Starburst. The Dell Lakehouse System software ensures life cycle management, while tailor-made compute hardware provides seamless integration. Notably, this platform is designed to support AI solutions with its AI-ready data platform. For storing and processing large datasets in open file and table formats, Dell's leading S3 storage platforms, ECS, ObjectScale, and PowerScale, deliver exceptional performance, reliability, and security.

The DDLH core lies the Dell Data Analytics Engine (DDAE), powered by Starburst, facilitating the discovery, querying, and processing of enterprise-wide data assets regardless of their physical location. By reducing data movement requirements and enhancing query efficiency, the DDLH sets a new benchmark in data platform optimization and performance.

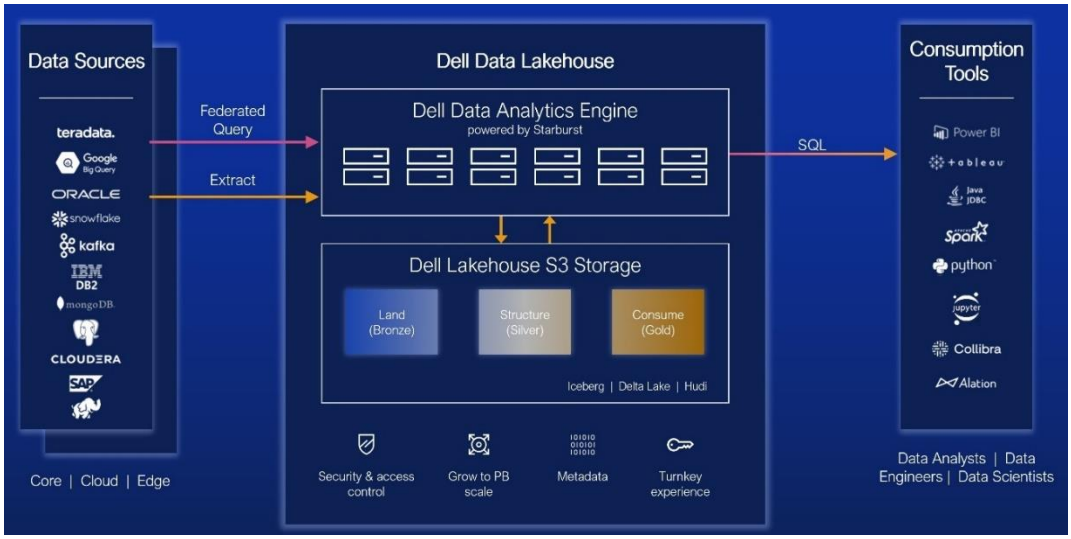


Figure 1. Dell Data Lakehouse diagram

Dell Data Analytics Engine

DDAE contains an analytics query engine powered by Starburst. It is a fully supported enterprise-grade distributed SQL query engine designed for high-performance analytics. It allows users to query large amounts of data stored in various data sources throughout an organization using standard SQL syntax.

One of the key features of DDAE is its ability to run queries across different data sources simultaneously, in the same query. These sources include relational databases, NoSQL databases, Object Storage systems, and more. Response times are fast enough to support real-time analysis. Users can use this query engine to process data across multiple systems and data sources.

With the integrated query engine, administrators can implement a layer on top of data that abstracts away details on location, connectivity, language variations, and API. This layer of abstraction is critical to simplify data analytics over a diverse set of data sources.

DDAE Metastore

DDAE Metastore is a dedicated Hive Metastore Server (HMS).

The Hive Metastore (HMS) is a central repository of metadata for Hive, Iceberg, Delta, and Hudi tables, and provides DDAE client access to this information using the Metastore service API. It is the building block for DDLH that uses the diverse world of open-source software, such as Apache Spark and DDAE's Trino.

Diskover

Diskover is a powerful solution designed to provide global visibility into unstructured data, helping organizations unlock the full potential of their digital assets. With features like global indexing, search, and analytics, Diskover enables businesses to efficiently manage data spread across diverse systems and locations. Its metadata inventory capabilities allow for seamless aggregation, organization, and contextualization of data, making it an invaluable tool for addressing the challenges of data complexity.

The platform enhances security by identifying and mitigating risks associated with uncontrolled or siloed data. The Diskover index provides visibility without actual file system access enabling Data Scientist the ability to find and request access to data desired for AI pipeline. By enabling efficient search and retrieval of relevant datasets, Diskover significantly reduces the time and resources required for data management. This not only minimizes operational costs but also boosts productivity across teams. Diskover's ability to integrate with multiple systems and environments, including business intelligence (BI) and AI pipelines, positions it as a versatile solution for enterprises.

A standout feature of Diskover is its ability to enrich data with metadata context, enabling advanced curation processes such as deduplication, tagging, and purging. This ensures that data pipelines are streamlined, improving the quality and relevance of datasets for AI and generative AI applications. By fostering smarter decision making and accelerating innovation, Diskover empowers businesses to harness their unstructured data effectively, driving success in the modern data-driven landscape.

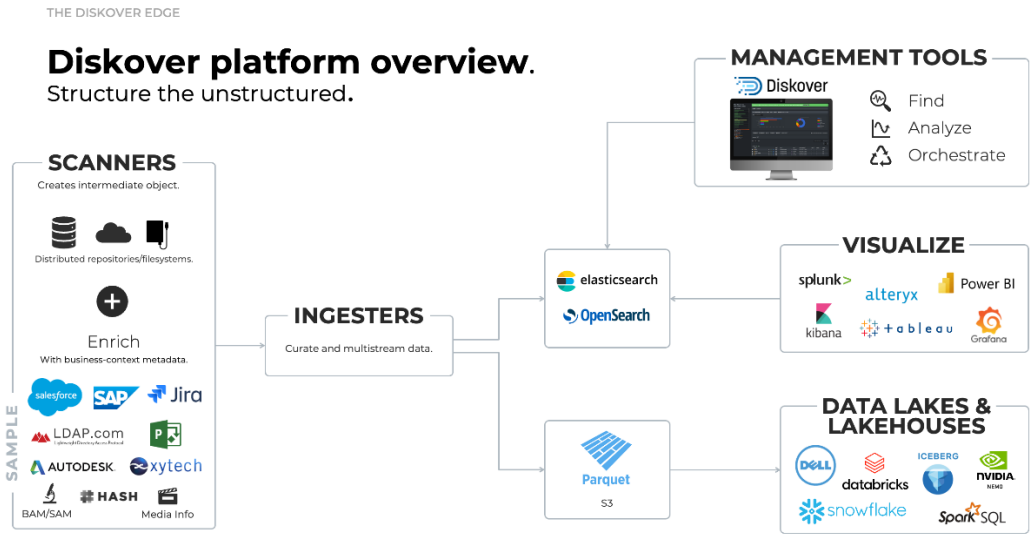
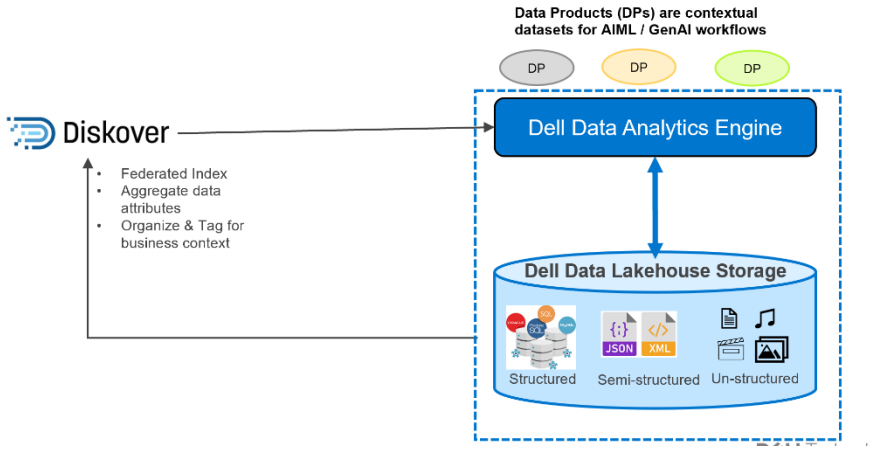


Figure 2. Diskover platform overview

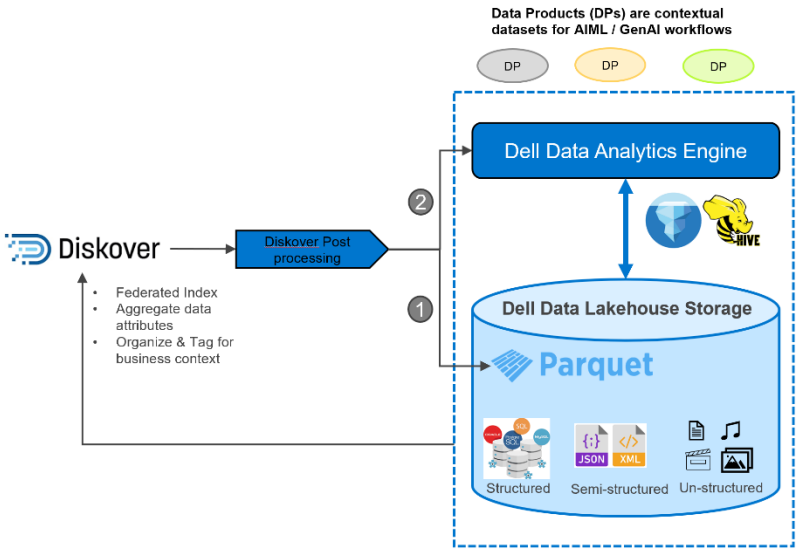
Solution architecture

The integration of Diskover and DDLH is designed to optimize metadata management, enabling enhanced organization, contextualization, and usability of unstructured data across AI and GenAI workflows. This architecture streamlines the flow of metadata for improved accessibility and scalability, ensuring seamless support for advanced analytics and decision making. While both solutions are fundamentally the same, they differ in their approach:

The first solution positions Diskover as the metadata source, where DDLH actively pulls the metadata inventory files for processing:



The other utilizes DDLH as the destination, with Diskover pushing metadata inventory files as partitioned Parquet files and registering the corresponding schema with DDLH.



The solutions within this framework are:

Solution 1: Diskover Metadata Inventory as a Federated Source for DDLH

Solution 2: DDLH as a Destination for Diskover Metadata Inventory Files

Solutions

Solution 1: Diskover Metadata Inventory as a Federated Source for DDLH

This solution demonstrates how Diskover and DDLH work together to streamline metadata management and make unstructured data more accessible and usable for AI and GenAI workflows. Here is a detailed step-by-step process:

Step 1: Metadata Inventory Creation

Diskover scans and indexes unstructured data sources across various systems and generates a comprehensive metadata inventory. This inventory provides critical information about each file, including attributes like file type, size, location, timestamps, and content-based metadata.

Step 2: Integration with ElasticSearch

The metadata inventory created by Diskover is stored in ElasticSearch, which acts as a central repository. This enables fast querying and retrieval of metadata, supporting high-performance integration workflows.

Step 3: Metadata Registration in DDLH

DDLH connects to the ElasticSearch repository and pulls metadata created by Diskover. By leveraging this metadata as a federated source, DDLH integrates it into its system, creating a unified view of metadata records.

Step 4: Exposing Metadata Inventory

Once the metadata is registered within DDLH, it is exposed as tabular datasets. These datasets are structured in a way that makes them easily accessible and interpretable by downstream applications. This ensures compatibility with various BI tools and AI workloads, facilitating seamless collaboration between teams.

Step 5: Leveraging Metadata for Advanced Workflows

With the metadata inventory now accessible as a federated source, businesses can utilize it to build contextual datasets. These datasets serve as the foundation for advanced data analysis and AI model development, enabling more informed decision making and optimized operations.

Integration and Data Flow

The interaction between Diskover and DDLH in this solution exemplifies a smooth data flow process:

- Diskover acts as the upstream source, generating enriched metadata.
- ElasticSearch bridges the storage and querying of metadata between the two systems.
- DDLH pulls metadata from ElasticSearch, contextualizing and organizing it for usage in various workflows.

Role of Metadata Management

This solution highlights the pivotal role of metadata management, providing businesses with the ability to:

- Organize unstructured datasets for better accessibility
- Contextualize data to enhance usability across AI workflows

- Streamline the preparation of data for BI and AI applications

By harnessing the power of metadata management through the integration of Diskover and DDLH, organizations can unlock new efficiencies and accelerate their AI and GenAI initiatives.

Solution 2: DDLH as a Destination for Diskover Metadata Inventory Files

This solution outlines how Diskover pushes its metadata inventory directly to DDLH, providing a structured and enriched foundation for BI and AI workflows. The following steps detail the process:

Step 1: Metadata Inventory Creation

Diskover scans unstructured data across various data sources and builds a detailed metadata inventory. This inventory captures file attributes such as size, type, location, timestamps, and content-based metadata, offering deep visibility into datasets.

Step 2: Partitioned Parquet File Generation

Diskover organizes the metadata inventory into partitioned Parquet files, an efficient and standardized file format. Partitioning ensures optimal data retrieval and processing performance by categorizing metadata based on logical criteria, such as time or file type.

Step 3: Metadata Transfer to DDLH

Diskover actively pushes these partitioned Parquet files to DDLH. By doing so, metadata files are directly delivered into the DDLH ecosystem, bypassing intermediary repositories like ElasticSearch.

Step 4: Schema Registration in DDLH

After transferring the metadata files, Diskover registers the corresponding schema with DDLH. This schema registration ensures that DDLH can effectively interpret and process the metadata files, enabling seamless integration with DDLH's querying and data transformation capabilities.

Step 5: Structuring Metadata for BI and AI Workloads

Within DDLH, the partitioned Parquet metadata files are made available as organized, tabular datasets. These datasets are optimized for querying and analysis, making them suitable for use in various BI tools and AI projects. This structured format allows teams to explore metadata insights and create sophisticated models tailored to their operational needs.

Integration and Data Flow

The integration process reflects the seamless handover of metadata between Diskover and DDLH:

- Diskover operates as the upstream system, processing and pushing metadata directly into DDLH.
- DDLH serves as the destination, where metadata is stored, interpreted, and made accessible for downstream workflows.

Role of Metadata Management

This solution underscores the importance of robust metadata management in enabling:

- Efficient data organization by leveraging partitioned file formats for scalability

- Seamless schema alignment between source and destination systems
- Enhanced usability of metadata for diverse data-driven applications, such as decision making, AI models, and predictive analytics

By utilizing DDLH as the destination, organizations gain a centralized, scalable, and actionable metadata repository. This integration empowers businesses to maximize their unstructured data's potential, driving success in AI and GenAI initiatives.

Key benefits

Improved Data Accessibility

Seamless access to metadata provides teams with actionable insights directly from organized datasets.

Enhanced Data Contextualization

Metadata integration ensures that data is not only accessible but also meaningful and relevant for analysis.

Streamlined AI and GenAI Workflows

Easy access to prepared, structured data accelerates the development and deployment of AI models.

Efficient Data Organization

Leveraging partitioned Parquet files and unified schemas ensures faster data retrieval and processing.

Scalability

The solutions support growing data volumes without compromising performance or accessibility.

Better Decision Making

Holistic and well-organized metadata lays the groundwork for better, data-driven insights and strategies.

Optimized Operations

Reduced complexity in cataloging and accessing metadata ultimately improves operational efficiency.

Accelerated AI Initiatives

Faster access to high-quality metadata supports rapid innovation in AI and GenAI projects.

These benefits collectively enable organizations to unlock the full potential of their unstructured data, driving success in data-intensive applications.

Solution validation

The validation process for both Solution 1 and Solution 2 focuses on ensuring seamless integration and functionality between Diskover and DDLH. Initially, the environment is set up for both Diskover and DDLH, making sure all required configurations are in place. The

integration is then established to enable Diskover to scan the DDLH storage for unstructured data and build a comprehensive metadata inventory. This metadata inventory is further processed and integrated with the DDLH Engine, where it is transformed into structured tabular datasets and contextual data sets that can be utilized as data products. These steps ensure that the solutions are fully optimized for BI and AI workloads, meeting the set objectives for accuracy and efficiency.

Solution 1: Diskover Metadata Inventory as a Federated Source for DDLH

Set up Dell Data Lakehouse

It is assumed the DDLH appliance and Dell PowerScale storage cluster is installed and configured.

Set up Dell Data Lakehouse storage

Configure Dell PowerScale as the primary storage cluster for the DDLH.

1. Log in to the DDLH system software.
2. Under **Storage** configure the Dell PowerScale S3 endpoint.

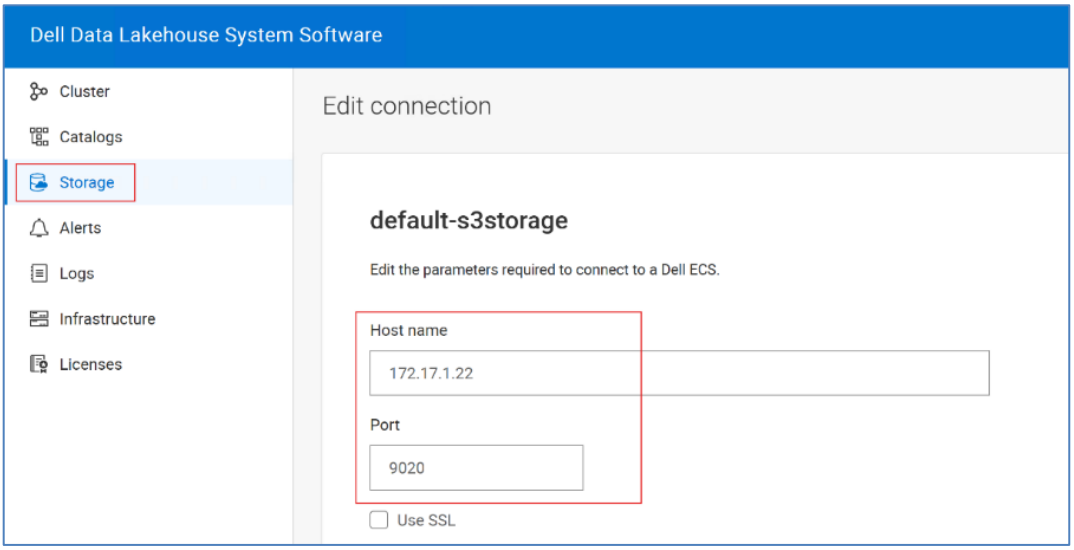


Figure 3. Add S3 storage endpoint

Setup Hive Catalog

Configure the Hive Catalog to storage Hive tables on the Dell PowerScale of the DDLH.

1. Log in to The DDLH system software
2. Under **Catalogs** select **Connect Catalog**.
3. Select **Properties**. For **Type** choose **Hive**. Enter the configuration parameters in the **Configuration** field.

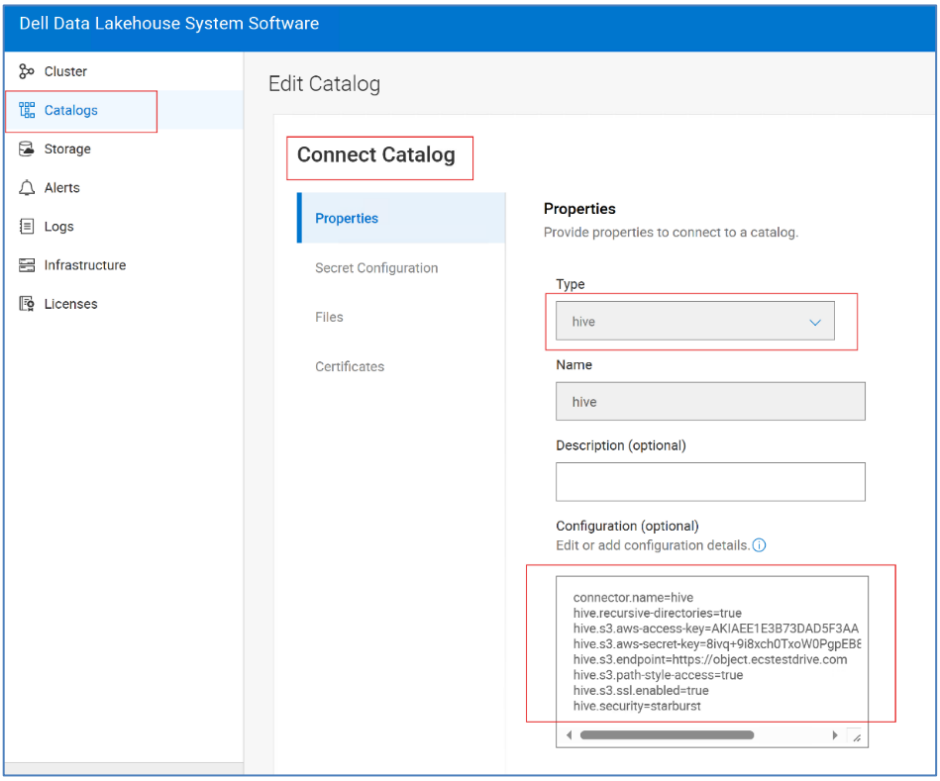


Figure 4. Add Hive Catalog

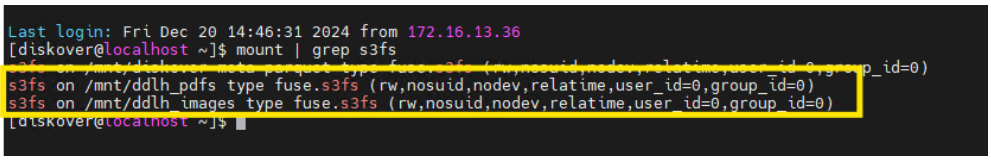
Setup Diskover

Diskover installation

On the utility node install Diskover. See here for further information. For more information, see Diskover documentation.

Diskover and DDL Storage S3 setup

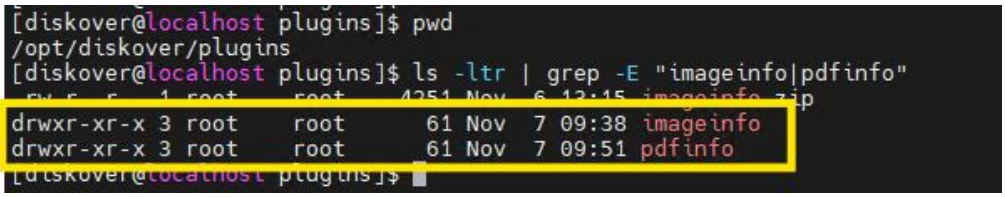
On the utility node where Diskover is installed, S3fs either as python library or Linux package. Mount the S3 bucket of DDLH storage to the Diskover client Linux file system as an S3fs mount point. This enables the Diskover plugin to run additional content metadata extraction of each unstructured object it scans.



If Dell PowerScale is used as the DDLH Storage, it is recommended to mount the PowerScale to Diskover as an NFS mount for better performance.

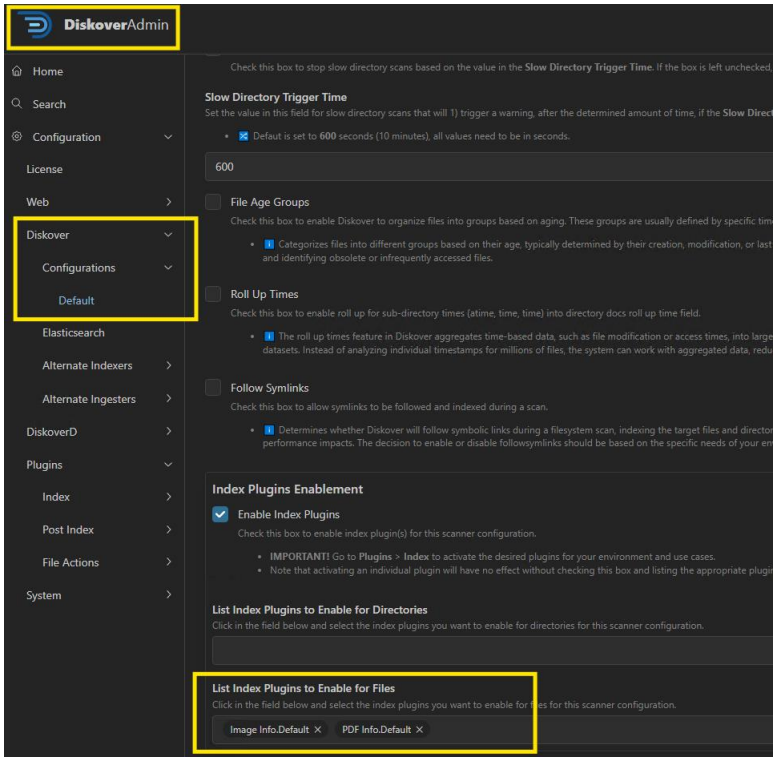
Diskover plugin configuration

Diskover's powerful extensibility enables the development of custom plugins by Diskover, third parties, or even end users to enrich metadata catalogs. For this solution validation, we focus on two specific plugins, PDFs and images. These plugins are enhanced to extract additional system metadata, including keywords, subject, and the absolute path within the S3 bucket. They also capture detailed content metadata, providing insights into unstructured files and contextual objects contained within them.



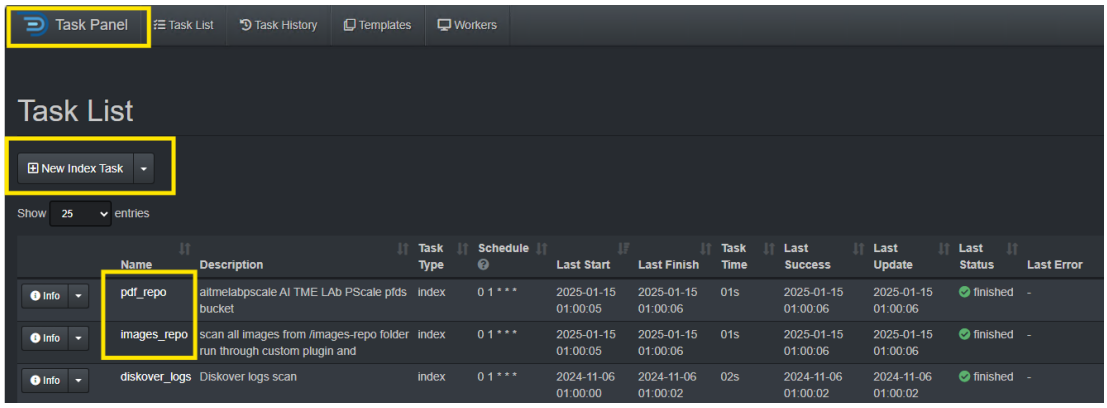
Diskover enable plugins

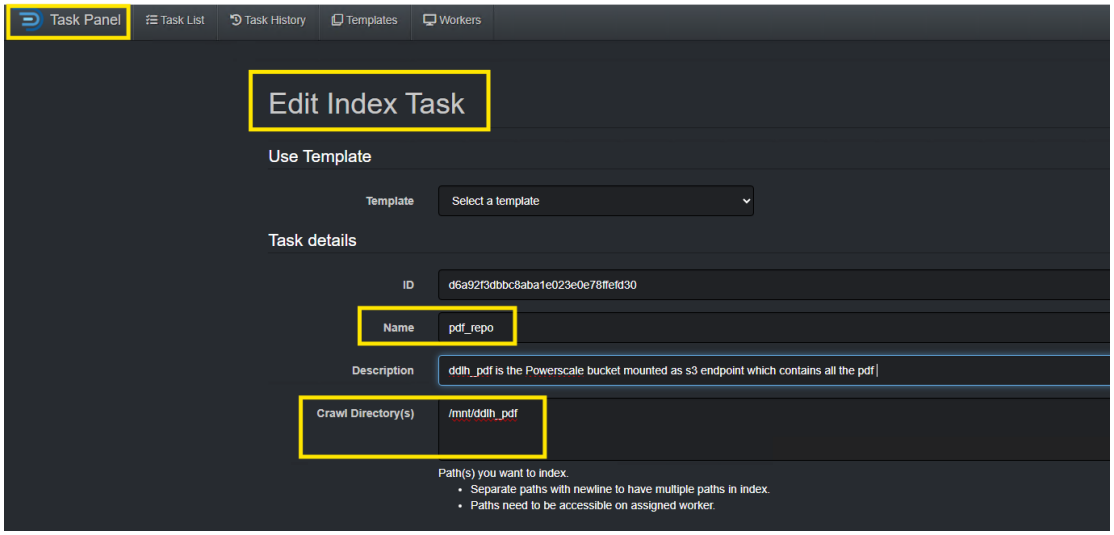
Navigate to **Diskover** > **Configurations** and enable the enriched plugins.



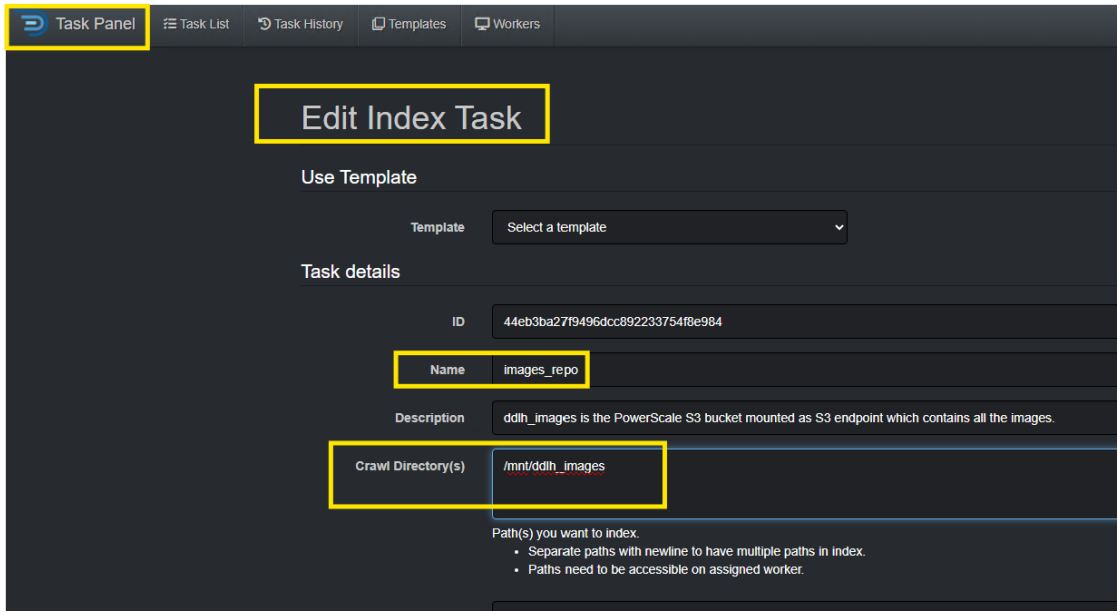
Diskover Add New Index Task

Select **Task Panel** and add two **New Index Task** items, one for *pdf_repo* and other for *images_ repo*.





New Index Task for *pdf_repo.*

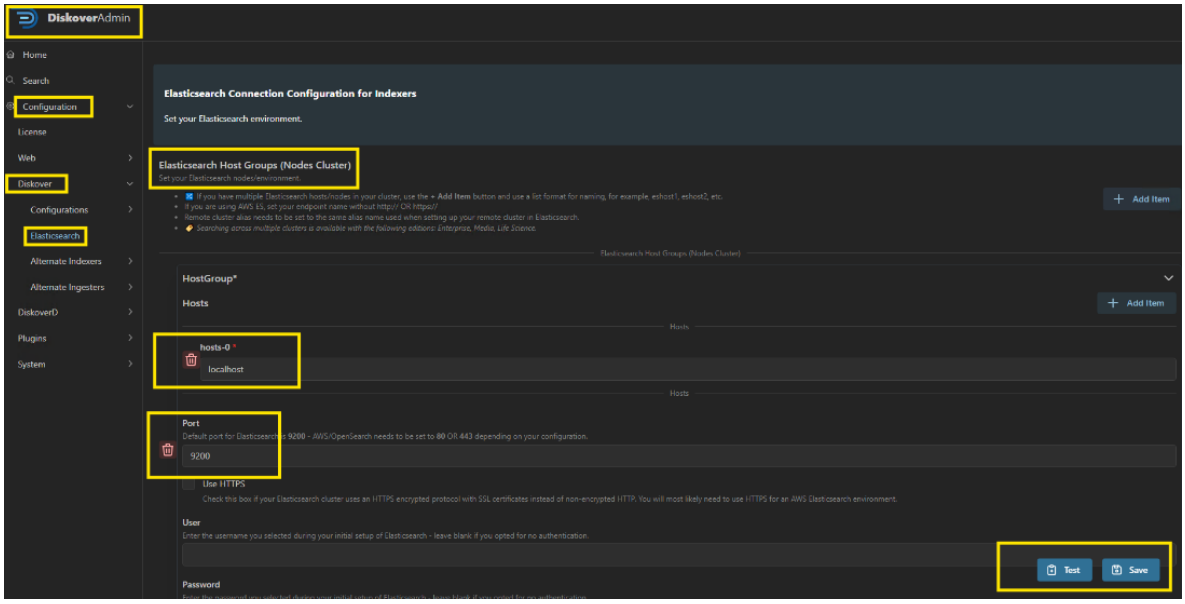


New Index Task for *image_repo.*

Set up ElasticSearch

To configure ElasticSearch in the Diskover Admin UI:

1. Log in and navigate to the Configuration section.
2. Locate the **ElasticSearch** configuration options and ensure the settings point to the default ElasticSearch instance included with Diskover. Verify that all configuration fields, such as host, port, and credentials match the default instance.
3. Save the settings and test the connection to confirm integration, ensuring seamless metadata indexing and search functionality.



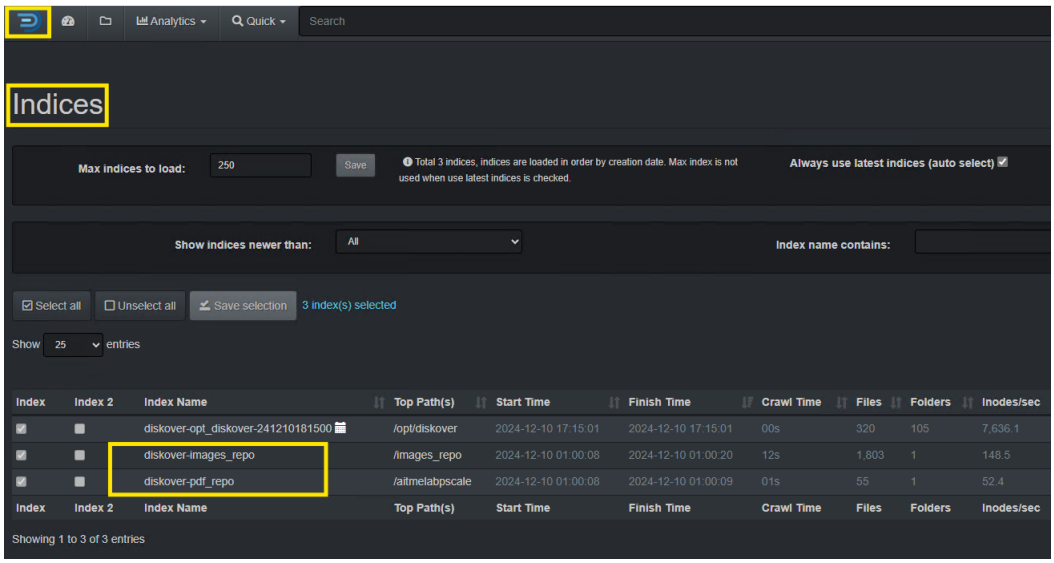
Set up DDLH and Diskover ElasticSearch Integration

To establish a connection between DDLH and the Diskover ElasticSearch instance, ensure network accessibility between the two systems and configure authentication credentials for secure communication. Verify proper metadata field mapping to maintain consistency during integration.

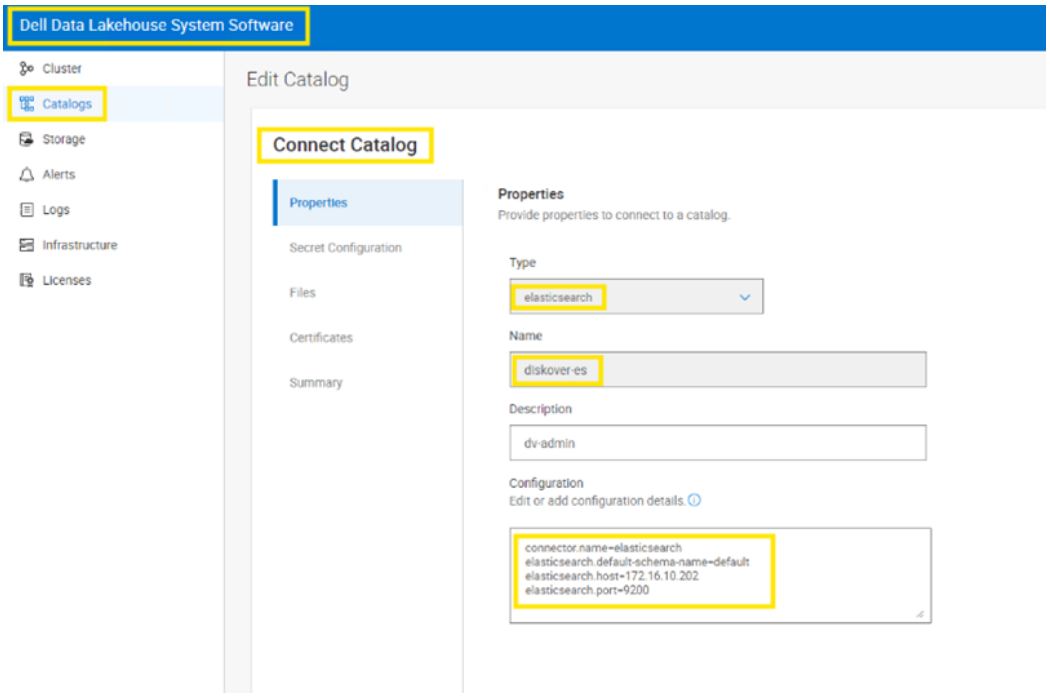
On the DDLH software UI, navigate to **Catalog** and create a new ElasticSearch catalog for Diskover Metadata Inventory to complete the setup.

Solution Validation

Validate metadata index tasks

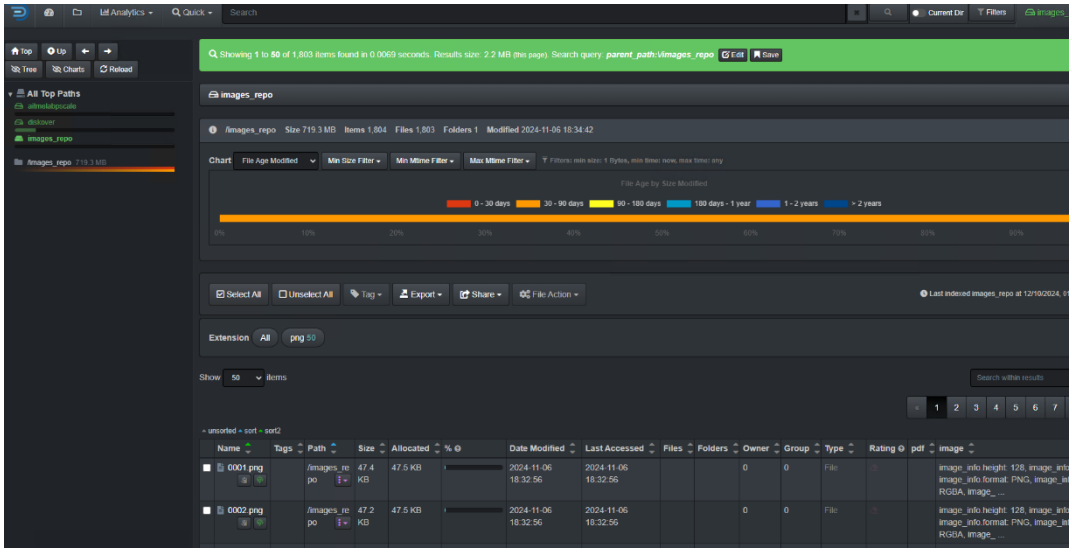


To verify that ElasticSearch is populated with metadata indices for the configured PDF and image paths, check that the metadata inventory has been successfully scanned, extracted, and stored as indices by Diskover. Confirm that the indices corresponding to these file paths are present in ElasticSearch, ensuring they are correctly populated. Once verified, these indices are ready for immediate use.



Validate ElasticSearch populated with Metadata inventory

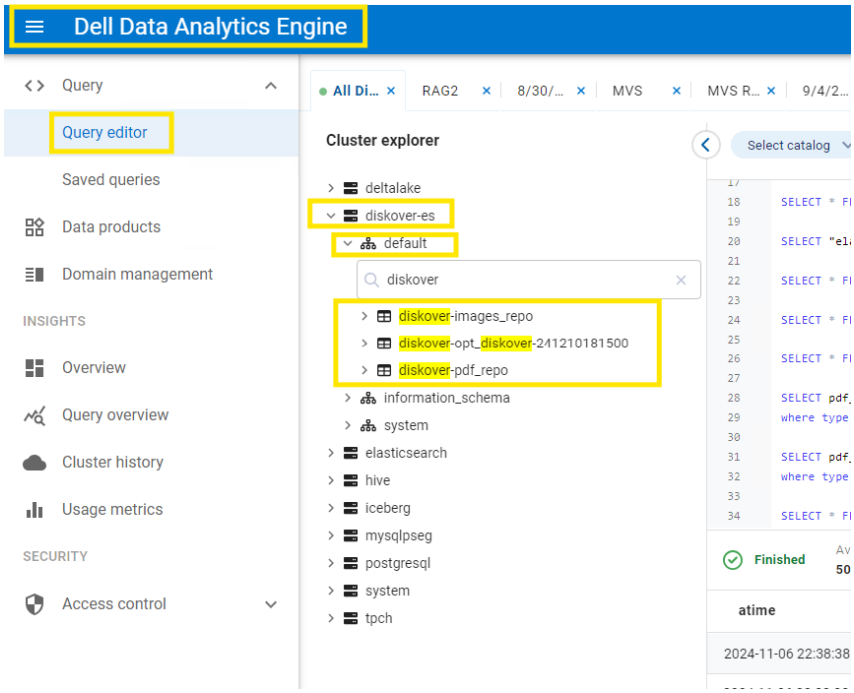
From the Diskover Web UI we can browse the metadata index, or Diskover can provide an analytics UI which populates the ElasticSearch Indices in more consumable dashboards and reports.



Validate DDLH Metadata with ElasticSearch Indices

Validate ElasticSearch Catalog schema

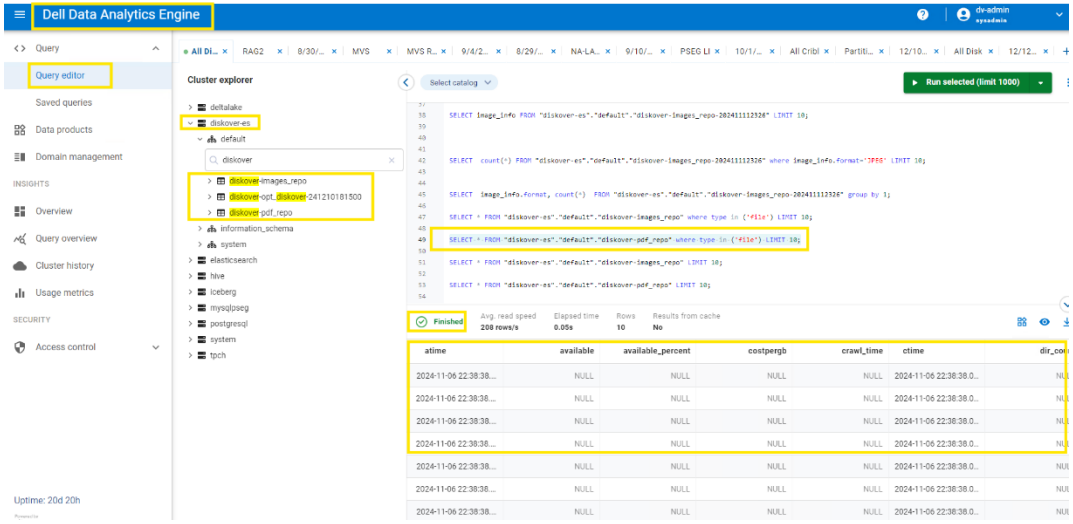
From the DDAE UI, browse the Diskover catalog previously set up. Under the default schema you will see the Diskover metadata indices the PDF and image repos.



Sample SQL on Diskover Metadata Indices

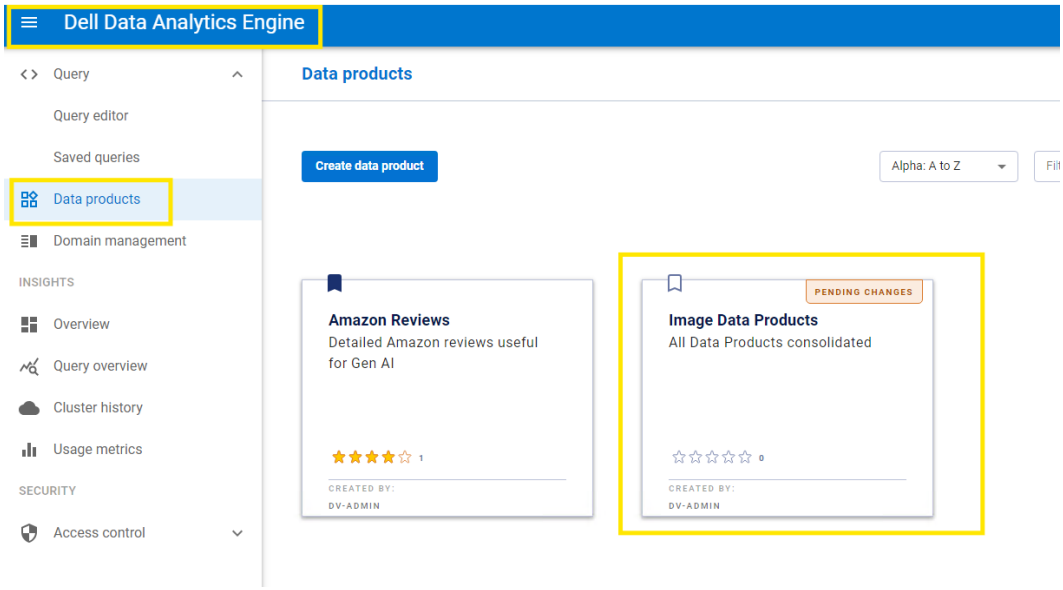
On the DDAE Query editor workspace, run a simple select SQL query to pull metadata information from Diskover ElasticSearch over the catalog connector that was established.

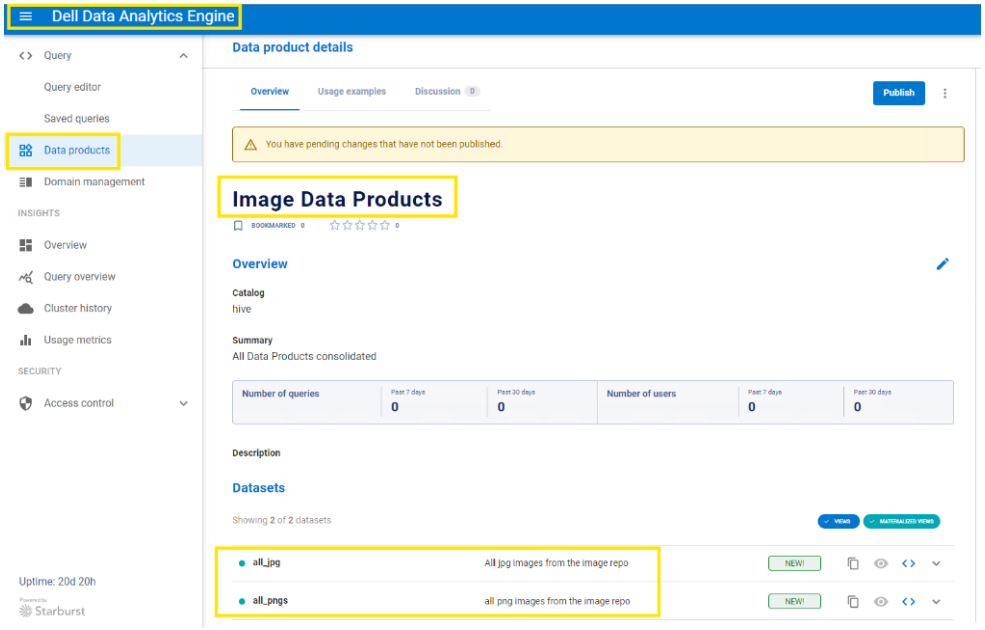
Create sample data product as contextual datasets



As the metadata is available in the DDLH, we can create contextual datasets as Data products. For more information on data products. See [Starburst Data Product Documentation](#).

From the DDAE UI, select **Data Product** and create new data product. For this validation we will create a separate image data product from the metadata inventory file from the Diskover Elasticsearch catalog. We will create two datasets under the data product: one for JPG files, and the other for PNG files.





Solution 2: DDLH as a Destination for Diskover Metadata Inventory Files

Setup Dell Data Lakehouse

These steps are same as described above in Setup Dell Data Lakehouse.

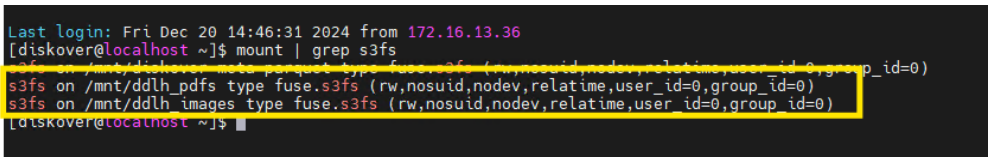
Setup Diskover

Diskover installation

On the utility node install Diskover. See here for further information. For more information, see Diskover documentation.

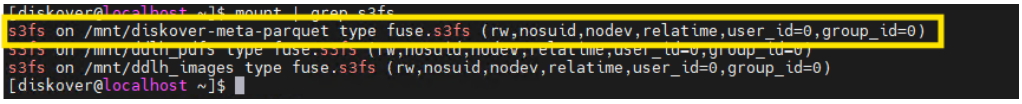
Diskover and DDLH Storage S3 setup

On the utility node where Diskover is installed, S3fs either as python library or Linux package. Mount the S3 bucket of DDLH storage to the Diskover client Linux file system as an S3fs mount point. This enables the Diskover plugin to run additional content metadata extraction of each unstructured object it scans.



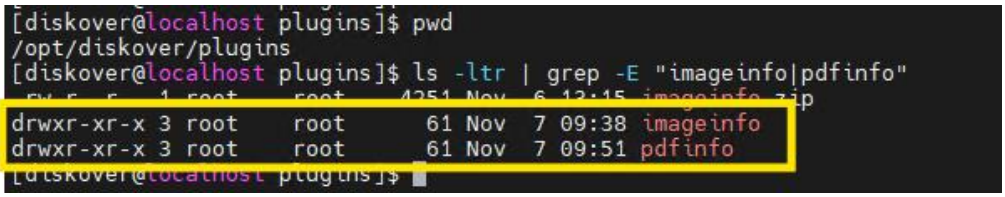
PowerScale S3bucket as the destination to Diskover Metadata inventory

Now we will mount the Dell PowerScale S3 bucket as the destination for Diskover to ingest Metadata inventory in the Parquet file format (open file format).



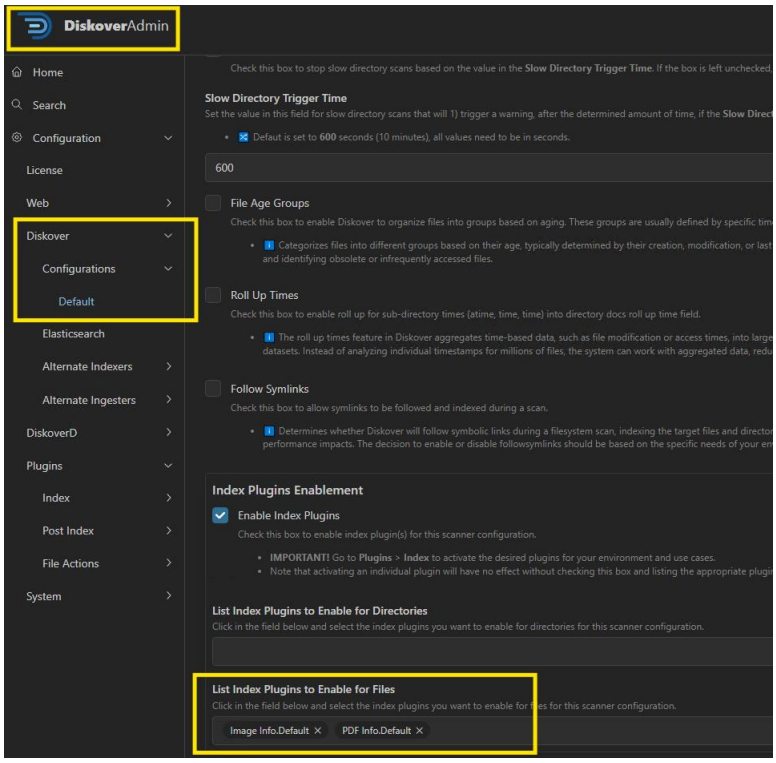
Diskover plugin configuration

Diskover's powerful extensibility enables the development of custom plugins by Diskover, third parties, or even end users to enrich metadata catalogs. For this solution validation, we focus on two specific plugins, PDFs and images. These plugins are enhanced to extract additional system metadata, including keywords, subject, and the absolute path within the S3 bucket. They also capture detailed content metadata, providing insights into unstructured files and contextual objects contained within them.



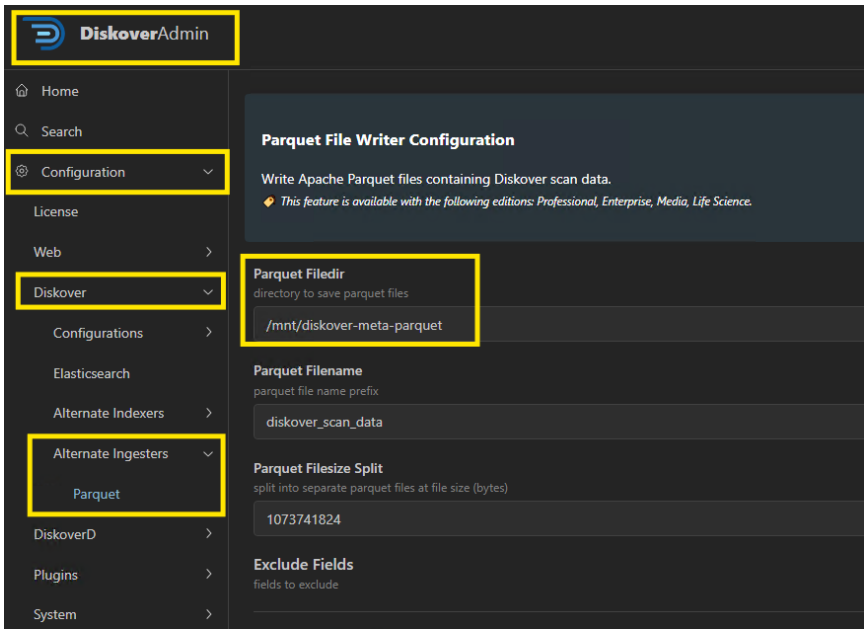
Diskover enable plugins

Under Diskover Admin UI, enable the enriched plugins:



Set up Diskover alternate ingesters to point to S3 bucket on Powerscale

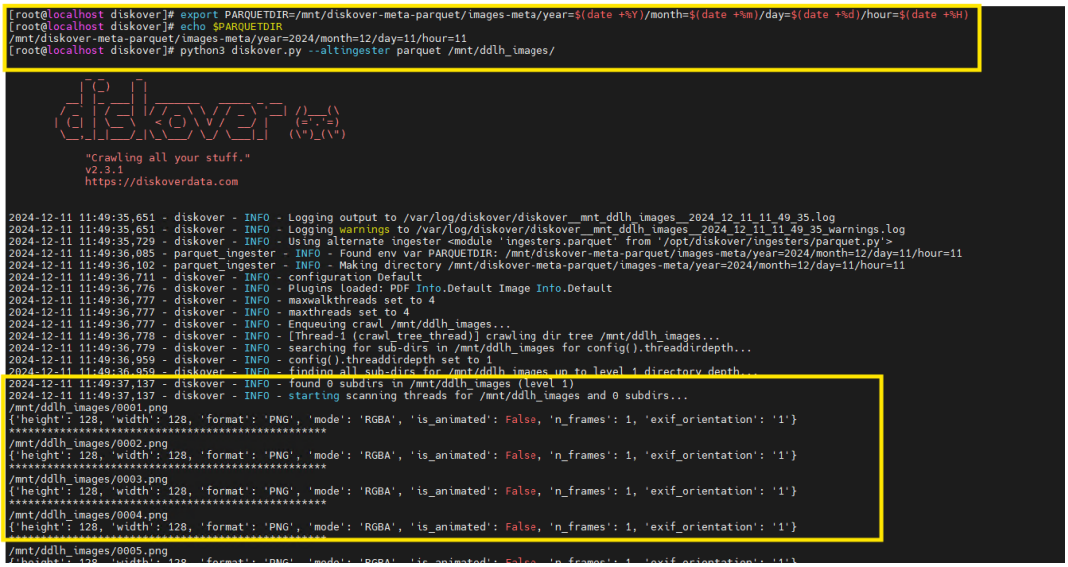
This is a crucial step where we enable alternate ingesters in Diskover and choose the Parquet file format. Diskover converts the metadata inventory into Parquet file format and pushes it into the S3 bucket mounted on the utility node running Diskover.

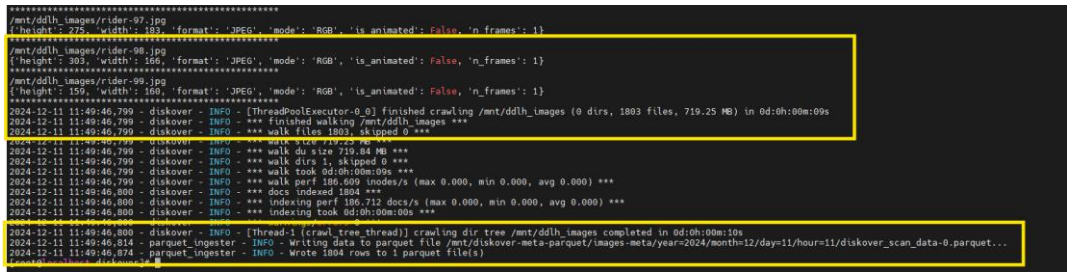


Solution Validation

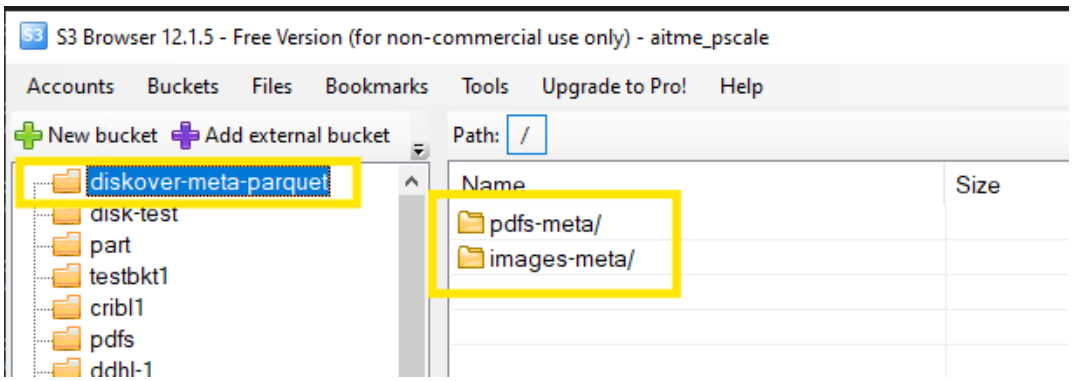
IMAGE Metadata Extraction and schema registration

1. Run the Diskover scanner on image repo.
2. On the utility node where Diskover is running we will run the Diskover scanner manually for the image repository. This is to scan the image repo S3 bucket on Dell Data Lakehouse storage that is mounted as S3 mount point to the utility node. The destination bucket will be again on the lakehouse storage S3 bucket. The metadata extracted needs to be partitioned as stored as parquet files.

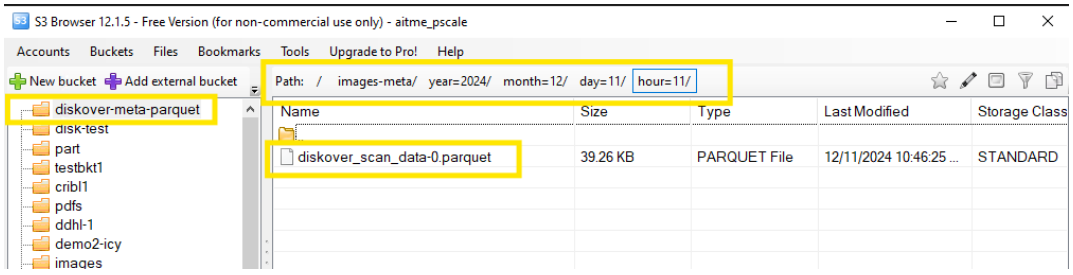




Verify the destination S3 bucket (using S3 Browser in this example) is populated with the image metadata inventory file as Parquet files:



Verify the proper partitioning is retained for the Image metadata repository:



On DDLH, run the schema extraction script on S3 bucket where Diskover saves image metadata in Parquet file format.

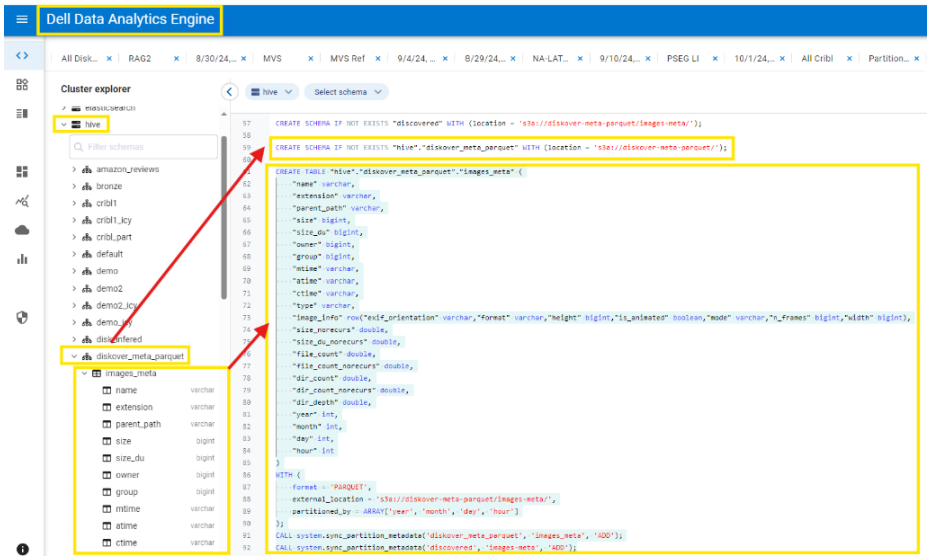
```
49
50  ----------------- Infer DDL schema from the image meta parquet file from s3 bucket
51  select sql from hive.schema_discovery.discovery
52  where
53  uri='s3a://diskover-meta-parquet/images-meta';
54
```

Finished | Avg. read speed: 16.9 rows/s | Elapsed time: 0.06s | Rows: 1 | Results from cache: No

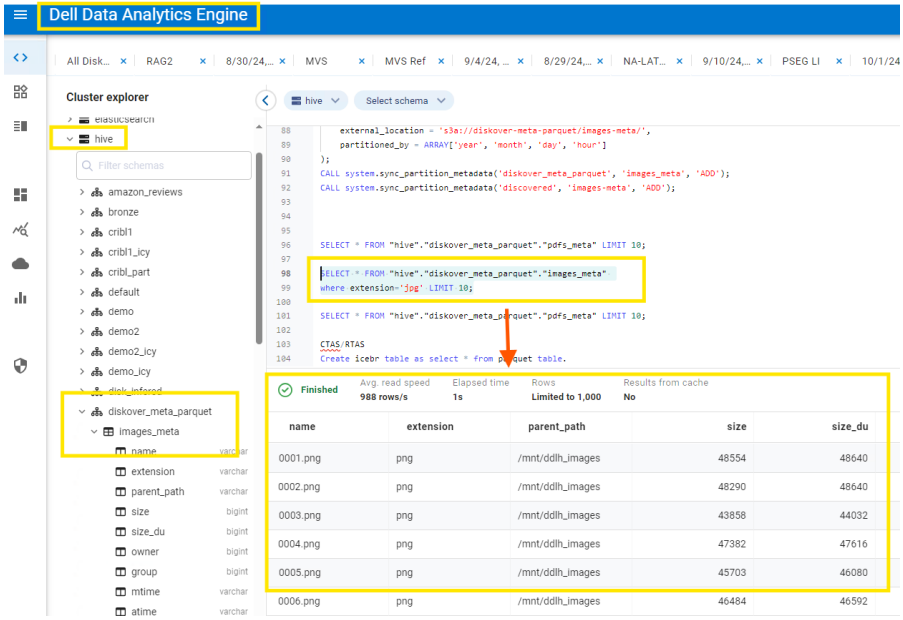
sql

```
CREATE SCHEMA IF NOT EXISTS "discovered" WITH (location = 's3a://diskover-meta-parquet/images-meta/');

USE "discovered";
CREATE TABLE "images-meta" (
   "name" varchar,
   "extension" varchar,
   "parent_path" varchar,
   "size" bigint,
   "size_du" bigint,
   "owner" bigint,
   "group" bigint,
   "mtime" varchar,
   "atime" varchar,
   "ctime" varchar,
   "type" varchar,
   "image_info" row("exif_orientation" varchar,"format" varchar,"height" bigint,"is_animated" boolean,"mode" varchar,"n_frames" bigint,"width" bigint),
   "size_norecurs" double,
   "size_du_norecurs" double,
   "file_count" double,
   "file_count_norecurs" double,
   "dir_count" double,
   "dir_count_norecurs" double,
   "dir_depth" double,
   "year" varchar,
   "month" varchar,
   "day" varchar,
   "hour" varchar,
   "year" int,
   "month" int,
   "day" int,
   "hour" int
)
WITH (
   format = 'PARQUET',
   external_location = 's3a://diskover-meta-parquet/images-meta/',
   partitioned_by = ARRAY['year', 'month', 'day', 'hour']
);

CALL system.sync_partition_metadata('discovered', 'images-meta', 'ADD');
```

Copy the schema discovery output, (DDL schema), and register it as a table under the schema *diskover_meta_parquet*. The new table created is called *images_meta*.

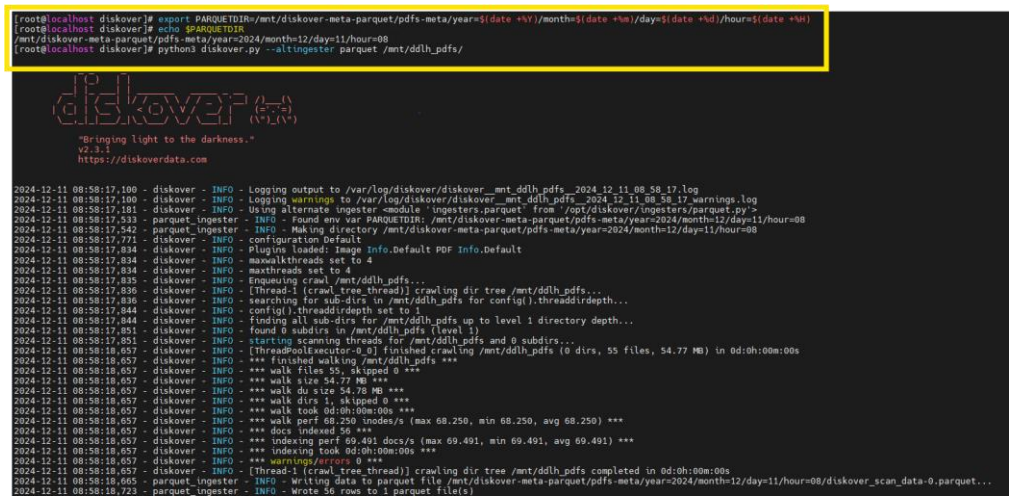


Run sample SQL queries on images metadata table:

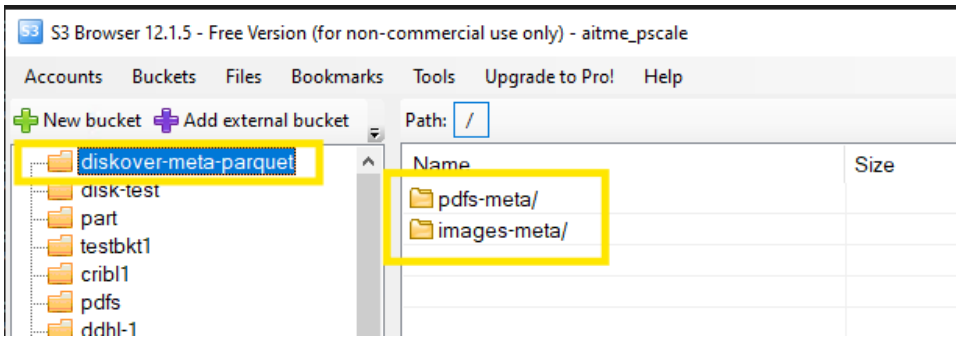


PDFs Metadata Extraction and schema registration

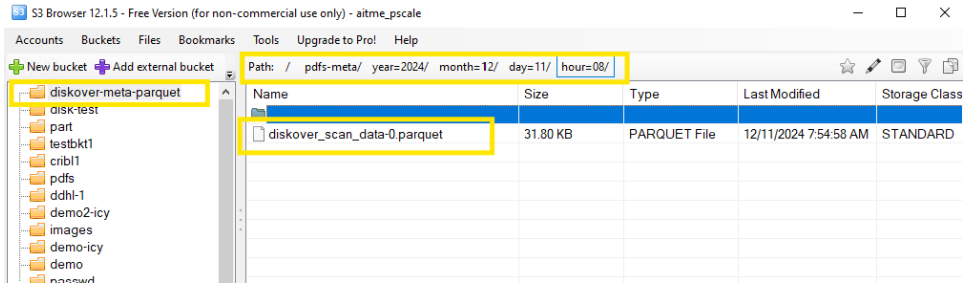
Run the Diskover scanner on PDF repo.



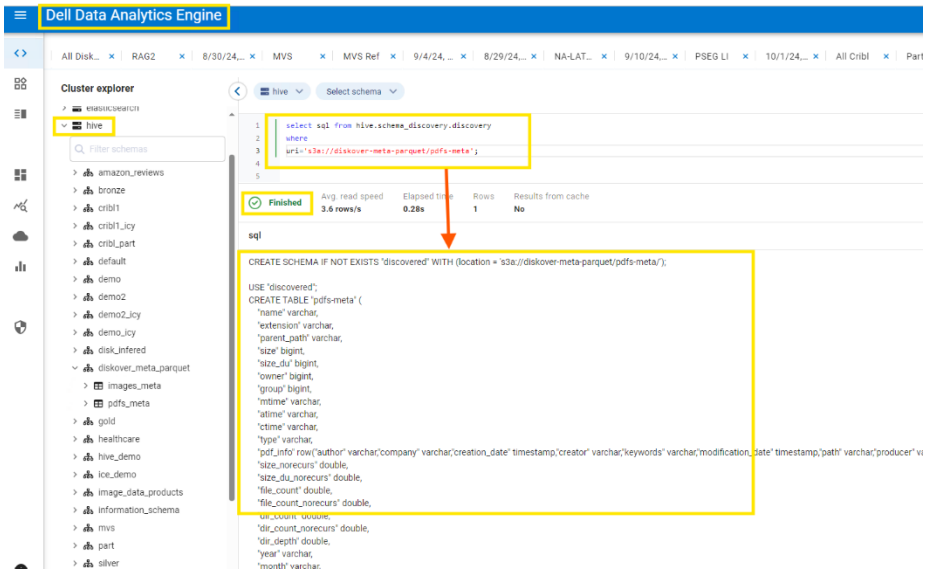
Verify the destination S3 bucket is populated with the PDF metadata inventory file as Parquet files.



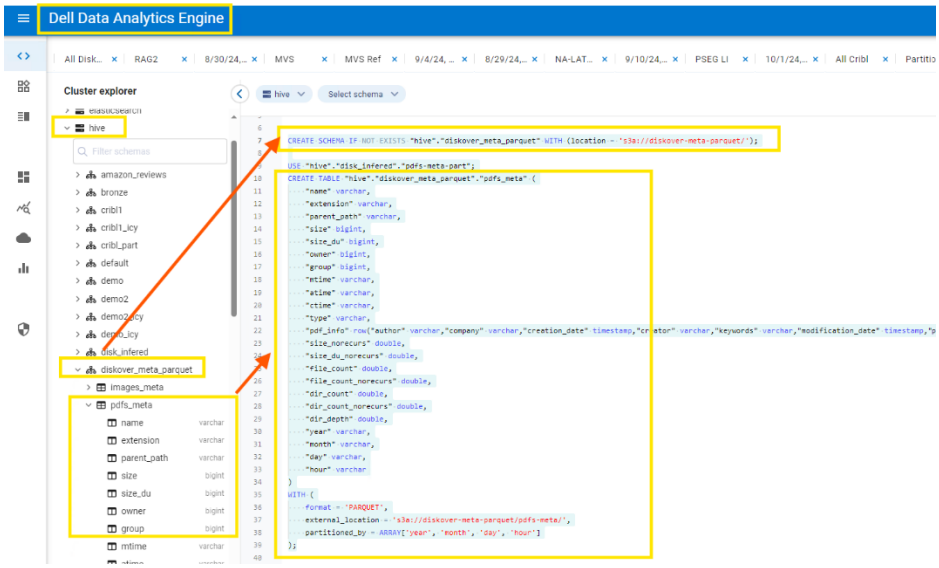
Verify the proper partitioning is retained for the PDF metadata repository.



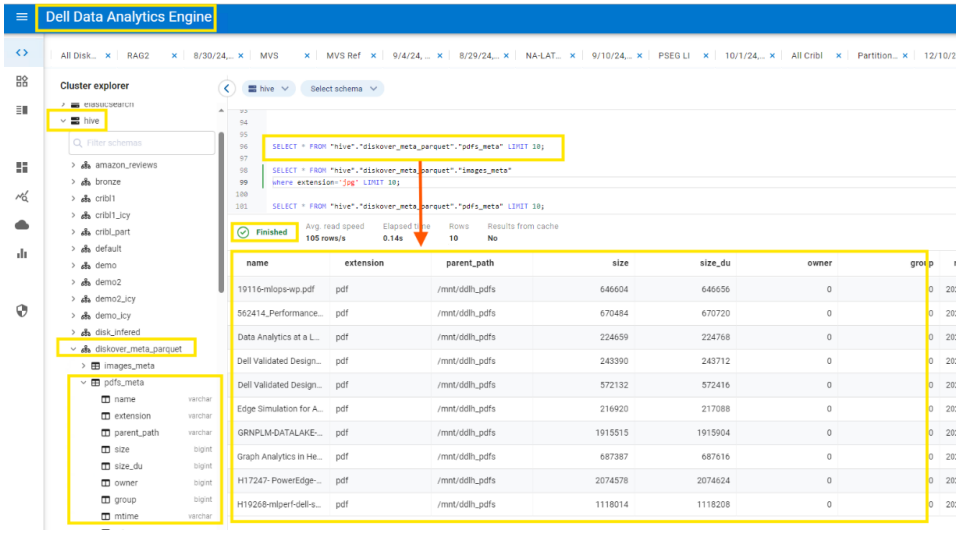
On DDLH, run the schema extraction script on the S3 bucket where Diskover saves PDF metadata in Parquet file format.



Register PDF metadata inventory file's schema as structured data in the DDLH catalog.



Run sample SQL queries on the PDF metadata table.



Conclusion

The integration of DDLH and Diskover offers a robust approach to optimizing metadata management and addressing the challenges of unstructured data. By validating two distinct solutions, this architecture proves its ability to enhance data usability and drive efficiency for AI and GenAI workflows.

Solution 1 positions Diskover as a federated metadata source, enabling DDLH to pull metadata inventory files from ElasticSearch, perform contextualization, and deliver actionable insights with its advanced analytics.

Solution 2 utilizes DDLH as the destination, where Diskover exports partitioned Parquet files and registers schemas for seamless ingestion and processing.

Both solutions highlight the synergy between Diskover's powerful metadata inventory capabilities and DDLH's scalable high-performance analytics. Diskover effectively captures, organizes, and structures metadata, creating a foundation for transforming raw data into meaningful, contextualized datasets. Meanwhile, DDLH supports large-scale data processing, essential for building AI pipelines and supporting demanding workloads. This collaboration enables faster model training, improved deployment outcomes, and reduced redundancies across workflows.

The benefits of this integration extend beyond efficiency. Scalable processing ensures the ability to handle vast data volumes as businesses grow, while automation reduces manual intervention, fostering productivity and innovation. By bridging the divide between data complexity and actionable insights, the combination of DDLH and Diskover unlocks unprecedented opportunities for AI and GenAI applications, positioning organizations to drive smarter operations and achieve technological advancements in an increasingly data-centric world.

References

Dell Technologies documentation

The following Dell Technologies documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

Additional information can be found on the Dell Technologies Info Hub for Data Analytics. If you need additional services or implementation help, contact your Dell Technologies sales representative.

Document type	Location
Dell Data Lakehouse	Dell Data Lakehouse Technical Solution Guide
	Dell Data Lakehouse Specification Sheet
	Dell Data Lakehouse Sizing and Configuration Guide
	Dell Data Lakehouse Solution Brief
Resilient Data Pipelines	Resilient Data pipelines on Dell Data Lakehouse
Data Governance solution	Privacera Platform with Dell Data Lakehouse
RAG Chatbot on DDLH	Multimodal RAG Chatbot Powered by Dell Data Lakehouse
Mainframe Data Analytics	Unlock mainframe data using Dell Data Lakehouse
Cribl Streaming with DDLG	Dell Data Lakehouse with Cribl Stream for Scalable Real-Time Data Processing and Analytics

Diskover documentation

The following Diskover documentation provides additional and relevant information.

Document type	Location
Diskover	Diskover documentation

Dell Technologies Info Hub

The Dell Technologies Info Hub is your one-stop destination for the latest information about Dell Solutions products. New material is frequently added, so browse often to keep up to date on the expanding Dell portfolio of cutting-edge products and solutions.

More information

For more information, including sizing guidance, technical questions, or sales assistance, email Analytics.Assist@dell.com, or contact your Dell Technologies or authorized partner sales representative.