May 2021

EDC 10.5 Discovery

EDC Product Management

Presenters: Siddharth, Krishna

Informatica®

Housekeeping Tips

- Today’s Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal

Bootstrap trial and POC Customers

Enriched Customer Onboarding experience

Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

Presentation , Demo & Q&A

❑ Platform Enhancements

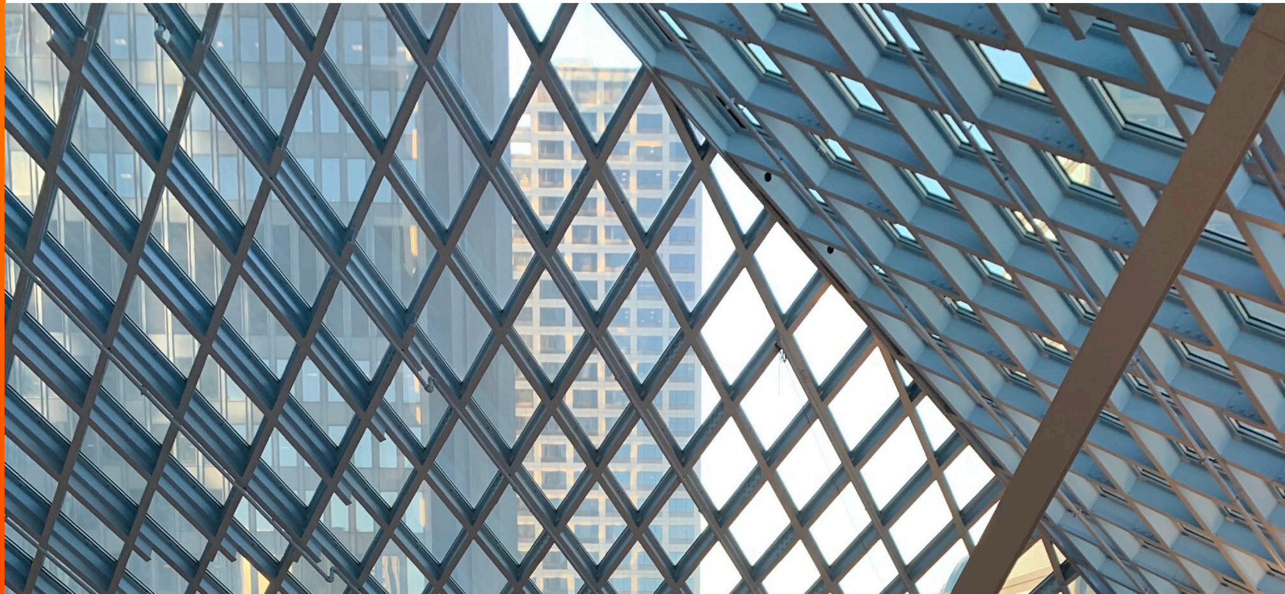
- Release Themes
- Architecture
- Enterprise Readiness

❑ Discovery Enhancements

- Column Similarity
- Profiling on Databricks Cluster

❑ Demo

- Column Similarity in EDC 10.5



Platform Enhancements

10.5 Release Themes

Enterprise Readiness

- **Security enhancements**
- **Non-Hadoop Platform**
- Profiling on Databricks
- Performance Enhancements
  - Relationship API
  - Lineage diagram rendering
  - Selective backup

User Experience

- Search result page enhancements
- BG Term bulk curation
- Data Flow Analytics (Preview)

CLAIRE

- Similarity and Discovery enhancements

Scanners/Connectivity

Product integration of Advanced Scanners

- Code: Oracle, SQL Server, Teradata, IBM DB2, Netezza, Sybase
- BI: SAS, Microsoft SSAS & SSRS,
- Legacy: Cobol, JCL
- ETL: Oracle Data Integrator, Talend DI, IBM DataStage, Microsoft SSIS

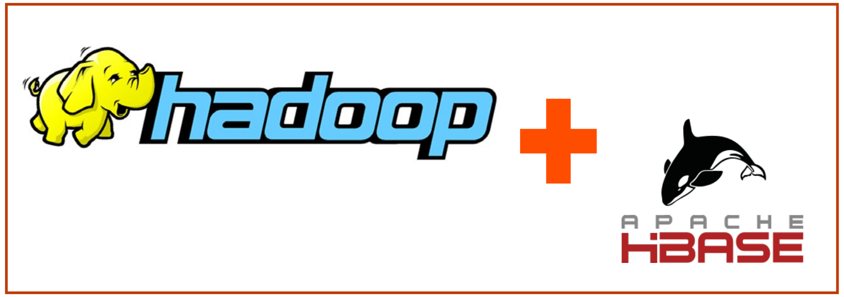
Standard Scanners

- S3 compatible filesystem (Scality)
- Enhanced Snowflake scanner
- SAP S/4 HANA (GA)

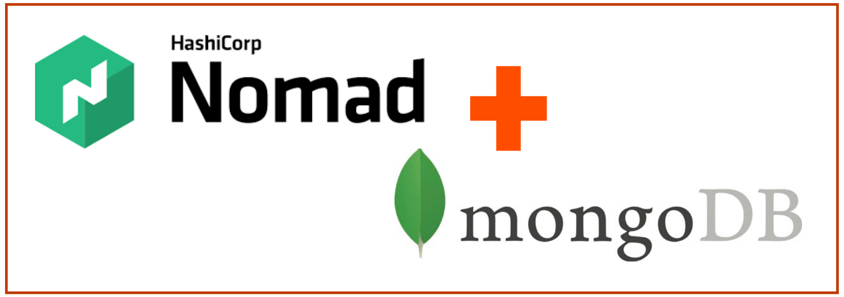
EDC Architecture Change in 10.5

Deprecated support for internal and external Hadoop clusters

Replacing



With



Motivations

- HDP going EOL end of 2021
  - Customers can continue to use old stack on 10.4.1 till March'22
- New architecture to keep-up with market trend

What to expect?

- No additional hardware or software prerequisites
- No functional loss, similar or improved performance and scale
- Seamless upgrade and content migration
- Continued and improved full support on EDC and all deployed services
  - Better support EDC customers on the longer term
  - Faster turn around for OS support, security patches

NOTE: EDC 10.4.x and earlier versions are going out of support by 31 March 2022

EDC 10.5 New Tech Stack

- For Storage:
  - MongoDB, MongoDB GridFS
  - PostgreSQL
  - SOLR
- For orchestration and security:
  - Nomad
- For compute:
  - Native Java (Scanners, ingestion service)

Note: Software license for the EDC platform components (MongoDB, PostgreSQL, SOLR, Nomad) are included with EDC

Store/Engine	Before (10.4.1)	After (10.5.0)
Asset Store	HBase	MongoDB
Graph Store	JanusGraph	MongoDB
Index Store	SOLR	SOLR
Event Store (DAA)	Relational (MRS)	Relational (MRS)
Stage Store	HBase	MongoDB
Scan Content Store	HDFS	MongoDB GridFS
Similarity Store	HBase	PostgreSQL
Config Store	Relational (MRS)	Relational (MRS)
Monitoring Store	Relational (MRS)	MongoDB
Compute	Native, Spark (YARN)	Native on Nomad

Enterprise Readiness

Security

- Authentication enable for all services (mTLS).
- Encryption of data in transit with TLS
- Encryption of data at rest with platform encryption mechanism (AES-256 in 10.5)

Highly available (HA)

- No more Single Point Of Failure (SPOF)
- Each component is deployed in HA mode across the node in the cluster

Disaster Recovery (DR) support

- Hot backup support for regular replication to the DR site

Discovery

Motive for changes to Similarity Discovery

- Moving away from Hadoop architecture in overall architectural changes.
- Similarity Discovery was executed across all resources in the catalog causing extreme capacity constraint on other jobs & services.
- Execution of similarity would take longer time and, in many cases it wouldn’t complete.
- Lot of false positives were computed for numeric & date data types.
- Similarity discovery was computed on all features (Name, Signature, Pattern, Value frequency), lot of customer wanted to do it based on few features.

What has changed in **Similarity Discovery?**

Similarity Discovery Execution

- Separate similarity discovery computation & storage from catalog to reduce impact on catalog service
- Dedicated Postgres SQL for similarity discovery store
- Faster execution & better performance.

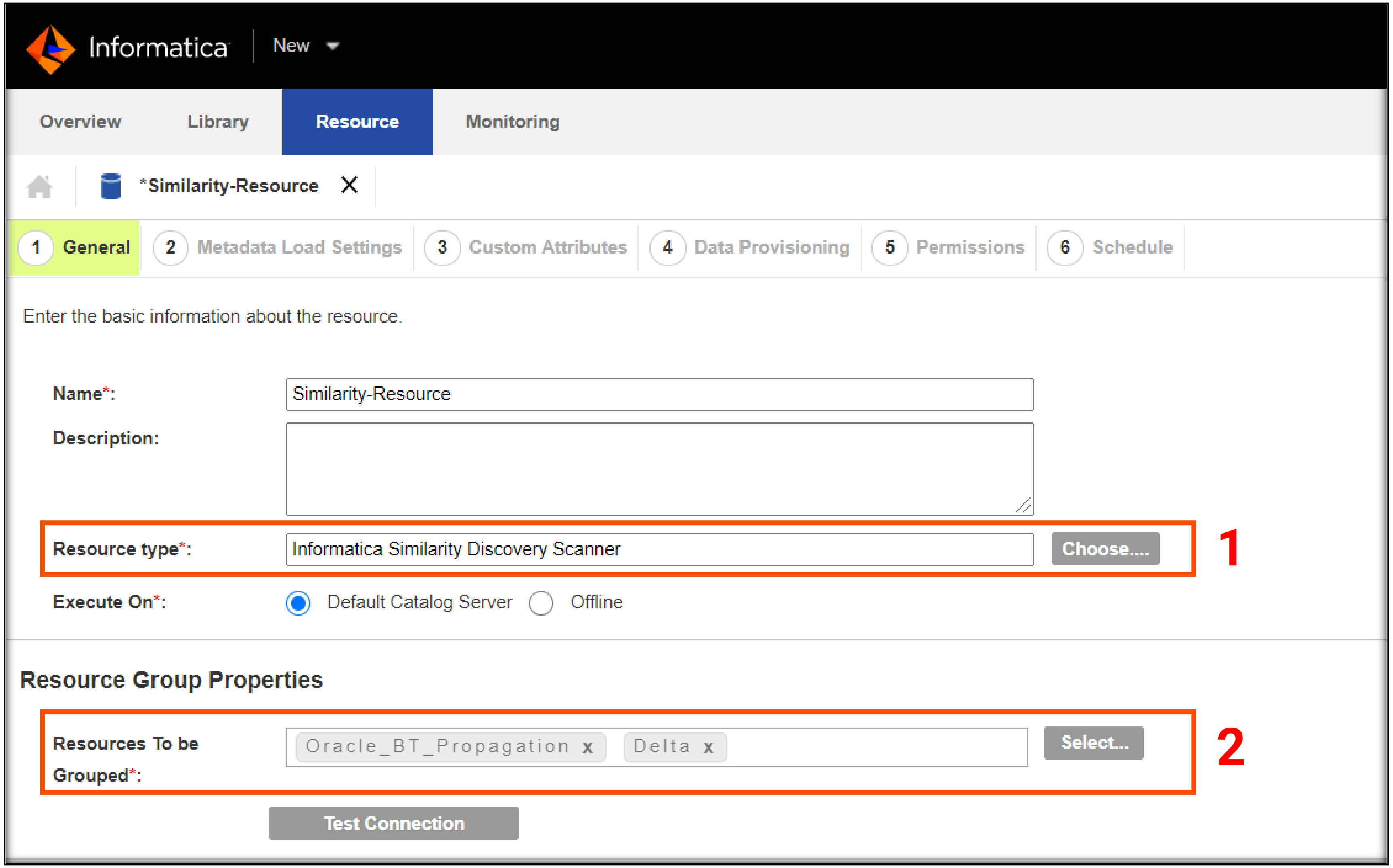
Resource Grouping with new Similarity Discovery Scanner for Similarity Computations

- Allow users to logically group resources for similarity computation
- Improves the performance and accuracy
- Goes through the scanner framework and lifecycle (create, edit, purge and delete)

Allow Similarity Computation based on Features Enabled

- Each resource group can be configured for the features against similarity is run
- Configurable features are column *Name*, *Pattern* distribution, and *Unique Values* distribution

Similarity Discovery Resource



Enter the basic information about the resource.

Name*: Similarity-Resource

Description:

Resource type*: Informatica Similarity Discovery Scanner

Execute On*: ● Default Catalog Server ○ Offline

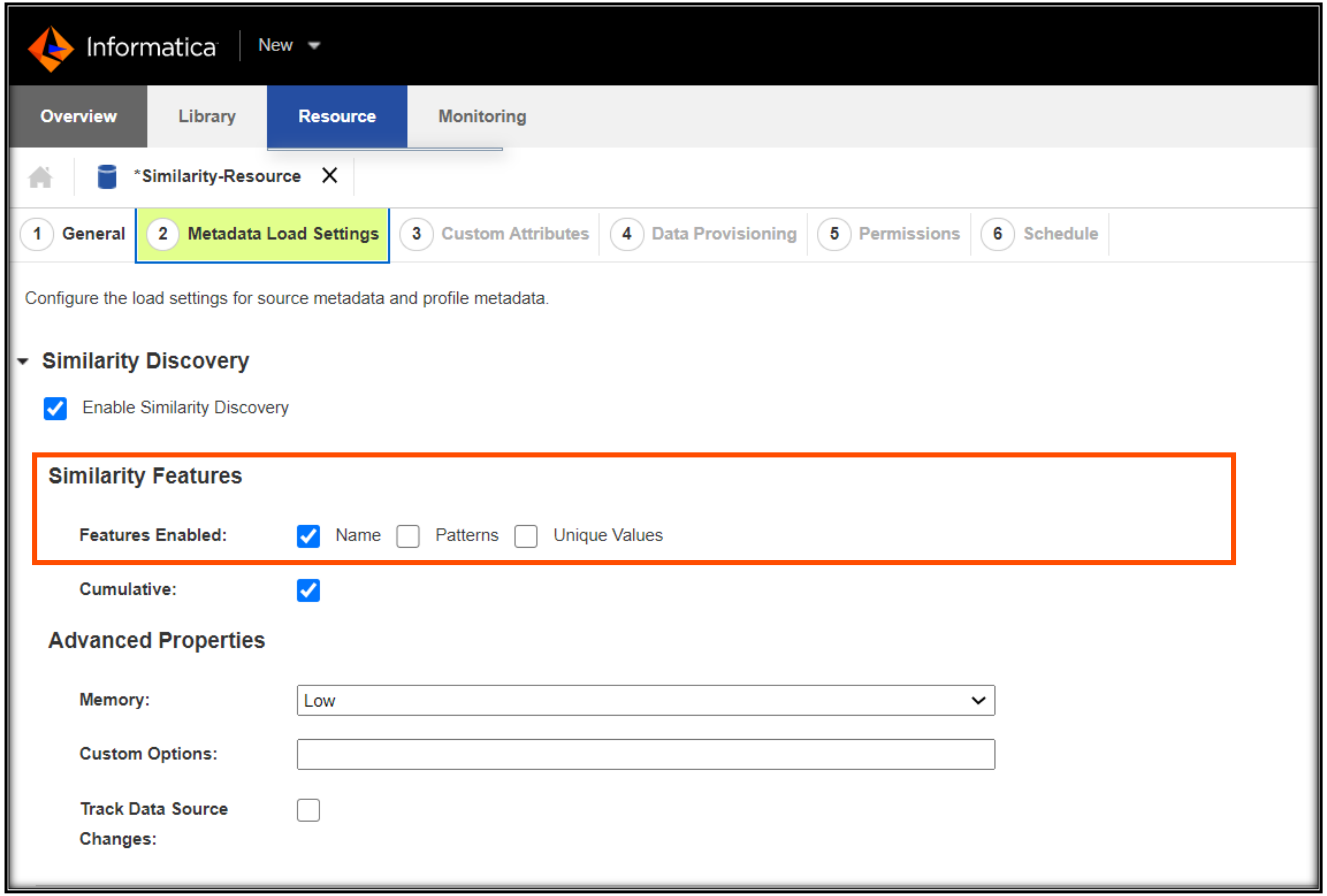
Resource Group Properties

Resources To be Grouped*: Oracle_BT_Propagation x Delta x

1 - Similarity Discovery Scanner Resource Type

2 - Restricting the # of resources for similarity discovery

Similarity Computation based on feature



Configure the load settings for source metadata and profile metadata.

Similarity Discovery

☑ Enable Similarity Discovery

Similarity Features

Features Enabled: ☑ Name □ Patterns □ Unique Values

Cumulative: ☑

Advanced Properties

Memory: Low

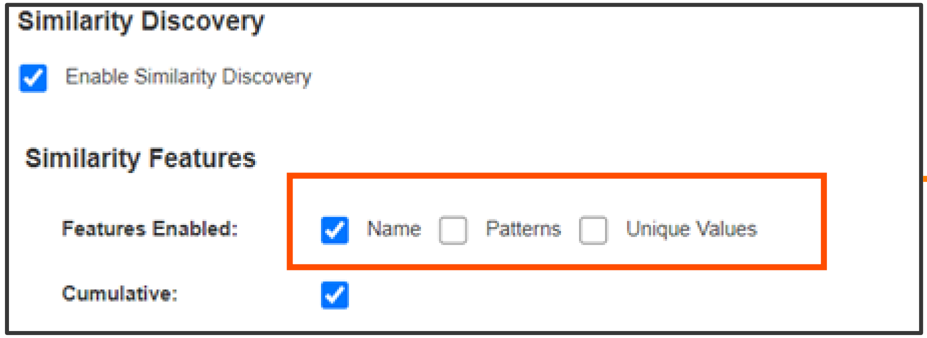
Custom Options:

Track Data Source Changes: □

Selecting similarity feature for computation, it can be any one or combinations.

Cumulative Similarity Feature Computation

Enabling only Name Similarity Feature



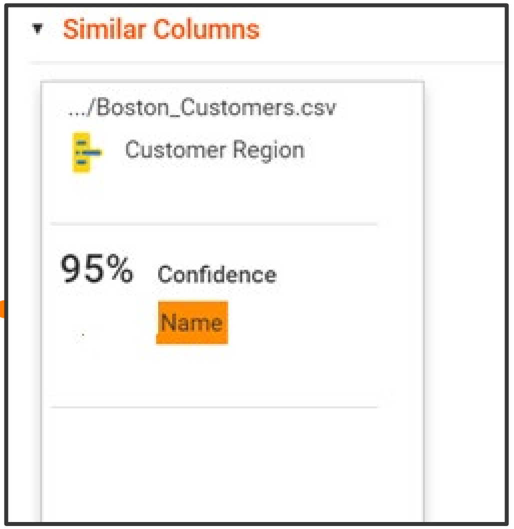
Similarity Discovery

☑ Enable Similarity Discovery

Similarity Features

Features Enabled: ☑ Name □ Patterns □ Unique Values

Cumulative: ☑



▾ Similar Columns

.../Boston_Customers.csv

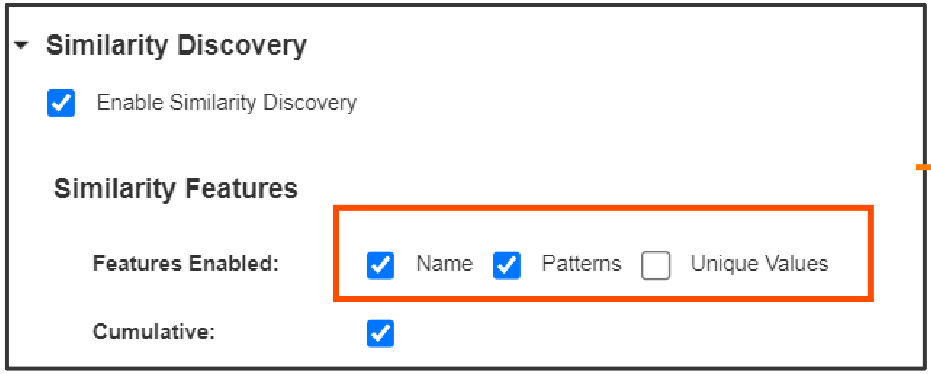
Customer Region

95% Confidence

Name

Similarity Computed on Name feature, Confidence score denotes percentage of conformance based on Name

Enabling only Name & patterns Similarity Feature



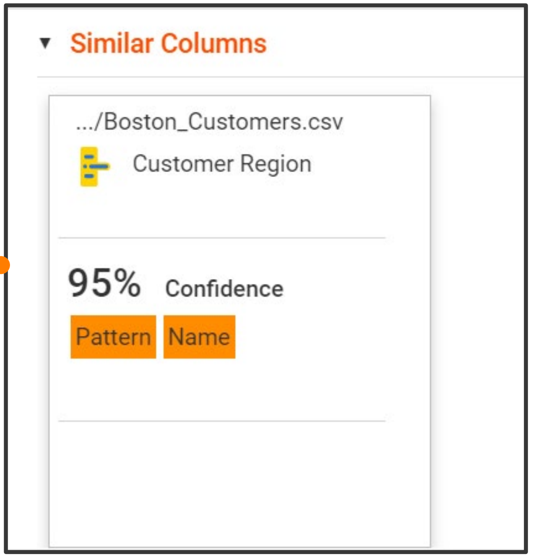
▾ **Similarity Discovery**

☑ Enable Similarity Discovery

Similarity Features

Features Enabled: ☑ Name ☑ Patterns □ Unique Values

Cumulative: ☑



▾ Similar Columns

.../Boston_Customers.csv

Customer Region

95% Confidence

Pattern Name

Similarity Computed on Name & Pattern feature, by enabling cumulative it will add on to existing selected feature.

Informatica

Profiling on Data bricks

Profiling on Databricks

- Databricks is a completely managed and optimized platform for running spark jobs available on Azure & AWS Clouds. It provides Infrastructure management, Security and comes with a whole bunch of tools support.
- Support Databricks Delta table for scanning & profiling – Delta is component of Databricks Unified Analytics Platform that provides a powerful transactional storage layer built on top of apache spark
- Enable profiling of resources on Azure Ecosystem, Support for running profiling of parquet files on ADLS Gen 2 for data discovery, domain discovery.
- Extend profiling of resources on AWS Ecosystem as like Databricks on Azure..

What is supported on Databricks Cluster

Sources

- Databricks Delta tables
- ADLS Gen 2 (Parquet - Partitioned)

Column Profiling

Data Domain Discovery

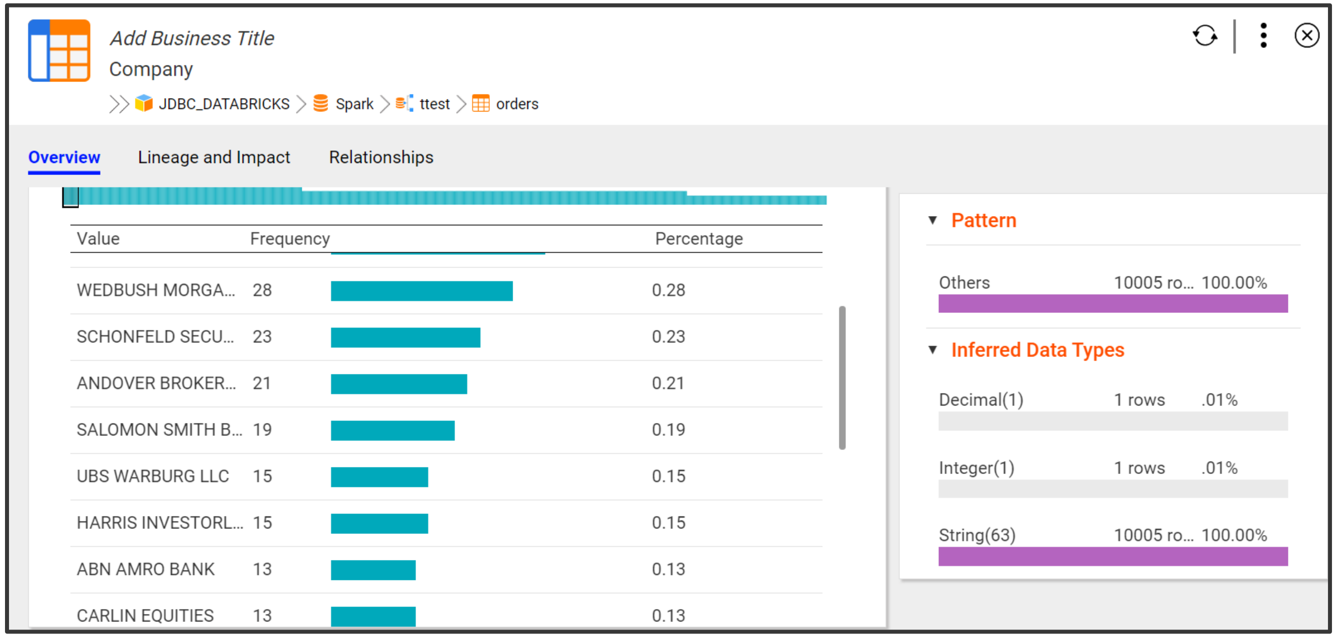
- Both Custom & OOTB are supported

Enterprise Data Discovery

Sampling Support (All rows, random N only)

Cloud

- Azure



Add Business Title

Company

JDBC_DATABRICKS > Spark > ttest > orders

Value	Frequency	Percentage
WEDBUSH MORGA...	28	0.28
SCHONFELD SECU...	23	0.23
ANDOVER BROKER...	21	0.21
SALOMON SMITH B...	19	0.19
UBS WARBURG LLC	15	0.15
HARRIS INVESTORL...	15	0.15
ABN AMRO BANK	13	0.13
CARLIN EQUITIES	13	0.13

▾ Pattern

Others	10005 ro...	100.00%

▾ Inferred Data Types

Decimal(1)	1 rows	.01%
Integer(1)	1 rows	.01%
String(63)	10005 ro...	100.00%

Demo

Similarity Discovery

Profiling on Databricks

Questions?

Thank You