Feb 22nd, 2022

Cloud Data Governance and Catalog – Databricks Notebook Case Study

Srinivasa Gopal - Principal Technologist, Customer Success

Sachin Jain - Principal Technologist, Customer Success

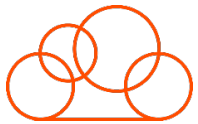
Informatica

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.

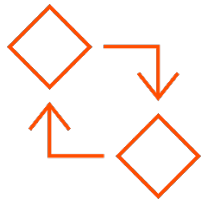
Informatica

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

Informatica

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

Informatica

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.

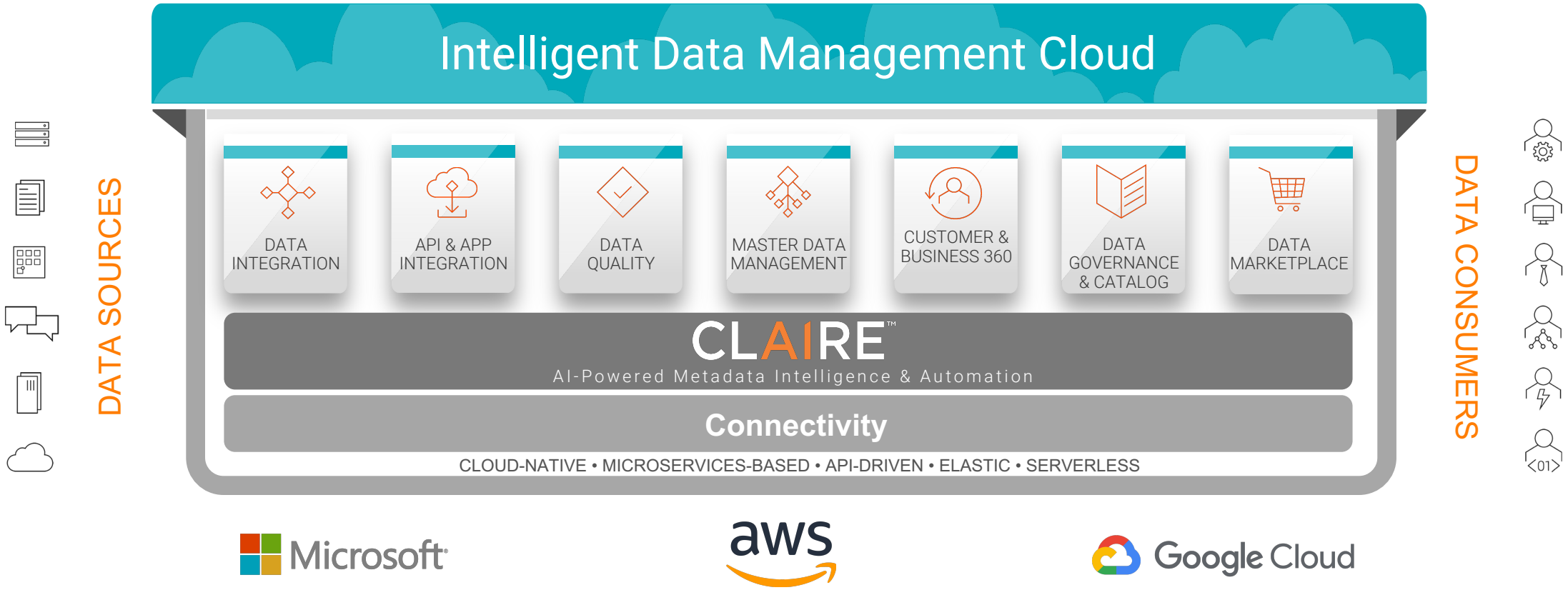
Informatica

Agenda

1. Overview of CDGC
2. Databricks Architecture Patterns
3. Databricks Scanners
4. Databricks Metadata
5. Demo
6. Q&A

Overview of CDGC

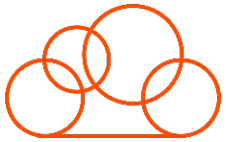
Intelligent Data Management Cloud



Informatica

Intelligent Data Management Cloud

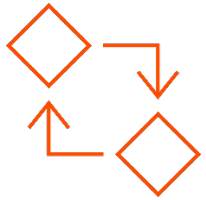
A SaaS solution that drives the success of your Cloud Data Lake and Data Warehouse implementation—provides data context, lowers privacy risks, builds trust and delivers insight



SaaS
Cloud Native

Data **Lineage**, Profiling, Discovery, Source Context

Glossary, **Business Context**, AI Models, DQ & Privacy Standards



Integrated with Cloud Catalogs, Integrated with CDQ

Business User, Navigation **Experience**

Informatica

Cloud Data Governance and Catalog

Core Solution Tenets

SERVERLESS

Serverless compute to work at any scale—there is no infrastructure to deploy or manage

MULTI-CLOUD

Works across all cloud ecosystems and on-premises data sources

INTEGRATED

Deep Integration with data management and data consumption tools

AI

Uses state of the art machine learning and intelligence to ensure end user productivity

OPEN & EXTENSIBLE

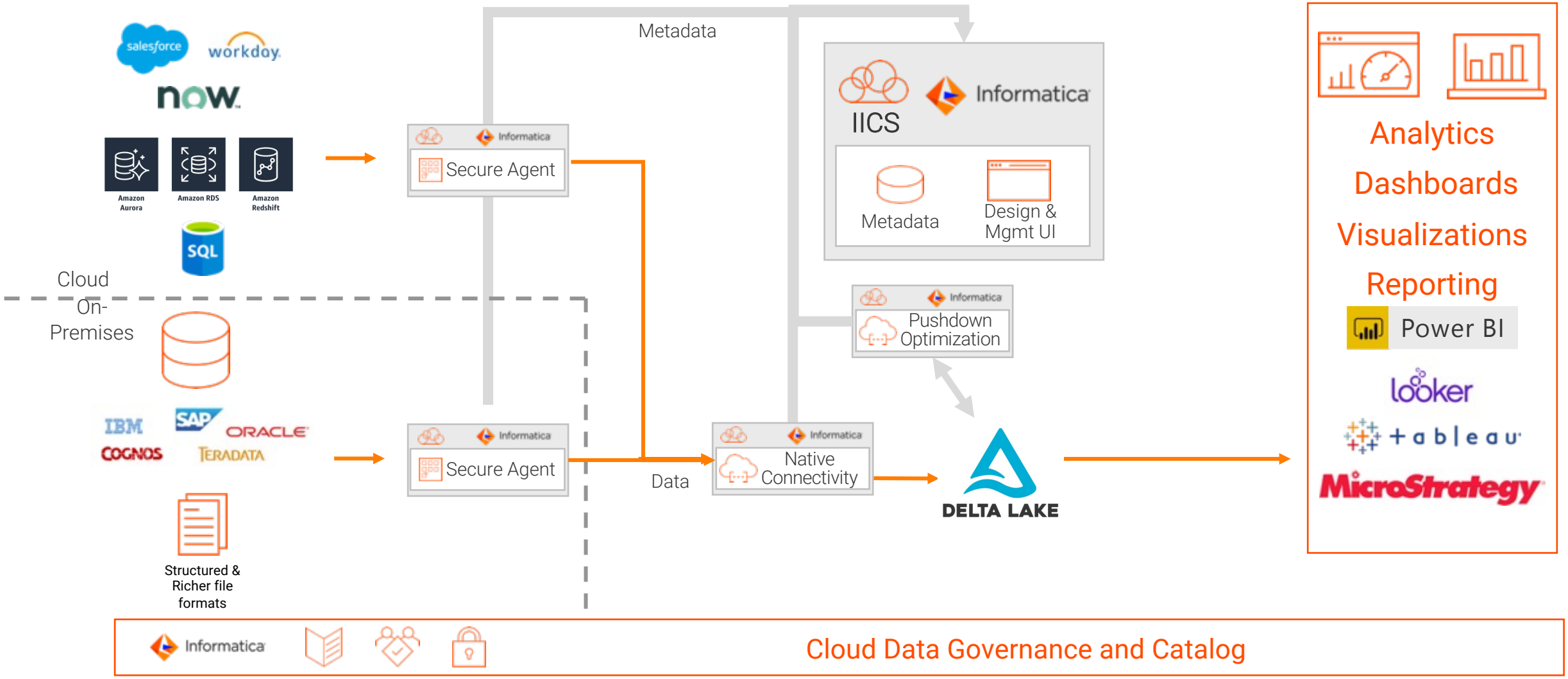
API-based access to metadata repository and functions

Informatica

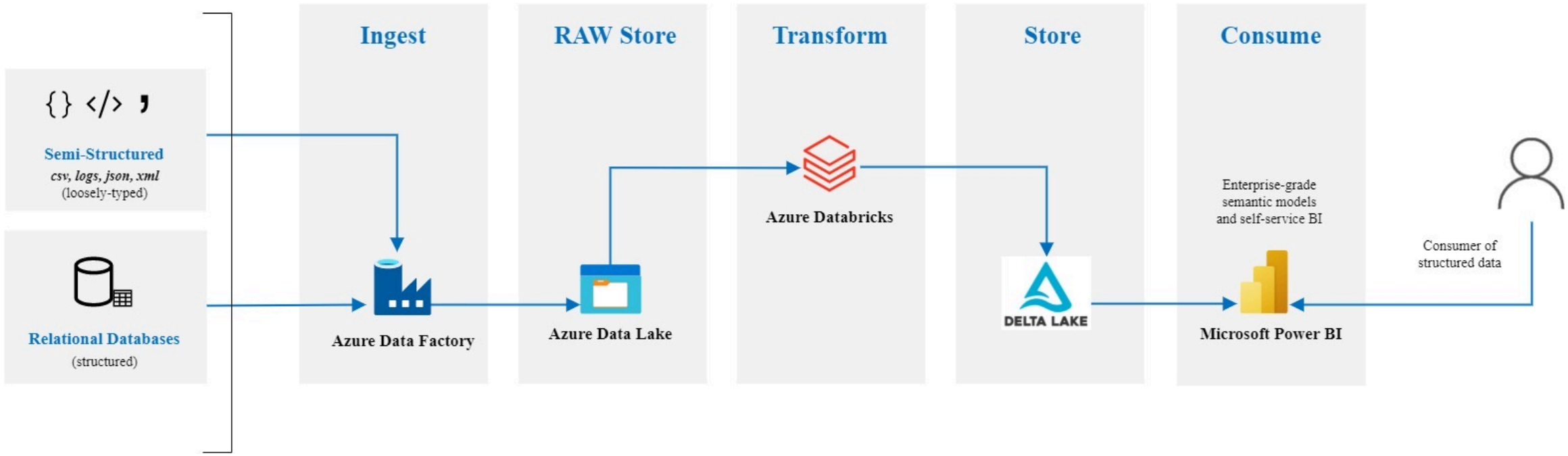
Databricks Architecture Patterns

Informatica

Databricks Delta Lake with IICS

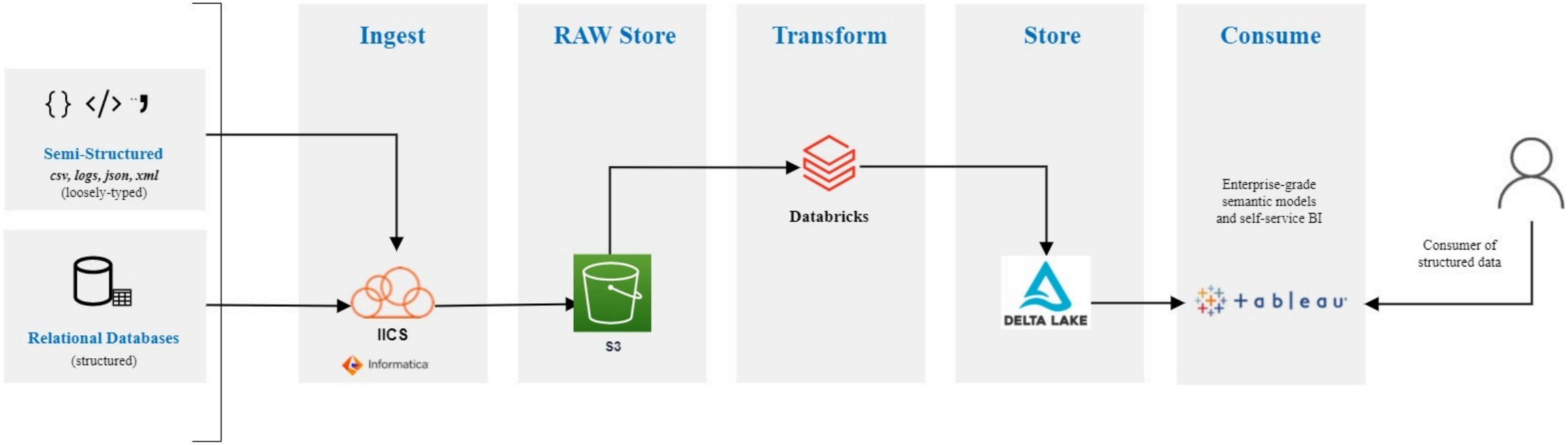


Azure Reference Architecture – Databricks



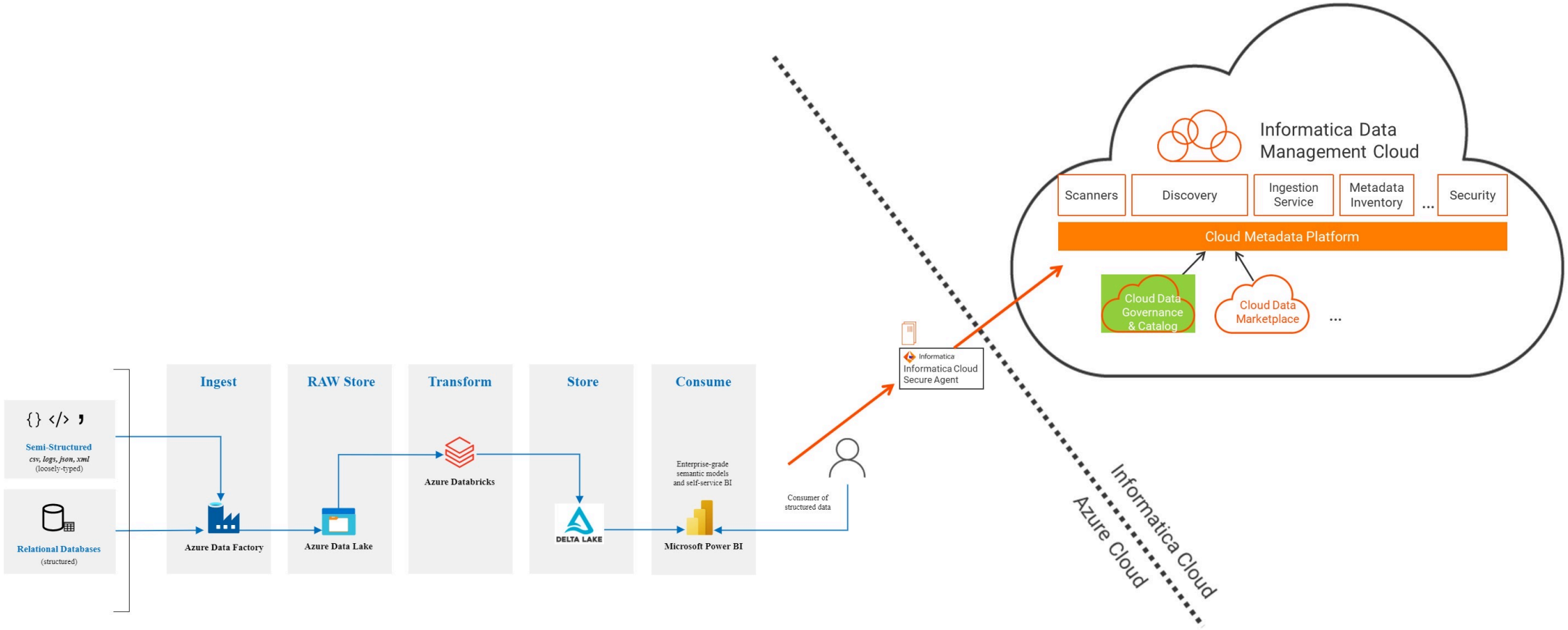
Informatica

AWS Reference Architecture – Databricks



Informatica

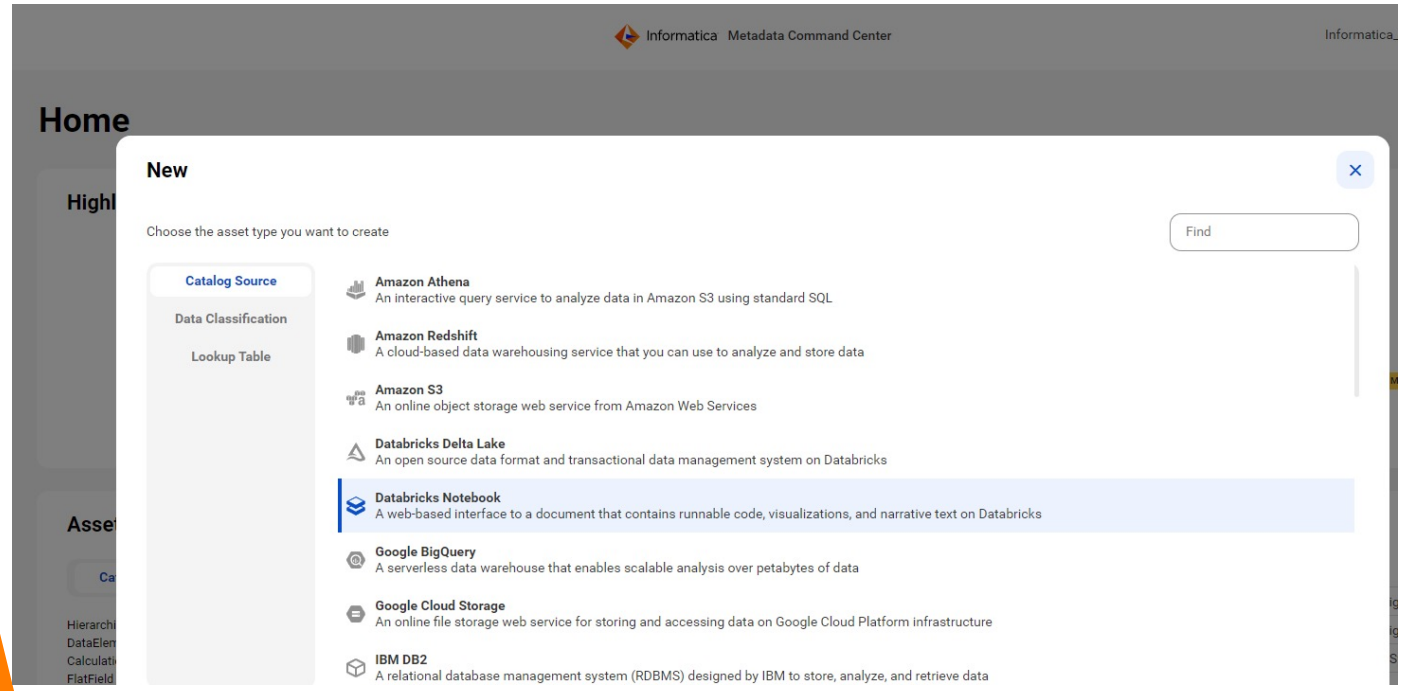
Databricks Delta Lake with CDGC



Informatica

Databricks Scanners

Databricks Notebook Scanner

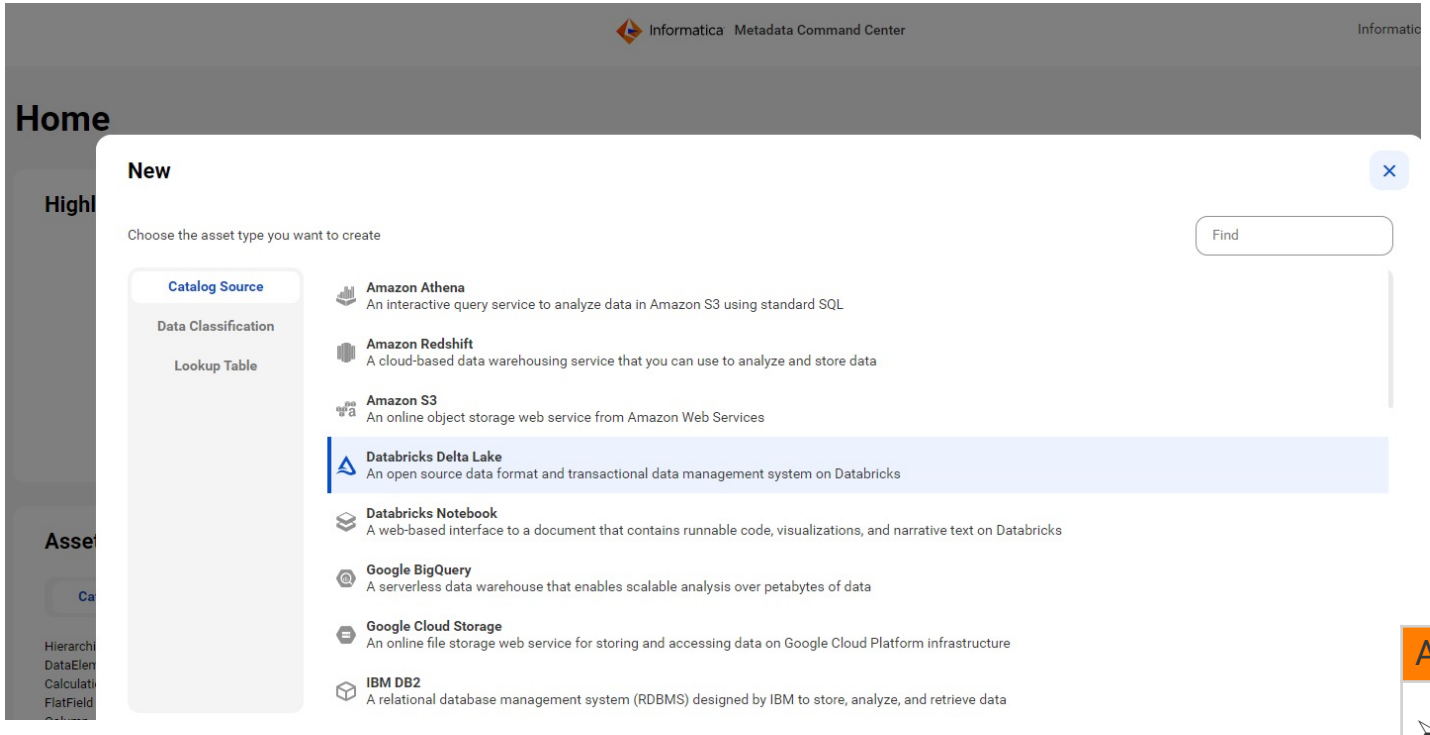


Databricks Notebook Scanner Config*

Asset	Attributes
Notebook	Notebook name object_type path language object_id
Folder	Folder name object_type path language object_id
Command	Command name Command Content

* Currently supports extract metadata from Notebook command types - Python and SQL. Not - Scala, R

Databricks Delta Lake Scanner



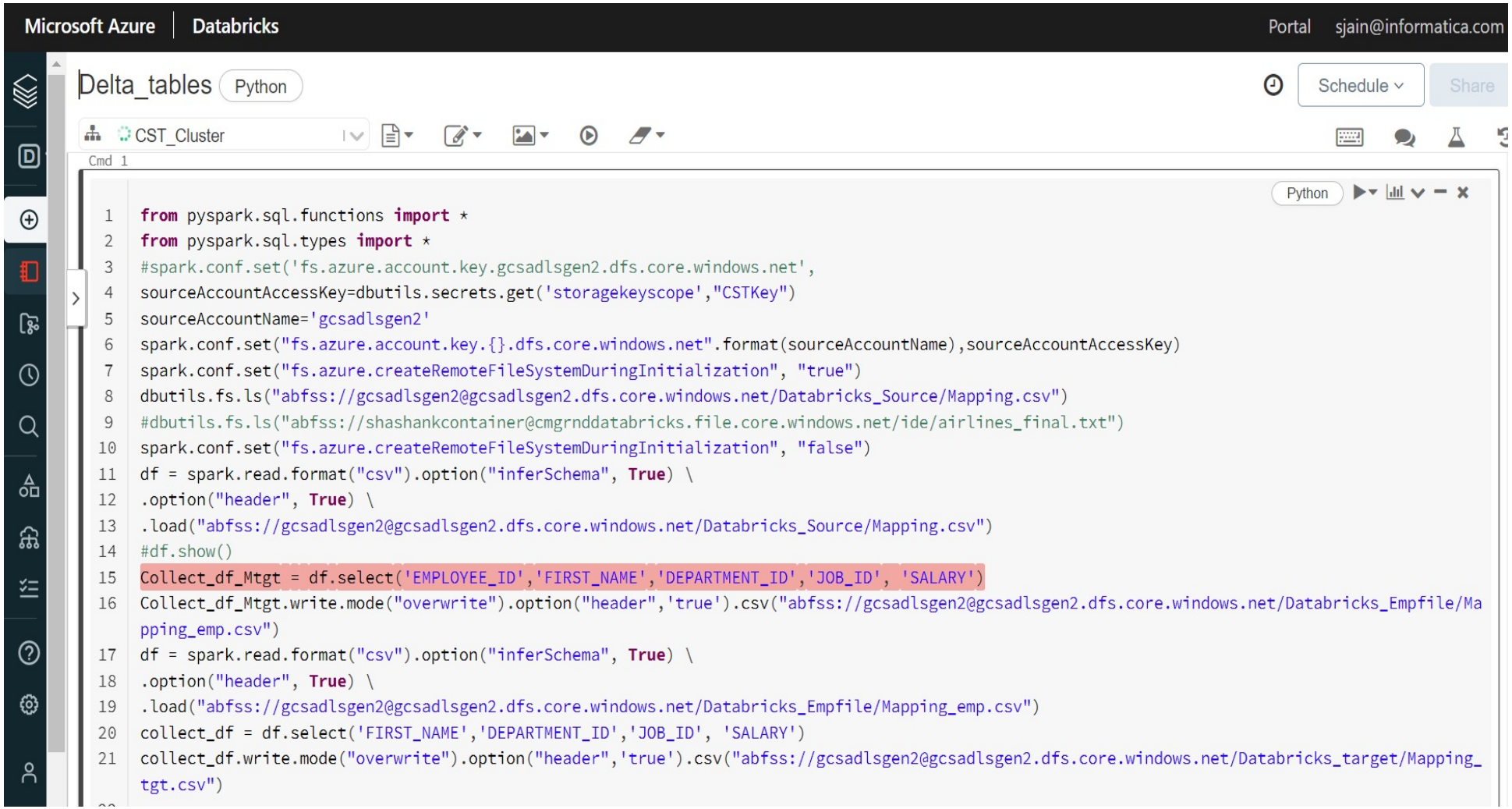
[Databricks Delta Table Scanner Config]

Asset	Attributes
➢ Database ➢ Schema ➢ Table ➢ Column ➢ View ➢ View Column	Name Type Description Path

Informatica

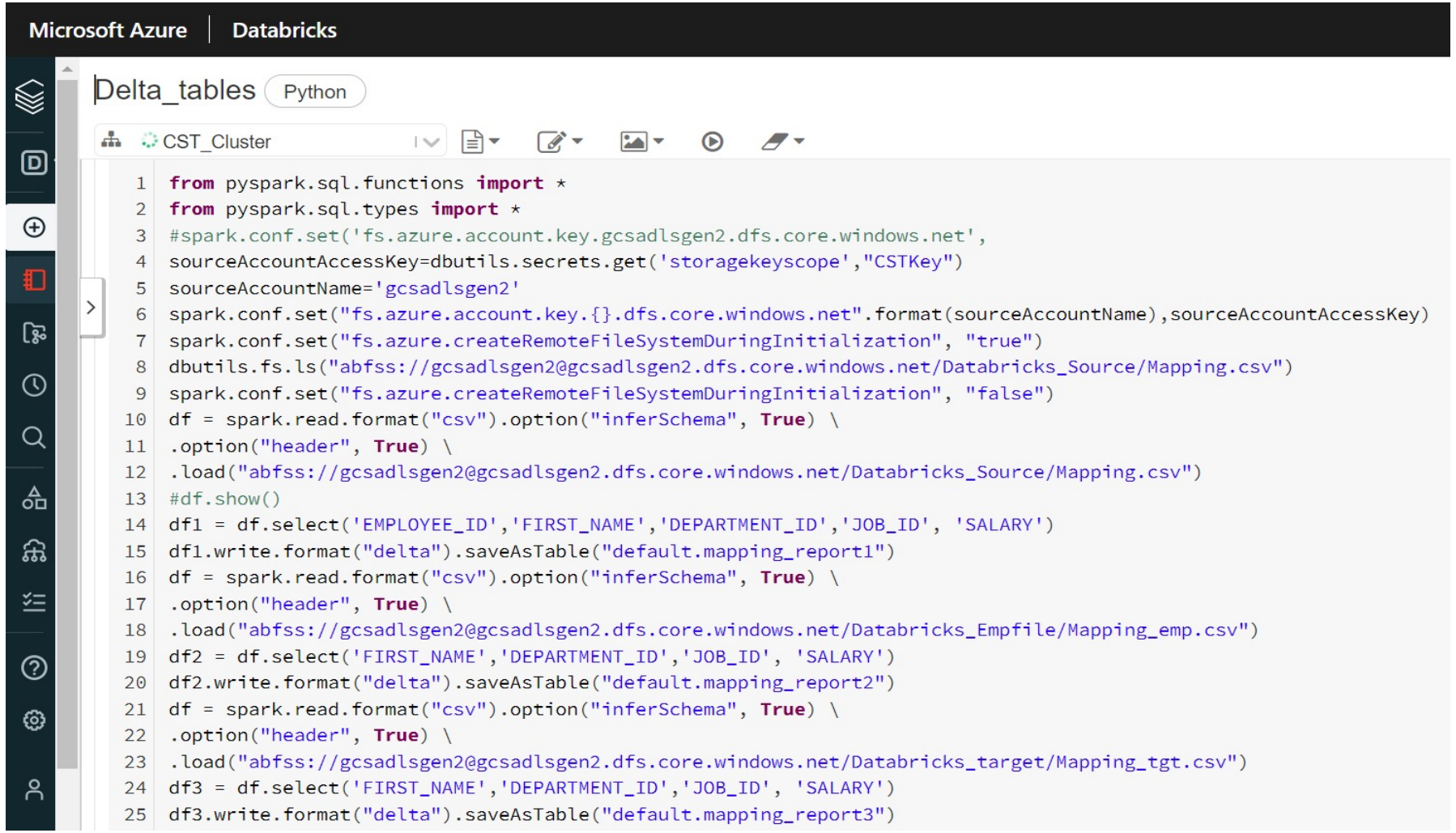
Databricks Metadata

Databricks Notebook



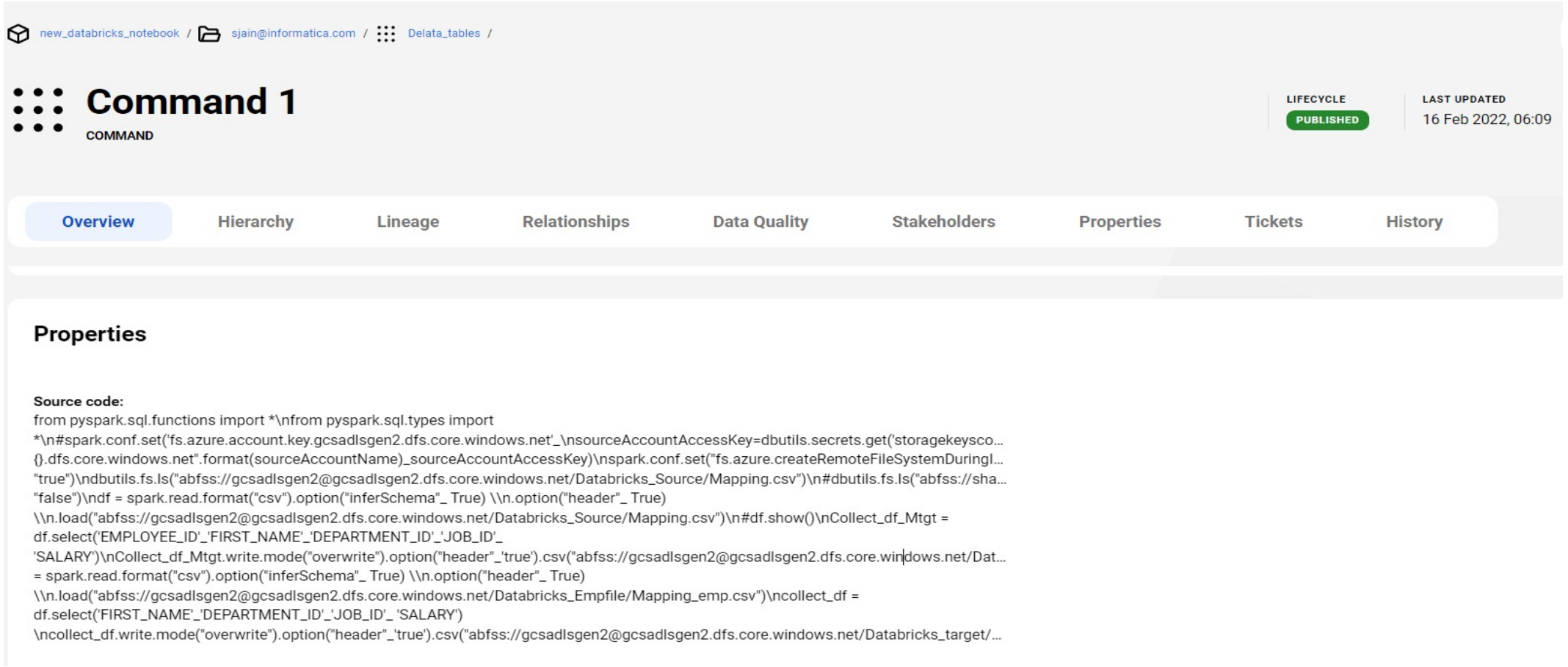
Informatica

Databricks Notebook



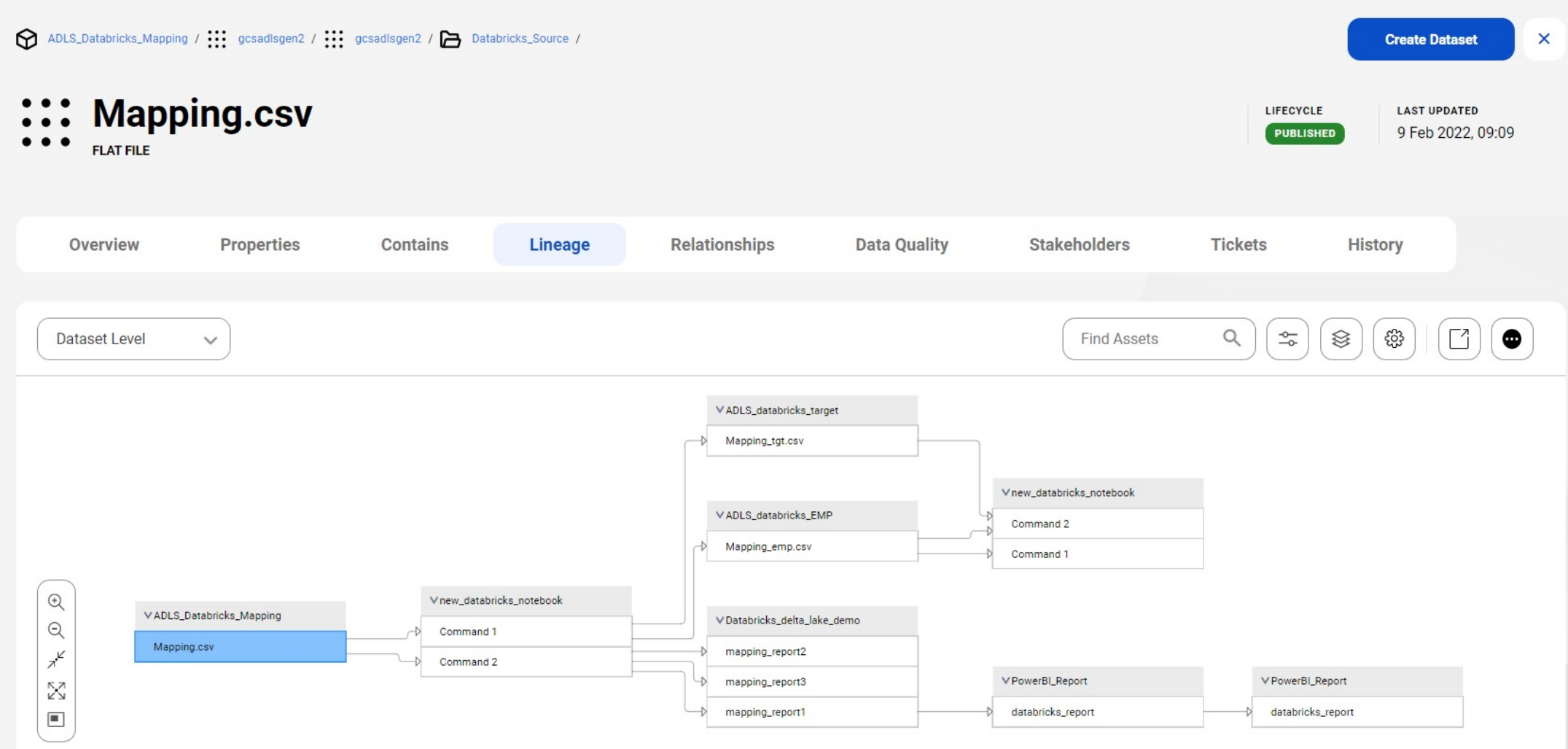

```python
from pyspark.sql.functions import *
from pyspark.sql.types import *
#spark.conf.set('fs.azure.account.key.gcsadlsgen2.dfs.core.windows.net',
sourceAccountAccessKey=dbutils.secrets.get('storagekeyscope',"CSTKey")
sourceAccountName='gcsadlsgen2'
spark.conf.set("fs.azure.account.key.{}.dfs.core.windows.net".format(sourceAccountName),sourceAccountAccessKey)
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization", "true")
dbutils.fs.ls("abfss://gcsadlsgen2@gcsadlsgen2.dfs.core.windows.net/Databricks_Source/Mapping.csv")
spark.conf.set("fs.azure.createRemoteFileSystemDuringInitialization", "false")
df = spark.read.format("csv").option("inferSchema", True) \
.option("header", True) \
.load("abfss://gcsadlsgen2@gcsadlsgen2.dfs.core.windows.net/Databricks_Source/Mapping.csv")
#df.show()
df1 = df.select('EMPLOYEE_ID','FIRST_NAME','DEPARTMENT_ID','JOB_ID', 'SALARY')
df1.write.format("delta").saveAsTable("default.mapping_report1")
df = spark.read.format("csv").option("inferSchema", True) \
.option("header", True) \
.load("abfss://gcsadlsgen2@gcsadlsgen2.dfs.core.windows.net/Databricks_Empfile/Mapping_emp.csv")
df2 = df.select('FIRST_NAME','DEPARTMENT_ID','JOB_ID', 'SALARY')
df2.write.format("delta").saveAsTable("default.mapping_report2")
df = spark.read.format("csv").option("inferSchema", True) \
.option("header", True) \
.load("abfss://gcsadlsgen2@gcsadlsgen2.dfs.core.windows.net/Databricks_target/Mapping_tgt.csv")
df3 = df.select('FIRST_NAME','DEPARTMENT_ID','JOB_ID', 'SALARY')
df3.write.format("delta").saveAsTable("default.mapping_report3")
```

Databricks Notebook – Overview



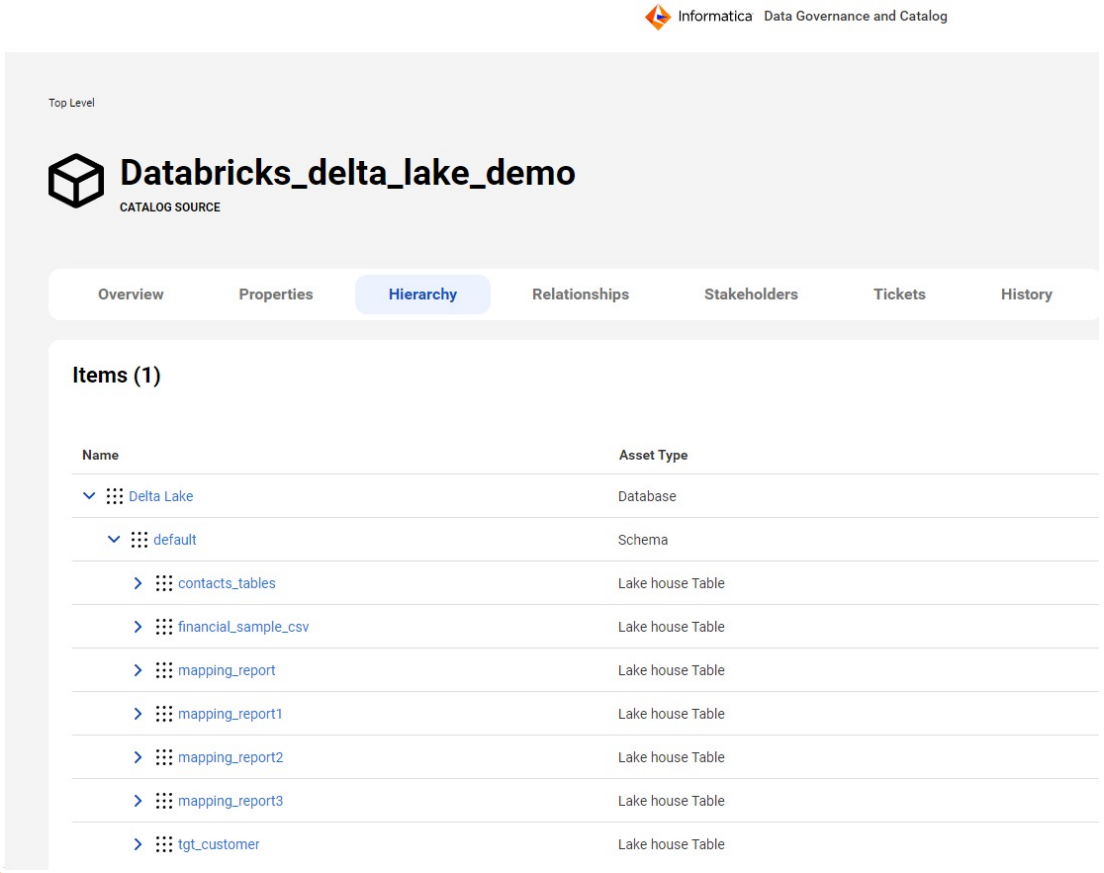
Informatica

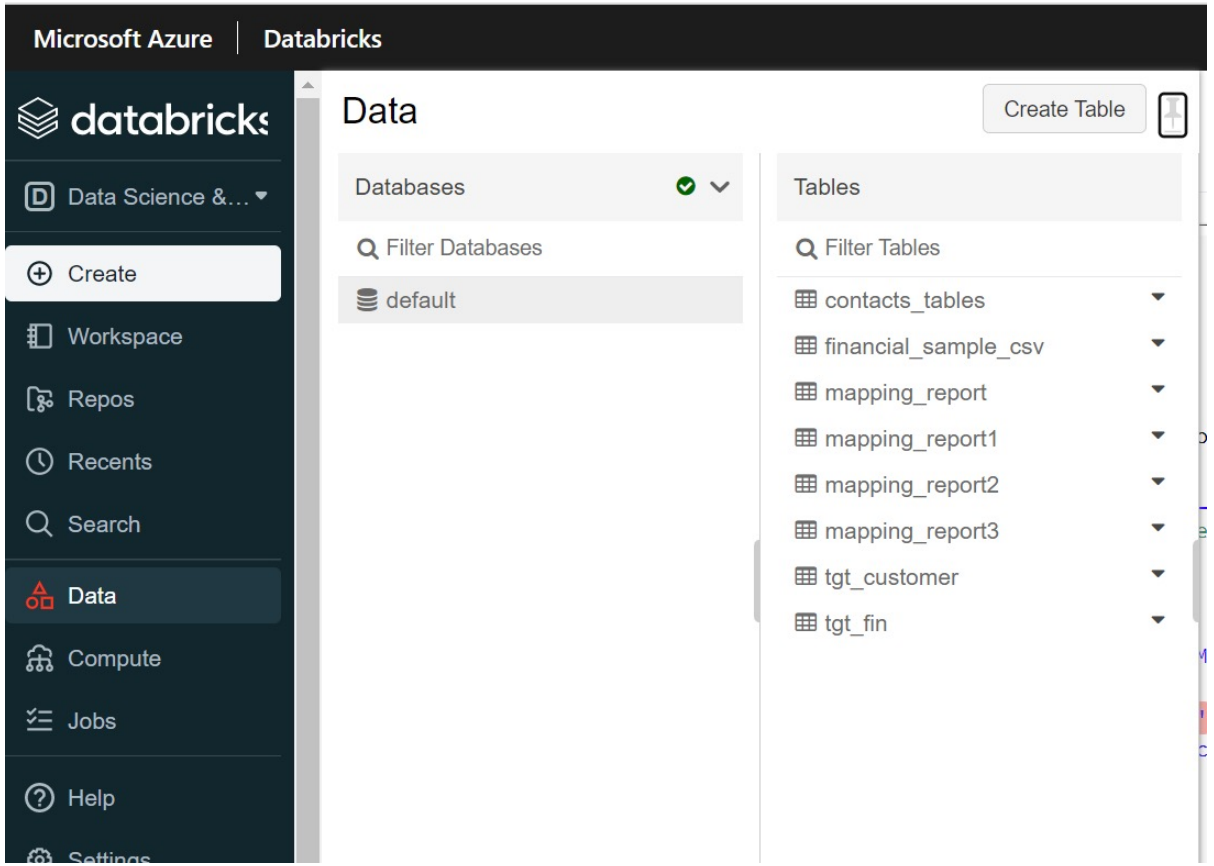
Databricks Notebook scan



Informatica

Delta Tables Scan





Informatica

DEMO

CDGC – Databricks CASE STUDY

Informatica

References

- Cloud Data Governance and Catalog: Click Here
- Introduction and Getting Started: Click Here
- Governance: Bulk Upload Business Metadata
- Metadata Command Center:
  - Catalog Scanner Configuration
  - Creating Catalog Source
  - Configure Runtime Environments