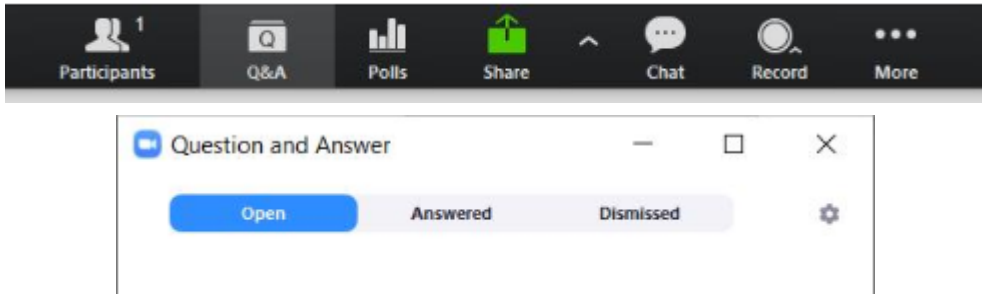
March 10, 2020

# An Introduction to Databricks and Informatica Data Engineering Integration

Stijn Carion, Associate Staff Engineer, Informatica GCS

# Housekeeping Tips

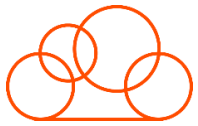



- Todays Webinar is scheduled to last 1 hour including Q&A
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available to view on our INFASupport YouTube channel and Success Portal. The link will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

# Success Portal

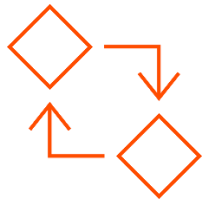
# https://success.informatica.com

## Learn. Adopt. Succeed.



Bootstrap product trial experience

Enriched Onboarding experience



FREE Product Learning Paths and weekly Expert sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations



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.

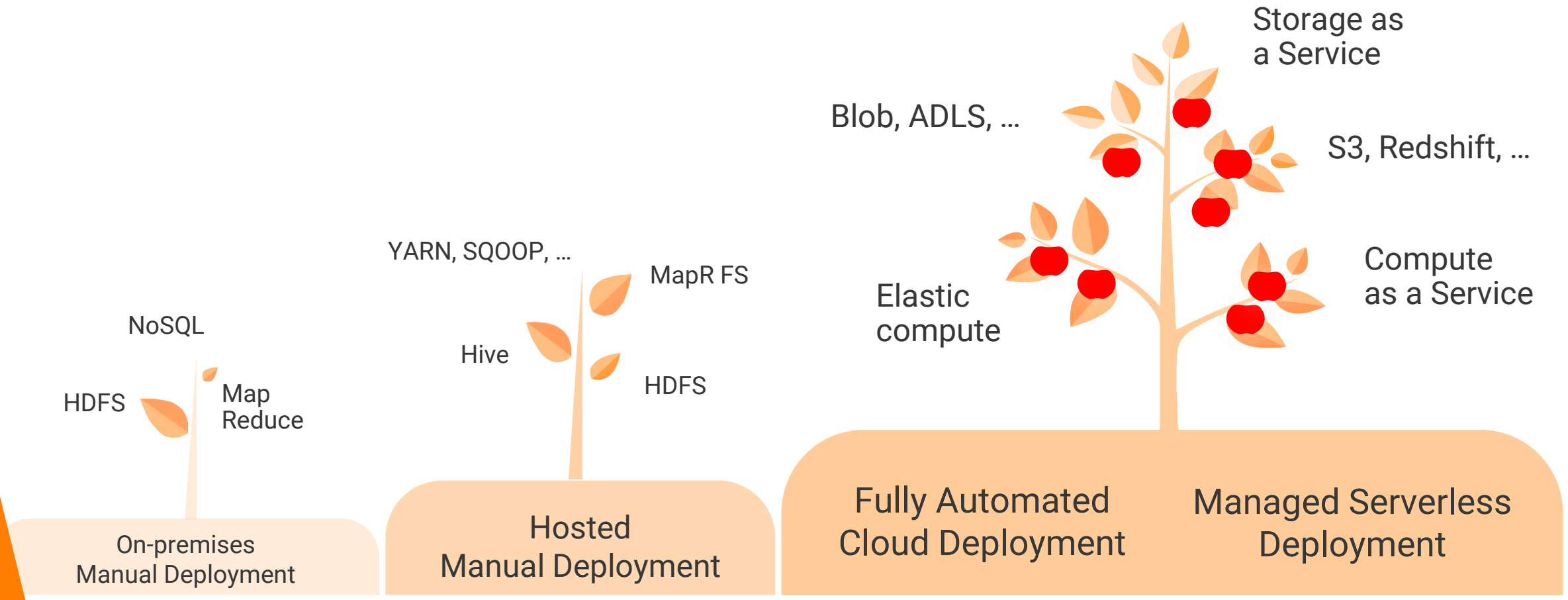
Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

# Agenda

- Databricks Introduction
- Databricks Delta Lake Introduction
- Informatica Data Engineering Configuration for Databricks
- Databricks Delta Lake Source/Targets in Informatica Data Engineering
- Demo
- Troubleshooting and self-service
- Q&A

# Databricks Introduction

# Ever-Evolving Big Data Technology





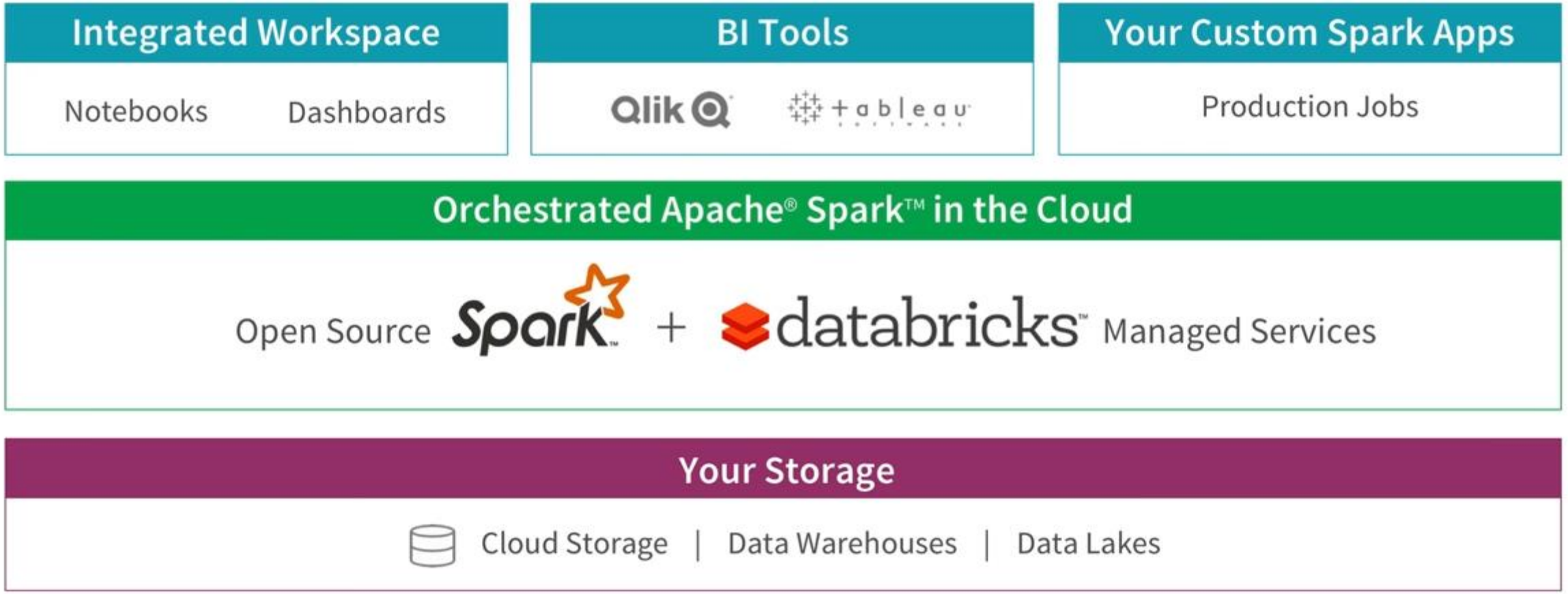


Informatica

# Databricks Introduction

- Company founded by the Creators of Apache Spark (late 2013)
- Cloud/Web based platform for working with Spark, providing automated cluster management
- Available on Azure and AWS as a service (screenshots/demo Azure-based)
- Start Spark Cluster in few clicks/minutes, allowing scaling on demand

# Databricks Introduction

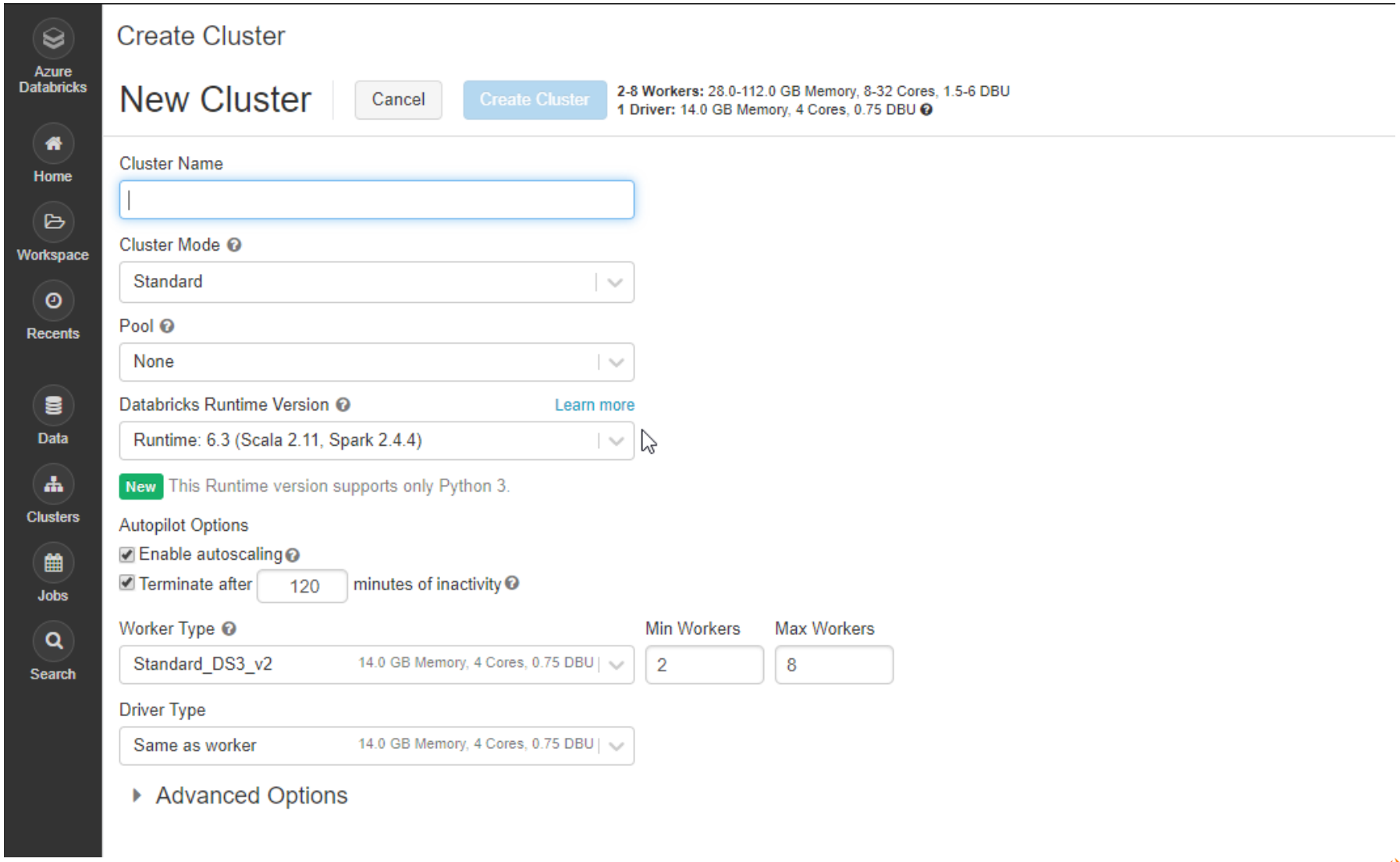






Informatica

# Databricks Introduction







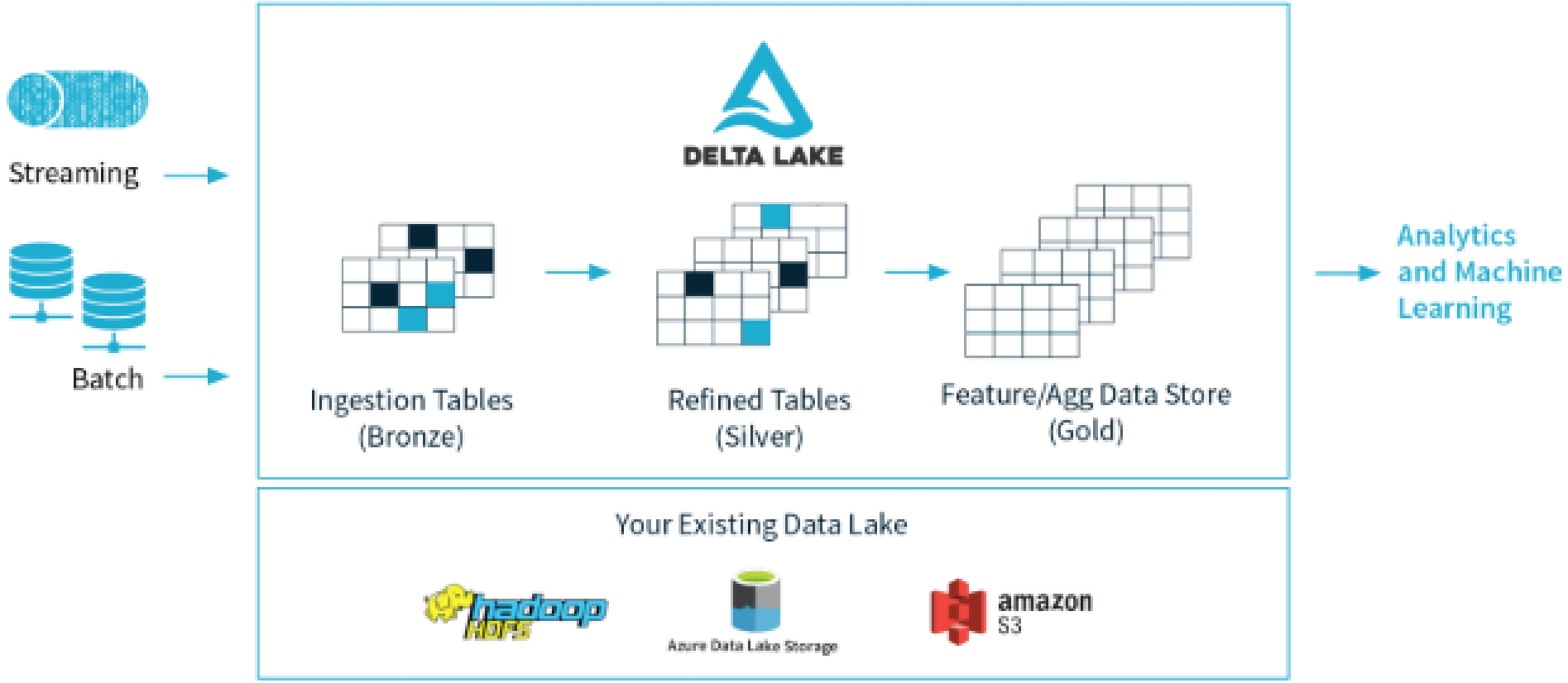
Informatica

# Databricks Delta Lake Introduction

# Databricks Delta Lake Introduction

- Sits on top of your existing Data Lake (HDFS/Cloud Storage)
- Delta Lake brings ACID transactions to your data lakes
  - Achieved by using a transaction log of all the commits made to the table
- Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.
- All data in Delta lake is store in Apache Parquet format

# Databricks Delta Lake Introduction







Informatica

# Informatica Data Engineering Configuration for Databricks

Informatica

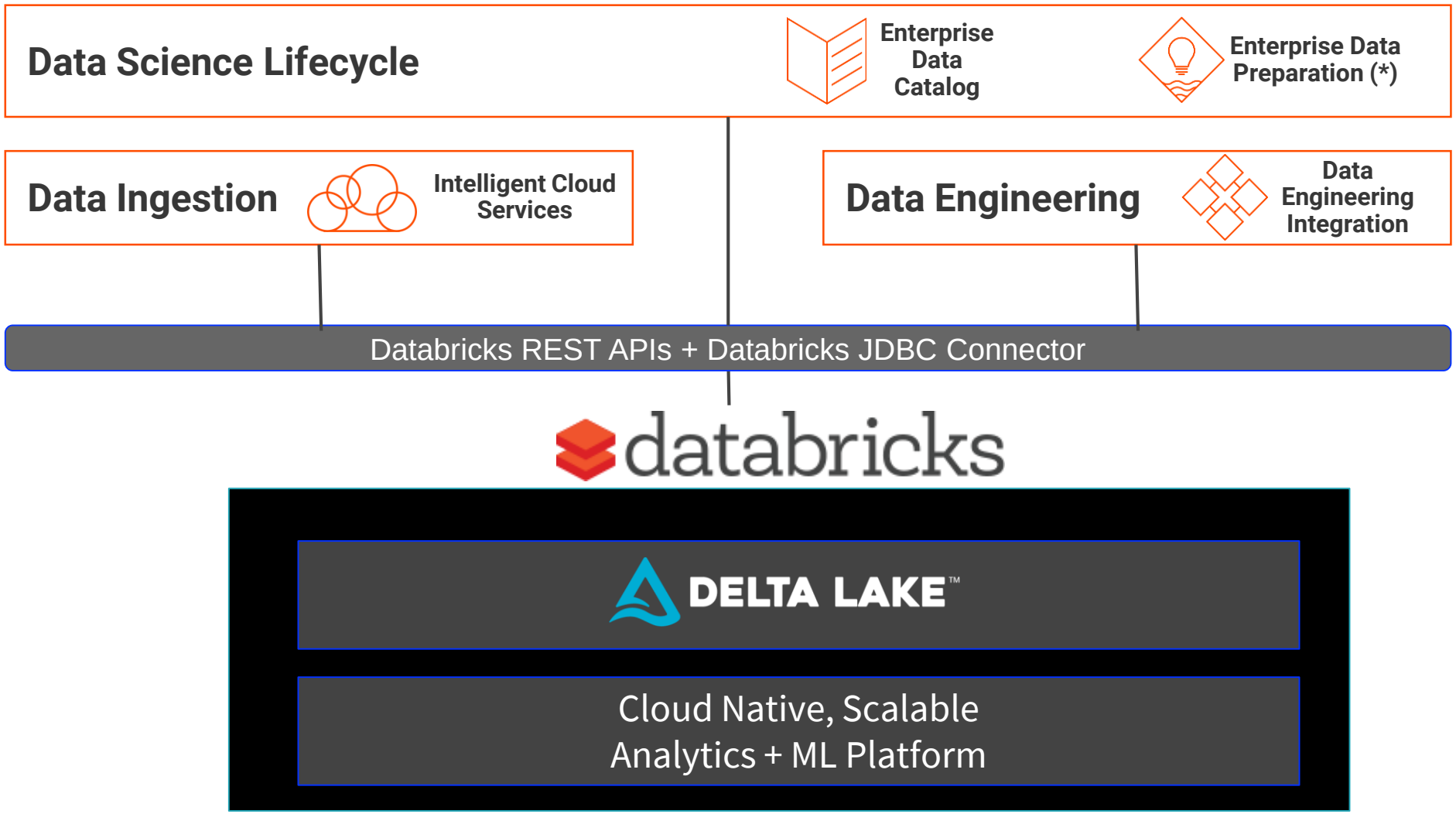
# Informatica Data Engineering

- Informatica 10.4.0 was released in December 2019
  - 10.4.0.1 available as of 28 February 2020
  - 10.4.0.2 scheduled for release in April 2020
  - 10.4.1.0 scheduled for release in June 2020
- Informatica 10.4.x currently supports Databricks 5.5
- Informatica 10.4 Docs: https://docs.informatica.com/data-engineering/data-engineering-integration/10-4-0.html

# Informatica Data Engineering

- Quick Overview on supported sources/targets
  - JDBC V2
  - Snowflake
  - Azure Blob / ADLS Gen1+Gen2 / AWS S3
    - Flat Files, Avro, Parquet, JSON
  - Azure Cosmos DB, Azure DW
  - Azure Event Hubs
  - AWS Redshift
- For Full Product Availability Matrix refer to https://network.informatica.com/docs/DOC-18443

# Three main product integrations

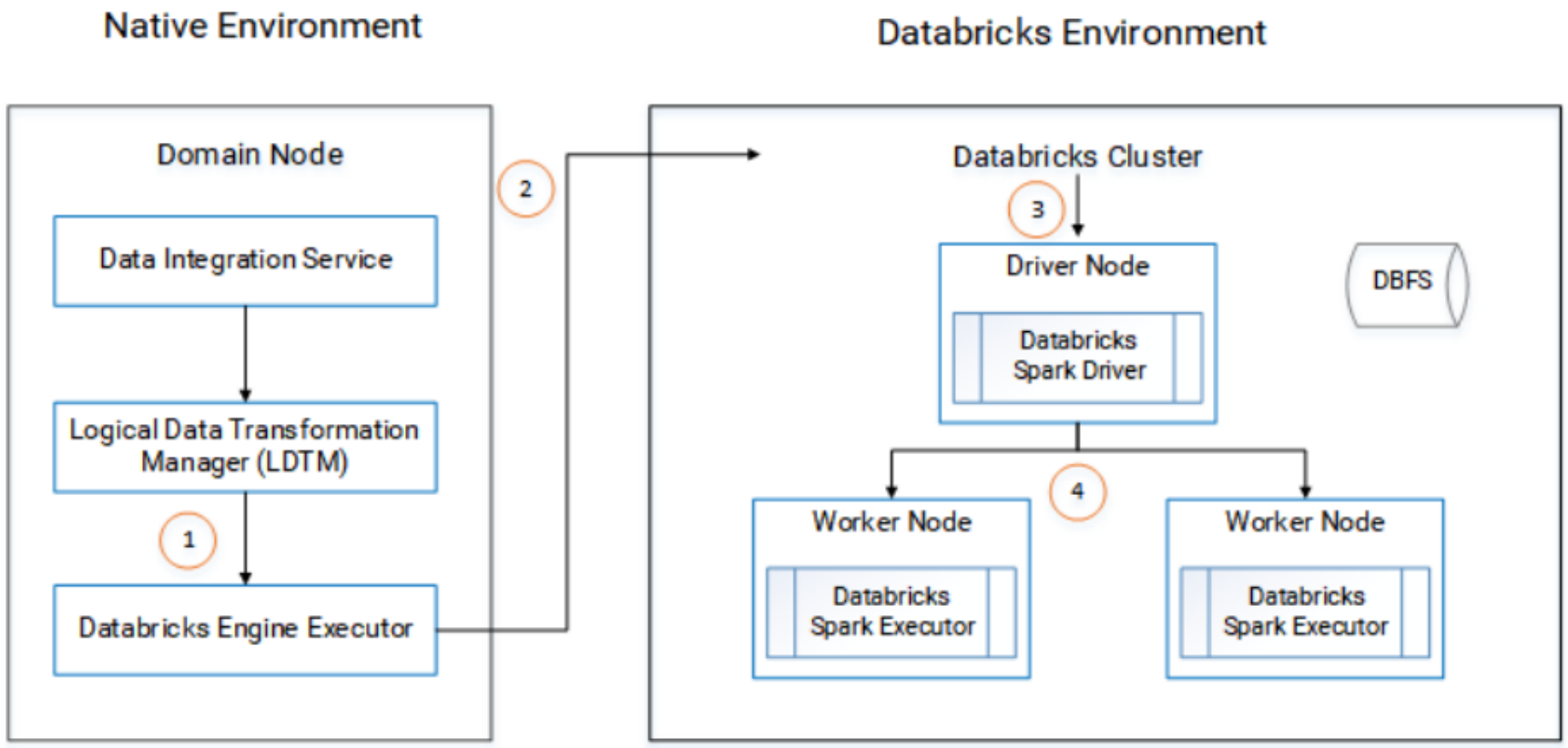



*(*) Roadmap*




Informatica

# Informatica Data Engineering Configuration for Databricks

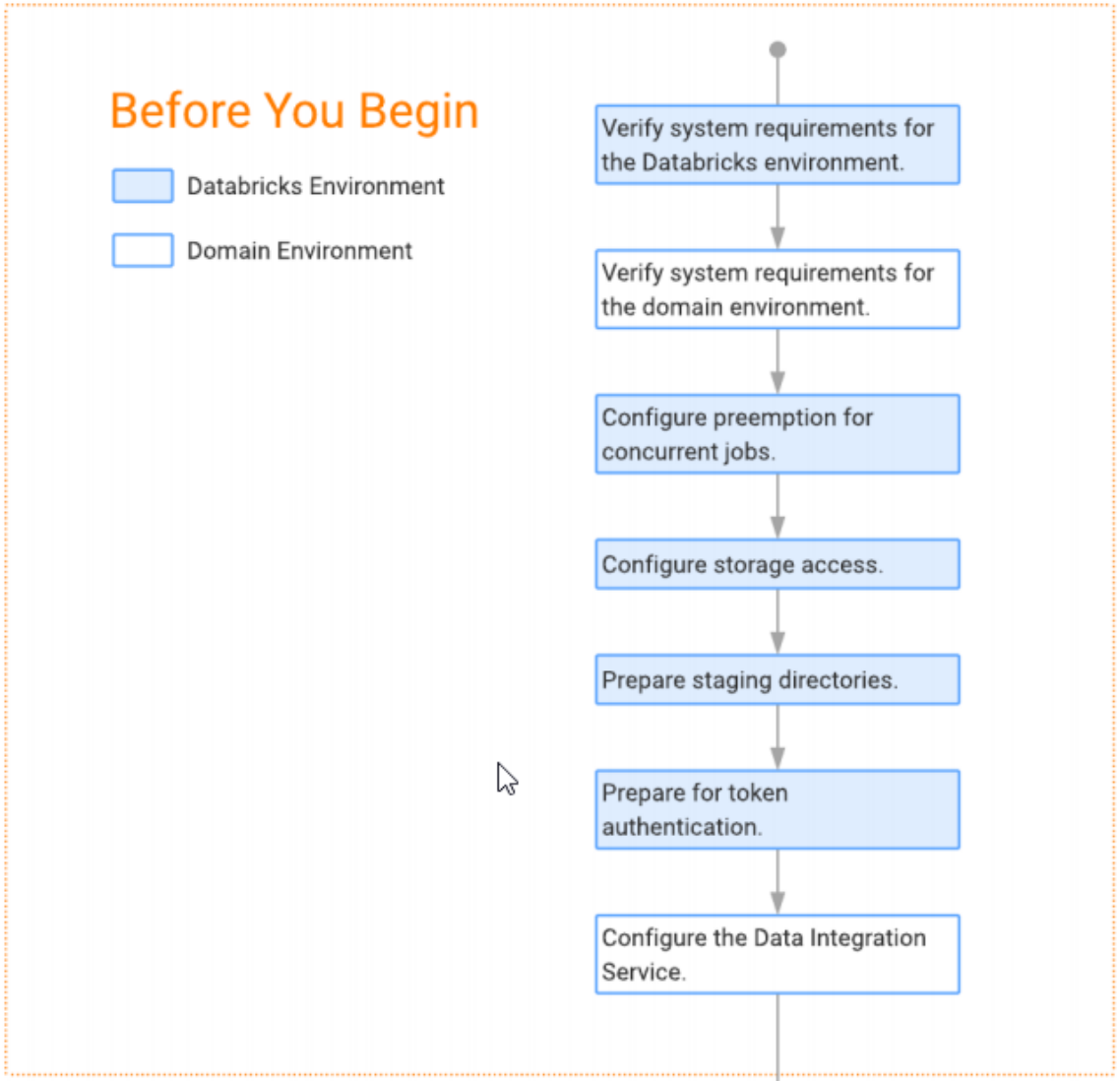


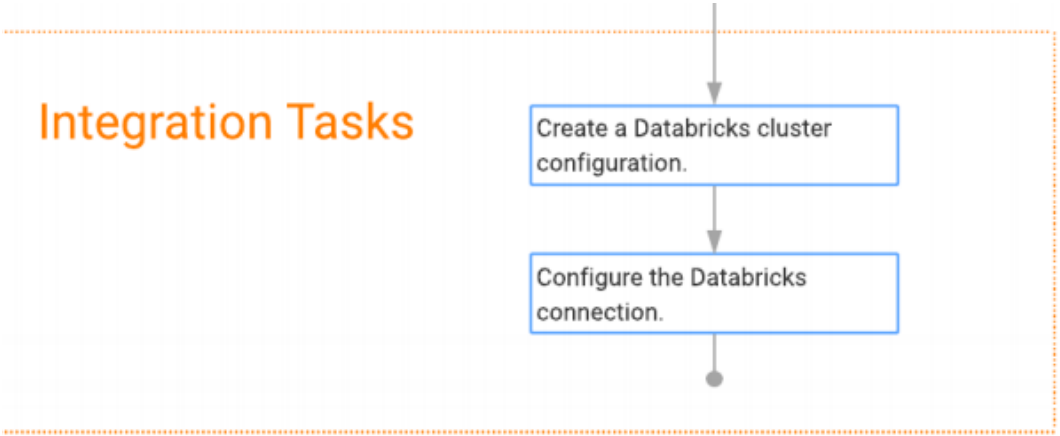




Informatica

# Informatica Data Engineering Configuration for Databricks



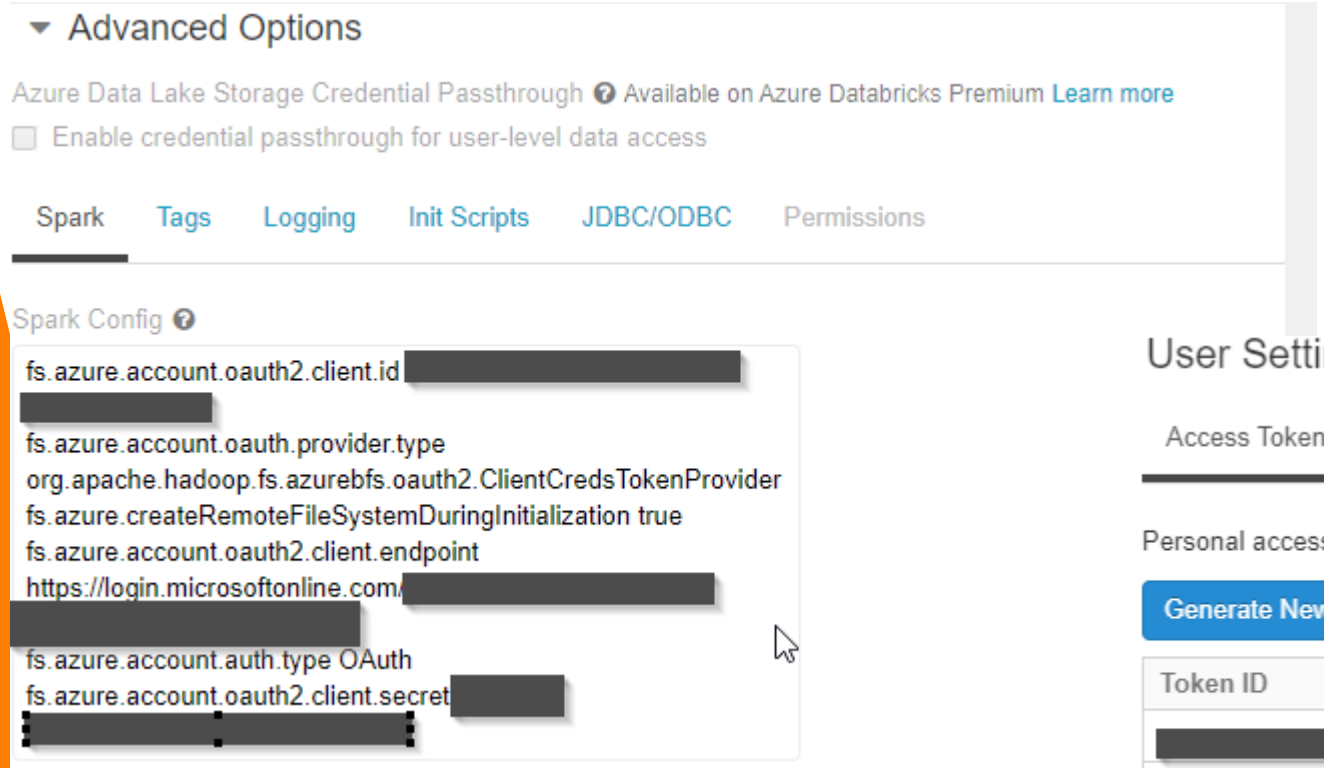



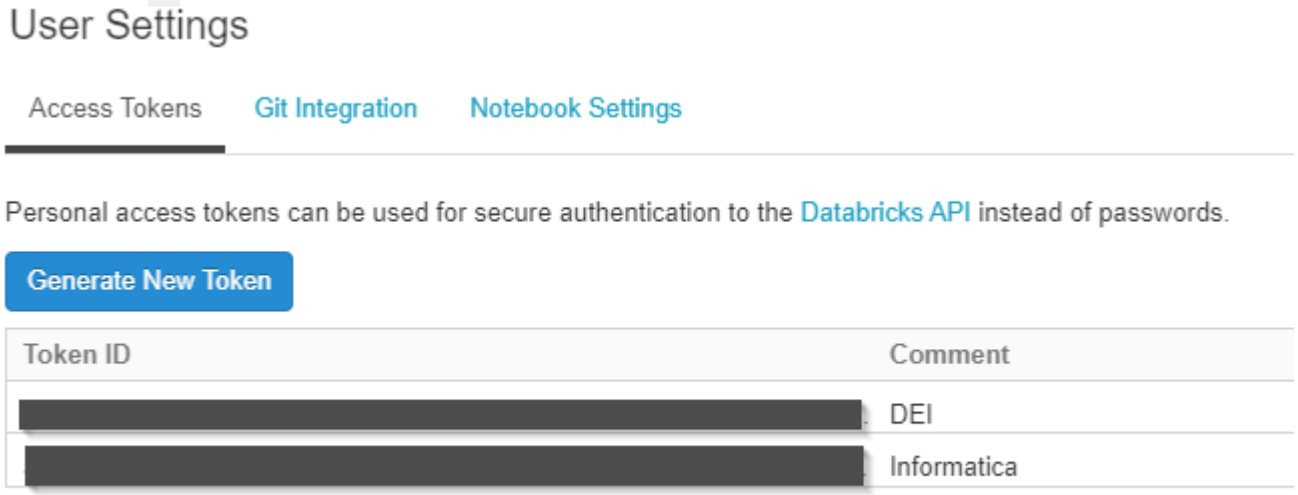




Informatica

# Informatica Data Engineering Configuration for Databricks



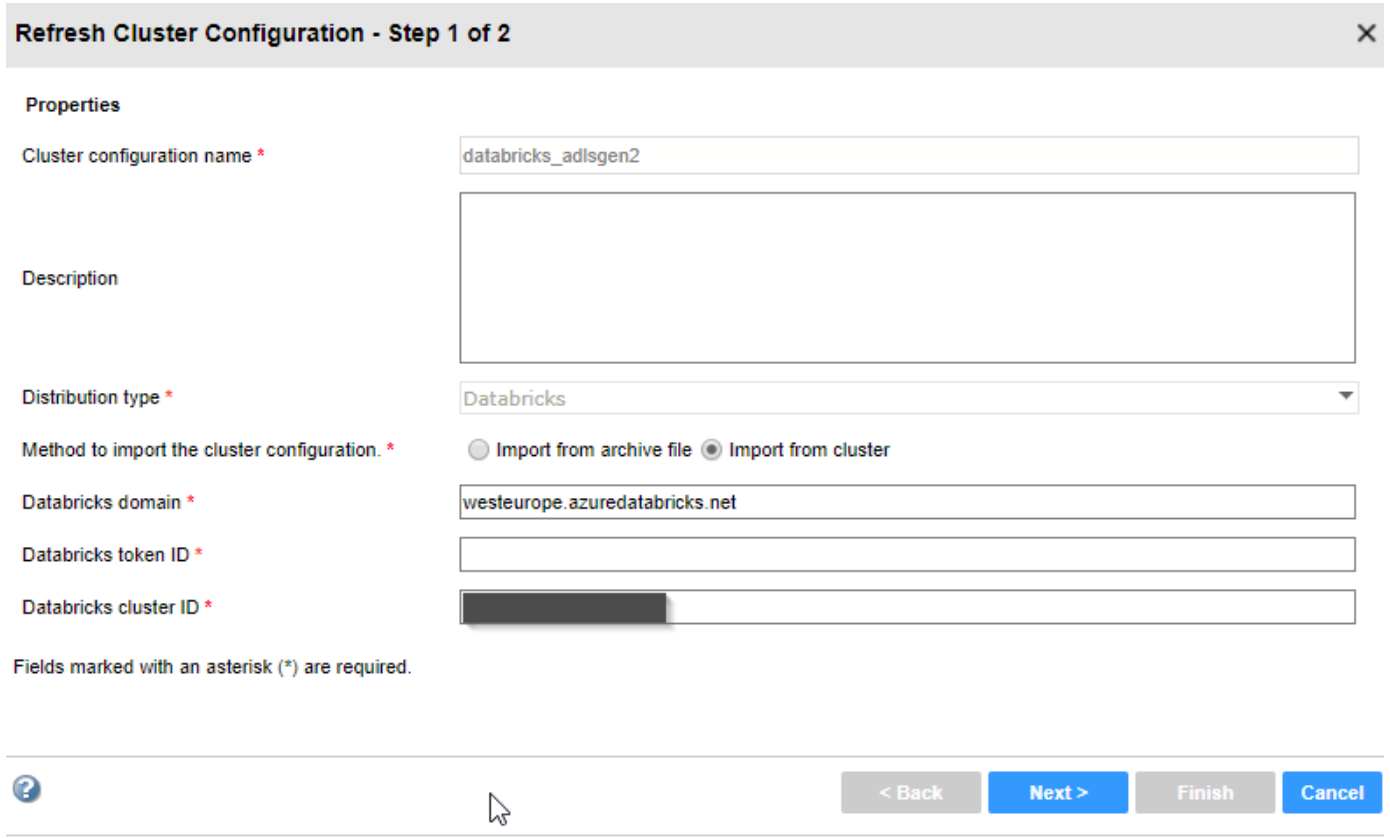







Informatica

# Informatica Data Engineering Configuration for Databricks







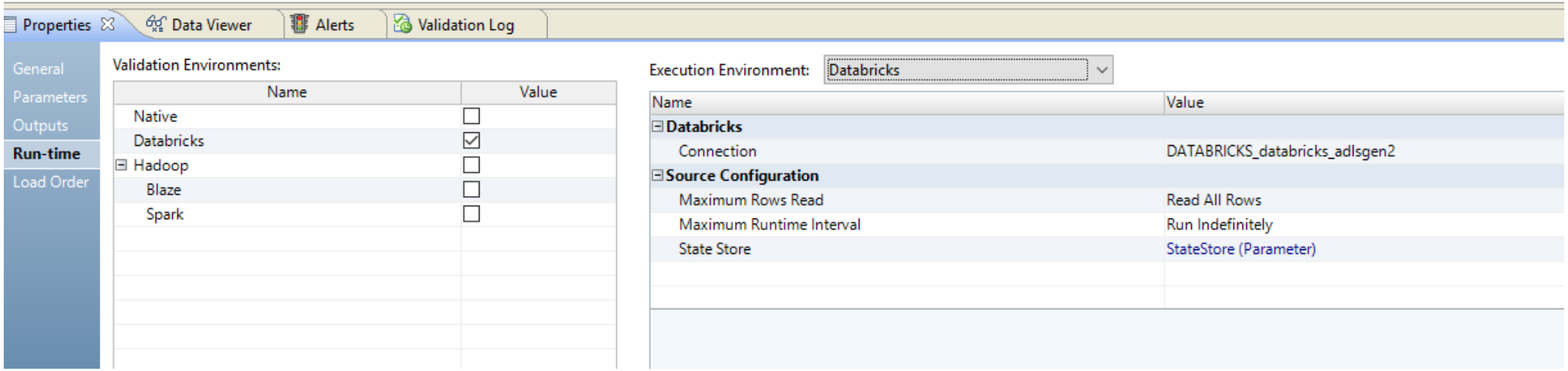
Informatica

# Informatica Data Engineering Configuration for Databricks

**DATABRICKS_databricks_adlsgen2**

**Cluster Properties**

| | |
|---|---|
| **Name** | DATABRICKS_databricks_adlsgen2 |
| **ID** | DATABRICKS_databricks_adlsgen2 |
| **Description** | |
| **Connection Type** | Databricks |
| **Cluster Configuration** | databricks_adlsgen2 |
| **Cloud Provisioning Configuration** | |



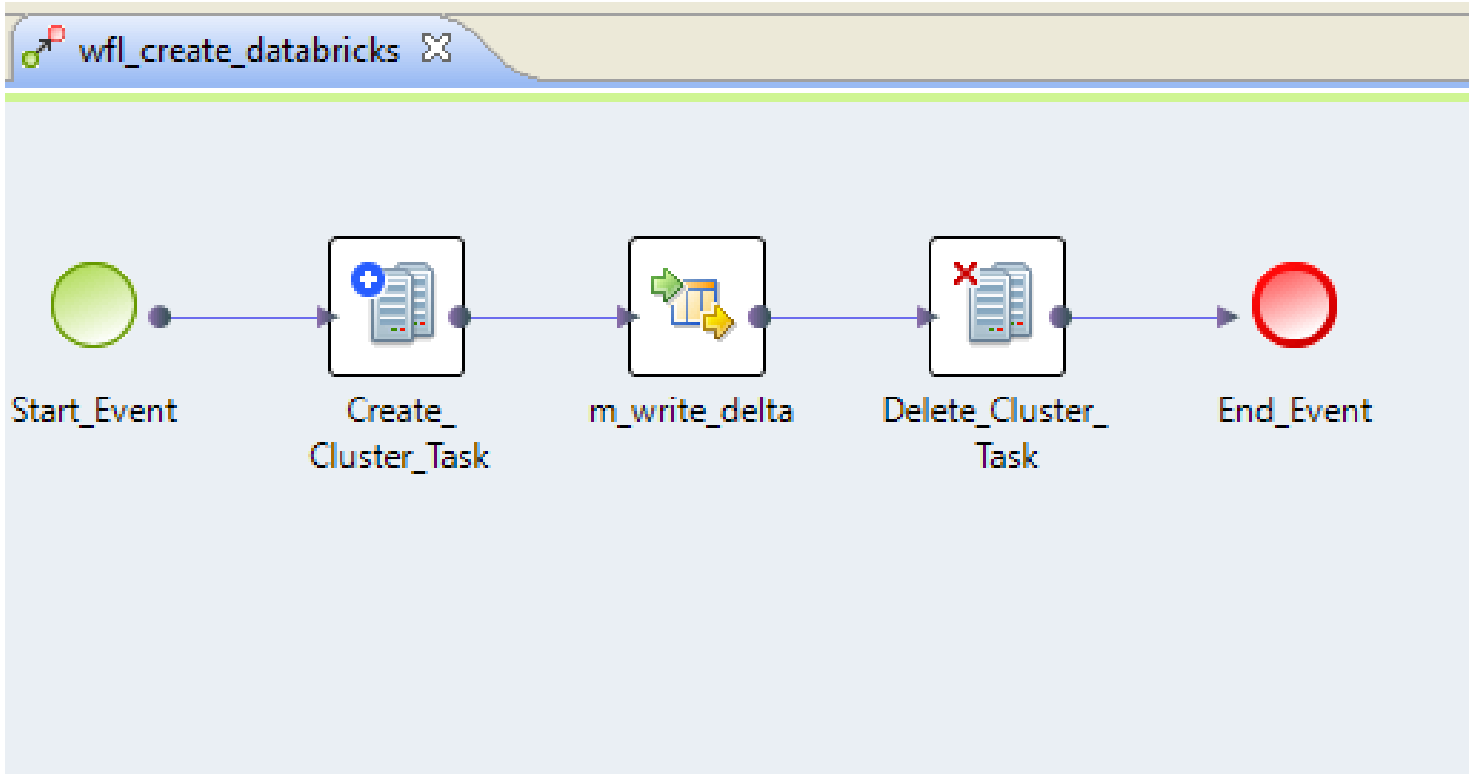




Informatica

# Informatica Data Engineering Cloud Provisioning Configuration

- Allows Cluster/Delete creation from within an Informatica Workflow

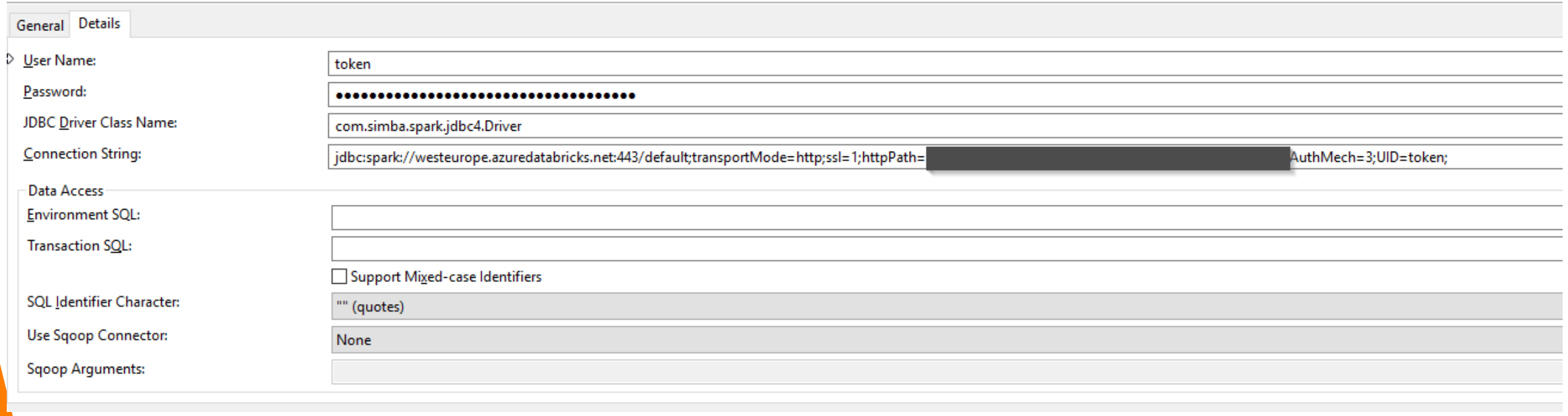


# Databricks Delta Source/Targets in Informatica Data Engineering

# Databricks Delta Source/Targets

- Connectivity is established using the Databricks Spark JDBC Driver. This needs to be present in externalJDBCJars directory on server+client
- Ability to read/write from/to Databricks Delta Lake
- Writing to Delta Lake is only supported while running the Mapping on the Databricks Engine.
- On the fly creation of Delta Lake Tables

# Databricks Delta Source/Targets




| Name | Value |
|---|---|
| Tracing Level | Normal |
| **Target** | |
| Load type | Normal |
| Update override | |
| Delete | ☑ |
| Insert | ☑ |
| Target Schema Strategy | CREATE - Create or replace table at run time |
| DDL query for create or replace | |
| Truncate target table | ☐ |




Informatica

# Demo

# Troubleshooting and self-service

Informatica

# Troubleshooting and self-service

- Informatica H2L / Knowledge Base
  - Integrating DEI 10.4.0.1 with Databricks and Delta Lake on the Azure Platform
  - Integrating DEI 10.4.0.1 with Databricks and Delta Lake on the AWS Platform
  - JDBC Connection to Databricks Delta is randomly failing with generic Simba SparkJDBC Driver Error
  - HOW TO: Enable Debug Tracing for the Simba Spark JDBC Driver connecting to Databricks Delta
  - HOW TO: Create Cluster Configuration Object (CCO) for Databricks cluster using infacmd command

# Troubleshooting and self-service

- Informatica Video KB:
  - Introduction to Azure Databricks - Part 1
  - How to Integrate Informatica BDM and Azure DataBricks Delta
  - Introduction to Azure Data Lake Storage Gen2 in DEI 10.4
  - Introduction to Databricks Transient and Ephemeral Cluster

# References

- Integrating DEI 10.4.0.1 with Databricks and Delta Lake on the Azure Platform
- Integrating DEI 10.4.0.1 with Databricks and Delta Lake on the AWS Platform
- Informatica® Data Engineering Integration - Integration Guide

# Q&A

Informatica

# Thank You

Stijn Carion