Talend Integration With Azure Databricks

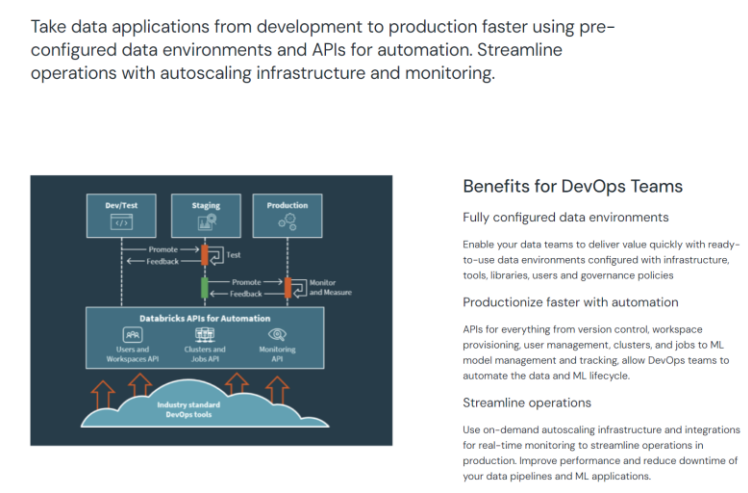
Talend Vs Databricks

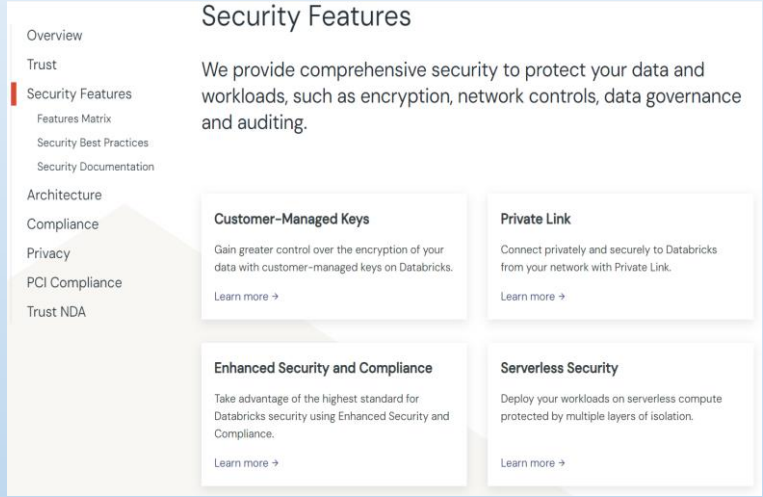
- ✓ Apache Spark - an open-source framework which helps process large chunks of semi-structured, unstructured and structured data for easy analysis. With Databricks (powered by Spark), the industry-leading, cloud-based unified platform, we can stay focused on your data science, data analytics, and data engineering tasks.
- ✓ Databricks supports multiple languages – Scala, Python, R, Java, SQL. In addition to that it can be easily integrated with data science frameworks and libraries including TensorFlow, PyTorch, and scikit-learn.
- ✓ It has scalable data processing capabilities, offering a wide range of built-in libraries and APIs for batch processing, streaming, graph processing, and machine learning. It can handle complex data transformations, aggregations, and analytics efficiently.
- ✓ It supports a wide range of data sources and formats, including Hadoop Distributed File System (HDFS), Apache Cassandra, Apache HBase, Apache Kafka, and many others.
- ✓ It has flexible deployment option in Azure (ARM), AWS (CloudFormation), GCP (Cloud Deployment Manager), HashiCorp Terraform. We can create workspace by calling REST API.

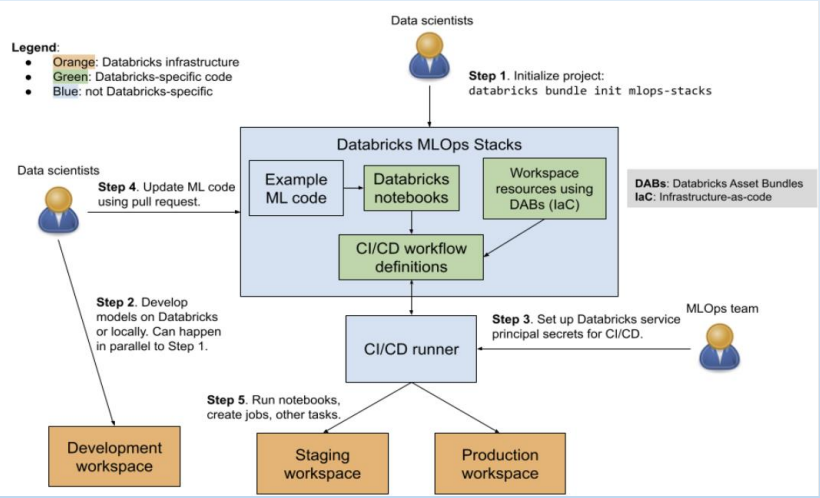
- ✓ Talend is an open-source data integration tool that provides a unified platform for designing, deploying, and managing various data integration processes. This ETL tool provides SQL templates to simplify the most common data query and update, schema creation and modification, and data access control tasks.
- ✓ It uses a Java-based programming language, although it also provides support for other languages through its components.
- ✓ It focuses more on data integration and ETL processes.
- ✓ It provides extensive connectivity options, although, its focus is primarily on data integration rather than the wide range of data sources supported by ADB.
- ✓ It has both on-premises and cloud deployment options, with support for various cloud platforms, such as AWS, Azure, and GCP.

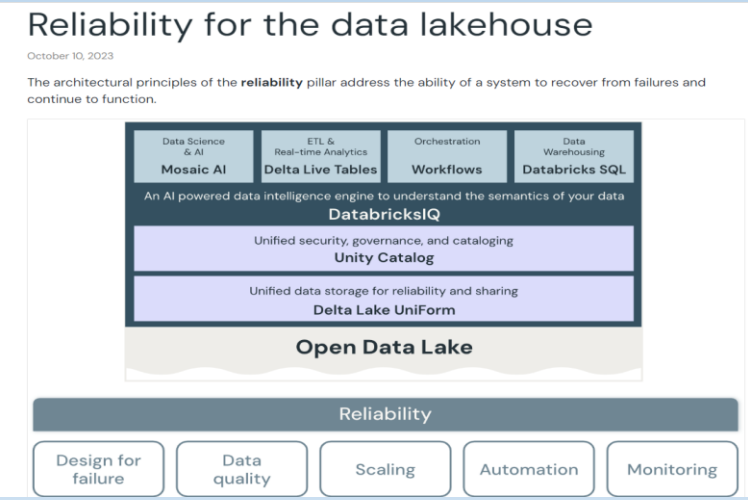
Problem Statement

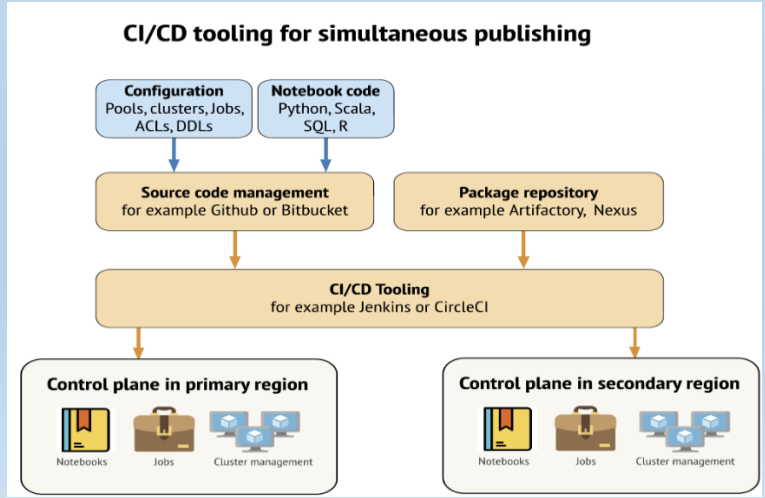
- ✓ As part of the project, we are trying to connect and run Azure Databricks notebook from Talend.
- ✓ This Talend job can be customized to build complex ETL pipelines.
- ✓ As DBS notebooks are Spark-SQL and codes, so it will be easy to migrate those codes to any on-premise and cloud environment.
- ✓ For easily deployable, scalable, secure, reliable and highly available code and data movement in ETL Big-data space, ADB is a great choice with Talend.

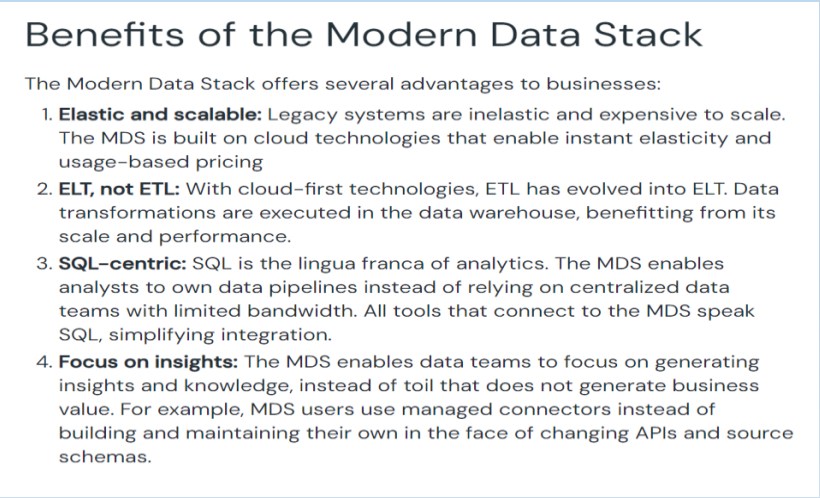






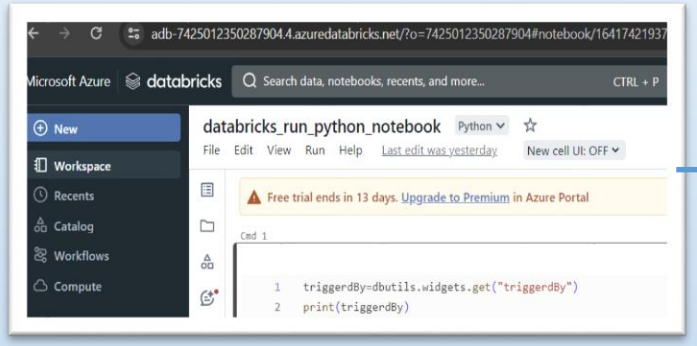


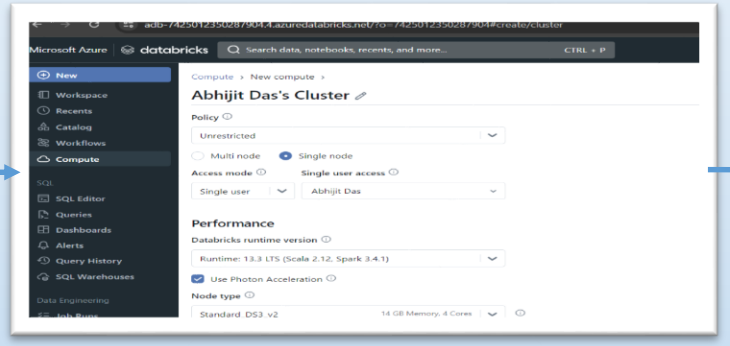


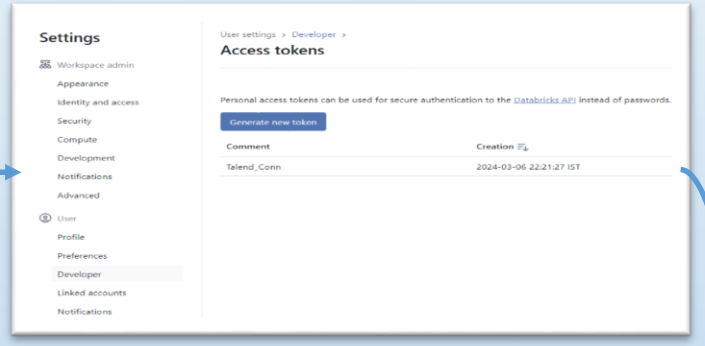


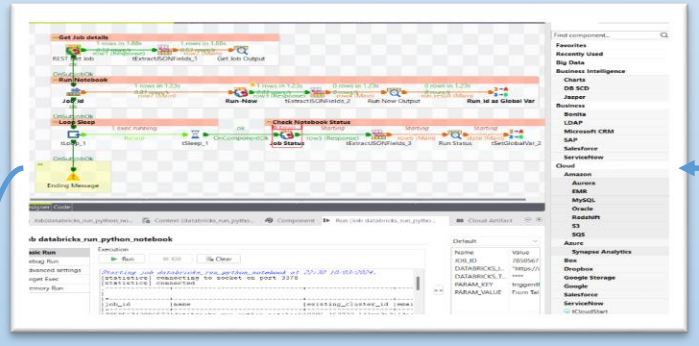
Steps To Integrate & Run Talend With Databricks

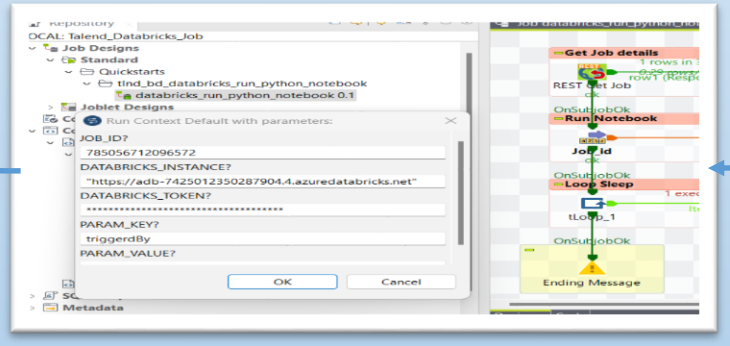
- ✓ Create Azure account and Databricks workspace with required permission.
- ✓ Create a notebook and add your scripts which you want to execute as part of your Talend job. This can be dynamic (parameterized).
- ✓ Create a Databricks cluster and attach to the notebook.
- ✓ Create & copy access token from User-Setting - > Developer.
- ✓ Create Talend job by using required components from Palette.
- ✓ Run job from Repository and pass – > Job-ID, DBS workspace URL and token.

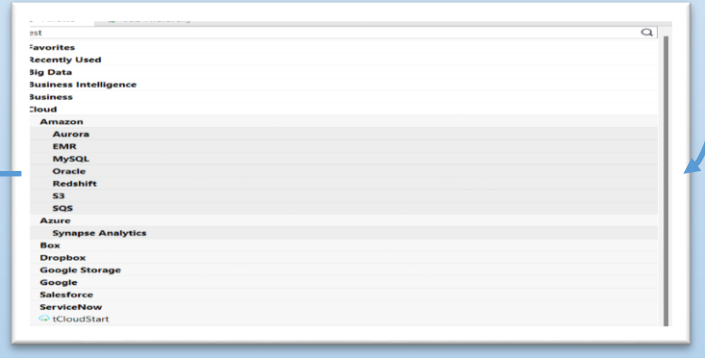


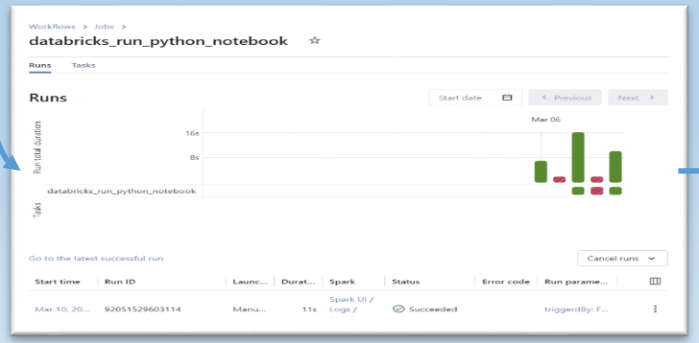


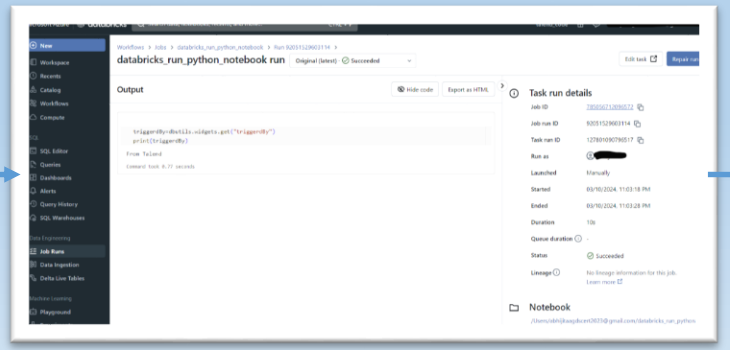


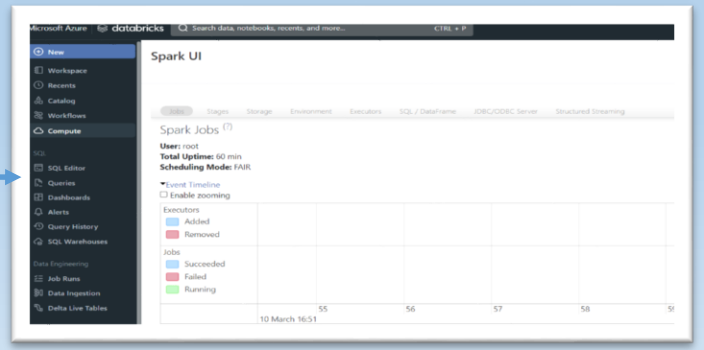












Thank You!