

Talend Integration With Azure Databricks



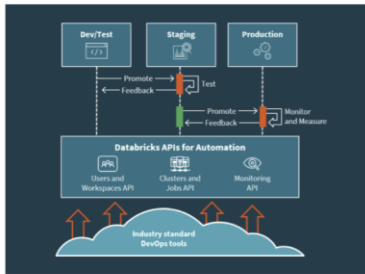
- ✓ Apache Spark - an open-source framework which helps process large chunks of semi-structured, unstructured and structured data for easy analysis. With Databricks (powered by Spark), the industry-leading, cloud-based unified platform, we can stay focused on your data science, data analytics, and data 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.

Take data applications from development to production faster using pre-configured data environments and APIs for automation. Streamline operations with autoscaling infrastructure and monitoring.



Benefits for DevOps Teams

Fully configured data environments

Enable your data teams to deliver value quickly with ready-to-use data environments configured with infrastructure, tools, libraries, users and governance policies

Productionize faster with automation

APIs for everything from version control, workspace provisioning, user management, clusters, and jobs to ML model management and tracking, allow DevOps teams to automate the data and ML lifecycle.

Streamline operations

Use on-demand autoscaling infrastructure and integrations for real-time monitoring to streamline operations in production. Improve performance and reduce downtime of your data pipelines and ML applications.

Security Features

Overview

Trust

Security Features

Features Matrix

Security Best Practices

Security Documentation

Architecture

Compliance

Privacy

PCI Compliance

Trust NDA

We provide comprehensive security to protect your data and workloads, such as encryption, network controls, data governance and auditing.

Customer-Managed Keys

Gain greater control over the encryption of your data with customer-managed keys on Databricks.

[Learn more →](#)

Private Link

Connect privately and securely to Databricks from your network with Private Link.

[Learn more →](#)

Enhanced Security and Compliance

Take advantage of the highest standard for Databricks security using Enhanced Security and Compliance.

[Learn more →](#)

Serverless Security

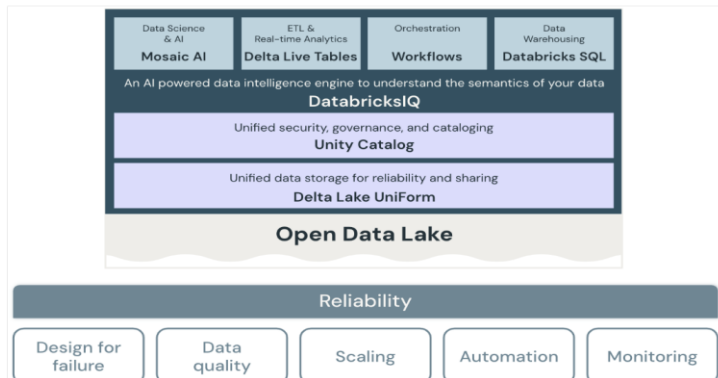
Deploy your workloads on serverless compute protected by multiple layers of isolation.

[Learn more →](#)

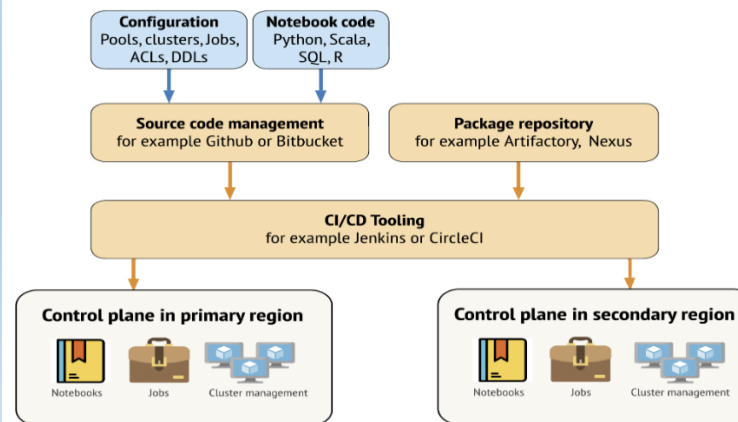
Reliability for the data lakehouse

October 10, 2023

The architectural principles of the **reliability** pillar address the ability of a system to recover from failures and continue to function.



CI/CD tooling for simultaneous publishing



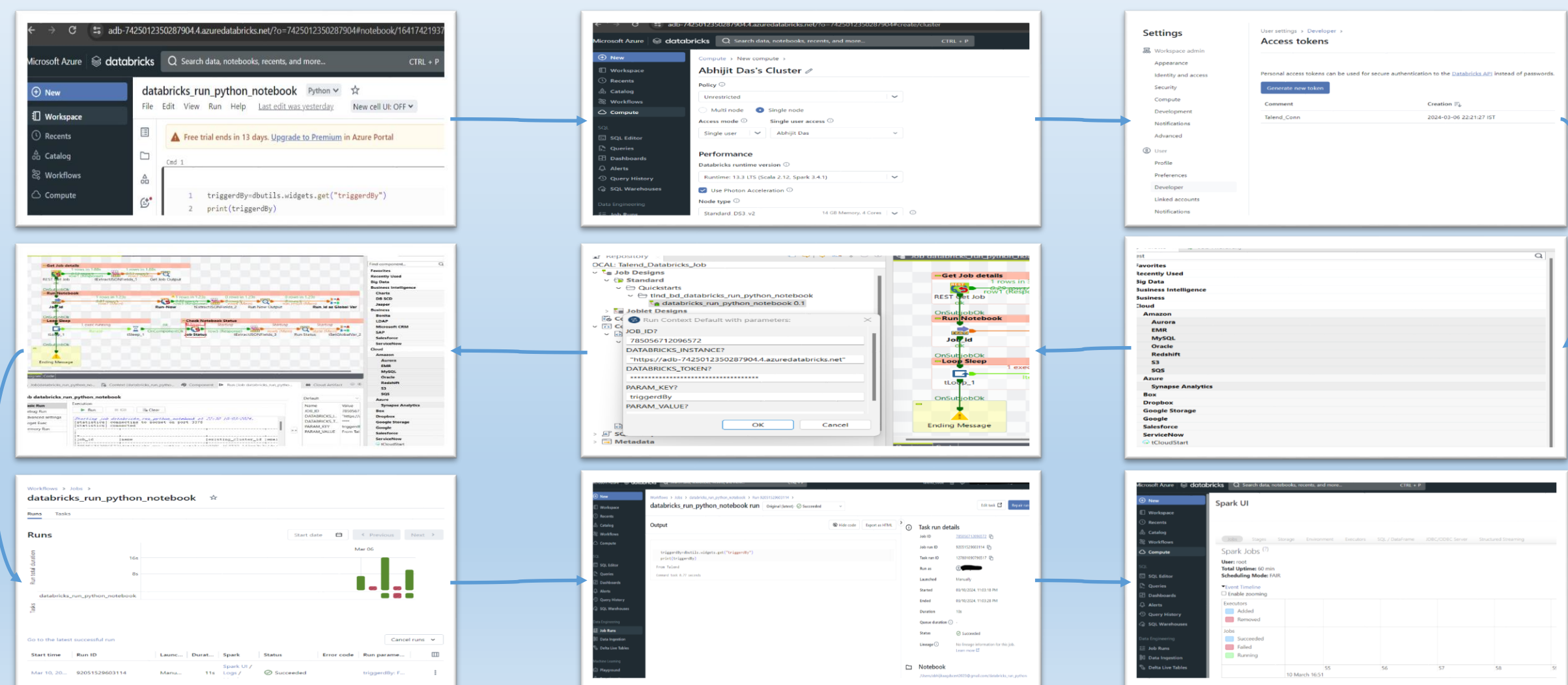
Benefits of the Modern Data Stack

The Modern Data Stack offers several advantages to businesses:

- 1. Elastic and scalable:** Legacy systems are inelastic and expensive to scale. The MDS is built on cloud technologies that enable instant elasticity and usage-based pricing
- 2. ELT, not ETL:** With cloud-first technologies, ETL has evolved into ELT. Data transformations are executed in the data warehouse, benefitting from its scale and performance.
- 3. SQL-centric:** SQL is the lingua franca of analytics. The MDS enables analysts to own data pipelines instead of relying on centralized data teams with limited bandwidth. All tools that connect to the MDS speak SQL, simplifying integration.
- 4. Focus on insights:** The MDS enables data teams to focus on generating insights and knowledge, instead of toil that does not generate business value. For example, MDS users use managed connectors instead of building and maintaining their own in the face of changing APIs and source schemas.

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!