Databricks for Data Engineering

Build Robust Production Data Pipelines at Scale

Data quality issues and the resultant data pipeline complexity to deal with them can be the bane of many data professionals' existence. Whether you are chartered with advanced analytics, developing new machine learning models, providing operational reporting or managing the data infrastructure, the concern with data quality is a common theme. Data engineers, in particular, strive to design and deploy robust data pipelines that serve reliable data in a performant manner so that their organizations can make the most of their valuable corporate data assets. Unified analytics is a modern approach that serves data engineering needs well.

Unified Analytics Powering Data Engineering

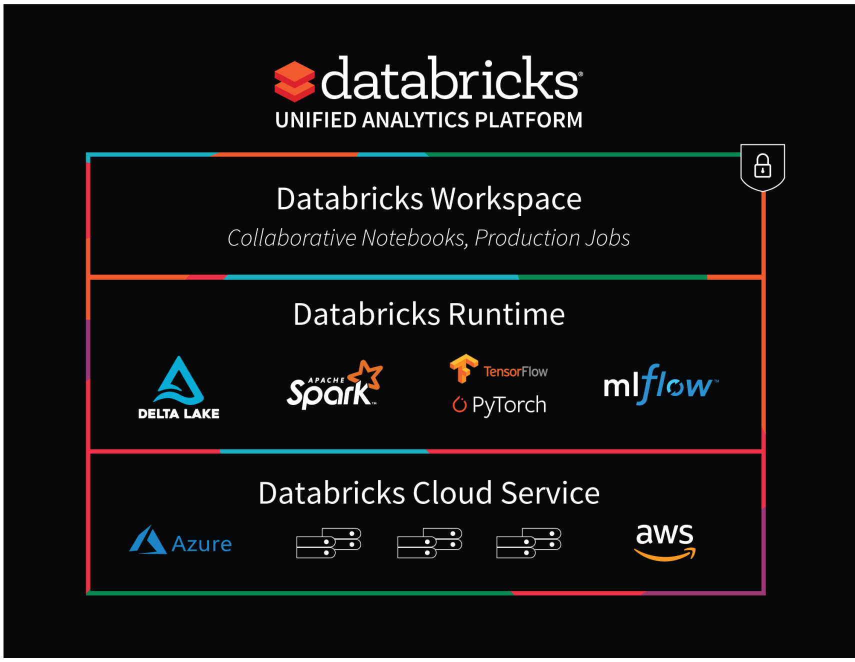
Founded by the team that originally created Apache Spark™, Databricks provides the **UNIFIED ANALYTICS PLATFORM** that accelerates innovation by unifying data science, engineering and business. The platform comprises three layers:

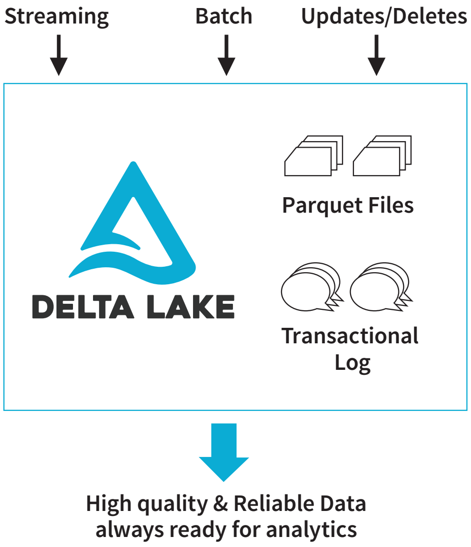
WORKSPACE
Provides notebook-based collaboration and automates pipelines through production jobs

RUNTIME
Unifies data and ML at scale through built-in ML frameworks and innovations like Delta Lake and MLflow

CLOUD SERVICE
Removes the complexity of managing the underlying cloud infrastructure

WITH THE PLATFORM DATA ENGINEERS CAN DEPLOY AND MANAGE ROBUST PRODUCTION DATA PIPELINES AT SCALE TO SUPPORT THEIR ORGANIZATIONS' ANALYTICS NEEDS.





Delta Lake is an open source storage layer that brings data reliability to data lakes. It provides ACID transactions, is compatible with Apache Spark APIs and includes streaming and batch readers and writers.

- ACID transactions
- Schema enforcement
- Data versioning
- Unified batch & streaming data
- Compatible with Apache Spark
- Schema evolution

SIMPLIFIED DEVOPS

Abstract the complexity of data infrastructure and DevOps, with auto-config and auto-scaling capabilities.

RELIABLE AND PERFORMANT INFRASTRUCTURE

Databricks Unified Analytics Platform offers 99.9% SLA and provides high-reliability for its pre-configured clusters, allowing teams to focus more on innovation rather than management of infrastructure.

RELIABLE DATA

Unify batch and real-time with **Delta** making reliable data ready for analytics and ML at massive scale as data flows from various storage sources into the processing engine.

HIGH PERFORMANCE

Benefit from built-in **Delta** performance optimizations and get reliable results 10-100x faster than open source Apache Spark™.

SECURE AND COMPLIANT

Databricks Unified Analytics Platform provides enterprise-grade security with encryption, auditing, role-based control and HIPAA and SOC 2 type 2 compliance.

PRODUCTION JOBS

Access all your data in one place, and automate the most complex data pipelines with jobs scheduling, monitoring, and workflows as notebooks or APIs, giving teams full flexibility to run and maintain data pipelines for ML at scale.

ECOSYSTEM INTEGRATION

Databricks supports SQL, R, Python, Java, and Scala and provides native integration with popular IDEs like RStudio, or BI tools with ODBC/JDBC connections, allowing data engineers and data scientists to use familiar languages and tools on Databricks.

SHARED NOTEBOOKS

Foster collaboration between data engineers and data scientists by unifying ETL, data wrangling, and model training, as part of interactive notebooks, APIs, or your favorite IDEs - backed with version history, tracking changes, and GitHub integration.

Data Engineering, Simplified

Databricks' Unified Analytics Platform removes the complexity of data engineering and accelerates data engineering tasks from data access to ETL, allowing engineers to build robust production data pipelines at scale.

Try building next-gen data pipelines today using the free trial or the Community Edition databricks.com/try

© Databricks 2019. All rights reserved. Apache, Apache Spark, Spark and the Spark logo are trademarks of the Apache Software Foundation.