03 Databricks Intro

Introduction to Databricks

tags: #databricks #spark #big-data #data-engineering #lecture-notes #ai #lakehouse

> **Prerequisites**
>
> For the best learning experience, attendees should have a basic understanding of:
>
> - Core data concepts (e.g., tables, databases)
> - Introductory Python or SQL
> - General cloud computing concepts (e.g., VMs, storage)

Overview of Apache Spark: The Foundation

Before diving into Databricks, it's essential to understand Apache Spark, as Databricks is built directly on it to simplify and enhance its use.

- **What is Apache Spark?**
  - A fast, general-purpose **distributed computing engine** designed for big data analysis and processing.
  - Enables **massive parallel processing** in a distributed environment, ideal for handling large-scale datasets (e.g., terabytes or petabytes) that can't fit on a single machine.
  - It operates on distributed data structures, most commonly the **DataFrame**, which is conceptually similar to a table in a relational database but is partitioned across many machines.

> **Core Concept: The Cluster**
>
> Spark distributes data and computations across a **cluster** of machines (nodes) to achieve speed and scalability.
>
> - **Cluster Setup**: Combines multiple computers (e.g., 10 laptops = 10 nodes) to process data faster than a single powerful machine.
> - **Nodes**:
>   - **Driver Node**: Acts as the coordinator—manages the application, schedules tasks, and communicates instructions to workers.
>   - **Worker Nodes**: Execute the actual computations on distributed data partitions.

- **Why is it Powerful?** Handles transformations (e.g., filtering, aggregating) and actions (e.g., saving results) in parallel, supporting languages like Python (**PySpark**), Scala, SQL, and R.
- **Use Cases**: ETL (Extract, Transform, Load), machine learning, real-time streaming, and interactive analytics on big data.
- **History**: Started as a research project at UC Berkeley's AMPLab around 2008-2009; open-sourced and donated to the Apache Software Foundation in 2013.
- **Challenges with "Vanilla" Spark**:
  - Manual cluster management: Provisioning/scaling nodes, handling failures, resource allocation.
  - Production complexities: Monitoring, security, optimization, and integration with cloud storage.
  - This is where Databricks comes in—created by the same team to eliminate these pain points.

What is Databricks?

- **Origin and Purpose**:
  - Founded in 2013 by the original creators of Apache Spark from UC Berkeley.
  - Databricks is a **unified data analytics platform** (now called the "**Data Intelligence Platform**") that provides a comprehensive management layer on top of Apache Spark.
  - **Core Goal**: To simplify big data and AI by automating infrastructure, enabling collaboration, and adding enterprise-grade features for data engineering, analytics, and machine learning.
  - Runs on major clouds (AWS, Azure, GCP) with serverless options for ultimate scalability.

The Lakehouse Architecture

The **Lakehouse Architecture** is the foundation of Databricks. It combines the best of both worlds:

- The low-cost, scalable, and flexible storage of a **Data Lake** (which holds raw data in any format).
- The performance, reliability, and ACID transactions of a **Data Warehouse** (which holds structured, governed data).
  This is achieved using an open-source storage format called **Delta Lake** as the backbone.

- **Why Use Databricks?**
  - Automates Spark's overhead: No need to manually manage VMs, clusters, or network connections—focus on code and insights.
  - Powers 10,000+ organizations (including 60%+ of Fortune 500) for faster innovation in data and AI.
  - **Key Benefits**:
    - **Scalability**: Auto-scales clusters from small dev environments to massive production workloads.

- **Collaboration**: Multi-language notebooks for teams (data scientists, engineers, analysts).
- **Cost Efficiency**: Pay-per-use with performance optimizations like the **Photon** engine.
- **Security & Compliance**: Built-in governance, encryption, and enterprise-grade security.

As of September 2025, Databricks continues to lead in Lakehouse innovation, with recent emphases on AI governance, predictive optimizations, and unified platforms like Databricks One.

Key Features of Databricks

Databricks enhances Spark with user-friendly tools and advanced capabilities.

1. Cluster Management

- **Description**: Automated provisioning, scaling, and optimization of Spark clusters.
  - Define cluster size, set auto-termination for idle clusters, and enable elasticity for dynamic workloads.
  - Supports **Photon Clusters**: A high-performance, C++ based engine that provides 2-8x faster performance on SQL, ETL, and ML tasks compared to standard Spark.
- **2025 Updates**: Serverless GPU compute with H100 accelerators (Beta) for AI workloads; predictive optimization (GA) uses AI to auto-manage data layout and cleanup, reducing manual tuning.
- **Benefits**: "Click a few boxes" to launch a powerful cluster—Databricks handles everything from driver/worker nodes to fault tolerance.

2. Notebooks

- **Description**: Interactive, web-based environment similar to Jupyter for writing code in cells.
  - Supports real-time collaboration, built-in visualizations (charts, tables), and debugging.
  - Seamlessly switch between languages in different cells using magic commands (e.g., `%sql`, `%python`, `%scala`, `%r`).
  - Integrates directly with Git for version control.
- **2025 Updates**: Enhanced with **Databricks Assistant Data Science Agent** (Beta)—an AI agent that automates tasks like data exploration, cohort analysis, and ML pipeline structuring. New SQL Editor (GA) includes multi-statement results, inline history, and AI integration.
- **Benefits**: Makes prototyping, development, and sharing insights seamless for diverse teams.

3. Unity Catalog

- **Description**: The unified, end-to-end **data and AI governance layer** for all assets within the Databricks Lakehouse. It manages and secures catalogs, databases/schemas, tables, views, models, and files.
  - Enforces a **three-level namespace** for all data assets: `catalog.schema.table`, providing a clear and unambiguous way to reference data.
  - Features centralized access control (e.g., row/column-level permissions), data **lineage** tracking, comprehensive auditing, and powerful data discovery.

    - Supports secure, multi-cloud data sharing via the open-source **Delta Sharing** protocol.
- **2025 Updates** (from Data + AI Summit):
    - **Intelligent Capabilities**: AI-powered data discovery and context.
    - **Full Apache Iceberg Support** (Public Preview): Read/write managed Iceberg tables; federate foreign Iceberg catalogs (e.g., AWS Glue, Snowflake) with Unity governance.
    - **External Lineage Metadata** (Public Preview): End-to-end lineage including external sources (e.g., Salesforce) and consumers (e.g., Tableau).
    - **Unity Catalog Metrics**: Define and govern business metrics (KPIs).
    - `SET MANAGED` **Command** (Public Preview): Convert external tables to managed for better governance/performance.
- **Benefits**: A cornerstone of the modern Databricks platform that solves critical governance challenges, ensures regulatory compliance (e.g., GDPR), and breaks down data silos.

4. Delta Live Tables (DLT) – Now part of Lakeflow Declarative Pipelines

- **Description**: A **declarative framework** for building reliable, maintainable ETL/ELT pipelines.
    - You define *what* to do with the data (transformations, quality rules) in SQL/Python; DLT handles *how* to do it (orchestration, scaling, error recovery).
    - Supports both batch and streaming data, Change Data Capture (CDC), and schema evolution.
    - Natively supports data quality enforcement through **expectations**. For example: `CONSTRAINT valid_email EXPECT (email IS NOT NULL) ON VIOLATION FAIL UPDATE.`
- **2025 Updates** (Rebranded as **Lakeflow** in June 2025):
    - **Lakeflow Declarative Pipelines**: Unified with Lakeflow Connect (ingestion) and Lakeflow Jobs (scheduling).
    - Enhanced streaming metrics and notebook-based development/debugging (Public Preview).
- **Benefits**: "Changed the game of data engineering"—automates data quality, lineage, and monitoring. It is the ideal tool for implementing the **Medallion Architecture** (Bronze/Silver/Gold layers).

5. Alerts

- **Description**: Set up notifications for data anomalies, query performance, or pipeline failures—similar to traditional warehouses.
- **2025 Updates**: Integrated with AI/BI tools; supports metric views for KPI alerts in dashboards/Genie spaces.
- **Benefits**: Enables proactive monitoring for production-grade reliability.

6. SQL Warehouses (Formerly SQL Endpoints)

- **Description**: Optimized compute clusters specifically for SQL and BI workloads. They are architecturally separate from the clusters used for data engineering or ML, preventing BI dashboards from competing for resources with ETL jobs.
    - Enables analysts to connect via BI tools (e.g., Tableau, Power BI) with high performance.
- **2025 Updates**:
    - **Predictive Optimization** (GA): AI automatically optimizes tables for faster queries.
    - New SQL Editor with real-time collaboration and Assistant integration.

- **Benefits**: "Make reports work faster"—up to 70% query speedup with Photon.

7. ETL Workflows (Now Lakeflow Jobs)

- **Description**: Orchestrate multi-step data pipelines with scheduling, dependencies, and retries.
  - Comparable to tools like **Apache Airflow**, **Azure Data Factory**, or **AWS Step Functions**, but deeply integrated into the Databricks ecosystem.
- **2025 Updates**: Unified under the **Lakeflow** umbrella; includes file arrival triggers for external storage (e.g., S3).
- **Benefits**: "Very powerful"—handles complex, production-grade ETL with robust monitoring and alerting.

Ecosystem and 2025 Advancements

> **Key Themes from Data + AI Summit 2025**
>
> The major trends are unification, simplification for business users, and deep integration of generative AI across the entire platform.

- **Core Integrations**: **Delta Lake** (storage backbone), **MLflow** for the machine learning lifecycle, **Mosaic AI** for building and deploying GenAI applications.
- **New in 2025**:
  - **Databricks One**: A unified portal for search, chat, and dashboards—simplifying self-service analytics for business users.
  - **Databricks Apps**: Managed containers (e.g., Streamlit) with Unity Catalog security for building custom data applications.
  - **Genie & AI/BI**: An AI-powered chat interface for natural language queries; metric views for governing KPIs.
  - **Serverless Enhancements**: Upgraded runtimes and expanded capabilities for serverless compute.
  - **Governance Focus**: Expanded governance for external data sources, Iceberg tables, and AI models.

Why Organizations Love Databricks

- Transforms Spark from a raw "computing engine" into a complete, enterprise-ready platform.

- Accelerates the development of reliable data pipelines and AI solutions, creating high demand for skilled Databricks developers.

✓ **Getting Started**

- **Databricks Community Edition**: A free, limited-feature version perfect for learning and hands-on practice. Note that it runs on a smaller, single-node cluster.
- **Databricks Free Trial**: A full-featured trial on your cloud provider of choice.

Further Reading

- Databricks Product Page
- Unity Catalog Documentation
- Lakeflow (DLT) Documentation