

Open in app ↗

Medium

 Search Write 18

Databricks Architecture Overview: Components & Workflow



AccentFuture

Follow

5 min read · Apr 1, 2025



Introduction

Databricks is a cloud-based data engineering platform that simplifies big data and artificial intelligence (AI) workloads. Built on Apache Spark, Databricks provides a unified analytics platform with robust data processing, machine learning, and business intelligence (BI) capabilities. It is widely used for large-scale data processing and advanced analytics.

In this article, we will explore the **Databricks architecture, its core components, and how it efficiently processes large datasets in cloud environments**. We will also explain the **Databricks architecture diagrams** in detail.

1. Databricks Standard Architecture

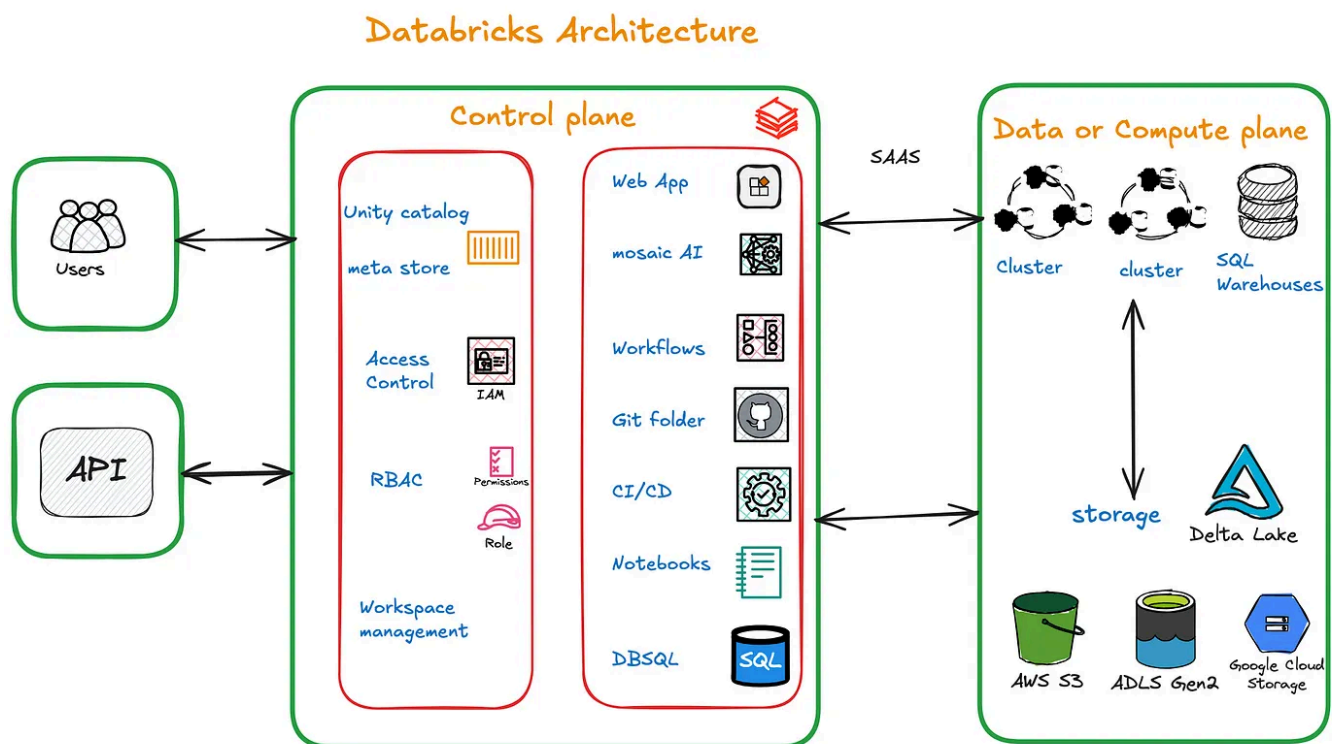
Databricks follows a **two-layer architecture**:

1. Control Plane

2. Data (or Compute) Plane

This architecture ensures **security, scalability, and flexibility** by separating data management and computation processes.

Standard Databricks Architecture with customer-managed compute



This is the **classic Databricks architecture**, where the **Control Plane** is fully managed by Databricks, while the **Compute Plane** is hosted in your cloud environment (AWS, Azure, or Google Cloud).

1.1 Control Plane

The Control Plane is responsible for managing user access, workspaces, job scheduling, and metadata storage. It operates as **Software as a Service (SaaS)** and is fully managed by Databricks.

Key Components:

1.Unity Catalog & Meta Store:

- Unity Catalog enables **data governance, access control, and lineage tracking**.
- The **Meta Store** stores metadata like table structures, schemas, and partitions.

2. Access Control & Security:

- **IAM (Identity and Access Management)**: Manages user identities and permissions.
- **RBAC (Role-Based Access Control)**: Assigns roles for **secure data access**.

3. Workspace Management:

- **Provides an interface** to manage notebooks, clusters, jobs, and assets.
- Organizes **projects and permissions** within Databricks.

4.Web Applications & AI Tools:

- **Mosaic AI**: A suite of AI/ML tools for advanced analytics.
- **Workflows**: Automates job execution for **data pipelines**.

5.Git & CI/CD Integration:

- Supports **Git repositories** for version control.

- Enables **CI/CD workflows** for deployment.

6. Notebooks & DBSQL

- Notebooks support Python, Scala, SQL, and R for **collaborative coding**.
- **DBSQL (Databricks SQL)**: A serverless SQL engine optimized for big data queries.

1.2 Compute Plane

The Compute Plane is where **actual data processing and storage** happens. Unlike the Control Plane (managed by Databricks), the Compute Plane is hosted **inside the customer's cloud environment** (AWS, Azure, or Google Cloud).

Key Components:

1. Compute Clusters:

- Databricks **processes large datasets efficiently** using clusters.
- Clusters **auto-scale** based on workload needs.
- Supports **Apache Spark** and **SQL Warehouses** for execution.

2. Storage & Data Lake: Databricks integrates with multiple **cloud storage solutions**.

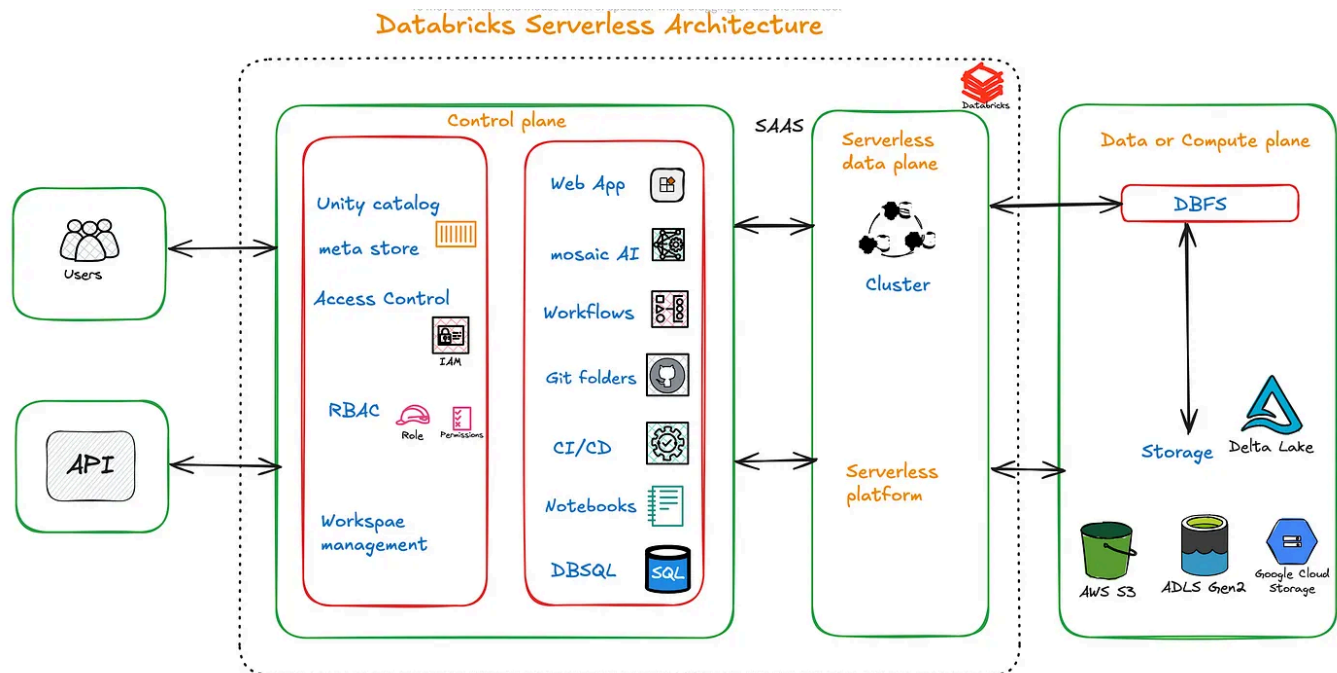
- **Delta Lake**: ACID transactions and time-travel queries.
- **AWS S3**: Cloud storage for structured/unstructured data.
- **Azure ADLS Gen2**: Microsoft's cloud data storage.

- **Google Cloud Storage:** Scalable storage for analytics.

2.Databricks Serverless Architecture

In serverless mode, Databricks manages both the control and compute planes. You don't need to handle clusters or storage infrastructure, as Databricks fully hosts the environment.

Databricks Serverless Architecture where compute is managed by Databricks



1.Serverless & Auto-Scaling Capabilities

Databricks automatically provisions resources based on demand, optimizing cost and performance.

- **Auto-scaling** adjusts compute power dynamically.

- **Serverless SQL** runs queries without managing infrastructure.

2.DBFS vs Cloud Object Storage

- In standard architecture, **Databricks File System (DBFS)** is used as an abstraction over cloud storage.
- In serverless mode, **cloud object storage (S3, ADLS, or GCS)** is directly used, improving efficiency.

3. Databricks Workflow: How It All Connects

This section explains the **Databricks workflow**, connecting all components.

1. **Users & APIs** interact with Databricks via the **web UI or API**.
2. The **Control Plane** manages **authentication, job scheduling, and metadata storage**.
3. The **Compute Plane** provisions clusters and processes data when jobs run.
4. The **results** are stored in **Delta Lake, AWS S3, ADLS, or Google Cloud Storage**.
5. Users analyze data using **SQL queries, Notebooks, or AI models**.

4. Benefits of Databricks Architecture

4.1 Scalability

- Supports auto-scaling of clusters based on workload.
- Handles massive datasets efficiently.

4.2 Unified Data Management

- Combines data engineering, machine learning, and business intelligence in one platform.

4.3 Security & Compliance

- Provides RBAC, IAM, and data governance with Unity Catalog.

4.4 Cost Efficiency

- Optimizes resources with serverless SQL & auto-scaling.
- Reduces operational overhead with managed infrastructure.

4.5 Multi-Cloud Support

- Works on AWS, Azure, and Google Cloud.
- Ensures flexibility in cloud adoption.

4.6 High-Performance Computing

- Uses distributed computing with Apache Spark.
- Provides in-memory caching for faster queries.

4.7 Simplified Collaboration

- Supports real-time teamwork via shared notebooks.
- Works with multiple languages (Python, Scala, SQL, R).

5. Use Cases of Databricks

Databricks is widely used in different industries:

1. Financial Services:

- Fraud detection using ML models.
- Real-time risk analysis in stock markets.

2. Healthcare & Life Sciences:

- Genomic data processing for research.
- Predictive analytics for disease detection.

3. Retail & E-Commerce

- Customer behaviour analysis and recommendation engines.
- Supply chain optimization.

4. Manufacturing

- IoT data processing for predictive maintenance.
- Quality control analysis using AI/ML.

5. Telecommunications

- Network performance monitoring.
- Churn prediction for customer retention.

6. Conclusion

Databricks provides a **robust and scalable architecture** for managing **big data workloads** efficiently.

- The **Control Plane** manages access, metadata, and job scheduling.
- The **Compute Plane** executes data processing using clusters and SQL Warehouses.
- **Cloud storage integration, AI tools, and SQL analytics** simplify big data processing.

Whether using **standard or serverless architecture**, Databricks offers a **scalable, secure, and cloud-native analytics solution**.

For data engineers, analysts, and AI/ML practitioners, Databricks is a **powerful tool** for handling complex data challenges.

Databricks Course | Databricks Training | Databricks Online Training — AccentFuture

Master Databricks with expert-led training on big data, AI, and ML, covering Apache Spark, real-time analytics, and cloud integration (AWS, Azure, Google Cloud). Gain hands-on experience and advance your career with our industry-focused [Databricks Training](#)!



Enroll Now: <https://www.accentfuture.com/enquiry-form/>



Call Us: +91-9640001789



Email Us: contact@accentfuture.com



Visit Us: [AccentFuture](#)

Related Posts:

1. [Building End-to-End Data Pipeline ADLS to Databricks to Snowflake](#)
2. [Revolutionize Data Ingestion with Databricks Auto Loader: Advanced Automation for Modern Data Engineer](#)
3. [Mastering Medallion Architecture: Build Scalable Real-Time Data Pipelines with Databricks](#)

Databricks

Architecture

Apache Spark

Cloud

Analytics



Written by AccentFuture

23 followers · 1 following

Follow

Master data engineering with real-time ETL, Apache Airflow, Spark, Databricks, cloud platforms like AWS, Azure, 24x7 support, and job assistance for career growth

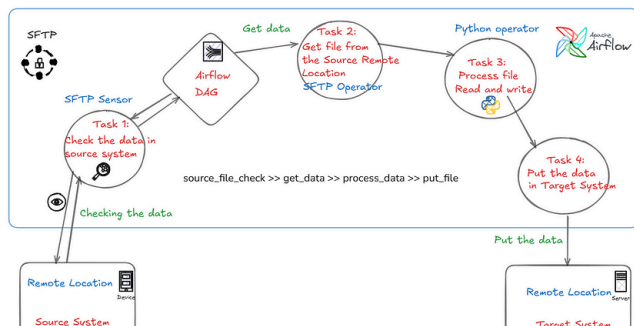
No responses yet




Abbey

What are your thoughts?

More from AccentFuture

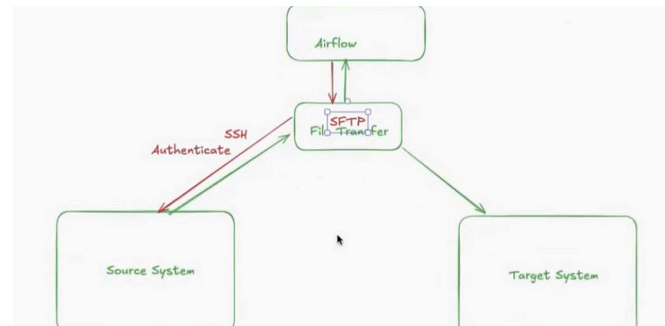



 AccentFuture

Automate File Transfers with Airflow and SFTP—Step-by-Step...

Workshop Agenda

Mar 31

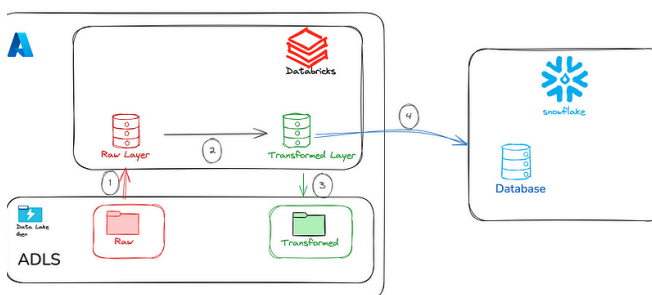


 AccentFuture

Hands-On Guide to Setting Up SSH Connectivity in Apache Airflow

Workshop Agenda

Feb 5

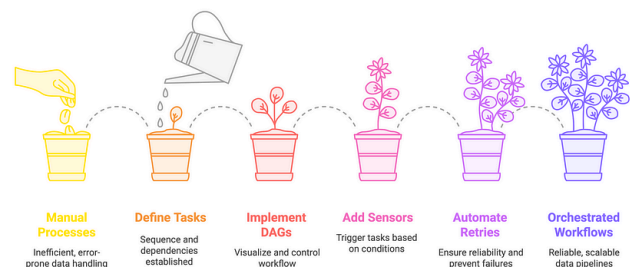



 AccentFuture

Building an End-to-End Data Pipeline: ADLS to Databricks to...

Workshop Agenda

From Chaos to Control: Mastering Data Workflows with Apache Airflow



 AccentFuture

Apache Airflow Explained: Workflow Orchestration for...

In today's data-driven world, managing complex workflows isn't just a backend task...

Dec 23, 2024

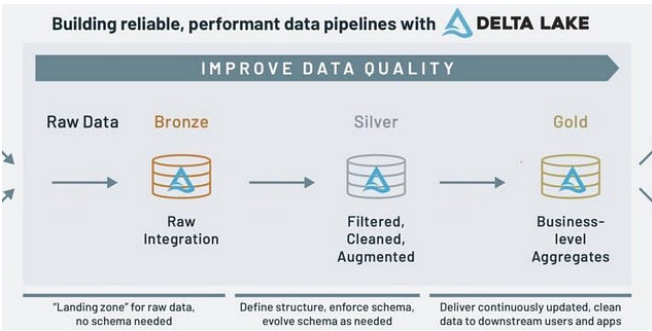
2

May 25

1

See all from AccentFuture

Recommended from Medium



 Badrish Davay

Medallion Architecture in Databricks

Implementing Medallion Architecture in Databricks: A Comprehensive Guide

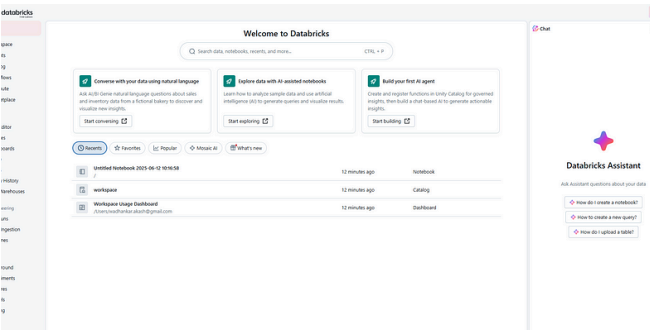
Jan 27


1

3d ago

20

2



 In Towards Data Engine... by THE BRICK LEARNI...

Databricks is Now Free for Learners—No Cloud Account...

In a major win for the data learning community, Databricks has just made its...

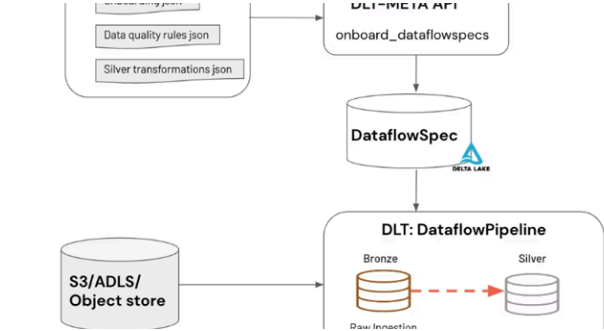


Venkat Dasari

Modular Ingestion on Databricks: Building a Scalable, Config-Drive...

A story about reusable pipelines, YAML-powered validation, and why configuration...

Apr 17 53

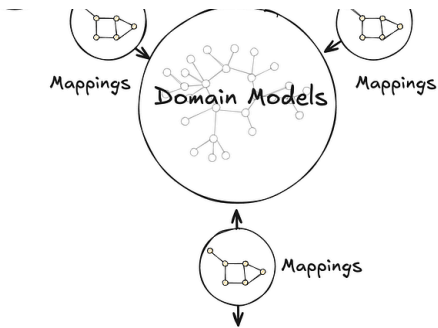


Varun Vemulapalli

Automating Data Pipelines at Scale: A Deep Dive into Metadata-...

In large enterprises, you'll often find dozens—if not hundreds—of similarly structured data...

Jun 1 1

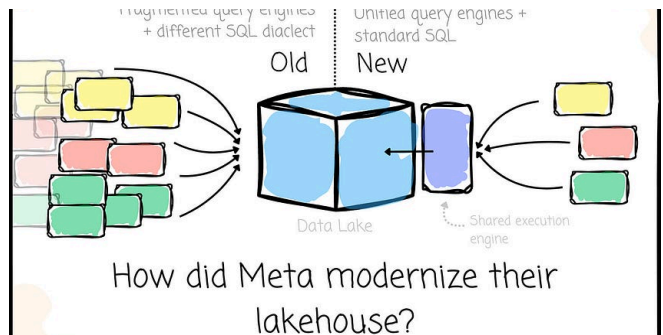


In Netflix TechBlog by Netflix Technology Blog

Model Once, Represent Everywhere: UDA (Unified Data...

Introducing UDA, the knowledge-graph-based architecture that translates conceptu...

3d ago 600 11



In Data Engineer Things by Vu Trinh

How did Meta modernize their lakehouse?

The new approach enabled Meta to innovate faster.

3d ago 198 1



See more recommendations