Databricks Architecture Deep Dive

karthick selvam

May 10, 2025

Contents

1 Databricks Architecture Overview

The Databricks platform architecture is fundamentally divided into two main components:

Core Architectural Components

- **Control Plane**: Handles backend services and platform management
- **Compute Plane**: Where actual data processing occurs

1.1 Control Plane Components

The Control Plane contains all essential backend services that power the Databricks platform:

Control Plane Elements

- **Databricks Web UI**: Browser-based interface for user interaction with workspaces, notebooks, and clusters
- **Cluster Manager**: Orchestrates creation, scaling, and termination of compute resources
- **Unity Catalog**: Centralized governance system managing data access controls and permissions
- **Control Plane Storage**: Stores workspace metadata including:
  - Notebook versions and revisions
  - Job run histories and configurations
  - Cluster settings and logs
  - Command execution histories

1.2 Compute Plane Options

Databricks offers two distinct compute models:

Classic Compute

- Provisioned within **customer's cloud subscription**
- Virtual machines deployed in your AWS/Azure/GCP account
- Requires network configuration between control plane and your VMs
- Customer shares infrastructure management responsibilities
- Typical startup time: 5-10 minutes

Serverless Compute (Introduced 2020)

- Runs on **Databricks subscription** resources
- Uses pre-allocated VM pools maintained by Databricks
- No customer-side infrastructure management
- Faster startup (typically under 1 minute)
- Currently supports specific workload types (SQL warehouses, etc.)

1.3 Workspace Cloud Storage

Automatically provisioned with each workspace:

Storage Characteristics

- **Location**: Customer's cloud subscription
- **Azure**: Azure Data Lake Storage Gen2
- **AWS**: S3 bucket
- **GCP**: Google Cloud Storage
- **Contents**:
  - System logs and diagnostics
  - Notebook revision history
  - Job execution details
  - Temporary processing data
- **Important**: Deleted automatically with workspace termination

2 Resource Location Summary

Component	Location	Management Responsibility
Control Plane	Databricks Subscription	Fully managed by Databricks
Serverless Compute	Databricks Subscription	Fully managed by Databricks
Classic Compute	Your Cloud Subscription	Shared (You + Databricks)
Workspace Storage	Your Cloud Subscription	Managed by Databricks
Business Data	Your Cloud Storage	Customer managed

Table 1: Resource Locations and Management

3 Azure-Specific Implementation Details

For Azure deployments, additional resource organization exists:

Managed Resource Group

- Automatically created per workspace (named `databricks-rg-<workspace-id>`)
- Contains:
  - Classic compute VMs (when used)
  - Workspace storage account (ADLS Gen2)
  - Azure Managed Identity
  - Unity Catalog Access Connector (if enabled)
  - Virtual Network + Security Group
- **Crucial Note**: Still in *your* subscription, just managed by Databricks

4 Practical Demonstration

To view these resources in Azure Portal:

1. Navigate to your Azure Databricks service
2. Locate the automatically created Managed Resource Group
3. Within this group you'll find:
    - Storage account for system data
    - Network security components
    - Identity management resources
4. Classic compute VMs appear here when clusters are created

5 Key Takeaways

Architecture Fundamentals

- Clear separation between control (Databricks) and compute (shared) planes
- Two compute models with different management and location characteristics
- Workspace storage is temporary/system-focused in your cloud account
- Business data should reside in separate, customer-managed storage
- Azure creates additional resource organization via Managed Resource Groups

End of Architecture Overview