BUSINESS INTELLIGENCE AND DATA ANALYTICS

Institute of Analytics Recognized Programme

University of Petra

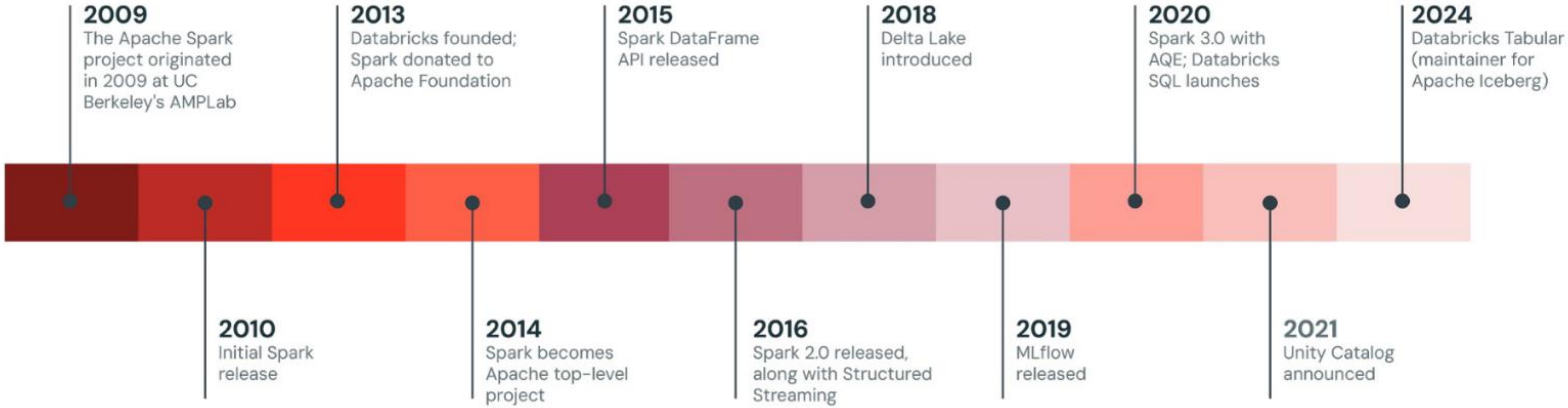
307401
Big Data and Data Warehouses

Introduction to Databricks

Introduction to Databricks

- Databricks is a cloud-based data and AI platform designed to unify data engineering, data science, and machine learning workflows.
- It was founded in 2013 by the original creators of Apache Spark and has since become one of the leading platforms for large-scale data processing and analytics.
- Built on top of Apache Spark, Databricks simplifies the process of working with big data by providing a collaborative workspace that integrates with popular cloud providers such as AWS, Azure, and Google Cloud.
- Its goal is to remove the operational complexity of big data and AI systems so that organizations can focus on generating insights and building innovative applications.
- The platform supports data ingestion, transformation, and advanced analytics through features like Delta Lake for reliable data lakes, MLflow for machine learning lifecycle management, and the Unity Catalog for data governance. Databricks provides a serverless and scalable environment where teams can run notebooks, SQL queries, streaming jobs, and machine learning models seamlessly.
- It is widely used by enterprises for building modern data architectures, enabling real-time analytics, and accelerating AI development.

Spark Timeline



2009 The Apache Spark project originated in 2009 at UC Berkeley's AMPLab

2010 Initial Spark release

2013 Databricks founded; Spark donated to Apache Foundation

2014 Spark becomes Apache top-level project

2015 Spark DataFrame API released

2016 Spark 2.0 released, along with Structured Streaming

2018 Delta Lake introduced

2019 MLflow released

2020 Spark 3.0 with AQE; Databricks SQL launches

2021 Unity Catalog announced

2024 Databricks Tabular (maintainer for Apache Iceberg)

Databricks Architecture

Databricks architecture is designed to run at cloud scale and is structured around three main layers, Storage, Compute, and Management.

1. **Storage Layer**: Databricks does not store data itself but integrates with cloud-native storage services like AWS S3, Azure Data Lake Storage (ADLS), and GCP GCS. Databricks File System (DBFS) is a virtual abstraction on top of these storages for seamless interaction. Delta Lake adds ACID transactions and schema enforcement, making the storage layer reliable.
2. **Compute Layer** – This layer consists of clusters that run Spark jobs. A cluster has a Driver node (coordinate execution) and Worker nodes (perform distributed tasks). Clusters can be on-demand, autoscaled, and optimized with Databricks Runtime. The Photon execution engine improves SQL performance.
3. **Management Layer** – This is where Databricks shines. It provides a web-based workspace for collaboration, security features like role-based access control, integration with Identity providers (AD, Okta), and tools like Jobs, MLflow, and Unity Catalog. It also provides notebooks for collaborative development and Jobs for production scheduling.

Overall, the architecture separates data storage from compute, ensuring scalability and flexibility. Users interact through the workspace, submit jobs to clusters, and store results in the storage layer. This modular design makes Databricks powerful for both batch and streaming analytics, machine learning pipelines, and ad-hoc exploration.

Databricks Platform

- **Workspace** – The Workspace is the central collaborative environment in Databricks where users organize notebooks, files, libraries, and folders. It supports team-based development by enabling shared access, versioning, and structured project organization across data engineering, analytics, and machine learning workflows.
- **Recents** – Recents provides quick access to notebooks, queries, dashboards, and other assets that were recently opened or modified. It improves productivity by reducing navigation time and helping users resume work efficiently.
- **Catalog** – Catalog (Unity Catalog) is the centralized governance layer for managing data and AI assets, including tables, views, volumes, models, and functions. It enforces fine-grained access control, lineage tracking, and auditing across workspaces.
- **Jobs & Pipelines** – Jobs & Pipelines is used to orchestrate production workloads such as scheduled jobs, batch processing, and continuous data pipelines. It enables reliable execution, monitoring, retries, and dependency management for end-to-end workflows.
- **Compute** – Compute manages the clusters and SQL warehouses that execute Databricks workloads. It allows users to configure performance, scalability, cost controls, and security for interactive and automated workloads.

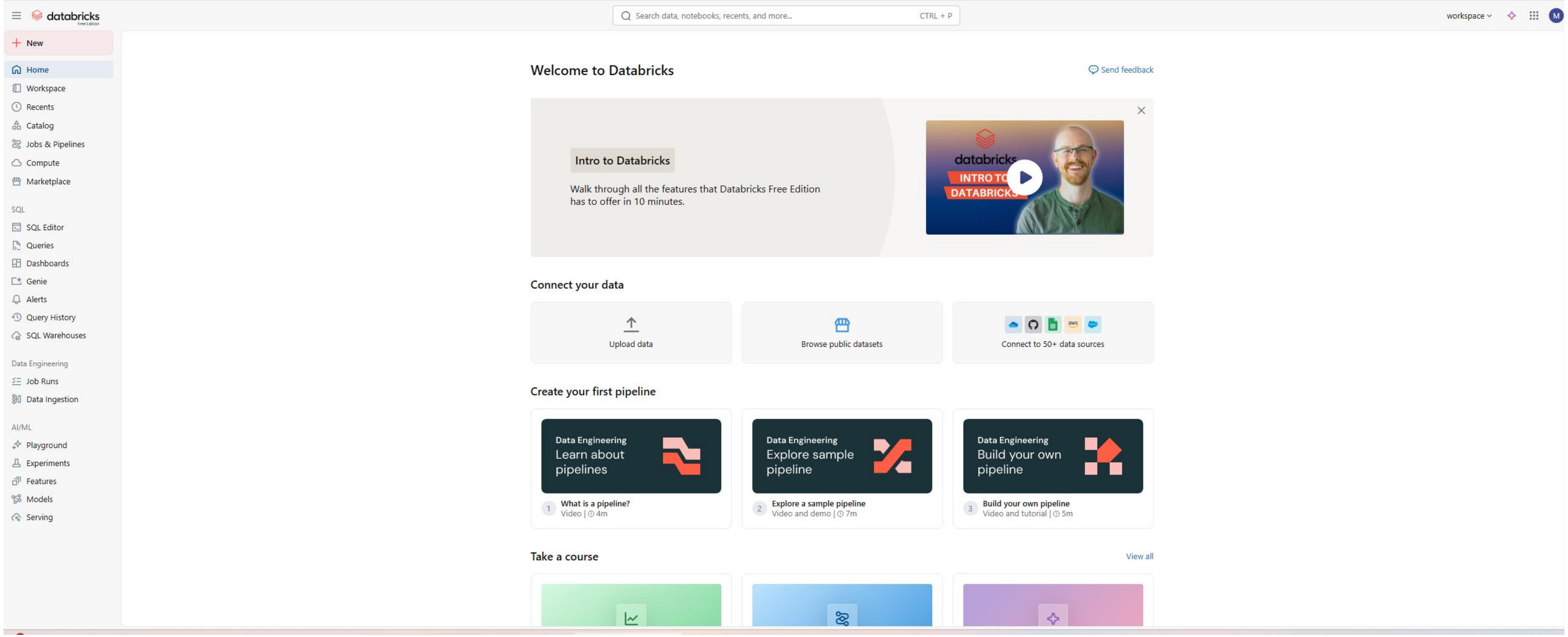
Databricks Platform

- **Marketplace** – Marketplace provides access to third-party and partner data products, machine learning models, and solution accelerators. It enables organizations to discover, evaluate, and integrate external assets securely.
- **SQL Editor** – SQL Editor is an interactive environment for writing and executing SQL queries against data stored in Databricks. It is optimized for analytics, exploration, and collaboration with built-in visualization support.
- **Queries** – Queries stores and organizes saved SQL queries for reuse and sharing. It enables teams to standardize analytics logic and maintain consistency across reports and dashboards.
- **Dashboards** – Dashboards allow users to create interactive, shareable visualizations based on SQL queries. They are designed for business intelligence use cases and can be shared with stakeholders securely.
- **Genie** – Genie is an AI-powered assistant that enables natural-language interaction with data. It helps users ask questions, generate queries, and explore insights without requiring deep SQL expertise.
- **Alerts** – Alerts monitor query results and data conditions, automatically notifying users when thresholds or rules are met. They support proactive data quality checks and operational monitoring.

Databricks Platform

- **Query History** – Query History provides a searchable log of executed SQL queries, including performance metrics and execution details. It is useful for debugging, optimization, auditing, and cost analysis.
- **SQL Warehouses** – SQL Warehouses are optimized compute resources dedicated to SQL analytics and BI workloads. They provide fast, scalable, and cost-efficient query execution with concurrency controls.
- **Job Runs** – Job Runs displays execution history and status for scheduled and triggered jobs. It provides detailed logs, metrics, and error information for troubleshooting and operational oversight.
- **Data Ingestion** – Data Ingestion tools support loading data from streaming and batch sources into Databricks. They enable scalable, reliable ingestion with built-in support for incremental processing and schema evolution.
- **Playground** – Playground is an experimental environment for testing and interacting with AI capabilities, including generative AI and large language models. It allows rapid prototyping without impacting production assets.
- **Experiments** – Experiments (MLflow Experiments) track machine learning runs, parameters, metrics, and artifacts. They enable reproducibility, comparison, and collaboration across ML development cycles.
- **Features** – Features (Feature Store) manages curated, reusable machine learning features. It ensures consistency between training and inference while supporting governance and lineage.
- **Models** – Models (MLflow Model Registry) provides centralized management of machine learning models, including versioning, staging, approvals, and lifecycle tracking. It supports collaboration and controlled deployment.
- **Serving** – Serving enables real-time and batch deployment of machine learning and foundation models as scalable endpoints. It provides low-latency inference, monitoring, and integration with applications.

Databricks Free Account



Welcome to Databricks

Intro to Databricks

Walk through all the features that Databricks Free Edition has to offer in 10 minutes.

Connect your data

Create your first pipeline

Data Engineering
Learn about pipelines

1 What is a pipeline?
Video | 4m

Data Engineering
Explore sample pipeline

2 Explore a sample pipeline
Video and demo | 7m

Data Engineering
Build your own pipeline

3 Build your own pipeline
Video and tutorial | 5m

Take a course

Databricks Free Edition

Databricks Free Edition (replacing Community Edition) — good for students, hobbyists, learning / prototyping.)

Databricks Free Trial — gives you temporary credits and access to more features. Valid for 14 days.

Steps to create a free Databricks account / workspace

Here are the general steps. Depending on your region or cloud provider, some details might differ.

1. Go to the Databricks website and find **Databricks Free Edition** or “Try Databricks for Free”.

2. Click **Sign up** for Free Edition, or start the Free Trial.

3. Choose your signup method:

You can often sign up just with an email (“Express signup”).

Or use your existing cloud provider account (e.g. AWS) if you want to link storage / compute resources.

4. Fill in registration details (name, email, password, etc.). If needed, verify your email.

5. (If Free Trial) You might need to enter payment method or link a cloud account so that after the trial expires, there is a way to transition if you decide to continue.

6. A workspace is created for you. In Free Edition, it’s serverless and quota-limited.

7. Once you’re inside the workspace, you can:

- Create notebooks
- Run code
- Explore tools like SQL editor, dashboards
- Work with data stored or uploaded into the workspace.

Limitations / things to watch out for

- Free Edition has **quotas** and **limitations** (e.g. on compute, storage, features).
- Free Trial gives you more access, but only for \~14 days. After that, if you don’t upgrade, some services might become unavailable.
- In some signup flows, especially via cloud providers, you may incur charges for cloud resources (storage, VM use) if those are outside what the credit allows. Always check what is free vs what might cost.

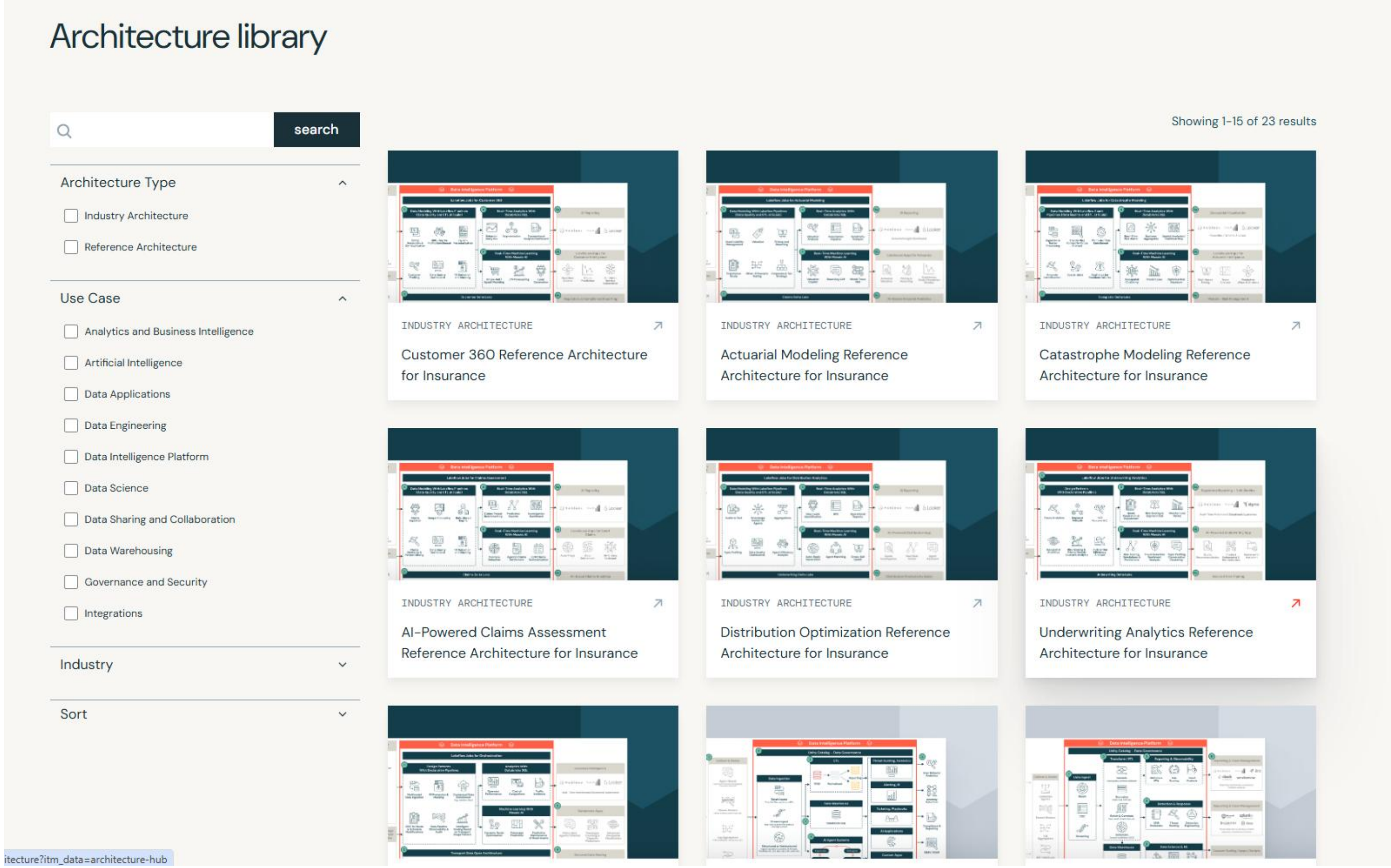
Important Databricks Resources

Databricks Training

- Databricks Fundamentals Course
- Getting Started with Lakehouse Architecture
- Get Started With Data Engineering on Databricks
- Redefine what’s possible with generative AI

Databricks Architectures Center

- https://www.databricks.com/resources/architectures



Architecture library

search

Showing 1-15 of 23 results

Architecture Type

☐ Industry Architecture
☐ Reference Architecture

Use Case

☐ Analytics and Business Intelligence
☐ Artificial Intelligence
☐ Data Applications
☐ Data Engineering
☐ Data Intelligence Platform
☐ Data Science
☐ Data Sharing and Collaboration
☐ Data Warehousing
☐ Governance and Security
☐ Integrations

Industry

Sort

INDUSTRY ARCHITECTURE ↗
Customer 360 Reference Architecture for Insurance

INDUSTRY ARCHITECTURE ↗
Actuarial Modeling Reference Architecture for Insurance

INDUSTRY ARCHITECTURE ↗
Catastrophe Modeling Reference Architecture for Insurance

INDUSTRY ARCHITECTURE ↗
AI-Powered Claims Assessment Reference Architecture for Insurance

INDUSTRY ARCHITECTURE ↗
Distribution Optimization Reference Architecture for Insurance

INDUSTRY ARCHITECTURE ↗
Underwriting Analytics Reference Architecture for Insurance

itecture?itm_data=architecture-hub