Top 30 Databricks FAQs

**What is Databricks?**

Databricks is a cloud-based unified data and AI platform that integrates data engineering, data science, machine learning, and analytics.

**How is Databricks different from Apache Spark?**

Databricks is built on Apache Spark but provides additional tools (like notebooks, MLflow integration, Delta Lake, etc.) and optimizations for enterprise use.

**Which cloud platforms support Databricks?**

Databricks is available on AWS, Microsoft Azure, and Google Cloud Platform.

**What is a Databricks Workspace?**

A workspace is a collaborative environment in Databricks for accessing notebooks, libraries, data, and managing clusters.

**What is a Databricks notebook?**

A collaborative document where you can run code (Python, SQL, R, Scala), visualize data, and document analysis interactively.

**What is Delta Lake?**

Delta Lake is an open-source storage layer that brings ACID transactions, schema enforcement, and versioning to data lakes.

**What is a Delta Table?**

A table stored in Delta Lake format that allows time travel, schema evolution, and optimized performance.

**What are Bronze, Silver, and Gold layers in Databricks?**

A medallion architecture where: Bronze = raw data, Silver = cleaned/filtered data, Gold = business-level aggregates.

**How does Databricks handle streaming data?**

Using Structured Streaming, Delta Live Tables (DLT), and Auto Loader for ingesting and processing data in near real-time.

**What is Auto Loader in Databricks?**

A feature that automatically detects new files in cloud storage and processes them incrementally and efficiently.

**What is MLflow?**

An open-source tool integrated with Databricks for tracking experiments, packaging models, and managing the ML lifecycle.

**Can you use scikit-learn, TensorFlow, or PyTorch in Databricks?**

Yes, Databricks supports all major ML libraries and frameworks.

**How does model serving work in Databricks?**

Databricks offers native Model Serving to deploy models as REST APIs.

**What is Feature Store in Databricks?**

A centralized repository to store, share, and reuse features for ML models across teams.

**What is Hyperopt in Databricks?**

A Python library for distributed hyperparameter tuning, integrated natively in Databricks notebooks.

**Can Databricks be used as a data warehouse?**

Yes, Databricks SQL provides a fully managed SQL experience for analytics, BI, and dashboarding.

**What is Databricks SQL?**

A SQL-native interface in Databricks for querying data, visualizing results, and creating dashboards.

**Can I connect Databricks to Power BI or Tableau?**

Yes, using JDBC/ODBC connectors and Databricks SQL endpoints.

**How do I optimize query performance in Databricks?**

Use Delta Lake, Z-Ordering, Caching, Photon engine, and query optimization techniques.

**What is the Photon engine?**

A vectorized query engine built in C++ to speed up SQL workloads on Databricks.

**How do you use Git with Databricks?**

You can connect your notebook or project with Git repositories (e.g., GitHub, GitLab) for version control.

**Does Databricks support CI/CD?**

Yes, using Databricks CLI, REST APIs, and integration with CI/CD tools (e.g., GitHub Actions, Azure DevOps).

**What is Unity Catalog?**

A unified governance solution for managing permissions, catalogs, schemas, and tables across all Databricks workspaces.

**Can I run scheduled jobs in Databricks?**

Yes, using Databricks Jobs, you can schedule notebooks, workflows, or pipelines to run at specified intervals.

**How do you monitor pipelines in Databricks?**

With job run logs, DLT monitoring, MLflow tracking, and Databricks SQL dashboards.

**What is a cluster in Databricks?**

A set of compute resources where you run notebooks, jobs, and applications.

**What is the difference between interactive and job clusters?**

Interactive clusters are used by notebooks. Job clusters are created automatically for running scheduled jobs.

**How does Databricks pricing work?**

Based on Databricks Units (DBUs) which depend on cluster size, type, and cloud provider costs.

**How do you secure data in Databricks?**

Using role-based access control, Unity Catalog, encryption, audit logs, and secure network configurations.

**Can Databricks be used for real-time fraud detection?**

Yes, with Structured Streaming, Delta Lake, and MLflow, Databricks can power real-time fraud detection systems.