Databricks & Spark Topics Covered and Overview

Databricks Overview

Topic	Description
Databricks	Unified analytics platform built on Spark for data, AI, and governance.
Editions	Different service tiers with varying security and feature sets.
High Level Components	Workspace, clusters, notebooks, jobs, and governance layers.
Core Components	Spark, Delta Lake, MLflow, Unity Catalog, and runtime.
Magic Commands	Notebook shortcuts for multi-language execution.
DB Utilities	Helper functions for file, secret, and workflow management.
GitHub Integration	Version control and CI/CD using GitHub repositories.

Spark Core

Topic	Description
About Spark	Distributed computing engine for large-scale processing.
Spark Installation	Local setup for development and testing.
RDD	Fault-tolerant distributed dataset.
RDD Operations	Transformations and actions on RDDs.
Narrow vs Wide Transformations	Operations with and without shuffling.
Shuffling	Data redistribution across partitions.
Spark Architecture	Driver, executors, cluster manager, and workers.
Driver	Controls job execution.
Executor	Runs tasks and stores data.
Executor Memory	Manages execution and storage memory.
Master/Worker/Cores	Cluster resource management.
Deployment Modes	Local, Standalone, and YARN.
Insurance Use Case	Real-time claim processing pipeline.

Databricks Spark & Delta Lake

Topic	Description
DataFrame	Structured distributed data abstraction.
Unity Catalog	Centralized governance system.
Binary Formats	Parquet, Avro, and ORC storage.
Partitioning	Improves query performance.
Temp Views	SQL-accessible temporary views.
API & SQL Mix	Combining DataFrame API with SQL.
Managed Tables	Databricks-managed storage.
External Tables	Externally stored tables.
Views	Virtual tables from queries.
Materialized Views	Physically stored query results.
Metadata Exploration	Schema and lineage inspection.
Delta Lake	ACID-compliant storage layer.
Schema Enforcement	Prevents invalid data.
Schema Evolution	Supports schema changes.
Time Travel	Accessing historical data.
Merge/Upsert	Efficient insert and update.
Transaction Log	Ensures consistency.

Optimization

Topic	Description
Optimize	Compacts small files.
Z-Order	Improves data skipping.
Vacuum	Deletes obsolete files.
CTAS	Creates optimized tables.
Clone	Creates deep or shallow copies.

Lakeflow Jobs

Topic	Description
Job Creation	Builds workflows.
Parameter Passing	Supplies runtime values.
Dependencies	Manages task order.
Parallel Execution	Runs tasks concurrently.
Branching & Loops	Supports conditional and iterative flows.

Databricks Management

Topic	Description
Workspace	Development environment.
Unity Catalog	Data governance system.
Security	Role-based access control.
Delta Sharing	Secure data sharing.
Service Principal	Secure Azure access.
DBUs	Billing unit measurement.
Policies	Resource governance rules.

Databricks UI Components

Topic	Description
SQL Editor	SQL development interface.
Dashboards	Visual analytics.
Genie	AI assistant.
Alerts	Automated notifications.
SQL Warehouse	Scalable SQL compute.

Spark Streaming

Topic	Description
Spark Streaming	Real-time data processing.
Auto Loader	Incremental file ingestion.
Micro-Batch	Batch-based streaming.
Watermarking	Late data handling.
Triggers	Execution scheduling.

Kafka & EventHub

Topic	Description
Kafka	Distributed messaging platform.
Kafka Components	Broker, producer, consumer, topic.
Kafka Architecture	Partitioned replicated logs.
Azure EventHub	Cloud streaming service.

Delta Live Tables

Topic	Description
DLT	Declarative ETL framework.
Live Tables	Streaming and batch tables.
Append Flow	Incremental ingestion.
CDC Flow	Automated change capture.
Validations	Built-in quality checks.
Fleet Use Case	Vehicle data processing.

CI/CD

Topic	Description
About CI/CD	Automates build, test, and deployment pipelines for faster and reliable releases.
Databricks Asset Bundle	Standardized packaging to version, deploy, and manage Databricks resources via CI/C
DAB Commands	CLI commands used to validate, deploy, and promote Databricks Asset Bundles across
GitHub Actions	Workflow automation tool to run CI/CD pipelines for Databricks using GitHub repositori