Azure Databricks Pocket Guide

Praveen Patel

Follow Me to Get Such More Like This

Workspace

The Azure Databricks Workspace is the collaborative environment where data scientists, data engineers, and analysts can work together. It contains notebooks, dashboards, libraries, and experiments. Users can organize their work in folders and manage access controls here.

Cluster

Clusters are groups of computation resources running Apache Spark. Clusters can be configured with different instance types, sizes, and autoscaling options. Users submit code to clusters, which execute jobs in parallel. You can have all-purpose clusters for interactive analysis or job clusters for scheduled workloads.

Notebook

Notebooks are interactive web-based documents that support multiple languages (Python, Scala, SQL, R). They allow combining live code, equations, visualizations, and narrative text. Notebooks facilitate exploration, visualization, and building ETL pipelines.

Job

Jobs are automated executions of notebooks or JARs, designed for running production workflows on a schedule or trigger. Databricks Jobs support retries, notifications, and dependency chaining for complex pipelines.

Databricks File System (DBFS)

DBFS is a distributed file system abstraction over Azure Blob Storage or Azure Data Lake Storage. It allows seamless access to files from notebooks and clusters with a unified namespace, supporting mount points for external storage.

Library

Libraries are packages or modules you install on your clusters to add functionality. This can include Python PyPI packages, Maven JARs, or custom wheel files. Libraries can be installed at cluster startup or attached dynamically.

Secret Management

Azure Databricks provides a secure mechanism for storing and managing secrets such as API keys, passwords, and tokens. Secrets can be accessed programmatically within notebooks without exposing them in code or logs.

Delta Lake

Delta Lake is an open-source storage layer that brings ACID transactions, scalable metadata handling, and data versioning to data lakes. It enables reliable and performant big data pipelines and supports time travel queries.

Jobs API

The Jobs API allows programmatic creation, management, and monitoring of Databricks jobs. It enables integration with CI/CD pipelines and custom orchestration tools.

Databricks Runtime

Databricks Runtime is an optimized Apache Spark environment customized by Databricks for performance, security, and compatibility with Azure services. Different runtime versions support different libraries and capabilities.

Databricks SQL

Databricks SQL is a serverless service designed to run SQL queries directly on your data lake with high concurrency and low latency, supporting BI and analytics workloads.

MLflow

MLflow is an open-source platform integrated into Databricks that manages the machine learning lifecycle including experimentation, reproducibility, deployment, and monitoring.

Widgets

Widgets are input controls (dropdowns, text boxes, etc.) in notebooks that allow parameterization and dynamic execution, enabling interactive reports and dashboards.

dbutils

dbutils is a set of utilities provided by Databricks to simplify common tasks such as file system access, secret retrieval, notebook workflow orchestration, and widget management.

Jobs Cluster

A job cluster is an ephemeral cluster spun up specifically to run a job and automatically terminated after the job completes, optimizing cost and resource usage.

Unity Catalog

A unified governance solution for all data and AI assets in Databricks, providing fine-grained access control, auditing, and data lineage across workspaces.

Autoloader

A Databricks feature for efficiently ingesting new data files from cloud storage in real time. It supports automatic file detection, schema inference, and incremental loading using Structured Streaming.

Job Clusters vs Interactive Clusters

Job clusters are ephemeral clusters spun up for jobs and terminated after completion, optimizing costs. Interactive clusters are long-running clusters for development and exploration.

Photon Engine

A native vectorized engine built by Databricks for faster query performance with improved CPU efficiency, available in Databricks Runtime.

Cluster Policies

Administrators can define cluster policies to enforce configurations and security standards, limiting what options users can set when creating clusters.

Pools

Pools reduce cluster start and auto-scaling times by maintaining a set of ready-to-use instances, improving overall efficiency.

Table Constraints

Constraints in Delta Lake such as NOT NULL, UNIQUE, and CHECK constraints that enforce data integrity.

Structured Streaming

A scalable and fault-tolerant stream processing engine built on Spark SQL that enables continuous data processing.

REST APIs

Databricks provides REST APIs for managing workspaces, clusters, jobs, secrets, libraries, and more, enabling automation and integration.

Cluster Init Scripts

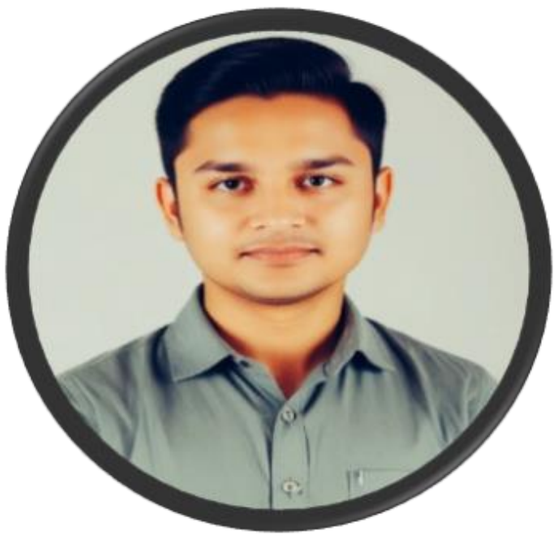
Scripts that run on cluster nodes at startup, used for custom configurations like installing system libraries or setting environment variables.

SQL Endpoints

Dedicated compute resources optimized for SQL workloads, used with Databricks SQL to run interactive and scheduled queries.

Delta Live Tables

Managed ETL framework that simplifies building reliable, maintainable, and testable data pipelines using declarative transformations.



Praveen Patel

Follow Me to Get Such More Post Like This