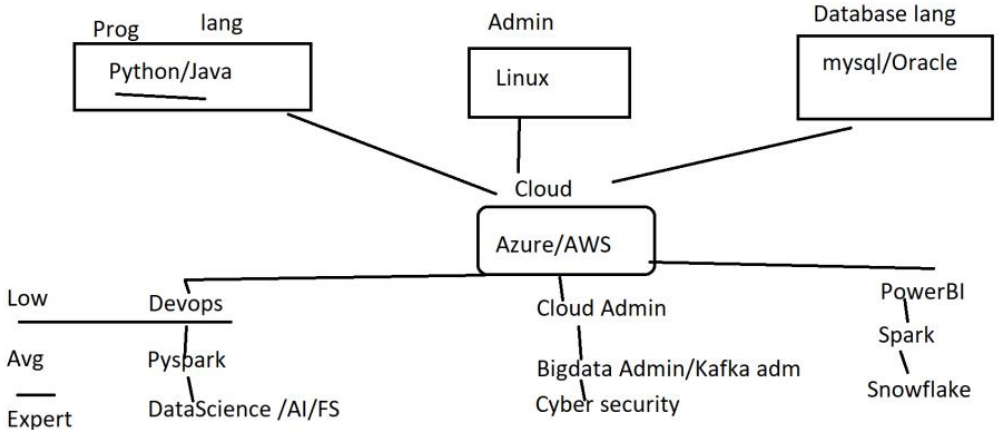
Mastering Databricks Data Engineering using AWS & Azure



Introduction to Big Data and Hadoop

- What is Big Data?
- What is Hadoop?
- What is Spark?
- What are NoSQL Databases?
- Difference Between Hadoop and Spark
- Common Big Data Problems
- Hadoop Ecosystem

AWS Introduction (40 Hours)

EC2

- Create Windows/Mac/Linux Servers
- Create a Sample Website
- Autoscaling
- Create and Use AMIs

Athena

- What is Serverless Computing?
- Process JSON and CSV Data with Athena
- Recommended Approaches

S3

- Store Data in S3
- Submit Commands in Client Mode
- Get Data from Various Sources and Store in S3
- S3 Bucket Policies

RDS

- Create Different Databases
- Create Sample Tables and Process Data
- Best Practices for Cost Optimization
- Practice Oracle and MySQL Using RDS

EMR

- Practice PySpark and Hive
- Create EMR Clusters and Process Data
- EMR vs EC2
- Hive Internals and Sample Programs
- Import Data from RDS to S3 Using Sqoop

Lambda & Boto3

- Access AWS Resources Using Boto3 from PyCharm
- Use Boto3 in Lambda Functions
- Integrate Lambda with Glue and Redshift
- Connect Boto3 with Services Like EC2, EMR, Glue, Redshift

CloudWatch

- How to Monitor Resources
- Debugging Application Failures
- Autoscaling Based on CloudWatch Metrics
- Usage Across AWS Services (EC2, RDS, Glue)

IAM (Identity and Access Management)

- Users, Groups, and Roles
- Custom Policies
- Importance of IAM Keys in Snowflake, Databricks, PyCharm Use Cases

Redshift

- Load and Process Data from S3
- SortKey and DistKey Optimization
- Redshift Architecture
- Compare Snowflake vs Redshift

Glue

- Process CSV and JSON Data Using Glue
- Retrieve Data from Athena Using Glue

- Use Crawlers and Execute PySpark/Scala Jobs
- Glue Architecture and Best Practices

Introduction to Spark

Spark Core

- Why Use Spark Instead of Hadoop?
- Importance of HDFS/YARN in Spark
- Spark Architecture
- Types of APIs: RDD, DataFrame, Dataset
- Use Cases for Spark
- Why Spark is Faster Than MapReduce
- In-Memory Processing in Spark

RDD Internals

- Properties of RDD: Immutability, Laziness, Fault Tolerance
- SparkContext, SQLContext, SparkSession Internals
- Create RDDs in Different Ways
- Transformations and Actions
- Debugging Transformations
- Spark Web UI

RDD Hands-On

- Map, FlatMap, Filter, Distinct
- ReduceByKey vs GroupByKey
- Spark-submit Examples
- 20 RDD Use Case Programs

Spark SQL

- Convert RDD to DataFrame
- Python DataFrame vs Spark DataFrame
- DataFrame Reader
- Processing Data in Different Formats: CSV, JSON, XML, Avro, ORC, Text, Parquet
- Database Integration: Oracle, MySQL, Sqoop vs Spark
- NoSQL Integration: HBase, Cassandra, MongoDB

PySpark Advanced Concepts

- Dataset API Importance
- Spark Memory Management
- Resource Optimization
- Spark Debugging with Client Mode and Web UI
- Automate Spark with Oozie and Airflow
- Spark-Snowflake Integration

Spark Streaming

Introduction to Spark Streaming

- Micro-Batch vs Stream Processing
- D-Stream API Internals
- Live Data Processing

Structured Streaming

- Real-World Examples
- Integration with Kafka
- Log Analysis
- Export to Databases
- Snowflake Integration

Apache Kafka

- Kafka Architecture
- Producer and Consumer APIs
- Integration with Spark
- End-to-End Workflow with AWS, Azure, Databricks, and Cloudera

Apache NiFi

- NiFi Internals
- Data Flow Examples (Local to S3, API to S3)
- Integration with Kafka and Spark
- Templates & most frequently used processors

Apache Airflow

- Airflow Installation in EC2
- Data Pipeline Creation
- DAG Management
- Airflow-Spark-Snowflake Integration

Introduction to Databricks

- Databricks vs Spark vs Snowflake
- Databricks Architecture
- Working in Databricks Workspace
- Using Databricks Notebooks

Databricks File System (DBFS)

- What is DBFS?
- DBFS Commands (mkdirs, cp, mv, head, put, rm, rmdir)
- Magic Commands (sh, fs, scala, python)

Databricks Utilities

- Credentials Utility
- FileSystem Utility
- Notebook Utility
- Secrets Utility
- Widgets Utility

Databricks Cluster Management

- Creating and Configuring Clusters
- Managing Clusters
- Starting, Terminating, and Deleting Clusters
- Cluster Information and Logs
- Types of Clusters: All-Purpose, Job Clusters
- Cluster Modes: Standard, High Concurrency, Autoscaling

Azure Overview

- Azure Databricks
- Azure VM & HDInsight vs EMR
- Azure Data Lake Storage (ADLS)
- Azure Blob Storage vs S3
- Azure SQL Database vs RDS
- Azure Active Directory vs IAM
- Azure Data Explorer
- Azure Stream Analytics vs SnowPipe
- Event Hub vs Kafka
- Azure Data Factory for Data Integration
- Azure Synapse vs Snowflake

Databricks Integration

- Integration with Azure Services:
- Blob Storage,
- Data Lake Storage Gen2,
- SQL Database, Synapse,
- Key Vault
- Triggers

Databricks Streaming API

- Introduction to Streaming
- Handling Bad Records, Regular Expression
- Streaming Data into Gen2 Lake and Tables

Databricks Lakehouse (Delta Lake)

- Data Lake vs Delta Lake

- Delta Lake Best Practices
- Delete, Update, Alter Tables
- Optimization Steps
- Handling SCD (Type 1 & Type 2)
- Deduplication and Streaming Data Handling

Databricks Unity Catalog

- Create Schema and Table Using Unity Catalog
- Access Controls, User Management, and Metastore
- Row-Level Access Control
- Masking Columns
- Roles, Users, and Groups
- Managing External Tables
- Lakehouse Federation

Databricks Workflows

- Introduction to Workflows
- Creating, Running, and Managing Jobs
- Scheduling and Monitoring Jobs
- Create Dependency Between Multiple Jobs

Delta Live Tables

- Introduction to Delta Live Tables
- Creating and Configuring Delta Pipelines
- Real-Time Streaming with Delta Live Tables
- Error Handling and Recovery in Delta Live Tables
- Delta Live Tables Best Practices