

Mastering Big Data Engineering

Hadoop • Hive • Spark • PySpark • Kafka • Airflow • Scala • AWS EMR

Technologies: HDFS • YARN • Hive • Sqoop • Apache Spark • PySpark • Scala • Spark SQL • Spark Streaming • Kafka • Oozie • Apache Airflow • AWS S3 • AWS EMR • AWS Glue • Amazon Athena • Parquet • ORC • Delta Lake

databrickstraining.in | Training & Placement Excellence
Phone: +91-8500002025 | **Email:** info@databrickstraining.in

★ ★ ★

COURSE OVERVIEW

This 90-hour industry-focused course gives you a complete hands-on journey through the Big Data ecosystem — from HDFS and Hive on Hadoop to high-performance PySpark and Scala pipelines, real-time Kafka streaming, workflow orchestration with Apache Airflow, and cloud-native Big Data on AWS. You will work through 18 structured modules, 3 real-world capstone projects, and an interview + certification preparation module.

Who Should Attend? Aspiring Big Data Engineers, Data Engineers moving from SQL/ETL backgrounds, Python developers transitioning into data, and professionals who want to modernise on-premise Hadoop skills with cloud-native approaches.

Prerequisites: Basic SQL knowledge, introductory Python programming, and comfort with Linux command line.

Training Format: Instructor-led sessions with hands-on labs, capstone projects, mock interviews, and certification guidance.

Module 1: Big Data Fundamentals & Hadoop Ecosystem | Duration: 3 Hours | Level: Beginner

What is Big Data — 5 Vs: Volume, Velocity, Variety, Veracity, Value

Big Data Use Cases — Real Industry Examples

- E-Commerce, Finance, Healthcare, Telecom, Social Media

Big Data Architecture — Batch vs Real-Time Processing

Hadoop Ecosystem Overview

- HDFS — Distributed Storage
- YARN — Resource Management
- MapReduce — Distributed Computation (Legacy)
- Hive — SQL on Hadoop
- Apache Spark — Unified Analytics Engine

Hadoop vs Traditional RDBMS — Key Differences

Modern Big Data Stack — Hadoop vs Cloud (AWS, Azure, GCP)

Big Data Career Roadmap — Roles & Skills in 2025

- Big Data Engineer, Data Architect, ML Engineer, Data Platform Engineer

Lab: Big Data Architecture Design — Real-World Use Case

Module 2: Hadoop — HDFS & YARN | Duration: 4 Hours | Level: Beginner-Intermediate

What is Hadoop & Hadoop Distributed File System (HDFS)

HDFS Architecture

- NameNode — Master metadata manager
- DataNode — Stores actual data blocks
- Secondary NameNode — Checkpoint helper

HDFS Data Blocks — Default Block Size (128MB) & Replication Factor (3)

HDFS Read & Write Pipeline — How Data Flows

HDFS High Availability — Active vs Standby NameNode (ZooKeeper-based)

HDFS Federation — Multiple NameNodes for scale

HDFS Commands — Essential CLI Operations

- `hdfs dfs -ls, -mkdir, -put, -get, -cat, -rm`
- `hdfs dfs -cp, -mv, -chmod, -chown`
- `hdfs fsck` — Check File System Health

YARN Architecture

- ResourceManager — Cluster resource master
- NodeManager — Per-node resource agent
- ApplicationMaster — Per-job orchestrator

YARN Resource Scheduling — FIFO, Capacity Scheduler, Fair Scheduler

HDFS Storage Policies — Hot, Warm, Cold, Archive

HDFS Balancer — Redistribute Data Across Nodes

Labs: HDFS CLI — Upload, Download, Manage Files; YARN — Submit Job, Monitor Resources

Module 3: Hive — SQL on Hadoop | Duration: 6 Hours | Level: Intermediate

What is Apache Hive & When to Use It

Hive Architecture — HiveServer2, Metastore, Driver, Compiler, Execution Engine

Hive Metastore — Central Schema Repository (MySQL/Derby backed)

HiveQL — SQL-Like Query Language

Hive Data Types — Primitive (INT, STRING, FLOAT) & Complex (Array, Map, Struct)

Managed Tables vs External Tables — Key Differences

Hive File Formats — TextFile, ORC, Parquet, Avro

ORC vs Parquet — When to Use What

Hive Partitioning — Static & Dynamic Partitioning

Hive Bucketing — Hash-Based Data Distribution

Hive Joins — Map Join, Reduce Join, Bucket Map Join

Hive Aggregations — GROUP BY, HAVING, ROLLUP, CUBE

Hive Window Functions — ROW_NUMBER, RANK, DENSE_RANK, LEAD, LAG

Hive Views & Materialized Views

Hive Transactions — ACID Tables (INSERT, UPDATE, DELETE)

Hive SerDe — Serialize & Deserialize Custom Formats

Hive Performance Optimization

- Partition pruning, Predicate pushdown
- Vectorized query execution
- Cost-Based Optimizer (CBO)
- TEZ execution engine vs MapReduce

Hive with ORC & Snappy Compression

HiveQL vs Spark SQL — When to Use What

Labs: Partitioned & Bucketed Hive Tables, Performance Optimization, ACID Transactions

Quiz: Hive Concepts — 20 Questions

Module 4: Sqoop — Data Ingestion from RDBMS | Duration: 3 Hours | Level: Intermediate

What is Apache Sqoop & Use Cases

Sqoop Architecture — Import & Export Flow

Sqoop Import — RDBMS to HDFS & Hive

- Full Table Import — `sqoop import`
- Custom SQL — `sqoop import --query`
- Incremental append — New records only
- Incremental lastmodified — Changed records
- Direct Hive import — `sqoop import --hive-import`
- Import as Parquet — `sqoop import --as-parquetfile`

Sqoop Export — HDFS to RDBMS

- Write back to database, Upsert pattern (`--update-mode`)

Sqoop Job — Save & Rerun Import Commands

Sqoop Metastore — Store Saved Jobs

Sqoop Parallel Import — Controlling Mappers (`-m` flag)

Sqoop Password Management — Secure Credentials

Sqoop with MySQL, PostgreSQL & Oracle

Sqoop Limitations & Modern Alternatives

- Sqoop vs Apache NiFi, Sqoop vs AWS DMS, Sqoop vs Spark JDBC

Labs: Full & Incremental Import MySQL → HDFS, Hive Partitioned Import, Export HDFS → MySQL

Module 5: Apache Spark — Core Fundamentals | Duration: 5 Hours | Level: Intermediate

What is Apache Spark & Why It Replaced MapReduce

- In-memory processing, DAG-based execution, 100x faster than MapReduce

Spark Architecture

- Driver — Orchestrates the job, hosts SparkContext
- Executors — Worker processes that run tasks
- Cluster Manager — YARN, Standalone, Kubernetes, Mesos

SparkContext vs SparkSession

RDD — Resilient Distributed Dataset

- Creating RDDs — `parallelize()` & `textFile()`
- Transformations — `map`, `filter`, `flatMap`, `reduceByKey`
- Actions — `collect`, `count`, `take`, `saveAsTextFile`
- Persistence — `cache()` & `persist()`

Lazy Evaluation & DAG — Directed Acyclic Graph

Spark Execution Model — Jobs, Stages, Tasks

Transformations vs Actions — Deep Dive

Spark Deployment Modes — Client vs Cluster Mode

`spark-submit` — Submit Jobs to Cluster

Spark Configuration — `spark-defaults.conf` & `SparkConf`

Labs: RDD WordCount & Log Analysis, `spark-submit` on YARN Cluster

Module 6: PySpark — DataFrame API | Duration: 6 Hours | Level: Intermediate

PySpark vs Scala Spark — When to Use What

SparkSession — Entry Point for DataFrame API

Creating DataFrames — From Files, RDD, Python Collections

Reading Data — CSV, JSON, Parquet, ORC, Avro

Writing Data — Overwrite, Append, Partitioned Writes

DataFrame Transformations

- select, filter, where, withColumn, drop, rename
- distinct, dropDuplicates, sort, orderBy
- groupBy, agg, pivot

DataFrame Actions — show, collect, count, take

Schema — StructType & StructField — Define Custom Schema

InferSchema vs Explicit Schema — Best Practices

Working with Nested JSON & Array Columns

- explode(), flatten(), struct()

Handling Null Values — dropna(), fillna(), coalesce()

String, Date & Math Functions

Conditional Expressions — when/otherwise

UDFs — User Defined Functions

Pandas UDFs — Vectorized UDFs

Working with Multiple DataFrames — union, unionByName

Reading & Writing to HDFS, S3, ADLS from PySpark

Labs: PySpark ETL Pipeline HDFS → Transform → Hive, Nested JSON Flatten → Parquet

Quiz: PySpark DataFrame — 20 Questions

Module 7: Spark SQL | Duration: 5 Hours | Level: Intermediate

What is Spark SQL & Unified Analytics Engine

Spark SQL Architecture — Catalyst Optimizer & Tungsten Engine

Creating Temporary Views & Global Temp Views

Running SQL Queries on DataFrames

Spark SQL vs HiveQL — Key Differences

Spark SQL with Hive Metastore Integration

Joins in Spark SQL — Inner, Left, Right, Full, Cross

Aggregations — GROUP BY, HAVING, ROLLUP, CUBE, GROUPING SETS

Window Functions in Spark SQL

- ROW_NUMBER, RANK, DENSE_RANK, NTILE

- LEAD, LAG — Offset Functions

- Running Totals & Moving Averages

Subqueries & CTEs — WITH Clause

Spark SQL with ORC & Parquet Files

Spark SQL on S3 — External Tables

Spark Catalog API — Manage Tables & Databases

Broadcast Variables & Accumulators

EXPLAIN Plan — Read & Understand Query Execution

Labs: Complex SQL with Window Functions, Hive Metastore External Tables, EXPLAIN Plan Optimization

Module 8: Spark Performance Optimization | Duration: 5 Hours | Level: Advanced

Partitioning Strategy — How Many Partitions & Optimal Sizing

Repartition vs Coalesce — When to Use Each

Broadcast Join — Avoid Shuffle for Small Tables

Sort Merge Join vs Broadcast Join

Data Skew — Causes & Salting Technique

Adaptive Query Execution (AQE)

- Auto Broadcast Join

- Auto Partition Coalescing

- Skew Join Optimization

Caching & Persistence — StorageLevel Options

- MEMORY_ONLY, MEMORY_AND_DISK, DISK_ONLY

Kryo Serialization — Faster than Java Serialization

Small File Problem — Causes & Solutions

Predicate Pushdown & Column Pruning

Reading Spark UI — Jobs, Stages, Tasks, Timeline

Spill to Disk — Causes & How to Fix

Memory Management — Executor Memory Tuning

GC Tuning — Reduce Garbage Collection Pauses

File Size Optimization — Compact Small Files

Labs: Fix Data Skew with Salting, Spark UI Bottleneck Analysis, AQE Before & After

Module 9: Scala for Spark | Duration: 5 Hours | Level: Intermediate

Why Learn Scala for Big Data Engineering

Scala vs Python for Spark — Performance & Use Cases

Scala Basics for Data Engineers

- val vs var — Immutable & Mutable Variables
- Data Types — Int, String, Double, Boolean
- Collections — List, Array, Map, Set, Tuple
- Functions — def, Anonymous Functions, Higher-Order Functions
- Pattern Matching — Scala's switch on steroids
- Case Classes — Typed Data Structures
- Option — Handle Null Values Safely
- For Comprehensions

Scala with Spark

- SparkSession in Scala
- DataFrame & Dataset API in Scala
- Typed Dataset — Compile-Time Type Safety
- Reading & Writing Files in Scala Spark
- Transformations & Actions in Scala
- Writing Spark Jobs in Scala — Best Practices

Building Spark JAR — sbt & Maven

Submitting Scala Spark JAR to YARN Cluster

Labs: PySpark ETL Rewritten in Scala (Side by Side), Build & Submit Scala JAR on YARN

Module 10: Spark Streaming — Real-Time Processing | Duration: 5 Hours | Level: Advanced

What is Spark Streaming & Real-Time Architecture

DStream API — Legacy Streaming (Overview Only)

Structured Streaming — Modern Streaming API

Structured Streaming Architecture

- Micro-Batch vs Continuous Processing
- Input Sources — Kafka, Files, Socket
- Output Sinks — HDFS, Kafka, Console, Delta

Reading Streaming Data from Kafka

Reading Streaming Data from Files — Auto Loader Style

Writing Streaming Output — Append, Update, Complete Mode

Trigger Modes — ProcessingTime, Once, Continuous, AvailableNow

Watermarking — Handle Late-Arriving Data

Stateful Streaming — Running Aggregations

Checkpointing & Fault Recovery

Streaming Joins — Stream-Stream & Stream-Static

Streaming Deduplication

Structured Streaming with Kafka — End to End

Kafka Architecture Recap — Topics, Partitions, Offsets

Exactly-Once Semantics — Kafka + Structured Streaming

Labs: Kafka → Spark → HDFS Pipeline, Watermarking Late Data, Windowed Word Count

Quiz: Spark Streaming — 20 Questions

Module 11: Apache Kafka — Real-Time Messaging | Duration: 3 Hours | Level: Intermediate

What is Apache Kafka & Event-Driven Architecture

Kafka Architecture

- Brokers — Kafka servers that store messages
- Topics — Named message streams
- Partitions — Parallel log segments
- ZooKeeper / KRaft — Cluster coordination

Producers & Consumers

Consumer Groups — Parallel Processing

Kafka Offsets — Track Message Position

Kafka Retention — Message Storage Policy

Producing Messages from Python — kafka-python library

Consuming Messages from Python

Kafka with Spark Structured Streaming

Kafka Connect — Import/Export Data Without Code

- Source connectors — MySQL, PostgreSQL, S3
- Sink connectors — HDFS, Elasticsearch, Snowflake

Kafka vs Pub/Sub vs Kinesis — Comparison

Labs: Produce & Consume Messages with Python, Kafka → Spark Structured Streaming → HDFS

Module 12: Apache Oozie — Hadoop Workflow Scheduler | Duration: 3 Hours | Level: Intermediate

What is Apache Oozie & Workflow Orchestration

Oozie vs Airflow — When to Use What

Oozie Workflow — Sequential & Parallel Actions

Oozie Coordinator — Schedule Workflows on Time & Data

Oozie Bundle — Group Multiple Coordinators

Oozie Actions

- Hive Action — Run HiveQL Script
- Sqoop Action — Run Sqoop Import/Export
- Spark Action — Submit Spark Job
- Shell Action — Run Shell Scripts
- Java Action — Run Java Program
- Email Action — Send Notifications

Oozie XML Workflow Definition

Oozie Web Console — Monitor Workflow Runs

Oozie with HDFS — Trigger on Data Arrival

Oozie Limitations & Why Airflow Replaced It

Labs: Oozie Workflow (Sqoop → Hive → Email Alert), Oozie Coordinator Daily Pipeline

Module 13: Apache Airflow — Modern Pipeline Orchestration | Duration: 5 Hours | Level: Intermediate-Advanced

What is Apache Airflow & Why It Replaced Oozie

Airflow Architecture — Webserver, Scheduler, Executor, Metadata DB

DAG — Directed Acyclic Graph Concepts

Writing DAGs in Python — Basics to Advanced

Airflow Operators for Big Data

- BashOperator — Shell Commands
- PythonOperator — Python Functions
- HiveOperator — Run HiveQL
- SparkSubmitOperator — Submit Spark Jobs to YARN
- SqoopOperator — Run Sqoop Commands
- SSHOperator — Remote Command Execution
- EmailOperator — Send Alerts on Failure
- HttpSensor — Wait for API Response
- FileSensor — Wait for File on HDFS/S3

Task Dependencies — set_upstream, set_downstream, >> operator

XComs — Pass Data Between Tasks

Airflow Variables & Connections — Store Config & Credentials

Airflow Hooks — Connect to Hive, HDFS, S3, MySQL

Trigger Rules — all_success, one_failed, all_done

Airflow Pools, SLA, Backfill

Airflow Monitoring — DAG Runs, Task Logs, Gantt Chart

Labs: Sqoop → Hive → Email DAG, SparkSubmitOperator on YARN, Dynamic DAGs from Config

Quiz: Airflow — 20 Questions

Module 14: AWS for Big Data Engineers | Duration: 4 Hours | Level: Intermediate

AWS Overview for Big Data Engineers

Amazon S3 — Object Storage for Big Data

- Buckets, Objects & Partitioning Strategy
 - S3 as Data Lake — Replacing HDFS
 - S3 + Spark — Reading & Writing Directly
- Amazon EMR — Managed Hadoop & Spark on AWS
- EMR Cluster Setup — Master, Core, Task Nodes
 - Submit PySpark Jobs on EMR
 - EMR with S3 — EMRFS (EMR File System)
 - EMR Serverless — No Cluster Management
 - Spot Instances — Cost Optimization

AWS Glue — Serverless ETL on AWS

- Glue Crawlers & Data Catalog
- Glue ETL Job with PySpark
- Glue vs EMR — When to Use What

Amazon Athena — Serverless SQL on S3

- Query S3 Data with SQL (no cluster needed)
- Athena with Parquet & ORC (columnar formats)

AWS IAM for Big Data — Roles & Policies

Migrating On-Premise Hadoop to AWS EMR

Labs: PySpark Job on AWS EMR, AWS Glue ETL S3 → S3, Athena Query on Parquet Data Lake

Module 15: Big Data Pipeline Architecture Patterns | Duration: 3 Hours | Level: Advanced

Pattern 1 — Traditional Hadoop Batch Pipeline

- Sqoop Import → HDFS → Hive Transform → Report

Pattern 2 — Modern Spark Batch Pipeline

- Sqoop/S3 → Spark ETL → Hive/Parquet → Reporting Tool

Pattern 3 — Real-Time Streaming Pipeline

- Kafka → Spark Structured Streaming → HDFS → Hive

Pattern 4 — Lambda Architecture

- Batch Layer (Spark) + Speed Layer (Spark Streaming)
- Serving Layer (Hive/HBase) — unify batch & real-time views

Pattern 5 — Cloud-Native Migration

- On-Premise Hadoop → AWS EMR + S3 + Glue

Pattern 6 — Airflow Orchestrated Pipeline

- Airflow DAG → Sqoop → Spark → Hive → Email Alert

Architecture Decision Factors

- Latency requirements, Data volume, Budget, Team skills
- Batch vs streaming, On-premise vs cloud

Lab: Design Full End-to-End Architecture for Given Real-World Use Case

Module 16: Big Data File Formats & Compression | Duration: 2 Hours | Level: Intermediate

File Format Comparison

- CSV — Simple but slow for big data, no schema
- JSON — Flexible but large footprint
- Avro — Best for row-based write-heavy workloads
- Parquet — Best for column-based analytics (Spark/Athena)
- ORC — Best for Hive workloads
- Delta — Best for ACID + Time Travel (Databricks/Spark 3.x)

Compression Codecs

- Gzip — High compression ratio, but NOT splittable
- Snappy — Fast, splittable, best for Spark
- LZO — Splittable, fast decompression
- Zstandard (ZSTD) — Best compression ratio with speed

Choosing Right Format + Compression Combination

- Batch analytics: Parquet + Snappy
- Hive: ORC + Snappy
- Streaming writes: Avro → Parquet on compaction

Converting CSV to Parquet & ORC using PySpark

Lab: File Format Benchmark — Query Speed Comparison across CSV, Parquet, ORC, Delta

Module 17: End-to-End Big Data Projects | Duration: 10 Hours | Level: Advanced

Project 1 — Retail Sales Analytics Pipeline

- Ingestion: Sqoop import from MySQL to HDFS
- Storage: HDFS with Hive external tables in ORC format
- Transform: PySpark — clean, aggregate, enrich sales data
- SQL: Spark SQL — sales by region, product, time period
- Orchestrate: Airflow DAG — daily scheduled pipeline
- Output: Hive Gold tables for BI reporting
- CI/CD: GitHub → deploy scripts to cluster

Project 2 — Real-Time E-Commerce Clickstream Pipeline

- Source: Python app publishing events to Kafka topics
- Processing: Spark Structured Streaming (micro-batch)
- Windowing: 5-minute aggregations with watermarking
- Storage: HDFS + Hive partitioned by date & hour
- Monitoring: Airflow SLA alerts on streaming lag
- Output: Real-time dashboard on aggregated metrics

Project 3 — Hadoop to AWS EMR Migration

- Source: Existing Hive tables on on-premise HDFS
- Migration: Export to S3 using PySpark jobs
- Transform: AWS Glue PySpark ETL jobs
- Query: Amazon Athena on S3 Parquet data lake
- Orchestrate: Airflow with AWS EMR & Glue operators
- Cost Optimisation: Spot instances + S3 Intelligent-Tiering

Module 18: Certification & Interview Preparation | Duration: 4 Hours | Level: Advanced

Certifications Overview

- Cloudera CDP Data Engineer Certification
- Databricks Certified Associate Developer for Apache Spark
- AWS Certified Data Engineer Associate — Big Data Topics

Top 60 Big Data Interview Questions & Answers

- Hadoop HDFS & YARN — Architecture, HA, Federation
- Hive Optimization — Partitioning, Bucketing, Skew, CBO
- Spark Performance — AQE, Skew, Broadcast, Persist
- Kafka & Streaming — Offsets, Consumer Groups, Exactly-Once
- Airflow Scheduling — DAGs, Operators, XCom, Backfill
- Real-world scenario & design questions

Resume Building for Big Data Engineer Roles

- Highlight cluster scale — 'Processed 5TB daily on 50-node YARN cluster'
- Quantify impact — 'Reduced pipeline runtime from 4 hrs to 25 min with AQE'
- Mention file formats & compression decisions

LinkedIn Profile Optimization

- Keywords: PySpark, Hive, Kafka, Hadoop, AWS EMR, Airflow

Salary Expectations

- India: ■8-12 LPA (Entry) | ■15-25 LPA (Mid) | ■25-35 LPA (Senior)
- US: \$110K-160K | Canada: CAD \$100K-140K | EU: €75K-115K

BONUS TOPICS: Advanced Big Data Mastery

1. Apache HBase — NoSQL on Hadoop

- HBase architecture — RegionServer, HMaster
- Row key design — single most important tuning factor
- Column families & TTL
- HBase vs Cassandra — use-case comparison
- HBase with Spark — Bulk load & scan

2. Delta Lake — ACID on Big Data

- Delta Lake vs Parquet — what changes
- Time Travel — query historical versions
- MERGE (upsert), UPDATE, DELETE on Delta
- Schema evolution & enforcement
- Optimize & Z-Order — compact & cluster Delta files

3. Apache Flink — True Stream Processing

- Flink vs Spark Streaming — true stream vs micro-batch
- DataStream API basics
- Event time & watermarks in Flink
- Flink on Kubernetes
- Flink SQL — unified batch & stream SQL

4. Databricks — Unified Analytics Platform

- Databricks architecture — clusters, notebooks, jobs
- Delta Live Tables — declarative pipelines
- Databricks AutoLoader — incremental ingestion from S3/ADLS
- Unity Catalog — enterprise data governance
- Databricks vs EMR — total cost of ownership

5. Data Lake Architecture Deep Dive

- Raw / Bronze / Silver / Gold zones
- Medallion architecture with Spark + Delta
- Data Lake vs Data Warehouse vs Lakehouse
- Partitioning strategies for petabyte-scale lakes
- Metadata management — Apache Atlas, AWS Glue Catalog

6. Apache NiFi — Data Flow Automation

- NiFi architecture — FlowFile, Processor, Connection
- NiFi vs Sqoop — when to choose NiFi
- NiFi for real-time data ingestion
- NiFi Registry — version control for flows
- NiFi with Kafka & HDFS

7. Big Data Security & Governance

- Kerberos authentication for Hadoop
- Apache Ranger — centralised authorisation
- HDFS encryption at rest
- Data masking & row-level security in Hive
- GDPR compliance patterns in Big Data pipelines

8. Hadoop Cluster Administration

- Cluster sizing — how many nodes & how much memory
- Capacity planning for growth
- Rolling upgrades — zero-downtime Hadoop version upgrades
- Ambari / Cloudera Manager — cluster management UI
- Alerting & monitoring — Ganglia, Nagios, Prometheus

9. Big Data Testing Strategies

- Unit testing PySpark with pytest & chispa
- Integration testing Spark pipelines
- Data quality testing — Great Expectations on Spark
- Schema validation & null-check patterns
- CI/CD for Spark jobs — GitHub Actions + EMR

10. Cloud-Native Big Data (Azure & GCP)

- Azure HDInsight — Managed Hadoop on Azure
- Azure Databricks + ADLS Gen2 pipeline
- GCP Dataproc — Managed Spark on GCP
- BigQuery — Serverless data warehouse
- GCP Pub/Sub — Kafka alternative on GCP

BIG DATA INTERVIEW PREPARATION & CAREER GUIDE

1. Technical Interview Strategy

Know HDFS + YARN architecture cold — every component

Explain Spark execution model: DAG → Stages → Tasks

Discuss file format decisions with trade-offs (Parquet vs ORC vs Delta)

Show Airflow DAG design skills — real pipeline questions

Demonstrate performance debugging — skew, spill, broadcast joins

2. Top Interview Questions

'Explain the difference between repartition and coalesce'

'How do you handle data skew in Spark?'

'When would you use a Broadcast join vs Sort Merge join?'

'Explain exactly-once semantics in Kafka + Spark Streaming'

'How do you optimise a slow Hive query?'

'Design a real-time fraud detection pipeline'

'What happens when a Spark executor runs out of memory?'

'How would you migrate a 10TB Hadoop cluster to AWS?'

3. Resume Bullet Point Upgrades

■ Built Spark pipelines

■ Engineered PySpark ETL processing 2TB daily sales data, cutting runtime from 6 hrs to 35 min using AQE + broadcast joins

■ Used Kafka

■ Designed Kafka → Spark Structured Streaming pipeline ingesting 500K events/sec with exactly-once semantics & 5-min watermarks

■ Worked with Hive

■ Redesigned Hive schema with dynamic partitioning + ORC + Snappy, reducing query time from 45 min to 4 min

4. Skills Checklist for Job Applications

■ Hadoop HDFS & YARN — hands-on CLI experience

■ Hive — partitioning, bucketing, performance tuning

■ PySpark — DataFrame API, UDFs, Spark SQL

■ Spark performance — AQE, skew, broadcast, Spark UI

■ Kafka — producer/consumer, consumer groups, streaming

■ Apache Airflow — DAG authoring, operators, scheduling

■ Cloud — AWS EMR / Glue / Athena (or Azure / GCP equivalent)

■ File formats — Parquet, ORC, Delta, Avro

5. Certifications Roadmap

Month 1-2: Databricks Certified Associate Developer for Apache Spark

Month 3-4: AWS Certified Data Engineer Associate

Month 5-6: Cloudera CDP Data Engineer

Month 7+: Azure Data Engineer DP-203 / GCP Professional Data Engineer

6. Salary Expectations (India)

Entry level (0-2 years): ■8-12 LPA

Mid-level (2-5 years): ■15-25 LPA

Senior (5+ years): ■25-40 LPA

Lead/Architect (7+ years): ■40-60+ LPA

7. Salary Expectations (Global)

US: \$110K-165K (Senior Big Data Engineer)

Canada: CAD \$100K-145K

UK: £70K-110K

EU: €75K-120K

Remote (India-based): 1.3-1.5x Indian market rates

RESUME & CAREER SKILL UPGRADE GUIDE

MUST-HAVE TECHNICAL SKILLS

- Linux & Shell scripting for cluster automation
- Python — pandas, PySpark, boto3
- SQL — advanced Hive/Spark SQL, window functions
- Apache Spark — PySpark + Spark SQL
- Hadoop HDFS + YARN — operations & tuning
- Apache Kafka — producers, consumers, Kafka Connect
- Apache Airflow — DAG authoring, scheduling
- At least one cloud platform (AWS EMR preferred)

DIFFERENTIATING SKILLS

- Scala for Spark — high-performance jobs
- Delta Lake — ACID, time travel, upserts
- Spark performance tuning — Spark UI, AQE, skew
- Real-time streaming — Spark Structured Streaming + Kafka
- Big Data security — Kerberos, Ranger
- Cloud migration — on-prem Hadoop → EMR/Glue

SKILLS PROGRESSION PATH

Year 1: Linux + SQL + HDFS + Hive + basic PySpark

Year 2: Advanced PySpark + Spark SQL + Kafka + Airflow

Year 3: Spark performance + Scala + Real-time streaming

Year 4+: Cloud migration + Architecture + Delta Lake + Leadership

PROJECT PORTFOLIO CHECKLIST

- End-to-end batch pipeline (Sqoop → Hive → PySpark → Report)
- Real-time streaming pipeline (Kafka → Spark → HDFS)
- Spark performance tuning case study (before vs after Spark UI)
- Airflow DAG orchestrating a full Big Data workflow
- AWS EMR / Glue cloud migration project
- Data skew solution with salting technique

TARGET JOB TITLES

Big Data Engineer (Junior / Senior / Lead)

Data Engineer (Spark / Hive / Kafka focused)

Analytics Engineer

Platform Engineer — Data Infrastructure

Cloud Data Engineer (AWS / Azure / GCP)

Data Architect (7+ years)

COURSE SUMMARY & OUTCOMES

WHAT YOU'LL MASTER

- ✓ Store and manage petabyte-scale data on HDFS + Hive
- ✓ Build high-performance PySpark & Scala Spark pipelines
- ✓ Optimise Spark jobs using AQE, broadcast joins, salting
- ✓ Process real-time events with Kafka + Spark Structured Streaming
- ✓ Orchestrate end-to-end pipelines with Apache Airflow
- ✓ Choose the right file format & compression for each workload
- ✓ Migrate on-premise Hadoop pipelines to AWS EMR + Glue
- ✓ Design Lambda & Kappa architectures for production
- ✓ Complete 3 enterprise-grade capstone projects
- ✓ Prepare for Databricks, AWS, and Cloudera certifications

COURSE STATISTICS

Total Duration: 90 hours (12-14 weeks)

18 Core Modules + 10 Bonus Topics

30+ Hands-On Labs

3 End-to-End Capstone Projects

200+ PySpark, HiveQL & Spark SQL examples

Mock certification exams & interview Q&A sessions

Lifetime Access to Course Materials

Job Placement Assistance & LinkedIn Review

ENROLLMENT & CONTACT

Website: databrickstraining.in

Email: info@databrickstraining.in

Phone: +91-8500002025

Training & Placement Excellence | Trusted by 5000+ Data Engineers