Mastering Big Data Pipeline: Hadoop, PySpark, DW, ETL & Power BI

Course Description

This intensive 5-day bootcamp equips participants with end-to-end expertise in building scalable big data pipelines. Dive deep into Hadoop's ecosystem for distributed storage and processing, Apache PySpark for advanced data transformations, data warehousing principles for structured analytics, robust ETL workflows, and Power BI for interactive visualization. Through hands-on labs, real-world case studies, and integration exercises, learners will design, deploy, and optimize production-grade data solutions—empowering digital transformation in enterprise environments.

Duration: 5 Days (40 hours)

Pre-requisites:

- Basic programming knowledge (Python or SQL preferred)
- Familiarity with databases and data concepts
- Comfort with command-line interfaces
- Laptop with admin rights (for local setups like Dockerized Hadoop/PySpark)

Learning Objectives:

- Architect distributed data systems using Hadoop and its ecosystem components.
- Master PySpark for scalable data processing, transformations, and machine learning pipelines.

- Design and optimize data warehouses for high-performance analytics.
- Build, automate, and troubleshoot ETL processes across batch and streaming data.
- Create dynamic, interactive dashboards and reports in Power BI for business intelligence.
- Integrate Hadoop, PySpark, ETL, data warehousing, and Power BI into unified big data workflows.
- Apply best practices for performance tuning, security, and scalability in production environments.

Content Coverage

Module 1: Big Data Fundamentals and Hadoop Overview

- Evolution of big data: 3Vs + Veracity, use cases
- Hadoop ecosystem intro: HDFS, YARN, Hive
- Cluster components: NameNode, DataNode, ResourceManager
- Installation and basic cluster setup
- Hands-on: Deploy single-node Hadoop
- Transition: Preparing HDFS for data ingestion

Module 2: HDFS for Distributed Storage

- HDFS architecture: Blocks, replication, fault tolerance
- Core operations: Write/read pipelines, shell commands
- Configurations: Rack awareness, high availability
- Performance basics: Block size tuning

- Hands-on: Ingest sample datasets into HDFS
- Transition: Querying HDFS data with Hive

Module 3: Hive for SQL Analytics on Hadoop

- Hive setup: Metastore, Thrift server, engines (Tez/Spark)
- HiveQL essentials: DDL/DML, partitioning, bucketing
- Optimization: Vectorization, CBO, file formats
- Advanced: Dynamic partitioning, SerDe
- Hands-on: Create Hive tables on HDFS data
- Transition: Scaling beyond Hive with PySpark

Module 4: Apache Spark and PySpark Foundations

- Spark advantages over Hive: In-memory processing
- PySpark basics: RDDs, DataFrames, Spark SQL intro
- Architecture: Driver/Executors on YARN
- Job submission and cluster integration
- Hands-on: Load Hive data into PySpark DataFrames
- Transition: DataFrame manipulations for ETL prep

Module 5: PySpark Data Processing and SQL

- DataFrame ops: Joins, aggregations, UDFs
- PySpark SQL deep dive: Windows, complex types
- Catalyst optimizer and caching strategies
- Broadcast joins, partitioning for performance

- Hands-on: Transform Hive datasets with PySpark SQL
- Transition: Streaming and ML extensions

Module 6: Advanced PySpark: Streaming and ML

- Structured Streaming: Sources, sinks, stateful ops
- MLlib: Pipelines, feature stores, models (regression/classification)
- Tuning: Cross-validation, persistence
- Integration with Hive/HDFS outputs
- Hands-on: Streaming pipeline + simple ML model
- Transition: Warehousing processed data

Module 7: Data Warehousing Concepts and Design

- DW fundamentals: Star/Snowflake schemas, SCD types
- Kimball vs. Inmon methodologies
- Hadoop DW: Hive tables, Delta Lake ACID
- Governance: Lineage, quality checks
- Hands-on: Model DW schema from PySpark outputs
- Transition: ETL to populate the warehouse

Module 8: ETL Principles and Patterns

- ETL/ELT flows: Extract from HDFS/Hive/PySpark
- Patterns: Cleansing, dedup, incremental loads
- Tools intro: Airflow orchestration, NiFi flows

- Error handling, monitoring basics
- Hands-on: Basic PySpark ETL job to Hive DW
- Transition: Advanced orchestration

Module 9: Advanced ETL Orchestration

- Airflow DAGs: Operators for PySpark/Hadoop tasks
- CDC, upsert strategies with PySpark
- Optimization: Predicate pushdown, partitioning
- Testing: Unit/integration for pipelines
- Hands-on: Airflow-scheduled ETL to DW
- Transition: Visualizing DW data in Power BI

Module 10: Power BI Basics and Big Data Connectivity

- Power BI Desktop: Import/DirectQuery from Hive/PySpark
- Data modeling: Relationships, DAX intro
- Gateways for Hadoop sources
- Incremental refresh, security setup
- Hands-on: Connect to DW and build initial reports
- Transition: Advanced DAX and visuals

Module 11: Power BI Advanced Analytics and Dashboards

- DAX mastery: Time intel, measures, variables
- Visuals: AI features, drill-through, custom
- Service features: Apps, workspaces, embedding

- Tuning: Aggregations, composite models
- Hands-on: Interactive dashboard from ETL/DW
- Transition: Full pipeline integration

Module 12: End-to-End Pipeline and Production Best Practices

- Pipeline assembly: Hadoop ingest → PySpark ETL → DW → Power BI
- Security: Kerberos, Ranger, RLS in Power BI
- Monitoring: Spark/YARN/Airflow dashboards
- Scaling: Cloud (Azure HDInsight), cost optimization
- Case study: E-commerce pipeline demo
- Hands-on: Capstone project deployment