

Big Data for Beginners with Hadoop and Spark

Programming foundations, cluster setup, ecosystem processing and security

FA-0758 · 7 days

Data & Analytics · DevOps, Cloud & Infrastructure · Software Development

Introduction

Explore the Hadoop ecosystem through programming refreshers, Linux practice, guided cluster setup and selected processing examples. Learn how storage, resource management, queries, event streaming and Spark fit together.

Seven days provide a beginner progression and prepared demonstrations across a broad ecosystem, not complete production administration expertise in every component. Core practical work uses compatible prepared environments; extended connectors and security are selected examples.

Learning outcomes

- Use introductory Python and Java concepts in small data-processing examples.
- Navigate Linux, SSH, scripts and files in an authorised training environment.
- Explain Hadoop components and follow a guided multi-node setup.
- Use selected Pig/Hive examples and inspect Kafka topics, replication and client behaviour.
- Process sample data with PySpark DataFrames, transformations, aggregation and joins.
- Explain RDDs, shared variables and a small iterative-processing example.
- Recognise Structured Streaming, non-relational integration patterns and basic Kerberos concepts.

Prerequisites

Basic operating-system, file-management and internet skills, plus introductory mathematics.

Access to three prepared Linux training nodes or equivalent virtual machines, compatible runtimes and approved network access. Programming fundamentals are introduced, but the broad ecosystem requires continued practice beyond the course.

Training outline

1. Day 1 – Programming foundations

Python

- Introduction
- IDE
- Variables
- Data Types
- Conditions
- Loops
- Functions
- Recursion
- File Systems
- Database

Java

- Introduction
- IDE
- Variables
- Data Types
- Conditions
- Loops
- Functions
- OOP
- Deployment

Selected exercises provide essential syntax and data-access foundations, not complete Python and Java language courses.

2. Day 2 — Linux and cluster foundations

- Terminal
- SSH
- Linux Commands
- Linux Scripts
- Linux file system
- Keys
- Bash and profiles
- Big-data problem framing and overview of Hadoop components.
- Follow a guided Hadoop multi-node setup in the prepared environment.
- MapReduce concepts and YARN resource allocation.
- Ambari provisioning/management/monitoring demonstration where supported by the selected stack; Tez is an execution engine, ZooKeeper coordination and Zeppelin a notebook interface.

3. Day 3 — Hadoop processing and Hive

- Pig installation or prepared environment, Pig Latin and a legacy processing use case.
- Hive environment, querying, partitioning, bucketing, UDF and SerDe concepts.
- Non-relational integration overview: HBase, Cassandra and MongoDB, with selected examples rather than three full database tracks.

4. Day 4 — Kafka ingestion and clients

- Prepared single-node and multi-node Kafka setups; topics and replication.
- Selected Python and Java producer/consumer examples.
- Kafka ingestion and the historical roles of Sqoop and Flume; Sqoop is retired and not a current installation requirement.

5. Day 5 — Spark structured processing

- Spark Execution Model Architecture
- Programming in Spark with python
- Structured API foundation
- Dataframes
- Sinks
- Transformations

- Aggregations
- Joins

Relate driver/executor execution to RDDs and the DataFrame API; source and sink compatibility depends on connectors and runtime versions.

6. Day 6 – Iterative and streaming examples

- A small breadth-first-search example and appropriate broadcast/accumulator use.
- Guided Structured Streaming input, transformation and output; checkpointing and restart considerations.
- Older Spark Streaming/DStreams as context. Latency and delivery semantics depend on the selected processing mode and connector.

7. Day 7 – Security and integrated review

- Introduction to Kerberos
- Kerberos on Hadoop
- Kerberos Terminology
- Demo: Enabling Kerberos

Distinguish Kerberos authentication from authorisation and data confidentiality. Review the guided pipeline and identify operational and security work needed before production use.