

Big Data Analytics

Course Content

Lesson 1 - Introduction to Big Data Concepts

- 3 Vs of Big Data
- Big Data processing paradigms
- Industry trends

Lesson 2 - Hadoop Ecosystem Overview

- HDFS fundamentals
- MapReduce concepts
- Hive and Pig basics

Lesson 3 - Introduction to Spark and PySpark

- Spark architecture
- RDDs vs DataFrames
- PySpark examples

Lesson 4 - Distributed Data Storage Systems

- Distributed file systems
- Object stores
- Scalability considerations

Lesson 5 - Data Ingestion Tools (Kafka, Flume)

- Streaming vs batch
- Event-driven pipelines
- Data stream processing

Lesson 6 - NoSQL Databases (MongoDB, Cassandra)

- Document vs columnar stores
- Sharding and replication
- CRUD operations

Lesson 7 - Big Data ETL Techniques

- Data wrangling at scale
- Schema-on-read vs schema-on-write
- Workflow orchestration

Lesson 8 - Data Lake vs Data Warehouse

- Concepts and differences
- Architecture best practices
- Tools and technologies

Lesson 9 - Real-time Analytics Applications

- Stream processing use cases
- Low latency dashboards
- Scalability patterns

Lesson 10 - Big Data Use Cases

- Retail personalization
- Fraud detection
- Healthcare analytics

Capstone Project

- Build an end-to-end big data pipeline on a public dataset.