Big Data Analytics Fundamentals

Course Description

This course introduces participants to the foundational concepts, tools, and techniques of Big Data Analytics. It covers data processing frameworks, storage systems, analytical methods, and visualization approaches, enabling learners to understand and work with large-scale datasets effectively.

Duration

5 Days (Instructor-led training with hands-on labs)

Pre-requisites

- Basic knowledge of databases and SQL
- Familiarity with programming concepts (Python/Java recommended)
- Understanding of data analysis fundamentals

Learning Objectives

By the end of this course, participants will be able to:

- Explain the core concepts of Big Data and its ecosystem
- Understand distributed storage and processing frameworks
- Apply data ingestion, transformation, and analysis techniques
- Perform exploratory analytics using Big Data tools
- Visualize and interpret analytical results for decision-making

Content Coverage

Day 1: Foundations of Big Data

Module 1: Big Data Overview

- Definition of Big Data and 5Vs (Volume, Velocity, Variety, Veracity, Value)
- Traditional vs Big Data approaches
- Big Data ecosystem components

- Data sources (structured, semi-structured, unstructured)
- Data lifecycle management

Module 2: Industry Applications

- Business use cases (finance, healthcare, retail, IoT)
- Challenges in adoption
- Data-driven decision making
- Ethics in Big Data
- Future outlook

Day 2: Storage & Architecture

Module 3: Distributed Storage

- HDFS architecture and components
- Replication and fault tolerance
- Data partitioning strategies
- Cluster setup basics
- Security in HDFS

Module 4: NoSQL & Data Lakes

- NoSQL databases (MongoDB, Cassandra)
- CAP theorem implications
- Data lakes vs Data warehouses
- Schema design considerations
- Cloud storage options

Day 3: Processing Frameworks

Module 5: Hadoop & MapReduce

- Hadoop ecosystem overview
- MapReduce programming model

- Job scheduling basics
- Limitations of MapReduce
- Hands-on MapReduce exercise

Module 6: Apache Spark

- Spark architecture and RDDs
- Spark SQL and DataFrames
- Structured Streaming concepts
- MLlib basics
- Performance optimization

Day 4: Analytics Techniques

Module 7: Exploratory Analytics

- Exploratory Data Analysis (EDA)
- Data profiling and cleansing
- Aggregation techniques
- Correlation and causation
- Case study on EDA

Module 8: Machine Learning & Predictive Models

- Classification algorithms
- Clustering techniques
- Regression models
- Recommendation systems
- Predictive analytics pipeline

Day 5: Visualization & Applications

Module 9: Visualization & Dashboards

- Visualization tools (Power BI, Tableau)

- Interactive dashboards
- Storytelling with data
- Scalability considerations
- Hands-on dashboard exercise

Module 10: Capstone & Future Trends

- End-to-end analytics workflow project
- Big Data in AI and IoT
- Edge computing applications
- Microsoft Fabric and Delta Lake
- Emerging trends