BIG DATA TECHNOLOGIES

1. HADOOP
2. SPARK

HADOOP

Learning Outcome Statements

- Learn how to make the most efficient use of Hadoop and its ecosystem
- Gain an insight into many of Hadoop libraries and packages
- Learn about Hadoop MapReduce and HDFS.
- Learn to control Hadoop ecosystem through various shell commands.
- Gain expertise in Hadoop technology and its related components.

Key Contents

- What Is Big Data
  - Characteristics Of 'Big Data'
  - Advantages Of Big Data Processing
- Introduction to Hadoop
  - Components of Hadoop
  - Features Of 'Hadoop'
  - Network Topology In Hadoop
- Hadoop Installation
- HDFS
  - Read Operation
  - Write Operation
  - Access HDFS using JAVA API
  - Access HDFS Using COMMAND-LINE INTERFACE
- Mapreduce
  - How MapReduce works
  - How MapReduce Organizes Work
  - Understanding MapReducer Code
- Counters & Joins In MapReduce
- Flume and Sqoop
  - Some Important features of FLUME
- Pig
  - Introduction to PIG
  - Pig Installation

SPARK

Learning Outcome Statements

- Use the core Spark APIs to operate on data
- Articulate and implement typical use cases for Spark
- Build data pipelines and query large data sets using Spark SQL and DataFrames
- Work with relational data using the GraphFrames APIs
- Understand how a Machine Learning pipeline works
- Understand the basics of Spark's internals

Key Contents

- Spark Overview

- In-depth discussion of Spark SQL and DataFrames, including:
  - The DataFrames/Datasets API
  - Spark SQL
  - Data Aggregation
  - Column Operations
  - The Functions API: date/time, string manipulation, aggregation
  - Joins & Broadcasting
  - User Defined Functions
  - Caching and caching storage levels
  - Use of the Spark UI to analyze behavior and performance

- In-depth discussion of Spark internals
  - Cluster Architecture
  - The Catalyst query optimizer
  - The Tungsten in-memory data format
  - How Spark schedules and executes jobs and tasks
  - Shuffling, shuffle files, and performance
  - How various data sources are partitioned
  - How Spark handles data reads and writes

- Spark Structured Streaming
  - Sources and sinks
  - Structured Streaming APIs
  - Windowing & Aggregation
  - Checkpointing & Watermarking
  - Reliability and Fault Tolerance
  - Kafka Integration

- Overview of Spark's MLlib Pipeline API for Machine Learning
  - Transformer/Estimator/Pipeline API
  - Perform feature preprocessing
  - Evaluate and apply ML models

- Graph processing with GraphFrames
  - Transforming DataFrames into a graph
  - Perform graph analysis, including Label Propagation, PageRank, and Shortest Paths