

Specialised Programme on Big Data Analytics – 2 Weeks

Pre-requisites for the course

- Basic knowledge of Linux

Aim

- The aim of this course is to understand and implement various Big Data technologies like Hadoop, Spark. Explain the impact of big data, including use cases, tools, and processing methods. Describe Apache Hadoop architecture, ecosystem, practices, and user-related applications

Objectives

Big Data Analytics involves the use of advanced tools and techniques to process, manage, and analyze massive volumes of data, with the goal of uncovering patterns, trends, and valuable information. Objectives of the course are

- Gain a comprehensive understanding of the fundamental concepts related to Big Data and the challenges associated with large-scale data processing
- Familiarize students with the tools and technologies commonly used in Big Data Analytics, such as Hadoop, Spark, Hive and NoSQL databases
- Learn how to efficiently process and manage large datasets, including data cleaning, transformation, and integration
- Develop skills in creating visualizations that effectively communicate insights derived from Big Data
- Cultivate problem-solving skills to address real-world challenges related to Big Data, such as identifying business opportunities, optimizing processes, or detecting anomalies

Course Contents

Introduction to Big data

Big data challenges, Big Data Applications, Types of Big Data Technologies

Introduction to Hadoop and Hadoop Architecture

What is Hadoop, Brief History and Evolution of Hadoop, Hadoop Distributions and Vendors. Hadoop Architecture, Core components of Hadoop

Hadoop Distributed File System

Core components of HDFS, The HDFS command line and web interfaces, Analyzing the Data with Hadoop, Scaling Out, HDFS – Monitoring & Maintenance. Demonstration to cloudera quickstart virtual machine. Hadoop Map Reduce paradigm

Big data analytics with Hive

Basics of Hive Query Language, working with Hive QL, Partitions and Buckets, Writing HQL scripts

Big data analytics with Spark

Initializing Spark, Spark Components and Architecture, RDD Operations, Data frames and datasets, Spark SQL, Spark MLlib

Big data analytics with MongoDB

Introduction to NoSQL, working with MongoDB (Installation, CRUD operations, Aggregation pipeline, Indexing, Data Modeling)