

Specialised Programme on Big Data Analytics – 2 Weeks

Pre-requisite

- Basic knowledge of Linux and Python

Aim

This course aims to equip learners with a thorough grasp of Big Data concepts, its ecosystem, practical tools and technologies utilized for Big Data analytics, such as Hadoop, Hive, Spark and MongoDB. Upon completion of the course, participants will possess the skills and knowledge necessary to successfully process, analyze and manage Big Data to address real-world challenges.

Objective

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

- Understand and explore Big Data ecosystem: The concept of Big Data and its significance in the modern digital era. Identify the challenges and opportunities associated with Big Data.
- Data Management with Hive: Employ Hive for searching and administering structured data stored in Hadoop. Compose HiveQL queries for data processing and analysis.
- Real-Time Analytics with Spark: Comprehend the fundamentals of Apache Spark and its essential principles. Execute real-time data processing and analytics with Spark.
- Overview of MongoDB: Examine MongoDB as a NoSQL database and its significance in Big Data. Execute CRUD operations and aggregation in MongoDB for data administration.
- Practical Applications: Engage in practical projects to examine and extract insights from Big Data. To formulate comprehensive solutions for data processing and analysis utilizing the acquired technologies. Address Big Data applications across several sectors.

Course Contents

Introduction to Big Data

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

Flagship Scheme of Government of India

A brief about Big Data Management Policy by CAG and National Data & Analytics Platform.

Introduction to Hadoop and Hadoop Architecture

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

Hadoop Distributed File System

Core components of HDFS, The HDFS command line and web interfaces, Analysing the Data with Hadoop, Scaling Out, HDFS – Monitoring & Maintenance

Big Data analytics with Hive

Basics of Hive Query Language, working with Hive QL, Managed vs External tables, Partitioning

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 Modelling)