Specialised Programme on Big Data Analytics – 3 Weeks

Aim

To explore the fundamental concepts of big data analytics with in-depth knowledge and understanding of the big data analytics domain. This course aims to present deep Knowledge of Various Big Data Technologies such as Hadoop, Map reduce, Hive and Spark. Emphasis on Hadoop Architecture and Environment using various case studies related to real world problems.

Objective

- To develop an understanding of Big Data technologies
- To implement and apply Big Data techniques on real-world requirements
- To design distributed systems that manage "big data" using Hadoop and related data engineering technologies
- To use HDFS and Map Reduce techniques for storing and analyzing data at scale
- To use HDFS and MapReduce for storing and analyzing data at scale
- To use Spark to create scripts to process data on a Hadoop cluster in more complex ways
- To analyze non-relational data using MongoDB

Course Contents

Fundamentals of Big Data

Introduction to big data platform, Structured and unstructured data, Big data use cases

Hadoop Architecture and programming model

Introduction to Hadoop, Hadoop Installation, Setting up a Hadoop Cluster, Cluster specification, Core components of Hadoop, Common Hadoop Shell command, HDFS Architecture overview, Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node, Hadoop distribution and basic commands, The HDFS command line and web interfaces, Analyzing the Data with Hadoop, Hadoop Map Reduce paradigm, Map and Reduce tasks, Anatomy of a Map Reduce Job run, Partitioners and Combiners, Map Reduce program structure.

Flagship Scheme of Government of India

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

Big Data Analytics using Apache Hive and Spark

The Hive Data-ware House, Basics of Hive Query Language, Working with Hive QL, Datatypes, Operators and Functions, Hive Tables, Partitions and Buckets, Storage Formats, Importing data, Altering and Dropping Tables. Querying with Hive QL, Querying Data-Sorting, Aggregating,

Joins, Views, Data manipulation with Hive, User Defined Functions, Writing HQL scripts, e-commerce case study using hive.
Overview, Initializing Spark, Resilient Distributed Datasets (RDDs), External Datasets, RDD Operations, Passing Functions to Spark, Working with Key-Value Pairs, Shuffle operations, RDD Persistence, Removing Data, Shared Variables, Working with Spark with Hadoop, Pyspark programming exercises, Spark MLlib , Spark SQL and DataFrame, use case of spark in healthcare

Introduction to MongoDB

Overview of SQL (DDL, DML, TCL), Introduction to NoSQL, Difference between SQL and NoSQL, working with MongoDB (Installation, CRUD operations, Aggregation pipeline, Indexing, Data Modeling)