Prof R. Madana Mohana

BIG DATA ANALYTICS

Course Outcomes | Syllabus

COMMON TO B.TECH / B.E. (CSE | IT | AI & ML | AI & DATA SCIENCE | IOT | CYBER SECURITY | COMPUTER SCIENCE & BUSINESS SYSTEMS) & M.E / M.TECH (CSE | AI & DATA SCIENCE)

Course Outcomes

Upon successful completion of this course, students will be able to:

1. Understand and analyze the processing of large datasets in Hadoop framework.
2. Apply MapReduce architecture to solve real world problems.
3. Understand NoSQL databases and create data models using MongoDB.
4. Develop scripts using Pig over large datasets and query using Hive.
5. Understand the fundamentals of the Scala programming and exercise Resilient Distributed Datasets (RDDs) for creating applications in Spark.

Syllabus | Unit-I

Introduction to Big Data:

- Importance of Big Data
  - When to consider Big Data as a solution
  - Big Data use cases:
    - IT for IT Log Analytics
    - The Fraud Detection Pattern
    - Social Media Pattern

Syllabus | Unit-I

The Hadoop Distributed Files System (HDFS):

- The Design of HDFS
- HDFS Concepts
  - Blocks
  - Name nodes and Data nodes
  - Block Caching
  - HDFS Federation
  - HDFS High Availability

Syllabus | Unit-I

The Hadoop Distributed Files System (HDFS):

- The Command-Line Interface
  - Basic File system Operations
- Hadoop File systems
  - Interfaces

Syllabus | Unit-I

The Hadoop Distributed Files System (HDFS):

- The Java Interface
  - Reading Data from a Hadoop URL
  - Reading Data Using the File System API
  - Writing Data
  - Directories
  - Querying the File system
  - Deleting Data
- Data Flow
  - Anatomy of a File Read
  - Anatomy of a File Write

Syllabus | Unit-II

MapReduce:

- What is Map reduce
- Architecture of map reduce

How MapReduce Works:

- Anatomy of a MapReduce Job Run
  - Job Submission
  - Job Initialization
  - Task Assignment
  - Task Execution
  - Progress and Status Updates
  - Job Completion

How MapReduce Works:

- Failures
  - Task Failure
  - Application Master Failure
  - Node Manager Failure
  - Resource Manager Failure
- Shuffle and Sort
  - The Map Side
  - The Reduce Side

Syllabus | Unit-II

MapReduce Types and Formats :

- MapReduce Types
  - The Default MapReduce Job
- Input Formats
  - Input Splits and Records
  - Text Input
- Output Formats
  - Text Output
- Developing a MapReduce Application

Syllabus | Unit-III

No SQL Databases:

- Review of traditional Databases
- Need for NoSQL Databases
- Columnar Databases
- Failover and reliability principles
- CAP Theorem
- Differences between SQL and NoSQL databases

Syllabus | Unit-III

Working mechanisms of Mongo DB:

- Overview
- Advantages
- Environment
- Data Modelling
- Create Database, Drop Database, Create collection, Drop collection, Data types, Insert, Query, Update and Delete operations, Limiting and Sorting records, Indexing, Aggregation

Syllabus | Unit-IV

Pig:

- Installing and Running Pig
- An Example
  - Generating Examples
- Comparison with Databases
- Pig Latin
- User-Defined Functions
- Data Processing Operators
- Pig in Practice

Syllabus | Unit-IV

Hive:

- Installing Hive
  - The Hive Shell
- An Example
- Running Hive
- Comparison with Traditional Databases
- HiveQL
- Tables
- Querying Data
- User-Defined Functions
  - Writing a User Defined Function
  - Writing a User Defined Aggregate Function

Syllabus | Unit - V

Spark:

- Importance of Spark Framework
- Components of the Spark unified stack
- Batch and Real-Time Analytics with Apache Spark
- Resilient Distributed Dataset (RDD)

Syllabus | Unit - V

Scala (Object Oriented and Functional Programming)

- Scala Environment Set up
- Downloading and installing Spark standalone,
- Functional Programming
- Collections

References

Text Books:

1. Tom White, "Hadoop: The Definitive Guide", 4th Edition, O'Reilly Media Inc, 2015.
2. Tanmay Deshpande, Hadoop Real-World Solutions Cookbook||, 2nd Edition, Packet Publishing 2016.

Suggested Reading:

1. Thilinagunarathne Hadoop MapReduce v2 Cookbook – 2nd Edition, Packet Publishing, 2015.
2. Chuck Lam, Mark Davis, Ajit Gaddam, -Hadoop in Action||, Manning Publications Company, 2016.
3. Alan Gates, "Programming Pig", O'Reilly Media Inc, 2011.

e-Resources

Web Resources:

1. https://www.datastax.com/what-is/nosql
2. https://www.iitr.ac.in/media/facspace/patelfec/16Bit/index.html
3. https://www.coursera.org/specializations/data-science
4. https://nptel.ac.in/courses/106104189
5. https://courses.cognitiveclass.ai/courses/course-v1:CognitiveClass+BD0101EN+v2/course/
6. https://www.edureka.co/masters-program/big-data-architect-training