II YEAR IV SEMESTER PAPER– IV

BIG DATA TECHNOLOGY

Objectives:

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

Outcome

1. Learn tips and tricks for Big Data use cases and solutions.
2. Acquire knowledge of HDFS components , Namenode, Datanode, etc.
3. Acquire knowledge of storing and maintaining data in cluster, reading data from and writing data to Hadoop cluster.
4. Able to maintain files in HDFS
5. Able to write MapReduce applications to access data present on HDFS
6. Able to read different formats of files into map-reduce application.
7. Able to develop MapReduce applications to analyze Big Data related to the real world use cases.

8. Able to write MapReduce applications that can take data from multiple datasets and join them

9. Able to optimize the performance of Map-Reduce application

Unit-I: Introduction to Big Data

Introduction –Distributed File System – Big Data and its importance, Characteristics of Big Data, Limitation of Conventional Data Processing Approaches, Need of big data frameworks, Big data analytics, Limitations of Big Data and Challenges, Big data applications

Unit-II

Hadoop: Basic Concepts of Hadoop and its features -The Hadoop Distributed File System (HDFS)- Anatomy of a Hadoop Cluster - Hadoop cluster modes - Hadoop Architecture, Hadoop Storage - Hadoop daemons (Name node-Secondary name node-Job tracker-Task tracker-Data node,etc) - Anatomy of Read & Write operations – Interacting HDFS using command-line (HDFS Shell and FS shell commands) -Interacting HDFS using Java APIs – Dataflow – Blocks –Replica - YARN.

Unit-III

Hadoop Ecosystem Components – Schedulers- Fair and Capacity, Hadoop 2.0 Vs Hadoop 3.0 and its new features.

Hadoop Cluster Setup – SSH & Hadoop Configuration –HDFS Administering – Monitoring & Maintenance.

Unit-IV

Hadoop MapReduce - Introduction - Phases in MapReduce Framework - Anatomy of MapReduce Job run - Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features. Understanding Basic MapReduce Pogram

(WordCount program): The Driver Code - The Mapper class - The Reducer class.

Unit-V:

Writing first MapReduce Program - Hadoop's Streaming API - Using Eclipse for Rapid Development – YARN Vs MapReduce Advanced MapReduce Concepts: Partitioner – Combiner – Joins – Map-side Join – Reduce-side Join - Case Study: Weblog Analysis done using Mapper, Reducer, Combiner, Partitioner, etc.

References

1. Boris lublinsky, Kevin t. Smith Alexey Yakubovich, "Professional Hadoop Solutions".
   Wiley, ISBN : 9788126551071, 2015.
2. Chris Eaton, Dirk Deroos et al., "Understanding Big Data", McGraw Hill , 2010.
3. Tom White, "HADOOP" : The definitive Guide", O Reilly 2012.
4. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", PACKT publishing, 2013.

Student Activity:

Case Study I: Centers for Medicare & Medicaid Services: The Integrity of Healthcare Data and Secure Payment Processing.

Case Study II: Movie Lens Data set Analysis

Case Study III: Web Server Log Analysis using MapReduce.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion

2. Try to solve MCQ’s available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills

6. Individual and group project reports like “Movie Lens Data Analysis”, "Youtube Click stream Data

Analysis", etc.

7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work

II YEAR IV SEMESTER PAPER– IV

BIG DATA TECHNOLOGY Through Hadoop LAB

1. Implement the following Data Structures in Java
   a) Linked Lists
   b) Stacks
   c) Queues
   d) Set
   e) Map

2. Hadoop Cluster Setup
   (i) Perform setting up and Installing Hadoop in its three operating modes: Standalone
   Pseudo
   distribute
   d Fully
   distribute
   d
   (ii) Use web based tools to monitor your Hadoop setup.

3. Implement the following file management tasks in Hadoop:
   - Adding files and directories, List the files and directories
   - Retrieving files
   - Deleting files
   - Copying files from one folder to another in HDFS
   - Copying files from Local File System to HDFS

4. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm
5. Write a Map Reduce program that mines weather data (NCDC). Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. Data available at: ftp://ftp.ncdc.noaa.gov/pub/data/noaa/.
   - Find average, max and min temperature for each year in NCDC data set
   - Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.
6. Implement Matrix Multiplication program with Hadoop Map Reduce.
7. Stop word elimination problem:
   Input:
   - A large textual file containing one sentence per line
   - A small file containing a set of stop words (One stop word per line)

Output:

- A textual file containing the same sentences of the large input file without the words appearing in the small file.

8. Write a MapReduce Application to implement Combiners
9. Write a MapReduce Application to implement Reduce-side Join
10. Write a MapReduce Application to implement Map-side Join

Outcome:

- Able to develop MapReduce applications to analyze Big Data related to the real world use cases.
- Able to setup, configure and manage Hadoop cluster on single node
- Able to access the Hadoop cluster through Web UI.
- Able to track the execution of MapReduce jobs through Web UI
- Able use Joins, partitioner, combiners as and when needed while developping MapReduce application to analyze the Big Data.