## II YEAR IV SEMESTER PAPER– V

## BIG DATA ACQUISITION AND ANALYSIS

### Objective

Learn to develop Hadoop applications for storing processing and analyzing data stored in Hadoop cluster. The course is mainly covering Big Data tools for Data Transformation (Apache PIG), Data Analysis (HIVE) and for handling unstructured data HBase. To Understand the complexity and volume of Big Data and their challenges. To analyse the various methods of data collection. To comprehend the necessity for pre-processing Big Data and their issues

### Outcome

1. Identify the various sources of Big Data
2. Able to collect and store Big Data from various sources
3. Able to write Pig Scripts- Extract, Transform and Load the data on HDFS
4. Able to write Hive Scripts- Extract, Transform, Load and Analyse the data present in HDFS
5. Able to write scripts to extract data from structured and un-structured data for analytics
6. Able to extract and process semi and un-structured data using HBase

### Unit- I

**Introduction To Big Data Acquisition:** Big data framework – fundamental concepts of Big Data Management and analytics – Current challenges and trends in Big Data Acquisition. Map Reduce Algorithm- Hadoop Storage [HDFS], Common Hadoop Shell commands - Anatomy of File Write and Read, NameNode, Secondary NameNode, and DataNode - Hadoop Configuration – Pig Configuration – Hive Configuration - HBase Configuration.

### Unit-II

**Data Collection And Transmission:** Big data collection – Strategies – Types of Data Sources – Structured Vs Unstructured data – ELT vs ETL – storage infrastructure requirements – Collection methods – Log files – sensors – Methods for acquiring network data (Libcap-based and zero-copy packet capture technology) – Specialized network monitoring softwares (Wireshark, Smartsniff and Winnetcap) – Mobile equipments, Transmission methods, Issues.

### Unit-III

**Apache Pig -** Introduction - Pig features - Pig Architecture - Pig Execution modes, Pig Grunt shell and Shell commands. Pig Latin Basics: Data model, Data Types, Operators - Pig Latin Commands - Load & Store , Diagnostic Operators, Grouping, Cogroup, Joining, Filtering, Sorting, Splitting - Built-In Functions, User define functions. Pig Execution Modes: Batch Mode – Embedded Mode – Pig Execution in Batch Mode –Use cases - Map Reduce programs with Pig – Pig Vs SQL

### Unit-IV

**Hive**: Introduction - Hive Features - Hive architecture -Hive Meta store - Hive data types -

Hive Tables - Table types - Creating database, Altering database, Create table, alter table, Drop table, Built-In Functions - Built-In Operators, User defined functions(UDFs),  View, Pig Vs Hive.

**HiveQL**–Introduction, HiveQL Select, HiveQL – MapReduce using HiveQL OrderBy, Group By Joins, LIMIT, Distribute By , Cluster By - Sorting And Aggregation – Partitioning: Static & Dynamic partitioning – Index Creation - Bucketing – Analysis of MapReduce execution – Hive Optimization – Setting Hiivng Parameters. Comparison between MapReduce,  Hive QL and SQL. UseCase: Implementation of MapReduce programs with HiveQL.

**Unit-V**

**Hbase :** HBasics, Features of HBase, Concepts, Clients, Example, Hbase Versus RDBMS, Limitations of HBase

**Big Data Privacy And Applications**: Data Masking – Privately identified Information (PII) – Privacy preservation in Big Data – Popular Big Data Techniques and tools –Applications- Social Media Analytics – Fraud Detection.

**References**

1. Bart Baesens, “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications’, John Wiley & Sons, 2014.
2. Tom White “ Hadoop: The Definitive Guide” Third Edit on, O’reily Media, 2012.
3. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
4. Min Chen. Shiwen Mao, Yin Zhang. Victor CM Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.
5. Michael Minelli, Michele Chambers Ambiga Dhiraj, “Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends”, John Wiley & Sons, 2013.
6. Raj. Pethuru “ Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global.

**Student Activity:**

**Case study I:** “BankAmeriDeals” provides cash-back offers to credit and debit-card customers based upon analyses of their prior purchases.

**Case Study II:** ***<u>GOOGLE</u>:*** Working with the U.S. Centers for Disease Control, tracks when users are inputting search terms related to flu topics, to help predict which regions may experience outbreaks.

**Case Study III**: Twitter data Analysis

**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, Twitter 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-V**

# Data Acquisition and Analysis Lab

1. Hadoop Cluster Setup
   - Perform setting up and Installing Hadoop in its three operating modes:
     - standalone
     - Pseudo distributed
     - Fully distributed
   - Use web based tools to monitor your Hadoop setup.

2. Install and Run Pig and also use Pig Shell commands to display the list of files in HDFS

3. Install and Run Hive and also use Hive Shell commands to display the list of files in HDFS

4. Install and Run HBase and also use HBase Shell commands to display the version and user of HBase

5. Use Hive to create, alter, and drop databases, tables, views, functions, and indexes

6. Write and execute Pig Script to Load data into a Pig relation without a schema

7. Write and execute Pig Script Load data into a Pig relation with a schema

8. Write a Pig script to find the word count in a text file

9. Write a Pig Script 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.

10. Write HiveQL command to create Weather table and to find the year-wise maximum temperature

11. Write a Pig Script to remove null and duplicate values from the given input file.

12. Write Pig scripts to implement filter, project, sort, group by, joins

13. Write Hive Query to create database, managed table, external table, join, index, view, etc

14. Create a table in HBase and insert the data into with Shell

15. Display the data present in a HBase table using Shell