KAKATIYA UNIVERSITY WARANGAL

Under Graduate Courses (Under CBCS with effect from Academic Year 2022-2023 onwards)

B.Sc. DATA SCIENCE

III Year: Semester-VI

Paper – VII (A): Big Data

[4 HPW :: 4 Credits :: 100 Marks (External:80, Internal:20)]

UNIT – I

Getting an overview of Big Data: Introduction to Big Data, Structuring Big Data, Types of Data, Elements of Big Data, Big Data Analytics, and Advantages of Big Data Analytics.

Introducing Technologies for Handling Big Data: Distributed and Parallel Computing for Big Data, Cloud Computing and Big Data, Features of Cloud Computing, Cloud Deployment Models, Cloud Services for Big Data, Cloud Providers in Big Data Market.

UNIT – II

Understanding Hadoop Ecosystem: Introducing Hadoop, HDFS and MapReduce, Hadoop functions, Hadoop Ecosystem. **Hadoop Distributed File System**- HDFS Architecture, Concept of Blocks in HDFS Architecture, Namenodes and Datanodes, Features of HDFS. MapReduce.

Introducing HBase- HBase Architecture, Regions, Storing Big Data with HBase, Combining HBase and HDFS, Features of HBase, Hive, Pig and Pig Latin, Sqoop, ZooKeeper, Flume, Oozie.

UNIT- III

Understanding MapReduce Fundamentals and HBase: The MapReduceFramework ,Exploring the features of MapReduce, Working of MapReduce, Techniques to optimize MapReduce Jobs, Hardware/Network Topology, Synchronization, File system, Uses of MapReduce, Role of HBase in Big Data Processing- Characteristics of HBase.

Understanding Big Data Technology Foundations: Exploring the Big Data Stack, Data Sources Layer, Ingestion Layer, Storage Layer, Physical Infrastructure Layer, Platform Management Layer, Security Layer, Monitoring Layer, Visualization Layer.

UNIT – IV

Storing Data in Databases and Data Warehouses: RDBMS and Big Data, Issues with Relational Model, Non – Relational Database, Issues with Non Relational Database, Polyglot Persistence, Integrating Big Data with Traditional Data Warehouse, Big Data Analysis and Data Warehouse.

NoSQL Data Management: Introduction to NoSQL, Characteristics of NoSQL, History of NoSQL, Types of NoSQL Data Models- Key Value Data Model, Column Oriented Data Model, Document Data Model, Graph Databases, Schema-Less Databases, Materialized Views, CAP Theorem.

Reference

1. BIG DATA, Black Book TM, DreamTech Press, 2016 Edition.

Suggested Reading:

2. Seema Acharya, SubhasniChellappan , "BIG DATA and ANALYTICS", Wiley publications, 2016
3. Nathan Marz and James Warren, "BIG DATA- Principles and Best Practices of Scalable Real-Time Systems", 2010

Practical – 7(A): Big Data (Lab)

[3 HPW:: 1 Credit :: 25 Marks]

Objectives:

- Installation and understanding of working of HADOOP
- Understanding of MapReduce program paradigm.
- Writing programs in Python using MapReduce
- Understanding working of Pig, Hive
- Understanding of working of Apache Spark Cluster

1. Setting up and Installing Hadoop in its two operating modes:
   - Pseudo distributed,
   - Fully distributed.
2. Implementation of the following file management tasks in Hadoop:
   - Adding files and directories
   - Retrieving files
   - Deleting files
3. Implementation of Word Count Map Reduce program
   - Find the number of occurrence of each word appearing in the input file(s)
   - Performing a MapReduce Job for word search count (look for specific keywords in a file)
4. Map Reduce Program for Stop word elimination:
   - Map Reduce program to eliminate stop words from a large text file.
5. Map Reduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather 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: https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all.
   - 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. Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.
7. Write a Pig Latin script for finding TF-IDF value for book dataset (A corpus of eBooks available at: Project Gutenberg)
8. Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes.
9. Install, Deploy & configure Apache Spark Cluster. Run apache spark applications using Scala.
10. Perform Data analytics using Apache Spark on Amazon food dataset, find all the pairs of items frequently reviewed together.