20CS006 BIG DATA & ANALYTICS

Hours Per Week :

L	T	P	C
3	-	2	4

Total Hours :

L	T	P
45	-	30

WA/RA	SSH/HSH	CS	SA	S	BS
15	30	-	5	5	-

Course Description and Objective:

This course gives an overview of Bigdata, i.e., storage, retrieval, and processing of bigdata. The focus will be on the "technologies", i.e., the tools/algorithms that are available for storage and processing of bigdata and variety of "analytics".

Course Outcomes

The student will be able to:

COs	Course Outcomes	POs
1	Understand Bigdata and its analytics in the real world	1
2	Use the Bigdata frameworks like Hadoop and NOSQL to efficiently store and process Bigdata to generate analytics	1
3	Design of algorithms to solve data intensive problems using MapReduce paradigm	3,5
4	Design and implementation of Bigdata analytics using Pig and Spark to solve data intensive and to generate analytics.	3,5
5	Analyze Bigdata using Hive	2

Skills:

- ✓ Build and maintain reliable, scalable, distributed systems with Apache Hadoop
- ✓ Develop MapReduce based applications for Bigdata
- ✓ Design and build applications using Hive and pig based Bigdata applications
- ✓ Learn tips and tricks for bigdata usecases and solutions

UNIT - I

Introduction to big data: Data, Characteristics of data and Types of digital data:, Sources of data, Working with unstructured data, Evolution and Definition of big data, Characteristics and Need of big data, Challenges of big data
Big data analytics: Overview of business intelligence, Data science and Analytics, Meaning and Characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment

UNIT - II

Introduction to Hadoop : Introducing Hadoop, need of Hadoop, limitations of RDBMS, RDBMS versus Hadoop, Distributed Computing Challenges, History of Hadoop , Hadoop Overview, Use Case of Hadoop, Hadoop Distributors, HDFS (Hadoop Distributed File System), Processing Data with Hadoop, Managing Resources and Applications with Hadoop YARN (Yet another Resource Negotiator), Interacting with Hadoop Ecosystem

UNIT - III

Introduction to MAPREDUCE Programming: Introduction, Mapper, Reducer, Combiner, Partitioner, Searching, Sorting , Compression, Real time applications using MapReduce, combiner, partitioner, matrix multiplication using MapReduce and page rank algorithm using MapReduce.

UNIT - IV

Introduction to Pig: Introduction to Pig, The Anatomy of Pig , Pig on Hadoop , Pig Philosophy , Use Case for Pig: ETL Processing , Pig Latin Overview , Data Types in Pig , Running Pig , Execution Modes of Pig, HDFS Commands, Relational Operators, Piggy Bank , Word Count Example using Pig , Pig at Yahoo!.

Introduction to Hive: Introduction to Hive, Hive Architecture , Hive Data Types, Hive File Format, Hive Query Language (HQL).

UNIT - V

Hive: Partitions and bucketing, RCFile Implementation, working with XML files, User-Defined Function (UDF) in Hive, Pig versus Hive.

Spark: Introduction, features of Spark, components of Spark, Programming with Resilient Distributed datasets (RDDS).

LIST OF EXPERIMENTS

LIST OF EXPERIMENTS **Total hours: 30**

1. HDFS basic command-line file operations.
2. HDFS monitoring User Interface.
3. WordCount Map Reduce program using Hadoop.
4. Implementation of word count with combiner Map Reduce program.
5. Practice on Map Reduce monitoring User Interface.
6. Implementation of Sort operation using MapReduce.
7. MapReduce program to count the occurrence of similar words in a file by using partitioner.
8. Design MapReduce solution to find the years whose average sales is greater than 30.

   input file format has year, sales of all months and average sales.

   Year Jan Feb Mar April May Jun July Aug Sep Oct Nov Dec Average.
9. MapReduce program to find Dept wise salary.

   Empno EmpName Dept Salary.
10. Install and Run Pig then write Pig Latin scripts to sort, group, join, project and filter the

data.

11. Implementation of Word count using Pig.
12. Creation of Database and tables using Hive query language.
13. Creation of partitions and buckets using Hive.
14. Practice of advanced features in Hive Query Language: RC File & XML data processing.
15. Implement of word count using spark RDDs.
16. Filter the log data using Spark RDDs.

TEXT BOOKS

1. Big Data Analytics, SeemaAcharya, SubhashiniChellappan, Wiley
2. Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, Andy Konwinski, Patrick Wendell, MateiZaharia, O'Reilly Media, Inc.

REFERENCE BOOKS :

1. Boris lublinsky, Kevin t. Smith, AlexeyYakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.
2. Chris Eaton,Dirkderooset al. , "Understanding Big data ", McGraw Hill, 2012.
3. Tom White, "HADOOP: The definitive Guide", O Reilly 2012.
4. VigneshPrajapati, "Big Data Analyticswith R and Haoop", Packet Publishing 2013.