17CS008 BIG DATA ANALYTICS

L	T	P	C
3	-	3	5

Course Description and Objectives:
This course gives an overview of Big Data, i.e. storage, retrieval and processing of big data. The focus will be on the “technologies”, i.e., the tools/ algorithms that are available for storage, processing of Big Data and a variety of “analytics”.

Course Outcome:
The student will be able to:

- Understand the theoretical issues involved in Big Data system design such as the curse of dimensionality.
- Familiarize with major approaches in Big Data Analytics.

Skills:
Upon completion of this course, students will be able to do the following:

- Students will be able to build and maintain reliable, scalable, distributed systems with Apache Hadoop.
- Students will be able to write Map-Reduce based Applications
- Students will be able to design and build applications using Hive and Pig based Big data Applications
- Students will learn tips and tricks for Big Data use cases and solutions

Activities:

- Install Hadoop and develop applications on Hadoop
- Develop Map Reduce applications
- Develop applications using Hive/Pig/Spark

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, Data serialization and Working with common serialization formats, Big data serialization formats

Unit-IV

Introduction to Hive: Introduction to Hive, Hive Architecture , Hive Data Types, Hive File Format, Hive Query Language (HQL), User-Defined Function (UDF) in Hive.

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!, Pig versus Hive

Unit-V

Spark: Introduction to data analytics with Spark, Programming with RDDS, Working with key/value pairs, advanced spark programming

Text Books

1. Big Data Analytics, Seema Acharya, Subhashini Chellappan, Wiley
2. Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, Inc.

Reference Books:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, ISBN: 9788126551071, 2015.
2. Chris Eaton, Dirk derooset al. , “Understanding Big data ”, McGraw Hill, 2012.
3. Tom White, “HADOOP: The definitive Guide”, O Reilly 2012.
4. Vignesh Prajapati, “Big Data Analytics with R and Hadoop”, Packet Publishing 2013.

LABORATORY EXPERIMENTS

Getting Hadoop Up and Running in a cluster:

1. Setting up Hadoop on standalone machine.
2. Wordcount Map Reduce program using standalone Hadoop.
3. Adding the combiner step to the Wordcount Map Reduce program.
4. Using HDFS monitoring UI
5. HDFS basic command-line file operations.
6. Setting Hadoop in a distributed cluster environment.
7. Running the WordCount program in a distributed cluster environment.
8. Practice on Map Reduce monitoring User Interface
9. Sort operation using MapReduce
10. Simple analytics using Map Reduce.
11. Creation of Database using hive.
12. Practice of Hive Query Language operations.
13. Basic operations in pig
14. Implementation of Word count using Pig.
15. Simple programs using Spark.
16. Implementation of WordCount using Spark

Text Books

1. Hadoop Map Reduce Cookbook, Srinath Perera & Thilina Gunarathne, 2013, PACKT PUBLISHING.
2. Learning Spark: Lightning – Fast Big Data Analysis, Holden Karau, Andy Konwinski, Patrick Wendell, MateiZaharia O’Reilly Media, Inc.
3. Tom White “ Hadoop The Definitive Guide” O’Reilly 2012