16IT445 BIG DATA ANALYTICS

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 Outcomes:

- 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 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

UNIT – I **9 Hrs**

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 **10 Hrs**

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 **8 Hrs**

Introduction to MAPREDUCE Programming: Introduction, Mapper, Reducer, Combiner, Partitioner , Searching, Sorting , Compression, Real time applications using MapReduce, Combiner, partitioner features in MapReduce Working with common serialization formats, Big data serialization formats

UNIT – IV 12 **Hrs**

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 **8 hrs**

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

List of experiments:

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. Creation of Database using hive.
11. Creation of partitions and buckets using Hive.
12. Practice of advanced features in Hive Query Language: RC File & XML data processing.
13. Install and Run Pig then write Pig Latin scripts to sort, group, join, project and filter the data.
14. Implementation of Word count using Pig.
15. Implement of word count using spark RDDs
16. Filter the log data using Spark RDDs.

Text Book:

1. **Big Data Analytics, Seema Acharya, Subhashini Chellappan, Wiley Publishers**

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. Vignesh Prajapati, “Big Data Analytics with R and Haoop”, Packet Publishing 2013.