Objective: Big data is an extremely useful area in the era of computing techniques as it aids in finding useful pattern from large datasets. Large datasets are so huge that they cannot be processed with traditional technologies. We require special computing system which can handle large data and tandem it with other important aspects like parallel processing, data failure, data pre-processing etc.

Credits Earned: 5 Credits

Course Outcomes: After completion of this course, student will be able to

- Gain Understanding about Big Data Technology and its Tools.
- Understand and apply extracting useful pattern from large datasets.
- Implementation of Big data mining techniques using different software.
- Understand how data analytics and data science maps to current industry.
- Understanding and implementing Algorithms in an optimized way using various Big Data Tools.

Pre-requisite of course: Basic Programming Knowledge.

### Teaching and Examination Scheme

<table>
<thead>
<tr>
<th>Teaching Scheme (Hours)</th>
<th>Credits</th>
<th>Theory Marks</th>
<th></th>
<th></th>
<th></th>
<th>Total Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory</td>
<td>Tutorial</td>
<td>Practical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ESE (E)</th>
<th>Mid Sem (M)</th>
<th>Internal (I)</th>
<th>Viva (V)</th>
<th>Term work (TW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
<td>20</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

150

Contents:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Topics</th>
<th>Contact Hours</th>
</tr>
</thead>
</table>
| 1    | Introduction to Big Data  
Introduction-Distributed file System, What is Big Data? Difference between traditional Distributed file system and Big Data Software, Big Data Analytics, Big data Applications. | 4 |
## Syllabus for Bachelor of Technology
### Computer Engineering

| 2 | **Introduction to Hadoop:**  
How Hadoop works? Hadoop Architecture, Explanation of Hadoop EcoSystem, Hadoop Basic commands. | 7 |
|---|---|---|
| 3 | **Hadoop Input and Output:**  
Data Integrity in Hadoop, Data Compression and Data Serialization in Hadoop, Avro, How Avro works? | 7 |
| 4 | **Hadoop MapReduce:**  
Mapper, Reducer, MapReduce YARN, Job Scheduling, Sorting and Shuffling in MapReduce, MapReduce Input Formats, MapReduce Output Formats, How to code in MapReduce program, analyze data using MapReduce. | 10 |
| 5 | **Hadoop Ecosystem/Environment: Pig, Hive, Hbase, ZooKeeper**  
Pig Latin Structures, Statements, Functions, User-Defined Function in Pig, Loading, Storing and Sorting Data in Pig, HiveQL, Tables in Hive, Querying Data, User-Defined Function in Hive, Introduction to HBase, HBASE vs RDBMS, What is ZooKeeper, Zookeeper Services, Build Application with ZooKeeper. | 12 |
| 6 | **Apache Spark:**  
Introduction to Apache Spark, pySpark, RDD, Working with Key-value pair, Loading and saving data in spark, Learning about Machine Learning Library in Spark. | 7 |
| 7 | **NoSql:**  
Introduction to NoSql, NoSql vs SQL, NewSql, Introduction to MongoDB, MongoDB Create-Drop Databases, Create-Drop Collection, CRUD operation in documents, MongoDB indexing, Aggregation, replication, sharding, Connect Java Application with MongoDB. | 5 |

### Total Hours

| Total Hours | 52 |

### References:

2. BIG Data and Analytics, Sima Acharya, Subhashini Chhellappan, Willey
4. Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau
Syllabus for Bachelor of Technology

Computer Engineering

**Suggested Theory distribution:**
The suggested theory distribution as per Bloom’s taxonomy is as per follows. This distribution serves as guidelines for teachers and students to achieve effective teaching-learning process.

<table>
<thead>
<tr>
<th>Remember</th>
<th>Understand</th>
<th>Apply</th>
<th>Analyse</th>
<th>Evaluate</th>
<th>Create</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>20%</td>
<td>30%</td>
<td>15%</td>
<td>10%</td>
<td>5%</td>
</tr>
</tbody>
</table>

**Suggested List of Experiments:**
1. Why we required Big Data Technology? Explain with details.
2. Installation of Hadoop in ubuntu using single node cluster requirement.
3. Run all HDFS commands in hadoop environment.
4. Run WordCount program using MapReduce algorithm in hadoop.
5. Run WordCount program using MapReduce algorithm using partitioner in hadoop.
6. Write a MapReduce program for finding out unique listeners per track by given data.
7. Write the MapReduce job to find the total count for each word.
8. Write the MapReduce job to reverse patent citation data.
10. Run all Hive related commands on given data.
11. Create an UDF in Hive to truncate blank space.
12. Install HBASE and Apply create table, insert data into table, alter table, update table queries.
13. Install MongoDB and try its basic commands in MongoDB Shell.
14. Connect MongoDB with java using IDE.
15. Install Scala and program in interactive mode and script mode.
16. Run wordcount job in Apache spark.

**Instructional Method:**
- The course delivery method will depend upon the requirement of content and need of students. The teacher in addition to conventional teaching method by black board, may also use any of tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.
Syllabus for Bachelor of Technology

Computer Engineering

➢ The internal evaluation will be done on the basis of continuous evaluation of students in the laboratory and class-room.

➢ Practical examination will be conducted at the end of semester for evaluation of performance of students in laboratory.

➢ Students will use supplementary resources such as online videos, NPTEL videos, e-courses, Virtual Laboratory

Supplementary Resources: