**Lecture Plan**

|  |  |
| --- | --- |
| **Faculty Name** | **Shreya Agarwal** |
| **Course** | **B. Tech** |
| **Session** | **2024-2025** |
| **Year** | **3rd** |
| **Branch** | **Computer Science** |
| **Section** | **A** |
| **Semester** | **6th** |
| **Subject Name** | **Big Data** |
| **Subject Code** | **KCS 061** |
| **Total Number of Students** | **58** |
| **Number of Lecture Proposed** | **39** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Unit No** | **Topic** | | **CO** | | **No of Lectures Required** | **Actual Date of Completion** | **Suggested Reference** |
|  | **1** | Types of digital data, history of Big Data innovation, introduction to Big Data platform, drivers for Big Data, Big Data architecture and  Characteristics, 5 Vs of Big Data. | |  | | **1** |  |  |
|  | Big Data technology components, Big Data importance  and applications, Big Data features – security, compliance, auditing and protection, | | **1** |  |  |
|  | Big Data privacy and ethics, Big Data Analytics. Challenges of conventional systems | | **1** |  |  |
|  | Intelligent data analysis, nature of data, analytic processes and tools, analysis vs reporting,  modern data analytic tools | | **1** |  |  |
|  |  |  | |  | |  |  |  |
| **No of Lectures Required to complete Unit 1** | | | | | | **4** | | | |
|  | **2** | | History of Hadoop, Apache Hadoop, the Hadoop Distributed File System- components of Hadoop, data format, analysing data with Hadoop, scaling out, Hadoop  Streaming, Hadoop pipes, Hadoop Echo System. |  | | **1** |  |  |
|  | Design of HDFS, HDFS concepts, benefits  and challenges, file sizes, block sizes and block abstraction in HDFS, data replication, how  does HDFS store, read, and write files. | **2** |  |  |
|  | Java interfaces to HDFS, command line interface,  Hadoop file system interfaces, data flow, data ingest with Flume and Scoop, Hadoop  Archives. | **1** |  |  |
|  | Hadoop I/O: compression, serialization, Avro and file-based data structures | **1** |  |  |
|  | Map Reduce framework and basics, how Map Reduce works, developing a  Map Reduce application. | **2** |  |  |
|  | Unit tests with MR unit, test data and local tests, anatomy of a  Map Reduce job run. | **1** |  |  |
|  | Failures, job scheduling, shuffle and sort, task execution using Map Reduce. | **1** |  |  |
|  | Map Reduce types, input formats, output formats, Map Reduce features, Real-world Map Reduce. | **1** |  |  |
|  |  | |  |  | |  |  |  |
| **No of Lectures Required to complete Unit 2** | | | | | | **10** | | | |
|  | **3** | | **Hadoop Environment:** Setting up a Hadoop cluster, cluster specification, cluster setup  and installation, Hadoop configuration. |  | | **2** |  |  |
|  | Security in Hadoop, administering Hadoop, HDFS Monitoring & maintenance, Hadoop benchmarks, Hadoop in the cloud. | **1** |  |  |
|  | **Hadoop Eco System Frameworks**: Applications on Big Data using Pig, Hive and HBase,NoSQL Databases:Introduction to NoSQL. | **1** |  |  |
|  | **MongoDB:** Introduction, data types, creating, updating and deleing documents. | **1** |  |  |
|  | Querying, introduction to indexing, capped collections in MongoDB. | **1** |  |  |
|  | **HBase** – Hbase concepts, clients, example, Hbase vs RDBMS, advanced usage, schema  design, advance indexing. | **2** |  |  |
|  | Zookeeper – how it helps in monitoring a cluster, how to build  Applications with Zookeeper. | **1** |  |  |
|  | IBM Big Data strategy, introduction to Infosphere, Big Insights and Big Sheets,  Introduction to Big SQL | **1** |  |  |
|  |  | |  |  | |  |  |  |
| **No of Lectures Required to complete Unit 3** | | | | | | **10** | | | |
|  | **4** | | **Hive** - Apache Hive architecture and installation, Hive shell, Hive services, Hive  metastore, comparison with traditional databases, |  | | **2** |  |  |
|  | HiveQL, tables, querying data and user defined functions, sorting and aggregating, Map Reduce scripts, joins & subqueries. | **1** |  |  |
|  | **Hadoop Eco System and YARN**: Hadoop ecosystem components, schedulers, fair and Capacity. | **1** |  |  |
|  | Hadoop 2.0 New Features – Name Node high availability, HDFS federation. | **2** |  |  |
|  | MRv2, YARN, Running MRv1 in YARN. |  | | **2** |  |  |
| **No of Lectures Required to complete Unit 4** | | | | | | **8** | | | |
|  |  | | **Pig** - Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases. | |  | **1** |  |  |
|  | Grunt, Pig Latin, User Defined Functions, Data Processing operators. | | **1** |  |  |
|  | **Spark:** Installing spark, spark applications, jobs, stages and tasks, Resilient Distributed | |  | **1** |  |  |
|  | Databases, anatomy of a Spark job run, Spark on YARN. | |  | **2** |  |  |
|  | **SCALA:** Introduction, classes and objects, basic types and operators, built-in control structures, functions and closures, inheritance. | |  | **2** |  |  |
| **No of Lectures Required to complete Unit 5** | | | | | | **7** | | |

|  |  |
| --- | --- |
| **Text Books & References** | |
| 1 | Big Data Analytics with R and Hadoop by Vignesh Prajapati |
| 2 | Hadoop The Definitive Guide by Tom White |

|  |  |  |
| --- | --- | --- |
| **Course Outcomes (COs)** | | |
| **At the end of this course students will demonstrate the ability to:** | | |
| **CO 1** | Identify Big Data and its business implications. |  |
| **CO 2** | Use various techniques for mining data stream |  |
| **CO 3** | List the components of Hadoop and Hadoop Eco-System. |  |
| **CO 4** | Apply Map Reduce programming model to access and process data on Distributed File System |  |
| **CO 5** | Manage job execution in Hadoop environment and develop Big Data solutions by applying  Hadoop Eco System components |  |

**Name of Faculty with Signature: Shreya Agarwal**

**Head of Department Dean Academics**