**PUNYASHLOK AHILYADEVI HOLKAR SOLAPUR UNIVERSITY, SOLAPUR**
**Final Year B.Tech. (COMPUTER SCIENCE & ENGINEERING)**
**SEMESTER - II**
**CS423A : ELECTIVE – IV : BIG DATA ANALYTICS**

| **Teaching Scheme** | **Examination Scheme** |
|---|---|
| Lectures: 3 Hours /Week, 3 credits | ESE - 70 Marks |
| Practical: 2 Hour/Week, 1 credit | ISE - 30 marks |
| | ICA - 25 marks |

---

**COURSE OUTCOMES:**

**At the end of this course, students will be able to**

1. Comprehend limitations of conventional DBMS and recognize need for Big Data Analytics.
2. Compare Big data processing technologies and choose appropriate one for a given scenario.
3. Use Various Big data technologies for Big data analytics
4. Write Map Reduce program to process Big Data.

---

**SECTION – I**

**Unit 1: Introduction to Types of Digital Data (4)**

Classification of Digital Data, Structured Data, Sources of structured data, Ease with Structured data, Semi-Structured data, sources of semi-structured data, Unstructured data, sources of unstructured data, Issues with terminology, Dealing with unstructured data, Place me in the basket.

**Unit 2: Introduction to Big Data (4)**

Big data, What is big data? Why big data?, Other characteristics of data which are not definitional traits of big data, Challenges with big data, Big data stack, Exercises - Puzzle, Fill in the blanks.

**Unit 3: Big Data Analytics (6)**

Big Data Analytics, Analytics 1.0, Analytics 2.0, Analytics 3.0, Traditional BI vs. Big Data Environment, Terminologies used in Big Data Environment, Big Data Technology Landscape, NoSQL Databases, NoSQL Vs. RDBMS, NewSQL, Hadoop, Hadoop 1.0 vs. Hadoop 2.0, Exercises, Data Science is multidisciplinary, Data Scientist - Your new best friend.

**Unit 4: Introduction to Hadoop (10)**

Introducing Hadoop, Why not RDBMS, Distributed Computing Challenges, A Brief History of Hadoop, Hadoop Overview, Hadoop Components, High Level Architecture of Hadoop, Hadoop Distributed File System, HDFS Architecture, Daemons Related to HDFS, Working with HDFS Command, Special Features of Hadoop, Processing Data With Hadoop, Introduction How Map Reduce Works, Map Reduce Example, Word Count Example using Java Managing Resources and Applications with YARN Introduction, Limitation of Hadoop 1.0, Hadoop 2: HDFS, Hadoop 2: YARN, Interacting with Hadoop EcoSystem Hive, Pig, HBase, Sqoop.

**SECTION – II**

**Unit 5: Introduction to MongoDB (4)**

Recap of NoSQL databases, MongoDB – CRUD, MongoDB- Arrays, Java Scripts, Cursors, Map Reduce Programming, Aggregations.

**Unit 6: Introduction to Cassandra (4)**

Features of Cassandra, CQLSH - CRUD, Collections, Counter, List, Set, Map, Tracing.

**Unit 7: Introduction to Hive (8)**

What is Hive? History of Hive and Recent Releases of Hive, Hive Features, Hive Integration and Work Flow, Hive Data Units, Hive Architecture, Hive Primitive and Collection Data Types, Hive

File Format, Hive Query Language(HQL)–Statements – DDL,DML. Hive Partitions – Bucketing, Views, Sub Query, Joins, Hive User Defined Function, Aggregations in Hive, Group by and Having, Serialization and Deserialization, Hive Analytic Functions.

**Unit 8: Introduction to Pig (4)**

Introducing Pig, History and Anatomy of Pig, Pig on Hadoop, Pig Philosophy, ETL Processing, Pig Latin Overview, Word count example using Pig.

**Internal Continuous Assessment (ICA) :**

- Objective of assignments should be to test students understanding and assess their ability to put into practice the concepts and terminologies learned.
- Assignments must be of nature, which require students to identify the use case scenarios for using technologies mentioned in syllabus.
- It should consist of the 08-10 practical based on following guidelines

1. Basic big data operations using NumPy, SciPy & Pandas.
2. Implementation of Plotting, Filtering and Cleaning a CSV File Data Using NumPy & Pandas.
3. Linear Regression using WEKA.
4. Implement multidimensional visualization by adding variables such as color, size, shape, and label by using Tableau.
5. Apply Filters on Dimensions and Measures for any dataset using tableau.
6. Apply K-means Clustering on iris dataset in tableau.
7. Integrate R with tableau for data visualization.
8. Simple MongoDB and its CRUD Operations
9. Performing import, export and aggregation in MongoDB.
10. Performing CRUD operations using Cassandra.
11. Store the login details of the user such as UserID and Password. The information stored should expire in a day's time using time to live (TTL).
12. Map-Reduce Programming examples
13. Partitioning and processing using Hive.
14. Perform group by, order by, sort by, cluster by, distribute by queries using Hive.
15. Find out frequency of each word (word count) using pig.

**Text Book :**


1. Big Data and Analytics, Seema Acharya, Subhashini Chellappan, - Wiley India Pvt. Ltd.
2. Hadoop: The Definitive Guide, 3rd Edition, Tom White , - O'reilly Media.
3. Programming Hive, Edward Rutherglen, Dean Wampler, Jason Rutherglen, Edward Capriolo. - O'reilly Media.
4. The Definitive Guide to MongoDB: A Complete Guide to Dealing with Big Data Using MongoDB (Definitive Guide Apress) 2e by David Hows, Eelco Plugge, Peter Membrey, Tim Hawkins.
5. Programming Pig, by Alan Gates - O'reilly Media.
6. Cassandra: The Definitive Guide, Eben Hewitt - O'reilly Media.


**Reference Book :**


1. Big Data For Dummies, Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, Wiley Brand.
2. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses (Wiley CIO), Michael Minelli, Michele Chambers, Ambiga Dhiraj : John Wiley & Sons.
3. Mining of Massive Datasets, Anand Rajaraman, Jure Leskovec, Jeff rey D. Ullman, Cambridge University Press.
4. Hadoop in Action, Chuck Lam, Dreamtech Press, ISBN : 978-81-7722-813-7.