# CLUSTER UNIVERSITY OF SRINAGAR
## Syllabus for PG Information Technology
### Batch 2023

3rd SEMESTER: M.Sc. (IT) Batch 2023

**Course Code:**
**Credits: 4 (L- 3, T- 0, P-1)**
**Contact Hrs: 75 (Theory: 45, Tutorial: 30)**

**Title: Big Data Analytics [DSE 1]     Max. Marks: 100**
**Theory External: 60; Min Marks: 24**
**Theory Internal (Continuous Assessment):15 Marks, Min Marks: 06**
**Practical Experimental Basis= 15, Min. Marks: 06**
**Practical Experimental (Continuous assessment) =10, Min. Marks: 04**

## Course Objectives:
- Understand the Big Data Platform and its Use cases.
- Provide an overview of Hadoop and Hadoop Architecture.
- Provide HDFS Concepts and Interfacing with HDFS Understand Map Reduce Jobs, Hadoop YARN, Hadoop2 Execution Model, Hadoop Ecosystem Tools, Hadoop Ecosystem.
- Provide concepts of NoSQL Big Data Management, Mongo DB

## Course Outcomes
After the completion of course, student must be able:
- To understand the building blocks of Big Data.
- To articulate the programming aspects of cloud computing (MapReduce etc).
- To understand the specialized aspects of Big Data with the help of different big data applications.
- To represent the analytical aspects of Big Data along with the knowledge of big data databases such as MongoDB and NoSQL.
- To know the recent research trends related to Hadoop File System, MapReduce and Google File.

## Syllabus

**UNIT – I: Introduction to Data Warehousing and OLAP Technology for Data Mining:**

Data Mining: Definition and uses, , KDD (Knowledge Discovery from Databases) Process, types of Data Can Be Mined?, Data Mining Functionality, Are all the patterns interesting? , Attribute Types, Data Warehouse: Data Warehouse Architecture, Data Cube: A multi-dimensional data model, Schemas for Multidimensional Data Models, OLAP Operations, Data Warehouse Usage (Applications), Data Mining Primitive, Architecture of Data Mining System.

**UNIT – II: Introduction Concept of Big Data**

Big Data: Definition, Web Data, Classification of Data: Structured, Semi-Structured, and Unstructured. Challenges of Conventional System Need of Big Data, Difference between Big Data and Small Data, Importance of Big Data. Big Data Characteristics (4V's Volume, Velocity, Variety, and Veracity), Big Data Types, Big Data Handling Techniques. Complexity of Big Data, Big Data Processing Architectures, Big Data Technologies, and Big Data Business Value. Big Data Analytics Application. Big Data Challenges and Future Scope.

13

**UNIT – III: Introduction to Hadoop and Hadoop Architecture**

Apache Hadoop & Hadoop EcoSystem: Hadoop Core Component, Features of Hadoop, The Hadoop Distributed File System: HDFS data storage, Hadoop Physical Organization, HDFS Commands, MapReduce Framework, MapReduce Programming Model, MapReduce Map Task, Reduce Task and MapReduce Execution, Hadoop YARN, Hadoop2 Execution Model, Hadoop Ecosystem Tools, Hadoop Ecosystem.

**UNIT – IV: NoSQL Big Data Management, Mongo DB**

NoSQL: Definition, uses, Types of NoSQL databases, need of NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NoSQL DataStore, NoSQL Data Architecture pattern, NoSQL to Manage Big Data. Data Base for the Modern Web: Introduction to MongoDB, Features of MongoDB, Data Types, MongoDB Query Language and Database Command.

**Practicals: Credit 01**
- o Implement a basic data mining algorithm (e.g., Decision Tree or K-Means Clustering) to extract patterns from a sample dataset.
- o Perform a data cleaning and preprocessing task on a given dataset (e.g., handling missing values, normalization).
- o Design a simple Data Warehouse schema using a star or snowflake schema for a given case study.
- o Perform OLAP operations (roll-up, drill-down, slice, and dice) on a multidimensional data model using a data cube.
- o Explore and classify a given dataset into structured, semi-structured, and unstructured data.
- o Analyze a real-world dataset to identify the 4V's of Big Data (Volume, Velocity, Variety, Veracity).
- o Implement a basic Big Data processing task using a sample dataset to showcase data handling techniques like filtering and aggregation.
- o Set up a Hadoop environment and perform basic HDFS commands like file creation, deletion, and directory management.
- o Write and execute a basic MapReduce program to process a large dataset (e.g., word count or log file analysis).
- o Develop and run a MapReduce program that performs a specific task (e.g., sorting, filtering) on a large dataset.
- o Explore the Hadoop YARN architecture and execute a job using YARN.
- o Set up a NoSQL database (e.g., MongoDB) and perform CRUD operations (Create, Read, Update, Delete) on a collection.

- o Compare SQL and NoSQL by implementing a simple database query in both SQL and MongoDB Query Language.
- o Design a data model using MongoDB for a given use case, incorporating various data types and creating appropriate collections.
- o Execute complex queries in MongoDB, including indexing, aggregation, and map-reduce functions.

**BOOKS RECOMMENDED:**

- *Tom White " Hadoop: The Definitive Guide" Third Edit on, O'reily Media, 2012.*
- *Analytics,* Raj Kamal and Preeti Saxena, McGraw Hill Education
- *Big Data: Black Book,* DT Educational Services, Dreamtech Press
- *Big Data Analytics,* Seema Acharya & Subhashini Chellappan, Wiley India
- *Big Data Analytics,* M. Vijayalakshmi & Radha Shankaramani, Wiley India

**References**

• Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007. • Jay Liebowitz, "Big Data and Business Analytics" Auerbach Publications, CRC press (2013)

• Tom Plunkett, Mark Hornick, "Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop", McGraw-Hill/Osborne Media (2013), Oracle press.

• Anand Rajaraman and Jef rey David Ulman, "Mining of Massive Datasets", Cambridge University Press, 2012.

• Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", John Wiley & sons, 2012.

• Glen J. Myat, "Making Sense of Data", John Wiley & Sons, 2007; Pete Warden, "Big Data Glossary", O'Reily, 2011.

• Michael Mineli, Michele Chambers, Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley Publications, 2013.

• ArvindSathi, "BigDataAnalytics: Disruptive Technologies for Changing the Game", MC Press, 2012

• Paul Zikopoulos ,Dirk DeRoos , Krishnan Parasuraman , Thomas Deutsch , James Giles , David Corigan , "Harness the Power of Big Data The IBM Big Data Platform ", Tata McGraw Hill Publications, 2012.