DSE-02 (a): Big Data

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Big Data	4	2	0	2	Class XII	DSC-04

***Learning objectives*:**

This course gives an overview of Big Data, i.e. storage, retrieval and processing of big data. In addition, it also focuses on the "technologies", i.e., the tools/algorithms that are available for storage, processing of Big Data. It also helps a student to perform a variety of "analytics" on different data sets and to arrive at positive conclusions.

Learning Outcomes:

1. *Perform data gathering of large data from a range of data sources.*
2. *Critically analyze existing Big Data datasets and implementations, taking practicality, and usefulness metrics into consideration.*
3. *Understand and demonstrate the role of statistics in the analysis of large of datasets*
4. *Select and apply suitable statistical measures and analyses techniques for data of various structure and content and present summary statistics*
5. *Understand and demonstrate advanced knowledge of statistical data analytics as applied to large data sets*
6. *Employ advanced statistical analytical skills to test assumptions, and to generate and present new information and insights from large datasets*

Unit-I **(5 hours)**

Introduction to big data: Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis – Nature of Data - Analytic Processes and Tools - Analysis vs. Reporting.

Unit-II **(5 hours)**

Mining data streams: Introduction to Streams Concepts – Stream Data Model and Architecture - Stream Computing - Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating Moments – Counting Oneness in a Window – Decaying Window - Real time Analytics Platform (RTAP) Applications - Case Studies – Real Time Sentiment Analysis- Stock Market Predictions.

Unit-III **(5 hours)**

Hadoop: History of Hadoop- the Hadoop Distributed File System – Components of Hadoop

Analyzing the Data with Hadoop - Scaling Out- Hadoop Streaming- Design of HDFS- Java interfaces to HDFS Basics- Developing a Map Reduce Application-How Map Reduce Works-Anatomy of a Map Reduce Job Run-Failures-Job Scheduling-Shuffle and Sort – Task execution - Map Reduce Types and Formats- Map Reduce Features Hadoop environment.

Unit-IV **(5 hours)**

Frameworks: Applications on Big Data Using Pig and Hive – Data processing operators in Pig – Hive services – HiveQL – Querying Data in Hive - fundamentals of HBase and Zoo Keeper - IBM Info Sphere Big Insights and Streams.

Unit-V **(10 hours)**

Predictive Analytics: Simple linear regression- Multiple linear regression- Interpretation of regression coefficients. Visualizations - Visual data analysis techniques- interaction techniques - Systems and applications.

References:

1. Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007.
2. Tom White "Hadoop: The Definitive Guide" Third Edition, O'reilly Media, 2012.
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", McGraw Hill Publishing, 2012.
4. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", CUP, 2012.
5. Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", John Wiley& sons, 2012.

***List of Practicals*:** **(60 hours)**

1. (i) Perform setting up and Installing Hadoop in its two operating modes:
a) Pseudo distributed,
b) Fully distributed.
(ii) Use web-based tools to monitor your Hadoop setup.
2. (i) Implement the following file management tasks in Hadoop:
a) Adding files and directories
b) Retrieving files
c) Deleting files
3. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.
a) Find the number of occurrences of each word appearing in the input file(s).
b) Performing a Map Reduce Job for word search count (look for specific keywords in a file).
4. Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.
5. Write a Pig Latin script for finding TF-IDF value for book dataset (A corpus of eBooks available at: Project Gutenberg).
6. Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions.