PARUL UNIVERSITY - Faculty of IT & Computer Science

Department of Computer Application

SYLLABUS FOR 2nd Sem M.Sc. (IT), MCA (A.Y.-II) 2020 PROGRAMME

Big Data Analytics - I (05201295)

Type of Course: M.Sc. (IT), MCA (A.Y.-II) 2020

Prerequisite: Basic Knowledge of Data Analytics and Computation Methods.

Rationale: The objective of this course is to provide Conceptual insight about Big Data Analytics, Installation and understanding of Hadoop Architecture and its ecosystems.

Teaching and Examination Scheme:

Teaching Scheme			Credit	Examination Scheme					Total
				External		Internal			
Lect Hrs/ Week	Tut Hrs/ Week	Lab Hrs/ Week		T	P	T	CE	P	
3	1	2	5	60	30	20	20	20	150

Lect - Lecture, **Tut** - Tutorial, **Lab** - Lab, **T** - Theory, **P** - Practical, **CE** - CE, **T** - Theory, **P** - Practical

Contents:

Sr.	Topic	Weightage	Teaching Hrs.
1	**Overview of Big Data**: Introduction to Big Data, Evolution of Big Data, Structuring of Big Data, Fundamentals of Big Data, Big Data Analytics, Career and Future in Big Data	10%	5
2	**Discovering the Use of Big Data in Business Context**: Big Data in Social Networking, Big Data in Preventing Fraudulent Activities, Big Data in Detecting Fraudulent Activities in Insurance Sector, Use of Big Data in Retail Industry.	10%	5
3	**Technologies for Handling Big Data**: Distributed and Parallel Computing for Big Data, Introducing Hadoop, Cloud Computing and Big Data, In-Memory Computing Technology for Big Data.	8%	4
4	**Understanding Hadoop Ecosystem**: History of Hadoop, Hadoop Ecosystem, Analysing Data with Unix tools, Analysing Data with Hadoop, Hadoop Streaming, Hadoop Echo System, IBM Big Data Strategy, Introduction to Infosphere BigInsights and Big Sheets.	12%	6
5	**HDFS(Hadoop Distributed File System)**: The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro and File-Based Data structures.	14%	6
6	**MapReduce Fundamentals**: Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features.	14%	6

7	**Understanding Big Data Technology Foundations, Storing Data in Databases and Data Warehouses**: Understanding Big Data Technology Foundations: Exploring the Big Data Stack, Virtualization and Big Data, Virtualization Approaches, Summary, Quick Revise Storing Data in Databases and Data Warehouses: RDBMS and Big Data, Non-Relational Database, Polyglot Persistence, Integrating Big Data with Traditional Data Warehouses, Big Data Analysis and Data Warehouse, Changing Deployment Models in Big Data Era.	16%	8
8	**Processing Your Data with MapReduce**: Recollecting the Concept of MapReduce Framework,Developing Simple MapReduce Application, Points to Consider while Designing MapReduce.	16%	8

***Continuous Evaluation:**

It consists of Assignments/Seminars/Presentations/Quizzes/Surprise Tests (Summative/MCQ) etc.

Reference Books:

1. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses
   Michael Minelli, Michelle Chambers, and Ambiga Dhiraj; Wiley, 2013
2. Big-Data Black Book
   DT Editorial Services; Wiley India
3. Hadoop Operations, Eric Sammer, O'Reilley.
   O'Reilley
4. Hadoop: The Definitive Guide by Tom White, Third Edition, O'Reilley.
   Tom White

Course Outcome:

After Learning the course the students shall be able to:

1. Work with big data platform and explore the big data analytics techniques business applications.
2. Understand the fundamentals of various big data analytics techniques.
3. Analyse the HADOOP and Map Reduce technologies associated with big data analytics.
4. Have knowledge on accessing, storing and manipulating the huge data from different resources.

List of Practical:

1. **Practical-1**

   (i)Perform setting up and Installing Hadoop in its two operating modes:
   - Pseudo distributed,
   - Fully distributed.

   (ii) Use web based tools to monitor your Hadoop setup.

2. **Practical-2**

   (i) Implement the following file management tasks in Hadoop:
   - Adding files and directories
   - Retrieving files
   - Deleting files

   ii) Benchmark and stress test an Apache Hadoop cluster

3. **Practical-3**

   Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.
   - Find the number of occurrence of each word appearing in the input file(s)
   - Performing a Map Reduce Job for word search count (look for specific keywords in a file)

4. **Practical-4**

To understand the overall programming architecture using Map Reduce API

5. **Practical-5**

Stop word elimination problem:

- Input:

A large textual file containing one sentence per line
A small file containing a set of stop words (One stop word per line)

- Output:

A textual file containing the same sentences of the large input file without the words appearing in the small file.

6. **Practical-6**

Write a Map Reduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. Data available at:
https://github.com/tomwhite/hadoopbook/tree/master/input/ncdc/all.

- Find average, max and min temperature for each year in NCDC data set?
- Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.

7. **Practical-7**

Store the basic information about students such as roll no, name, date of birth , and address of student using various collection types such as List, Set and Map

8. **Practical-8**

Basic CRUD operations in MongoDB

9. **Practical-9**

Retrieve various types of documents from students collection

10. **practical-10**

To find documents from Students collection

11. **Practical-11**

Develop Map Reduce Work Application

12. **Practical-12**

Creating the HDFS tables and loading them in Hive and learn joining of tables in Hive