CVM UNIVERSITY
(Established under Gujarat Private Universities (Second Amendment) Act : 2019 Gujarat Act No. 20 of 2019)

FACULTY OF ENGINEERING & TECHNOLOGY

First Year Master of Engineering

Semester I

Course Code: 102340102

Course Title: Big Data

Type of Course: Program Elective I/Program Elective II

Course Objectives: To understand basics of Big Data. To understand various Big Data Tools.

Teaching & Examination Scheme:

Contact hours per week			Course Credits	Examination Marks (Maximum / Passing)				
				Internal		External		Total
Lecture	Tutorial	Practical		Theory	J/V/P*	Theory	J/V/P*	
3	0	2	4	30 /15	20 /10	70/35	30/15	150 /75

* **J**: Jury; **V**: Viva; **P**: Practical

Detailed Syllabus:

Sr.	Contents	Hours
1	INTRODUCTION TO BIG DATA Introduction– Distributed file system–Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications.	2
2	NoSQL and MongoDB What is it? Where It is Used Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL Introduction to MongoDB - Architecture and Installation, Schema Design and Data Modelling, MongoDB query language - CRUD Operations, Indexing and Aggregation	6
3	INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE, HDFS Big Data – Apache Hadoop & Hadoop Ecosystem, Moving Data in and out of Hadoop MapReduce: Architecture, Map-reduce job execution workflow, Data Serialization, Input Formats, output formats, Counters, Sorting, Joins HDFS-Overview, Installation and Shell, Data flow, Setting up Hadoop cluster and, Administrating Hadoop YARN: Architecture, Scheduling	9
4	FLUME AND SQOOP Flume: Architecture, Data flow, Configuration, Fetching twitter data Sqoop: Import, Export	3
5	HIVE, PIG, HBase, Oozie, Zookeeper	10

	Hive: Hive vs MapReduce, Hive DDL – Create/Show/Drop Tables, Internal and External Tables, Hive DML – Load Files & Insert Data, Hive Architecture & Components, Difference between Hive and RDBMS, Partitions in Hive, HiveQL PIG: PIG vs MapReduce, PIG Architecture & Data types, Shell and Utility components, PIG Latin - Relational Operators, File Loaders and UDF, Limitations of PIG. HBase: HBase concepts, Advanced Usage, Schema Design, Advance Indexing, Oozie: Oozie – Simple/Complex Flow, Oozie Workflow, Oozie Components, Demo on Oozie Workflow in XML Zookeeper: What is Zookeeper? Features of Zookeeper, Zookeeper Data Model	
6	SPARK Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine Learning with MLlib.	**5**
7	SPLUNK Splunk Basics, User Management & Splunk Configuration Files, Data Ingestion, Splunk Search, and Reporting Commands, Splunk Alerts, Visualizations, Reports, & Dashboards	**5**

Suggested Specification table with Marks (Theory) (Revised Bloom's Taxonomy):

Distribution of Theory Marks						**R**: Remembering; **U**: Understanding; **A**: Application, **N**: Analyze; **E**: Evaluate; **C**: Create
R	**U**	**A**	**N**	**E**	**C**	
20%	20%	15%	20%	15%	10%	

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

1	Boris lublinsky, Kevin t. Smith, AlexeyYakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.
2	Chris Eaton,Dirk derooset al. , "Understanding Big data ", McGraw Hill, 2012.
3	BIG Data and Analytics , Sima Acharya, Subhashini Chhellappan, Willey.
4	MongoDB in Action, Kyle Banker,Piter Bakkum , Shaun Verch, Dream tech Press.
5	Tom White, "HADOOP: The definitive Guide", O Reilly 2012.
6	VigneshPrajapati, "Big Data Analyticswith R and Haoop", Packet Publishing 2013.
7	Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau.

Course Outcomes (CO):

Sr.	Course Outcome Statements	%weightage
CO-1	Students will to build and maintain reliable, scalable, distributed systems with Apache Hadoop.	**10%**
CO-2	Students will be able to write Mapreduce based Applications	**20%**
CO-3	Students will be able designed and build big data applications using pig,hive and hbase.	**30%**
CO-4	Students will learn difference between conventional SQL query language and NoSQL basic concepts	**10%**
CO-5	Students will learn tips and tricks for Big Data use cases and solutions.	**10%**
CO-6	Students will be able to design and build Spark based Big data Applications.	**20%**

List of Practicals / Tutorials:

1	Installation of MongoDB and try following operation: CURD Operation, Aggregation and indexing operation
2	Configure Hadoop cluster in pseudo distributed mode. Try Hadoop basic commands
3	Write Map Reduce code for following A. Count frequency of word from a large file. B. Find year wise maximum temperature using whether data set. C. find out what are the top 5 categories with maximum number of videos uploaded from youtube dataset.
4	Write a Word Count program using partitioner and combiner. Configure multimode Hadoop Cluster.
5	Configure flume and load Twitter data into HDFS
6	Configure Sqoop and try basic Sqoop Commands to load data to and from MySql database to hadoop.
7	Configure Hive and Write a hive commands for following for Olympic dataset. 1. Create table 2. Load data 3. list the total number of medals won by each country in swimming. 4. Display real life number of medals India won year wise. 5. Find the total number of medals each country won display the name along with total medals. 6. Find the real life number of gold medals each country won. 7. Which country got medals for Shooting, year wise classification?
8	The dataset is a simple text (movies_data.csv) file lists movie names and its details like release year, rating and runtime. Write HIVEQL for the followings. A sample of the dataset is as follows: 1,The Nightmare Before Christmas,1993,3.9,4568 2,The Mummy,1932,3.5,4388 3,Orphans of the Storm,1921,3.2,9062 4,The Object of Beauty,1991,2.8,6150 5,Night Tide,1963,2.8,5126 A. Load the data B. List the movies that having a rating greater than 4 C. Store the result of previous query into file D. List the movies that were released between 1950 and 1960 E. List the movies that have duration greater that 2 hours F. List the movies that have rating between 3 and 4 G. List the movie names its duration in minutes H. List all the movies in the ascending order of year. I. List all the movies in the descending order of year. J. list the distinct records present movies_with_dups: K. Use the LIMIT keyword to get only a limited number for results from relation. L. Use the sample keyword to get sample set from your data. M. To view the step-by-step execution of a sequence of statements use the ILLUSTRATE command.

9	Configure Pig. The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. find the following from airline databse. Which month has seen the most number of cancellations due to bad weather? Top 10 route (origin and dest) that has seen maximum diversions? Top 5 visited destination.
10	Configure Hbase and Try different command
11	Configure Hb Setting up Splunk environment and search and navigate in Splunk, use fields, get statics from your data, create reports, dashboards, lookups, and alerts,creating workflow actions and data models ase and Try different command
12	Configure Spark and implement collaborative filtering recommendation using spark.
13	Demonstrate use of Spark Machine learning library.

Supplementary learning Material:	
1	http://www.bigdatauniversity.com/
2	https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics
3	https://www.analyticsvidhya.com/blog/2015/07/big-data-analytics-youtube-ted-resources/

Curriculum Revision:	
Version:	**1**
Drafted on (Month-Year):	Apr-20
Last Reviewed on (Month-Year):	Jul-20
Next Review on (Month-Year):	Apr-22