Krantiguru Shyamji Krishna Verma Kachchh University
Master of Science (Information Technology)
Semester: III

Paper Code: CCCS306		**Total Credit :** 4
Title of Paper: Data Science		**Total Marks :** 70 **Time :** 3 Hrs
Unit	**Description**	**Weighting**
I	**An Introduction to Big Data** Challenges, Managing varieties of Data, The Emerging Big Data Stack, Gartner hype cycle for Big Data emerging technologies, Big Data life Cycle, Types of Data (Unstructured, Structured, semi-structured) Opportunities in Big Data. Introduction to NoSQL: Difference between RDBMS and NoSQL, CAP Theorem for NoSQL, Features / Advantages of NoSQL, Types of NoSQL (Document, Key-Value, Columnar, Graph)	**20%**
II	**Apache Hadoop** Introduction, Hadoop eco-System, High Level Architecture: Component Level Architecture: MapReduce with Yarn, HDFS/ HDFS2, introduction to Yarn, Features of Yarn , Intro to Tez, Features of Tez, Introduction and Features : Pig, Hive, Hbase. Distributed publish – subscribe Messaging: Apache Kafka Distributed MapReduce: Introduction to Apache Spark	**20%**
III	**Hadoop Distributed File System** HDFS Architecture, HDFS Read / Writes processes, HDFS Performance tuning: Overview of HDFS Access, API's & Applications. HDFS Commands, Native Java APIs, Rest APIs.	**20%**
IV	**An Introduction to MapReduce** Introduction to Map-Reduce, Map-Reduce Hands-on with Hadoop streaming. Introduction to Hbase, Hbase vs HDFS, Features/Adv. Of Hbase, Hbase Data Model best practices. [Hands-on]: setup single node Hbase cluster on Ubuntu, configuration setup. Introduction to Hive, how Hive works? Component level architecture: Hive, Hive Commands, Hive Query Language.	**20%**
V	**Distributed MapReduce Computing with Apache Spark** An introduction to Apache Spark, features / advantages of Spark, component level architecture, Resilient Distributed Datasets (RDDs), Parallelized Collections, External Datasets, RDD Operations, Passing functions to Spark, Understanding closures, Printing elements of an RDD, Working with Key-Value Pairs, Transformations, Actions, Shuffle operations, RDD Persistence, Removing Data, Shared Variables, Broadcast Variables, Accumulators. Map-Reduce on file / streaming with spark, Machine Learning with Spark Mlib – Clustering, Regression, Recommender, Graph Analytics: Introduction to Graphx, Features of Graphx, Basic path analytics algorithm with Graphx, Implement Dijkstra Algorithm with GraphX. Data Visulization: An Introduction to Data Viz., Various BI tools, Data Visualization with Tableau.	**20%**
Basic Text & Reference Books :-		
1.	Hadoop: The Definitive Guide, 3rd Edition By Tom White, O'Reilly	
2.	Learning Spark: Lightning-Fast Big Data Analysis by Andy Konwinski, Holden Karau, and Patrick Wendell, O'Reilly	

Krantiguru Shyamji Krishna Verma Kachchh University
Master of Science (Information Technology)
Semester: III

Paper Code: CCCS306			**Total Credit :** 4
Title of Paper: Data Science			**Total Marks :** 70 **Time :** 3 Hrs
Unit	**Description**		**Total Marks**
I	Q.1 (A) Answer the Following. (Definitions, Blanks, Full Forms, True/False, Match the Following)	06	14
	Q.1 (B) Medium / Long Questions. (With Internal Option)	08	
II	Q.2 (A) Answer the Following. (Definitions, Blanks, Full Forms, True/False, Match the Following)	06	14
	Q.2 (B) Medium / Long Questions. (With Internal Option)	08	
III	Q.3 (A) Short / Medium Questions (With Internal Option)	06	14
	Q.3 (B) Medium / Long Questions. (With Internal Option)	08	
IV	Q.4 (A) Short / Medium Questions (With Internal Option)	06	14
	Q.4 (B) Medium / Long Questions. (With Internal Option)	08	
V	Q.5 (A) Short / Medium Questions (With Internal Option)	06	14
	Q.5 (B) Medium / Long Questions. (With Internal Option)	08	