PROGRAM	**Master of Business Administration (Business Analytics)**
SEMESTER	**4**
COURSE TITLE	**Hadoop & Apache Spark**
COURSE CODE	**04MB0456**
COURSE CREDITS	**2**
COURSE DURATION	**28 Sessions**

COURSE OUTCOMES:

- Break down the fundamentals of Hadoop and Spark, the technology that is revolutionizing the Analytics and Big Data world.
- Judge it, performs faster than Map Reduce for Iterative Algorithms or Interactive Data Mining.
- Integrate and Analyze the in-memory cluster computing for lightning fast speed and supports Python, R, and Scala APIs for ease of development.
- Integrate a wide range of data processing scenarios by combining SQL, streaming and complex analytics together seamlessly in the same application.
- Analyze and Judge on top of Hadoop, Mesos, standalone, or in the cloud, diverse data sources such as HDFS, Cassandra, HBase, or S3.

COURSE CONTENTS:

Unit No	Unit / Sub Unit	Sessions
I	Introduction, Architecture, HDFS, Cluster, Ecosystems, Map Reduce: Introduction:Big Data concepts, Data Analytics Architecture, Compare Hadoop vs traditional systems, Attributes of Big Data, Types of data, other technologies vs Big Data. Hadoop Architecture and HDFS:What is Hadoop? Hadoop History, Distributing Processing System, Core Components of Hadoop, HDFS Architecture, Hadoop Master – Slave Architecture, Daemon types - Learn Name node, Data node, Secondary Name node. Hadoop Clusters and the Hadoop Ecosystem: What is Hadoop Cluster? Pseudo Distributed mode, Type of clusters, Hadoop Ecosystem, Pig, Hive, Oozie, Flume, SQOOP. Hadoop Map Reduce Framework: Overview of Map Reduce Framework, Map Reduce Architecture, Learn about Job tracker and Task tracker, Use cases of Map Reduce, Anatomy of Map Reduce Program.	8
II	Hive and HiveQL, PIG, Apache SQOOP, Flume, NoSQL, Oozie and Zookeeper: Hive and HiveQL: What is 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 PIG: PIG vs MapReduce, PIG Architecture & Data types, Shell and Utility components, PIG Latin Relational Operators, PIG Latin: File Loaders and UDF, Programming structure in UDF, PIG Jars Import, limitations of PIG. Apache SQOOP, Flume: Why and what is SQOOP? SQOOP Architecture, Benefits of SQOOP, Importing Data Using SQOOP, Apache Flume Introduction, Flume Model and Goals, Features of Flume, Flume Use Case. NoSQL Databases: What is HBase? HBase Architecture, HBase Components, Storage	12

	Model of HBase, HBase vs RDBMS. Oozie and Zookeeper: Oozie – Simple/Complex Flow, Oozie Workflow, Oozie Components, Demo on Oozie Workflow in XML, What is Zookeeper? Features of Zookeeper, Zookeeper Data Model.	
III	Introduction to Spark, Resilient Distributed Dataset and Data Frames: Introduction to Spark: What is Spark and what is its purpose?, Components of the Spark unified stack, Resilient Distributed Dataset (RDD), Downloading and installing Spark standalone, Scala and Python overview, Launching and using Spark's Scala and Python shell. Resilient Distributed Dataset and Data Frames: Understand how to create parallelized collections and external datasets, Work with Resilient Distributed Dataset (RDD) operations, Utilize shared variables and key-value pairs. Introduction to Spark libraries: Understand and use the various Spark libraries. Spark configuration, monitoring and tuning: Understand components of the Spark cluster, Configure Spark to modify the Spark properties, environmental variables, or logging properties, Monitor Spark using the web UIs, metrics, and external instrumentation, Understand performance tuning considerations. Graphs and Social Media data and construct recommendation systems	8
	Practical: • Optimization of Geographic Cluster • Minimize Network traffic among Cluster • Word Count Mapper examples • Program applications using tools like Hive, pig, NO SQLfor Big data Applications • Construct scalable algorithms for large Datasets using Map Reduce techniques • Event deduction model • Implement algorithms forClustering, Classifying and finding associations in Big Data • Design and implement algorithms to analyze Big data like streams, Web	

EVALUATION:

The students will be evaluated on a continuous basis and broadly follow the scheme given below:

	Component	**Weightage**
A	Continuous Evaluation Component (Assignments / Presentations/ Quizzes / Class Participation/ Practical Record / Practical Examination etc.)	20% (C.E.C)
B	Internal Assessment	30% (I.A.)
C	End-Semester Examination (Practical / Viva)	50% (Practical/Viva) (External Assessment)

SUGGESTED READINGS:

Text Books:

Sr. No	**Author/s**	**Name of the Book**	**Publisher**	**Edition and Year**
T-01	Garry Tarkington	Hadoop Beginner's Guide	PACKT Publishing Ltd.,	1 st Edition, 2015
T-02	Muhammad Asif Abbasi	Learning Apache Spark 2	PACKT Publishing Ltd.,	1 st Edition, 2017

Reference Books:

Sr. No	**Author/s**	**Name of the Book**	**Publisher**	**Edition and Year**
R-01	Benjamin Bengfort, Jenny Kim	Data Analytics with Hadoop: An Introduction for Data	O' Reilly Media Inc.,	1 st Edition , 2016.
R-02	Jonathan R. Owens, Brain Femiano, Jon Lentz	Hadoop Real-World solution's Cook book	PACKT Publishing Ltd.,	1 st Edition , 2016.
R-03	Bill Chambers, Matei Zahria	Spark: The Definitive Guide	O' Reilly Media Inc.	1 st Edition , 2018.