Course Code:	**Subject Title: Insights of Big Data**
	Year and Semester: IV Year I Semester

Course Objectives:

1. To understand the complexity and volume of Big Data and their challenges.
2. To analyze the various methods of data collection.
3. To comprehend the necessity for pre-processing Big Data and their issues.
4. To understand predictive analytics and descriptive analytics.
5. To understand and implement Big Data Analytics with data convergence and Business Maturity Model.

Course Outcomes: At the end of the course student will be able to:

COS' STATEMENT

CO1 Identify the various sources of Big Data.

CO2 List the components of Hadoop and Hadoop Eco-System and Analyze file systems such as GFS and HDFS.

CO3 Apply map reduce concepts for desired applications.

CO4 Demonstrate the Pig architecture and evaluation of pig scripts.

CO5 Implement Big Data Activities using Hive

UNIT - I **9 Hours**

INTRODUCTION TO BIG DATA: Data, Characteristics of data and types of digital data, Sources of data, Working with unstructured data, Evolution and definition of big data, Characteristics and need of big data, Challenges of big data.

BIG DATA ANALYTICS: Overview of business intelligence, Data science and analytics, Meaning and characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment.

UNIT - II **9 Hours**

INTRODUCTION TO HADOOP : Introducing hadoop, Need of hadoop, Limitations of **RDBMS, RDBMS** versus hadoop, Distributed computing challenges, History of hadoop , Hadoop overview, Use case of hadoop, Hadoop distributors, HDFS (Hadoop distributed file system), Processing data with hadoop, Managing resources and applications with hadoop YARN (yet another resource negotiator), Interacting with hadoop ecosystem.

UNIT - III **9 Hours**

INTRODUCTION TO MAPREDUCE PROGRAMMING: Introduction-mapper, reducer, combiner, partitioner, searching, sorting, compression, real time applications using mapreduce, combiner, partitioner, matrix multiplication using mapreduce and page rank algorithm using mapreduce.

UNIT - IV **9 Hours**

INTRODUCTION TO PIG: The anatomy of pig, Pig on hadoop, Pig philosophy, Usecase for pig, ETL processing, Pig latin overview, Data types in pig, Running pig, Execution modes of pig, HDFS commands, Relational operators, Piggy bank, Word count example using pig, Pig at Yahoo.

UNIT - V **9 Hours**

INTRODUCTION TO HIVE: Introduction to hive, Hive architecture, Hive data types, Hive file

format, Hive query language (HQL).
HIVE: Partitions and bucketing, RCFile Implementation, working with XML files, User-defined Function (UDF) in Hive, Pig versus Hive.

Text Books:

1. Seema Acharya, Subhashini Chellappan, "Big Data and Analytics", 1st edition, Wiley, Publishers, 2015.

Reference Books:

1. Boris lublinsky, Kevin t. Smith, AlexeyYakubovich, "Professional Hadoop Solutions",
2. 1st edition, Wiley, 2015.
3. Chris Eaton, Dirkderoosetal, "Understanding Big data ", 1st edition, McGraw Hill, 2012.
4. Tom White, "HADOOP: The definitive Guide", 1st edition, O Reilly 2012.
5. Vignesh Prajapati, "Big Data Analytics with R and Hadoop",1st edition, Packet Publishing,
6. 2013.

Software Links:

1. **Hadoop:** http://hadoop.apache.org/
2. **Hive:**https://cwiki.apache.org/confluence/display/Hive/HomePigLatin: http://pig.apache.org/docs/r0.7.0/tutorial.html

Micro Syllabus of Insights of Big Data

T I : INTRODUCTION TO BIG DATA: Data, Characteristics of data and types of digital data, Sources of data, Working with unstructured data, Evolution and definition of big data, Characteristics and need of big data, Challenges of big data.

DATA ANALYTICS: Overview of business intelligence, Data science and analytics, Meaning and characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment.

Unit	Module	Micro Content
UNIT I	Introduction to Big Data	Data
		Characteristics of data
		Types of digital data
		Sources of data
		Working with unstructured data
		Evolution and Definition of big data
		Characteristics and Need of big data
		Challenges of big data
		Evolution and Definition of big data
		Characteristics and Need of big data
	Big Data Analytics	Overview of business intelligence
		Data science and Analytics
		Meaning and Characteristics of big data analytics
		Need of big data analytics
		Classification of analytics

		Challenges to big data analytics
		Importance of big data analytics
		Basic terminologies in big data environment

T – II: INTRODUCTION TO HADOOP : Introducing hadoop, Need of hadoop, Limitations of RDBMS, RDBMS versus hadoop, Distributed computing challenges, History of hadoop , Hadoop overview, Use case of hadoop, Hadoop distributors, HDFS (Hadoop distributed file system), Processing data with hadoop, Managing resources and applications with hadoop YARN (yet another resource negotiator), Interacting with hadoop ecosystem.

Unit	Module	Micro Content
UNIT II	**Basics of Hadoop**	Introducing Hadoop
		Need of Hadoop
		Limitations of RDBMS
		RDBMS versus Hadoop
		Distributed Computing Challenges
		History of Hadoop
	Hadoop Overview	Hadoop Overview
		Hadoop distributors
		Use case of hadoop
		Processing data with hadoop
		Hadoop distributed file system
		Hadoop Eco System
		YARN Yet Another Resource Negotiator

T – III : INTRODUCTION TO MAPREDUCE PROGRAMMING: Introduction-mapper, reducer, combiner, partitioner, searching, sorting, compression, real time applications using mapreduce, combiner, partitioner, matrix multiplication using mapreduce and page rank algorithm using mapreduce.

Unit	Module	Micro Content
UNIT III	**MapReduce Programming**	Introduction
		Maper
		Reducer
		Combiner
		Partitioner
		Searching
		Sorting
		Compression
		real time applications using mapreduce
		matrix multiplication using mapreduce
		page rank algorithm using mapreduce

T - IV : INTRODUCTION TO PIG: The anatomy of pig, Pig on hadoop, Pig philosophy, Usecase for pig, ETL processing, Pig latin overview, Data types in pig, Running pig, Execution modes of pig, HDFS commands, Relational operators, Piggy bank, Word count example using pig, Pig at Yahoo.

	Module	Micro Content
UNIT IV	**Introduction and Overview of Pig**	PIG
		Motivation of using pig
		The anatomy of pig
		Key Features of Pig
		Pig on Hadoop
		Pig Philosophy
		Usecase for pig -ETL Processing
		Pig Latin Overview
		Pig Data Types
		Running Pig
		Execution Modes of Pig
		HDFS Commands
		Pig Operations
		Piggy Bank
		Word count example using pig
		Pig at Yahoo

T V : INTRODUCTION TO HIVE: Introduction to hive, Hive architecture, Hive data types, Hive file format, Hive query language (HQL)

E: Partitions and bucketing, RCFile Implementation, working with XML files, User-defined Function (UDF) in Hive, Pig versus Hive.

Unit	Module	Micro Content
UNIT V	**Hive**	Introduction to Hive
		Hive architecture
		Hive Data Types
		Hive File Foramt
		Hive Query Language (HQL)
		Hive Partitions and bucketing
		RCFile Implementatiom
		Working with XML files
		User defined Function (UDF) in Hive
		Pig versus Hive