19CS520	BIG DATA ANALYTICS (COMMON TO CSE,BME, ECE)	L	T	P	C
		3	0	0	3

PREAMBLE

This course focuses on big data technologies used for storage, analysis and manipulation of data. The student will learn about fundamentals of Hadoop, MapReduce, Pig, Hive, R and have hand on training on the same It also help to develop projects and apply existing data analytics tools to gain comprehensive knowledge on Data analytics.

PREREQUISITE

19CS411- Virtualization and Cloud Computing / 19CS519 – Data Science

COURSE OUTCOMES

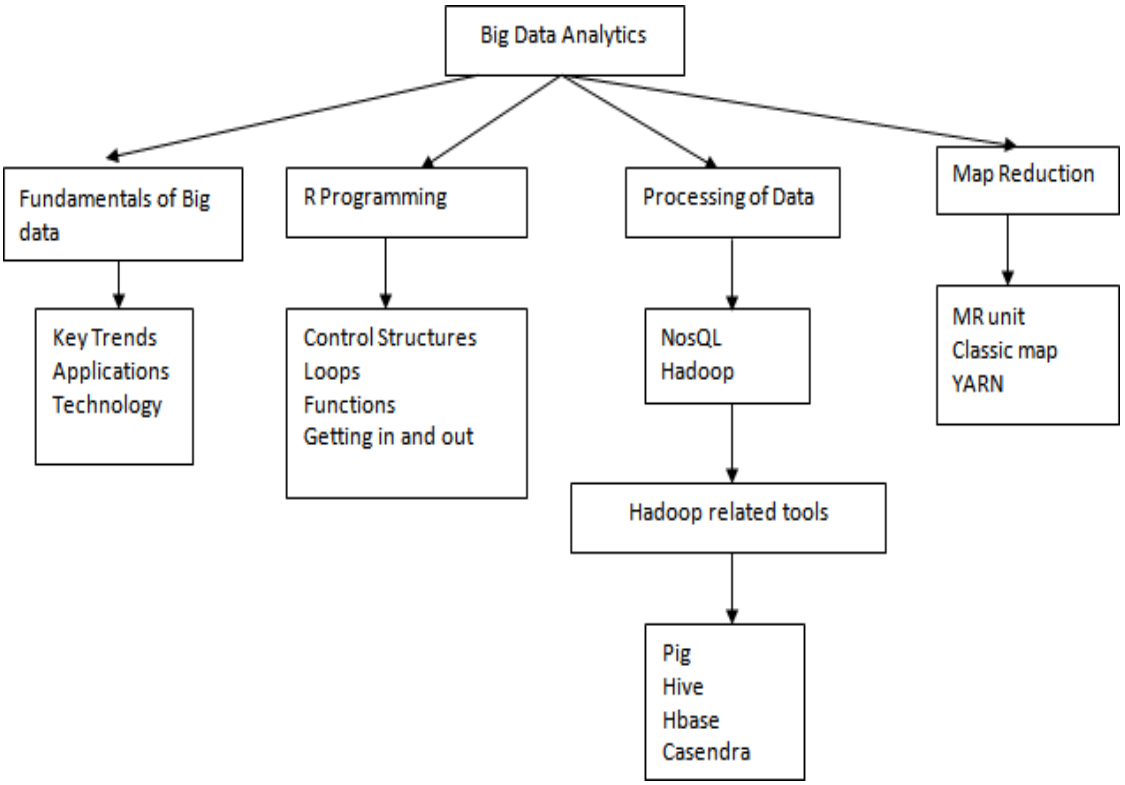
At the end of the course learners will be able to		
CO1	Outline the basic big data concept.	Remember
CO2	Categorize and summarize the processing in Big Data and its importance.	Understand
CO3	Simulate various Big data technologies like Hadoop MapReduce, Pig, Hive, Hbase and No- SQL.	Apply
CO4	Determine tools and techniques to analyze Big Data.	Analyse
CO5	Resolve problems associated with big data with the features of R programming	Create

MAPPING OF COs WITH POs & PSOs

COURSE OUT COMES	PROGRAMME OUTCOMES												PROGRAMME SPECIFIC OUTCOMES			
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO 10	PO 11	PO 12	PSO1	PSO2	PSO3	PSO4
CO1	-	3	-	-	3	-	-	-	-	-	-	-	-	3	-	3
CO2	2	3	1	3	2	-	-	-	-	-	-	-	2	3	-	3
CO3	1	2	3	2	2	2	-	-	-	-	-	-	2	2	-	2
CO4	-	2	2	-	3	-	-	-	-	-	-	-	1	2	2	-
CO5	1	2	-	-	3	2	-	-	2	-	-	3	1	3	-	2

1. LOW 2. MODERATE 3. SUBSTANTIAL

CONCEPT MAP



SYLLABUS

UNIT I FUNDAMENTALS OF BIG DATA 9

Introduction to big data - convergence of key trends – Applications of big data in web analytics, marketing, fraud, risk, credit risk management, algorithmic trading, advertising, healthcare and medicine – big data technologies – introduction to open source technologies – Hadoop, cloud and big data – mobile business intelligence – Crowd sourcing analytics – inter and trans firewall analytics - Information Management .

UNIT II PROCESSING BIG DATA 9

Introduction to NoSQL – aggregate, key-value and document data models – relationships – graph, schema less databases – materialized views – distribution models – sharding – master-slave and peer-peer replication – consistency & relaxing – version stamps – map reduce – partitioning and combining – composing map-reduce calculations- Analysis of Data format with Hadoop – scaling out – Hadoop streaming & pipes – Design of Hadoop distributed file system (HDFS) and related concepts- Selecting appropriate execution modes: local, pseudo-distributed, fully distributed.

UNIT III MAP REDUCE APPLICATIONS 9

Employing Hadoop Map Reduce - Creating the components of Hadoop Map Reduce jobs & workflows – unit tests with MRUnit – test data and local tests – anatomy of MapReduce job run – classic Map-reduce – Distinguishing Hadoop daemons - YARN – failures in classic Map-reduce and YARN – job scheduling – shuffle and sort – task execution – MapReduce types – input & output formats

UNIT IV HADOOP RELATED TOOLS **9**

Hbase – data model and implementations – Hbase clients – Hbase examples – praxis.Cassandra – cassandra data model – cassandra examples – cassandra clients – Hadoop integration. Pig – Grunt – pig data model – Pig Latin – developing and testing Pig Latin scripts. Hive – data types and file formats – HiveQL data definition – HiveQL data manipulation – HiveQL queries.

Compute the sum of uploaded ,downloaded total transmitted bytes using Pig.Find the minimum value of client MSS.

UNIT V R PROGRAMMING **9**

R Nults and Bolts -Getting Data In and Out of R- Control Structures and Functions- Loop Functions- Data Manipulation (dplyr, reshape2 packages)- String Operations (stringr package)- Packaging, Debugging and Object Oriented Programming- Graphics and Data Visualization (ggplot2 package) – Clustering- Regression and Classification - swirl Programming.

TOTAL : 45 PERIODS

TEXT BOOKS

1. Michael Minelli, Michelle Chambers and AmbigaDhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
2. ArvindSathi, "Big Data Analytics: Disruptive Technologies for Changing the Game", 1st Edition, IBM Corporation, 2012.

REFERENCES

1. Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", 1st Edition, Wiley and SAS Business Series, 2012.
2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
4. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
5. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.

COURSE DESIGNER

1. Dr.P.Valarmathie - valarmathie@saveetha.ac.in