DJ19CEEC6011

Big Data Infrastructure

Dr. Nilesh M. Patil

Associate Professor

Computer Engineering Department

SVKM’s D J Sanghvi College of Engineering

Syllabus Scheme

Program: Third Year B.Tech. in Computer Engineering									Semester : VI			
Course : Big Data Infrastructure									Course Code: DJ19CEEC6011			
Course : Big Data Infrastructure Laboratory									Course Code: DJ19CEEL6011			
Teaching Scheme (Hours / week)				Evaluation Scheme								
				Semester End Examination Marks (A)			Continuous Assessment Marks (B)				Total marks (A+ B)	
Lectures	Practical	Tutorial	Total Credits	Theory			Term Test 1	Term Test 2	Avg.			
				75			25	25	25		100	
				Laboratory Examination			Term work					
3	2	-	4	Oral	Practical	Oral &Practical	Laboratory Work	Tutorial / Mini project / presentation/ Journal	Total Term work		50	
				25	-	-	15	10	25			

Pre-requisite: Databases, Python ,Java,R, Linux OS

Objectives:

1. To define big data solutions for business intelligence.
2. To analyze business case studies for big data analytics.
3. To develop map-reduce analytics using Hadoop and related tools.
4. To perform data storage and management using NoSQL.
5. To perform real-time analysis on streaming data.

Outcomes: On completion of the course, the learner will be able to:

1. Describe big data and use cases from selected business domains.
2. Perform map-reduce analytics using Hadoop.
3. Use Hadoop-related tools such as HBase, Cassandra, Pig, and Hive for big data analytics.
4. Build and maintain reliable, scalable, distributed systems using Apache Spark.
5. Design and build MongoDB-based Big data Applications and learn MongoDB query language.
6. Use streaming tools for real-time analysis of bigdata.

UNIT 1
INTRODUCTION TO BIG DATA AND HADOOP

- Introduction to Big Data
- Distributed file system
- Big Data characteristics, Drivers, types of Big Data,
- Traditional vs. Big Data business approach,
- Case Study of Big Data Solutions.
- Big data Applications
- Societal and Ethical issues associated with the use of big data analytics
- The key privacy issues.
- **2 Hours**
- **Marks: 10 (approx.)**

Unit 2

INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE

- Big Data – Apache Hadoop & Hadoop EcoSystem
- Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce Concept of Hadoop
- HDFS Commands
- MapReduce-The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution
- **8 Hours**
- **Marks: 20 (approx.)**

Unit 3
HDFS, HIVE AND HIVEQL, HBASE

- HDFS-Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting, and Aggregating,
- Map Reduce Scripts, Joins & Subqueries
- HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIGGrunt – pig data model – Pig Latin – developing and testing Pig Latin scripts
- Zookeeper , how it helps in monitoring a cluster
- Build Applications with Zookeeper and HBase
- **12 Hours**
- **Marks: 30 (approx.)**

Unit 4
SPARK

- Introduction to Data Analysis with Spark
- Downloading Spark and Getting Started
- Programming with RDDs
- Machine Learning with MLlib.
- **6 Hours**
- **Marks: 15 (approx.)**

Unit 5
NoSQL

- Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL,
- Introduction to MongoDB key features
- Core Server tools, MongoDB through the JavaScript's Shell, Creating and Querying through Indexes, Document-Oriented, principles of schema design, Constructing queries on Databases, collections and Documents, MongoDB Query Language.
- **8 Hours**
- **Marks: 20 (approx.)**

Unit 6

PROCESSING OF REAL-TIME DATA AND STREAMING DATA

- Data Streams: Introduction and Ingestion
- Kafka
- Storm & Storm Assignment
- Spark Streaming
- **8 Hours**
- **Marks: 15 (approx.)**

Resources

Books Recommended:

Text Books

1. Understanding Big data - Chris Eaton,Dirk derooset al. McGraw Hill
2. MongoDB in Action - Kyle Banker,Piter Bakkum, Shaun Verch, Dream tech Press
3. Beginning Apache Pig-Big Data Processing Made Easy-Balaswamy Vaddeman, Apress'
4. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
5. Eric Sammer, "Hadoop Operations", Reilly, 2012.

Reference Books

1. Paul Zikopoulos, Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Understanding *Big Data: Analytics for Enterprise Class Hadoop and streaming Data*,The McGraw-Hill Companies, 2012.
2. Vignesh Prajapati, Big data analytics with R and Hadoop, SPD 2013.
3. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
4. Alan Gates, "Programming Pig", O'Reilley, 2011