SILVER OAK UNIVERSITY

Engineering and Technology (M.Tech.)

Computer Engineering (Software Engineering)

Subject Name:Big Data Systems
Subject Code:
Semester: II

Prerequisite:

Objective: The main goal of the course is to help students learn, understand, and practice big data systems and its analytics approaches, which include the study of modern computing big data technologies and scaling up machine learning techniques focusing on industry applications. It also helps to conceptualize and summarize the trivial data versus big data, big data computing technologies, machine learning techniques, and scaling up machine learning approaches.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Evaluation Scheme				Total Marks
L	T	P	C	Internal		External		
				Th	Pr	Th	Pr	
3	0	2	4	40	20	60	30	150

Content:

Unit No.	Course Contents	Teaching Hours	Weightage %
1	**INTRODUCTION TO BIG DATA** **Big Data –** Definition, Characteristic Features – Big Data Applications - Big Data vs Traditional Data - Risks of Big Data - Structure of Big Data - Challenges of Conventional Systems - Web Data – Evolution of Analytic Scalability - Evolution of Analytic Processes, Tools and methods - Analysis vs Reporting - Modern Data Analytic Tools.	5	15
2	**Mining data streams** Introduction to Streams Concepts – Stream Data Model and Architecture - Stream Computing - Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating Moments – Counting Oneness in a Window – Decaying Window - Real time Analytics Platform (RTAP) Applications – Case Studies - Real Time Sentiment Analysis- Stock Market Predictions.	11	25

3	**Hadoop** History of Hadoop- the Hadoop Distributed File System – Components of Hadoop Analyzing the Data with Hadoop- Scaling Out- Hadoop Streaming- Design of HDFS-Java interfaces to HDFS Basics- Developing a Map Reduce Application-How Map Reduce Works-Anatomy of a Map Reduce Job run-Failures-Job Scheduling-Shuffle and Sort – Task execution - Map Reduce Types and Formats- Map Reduce Features- Hadoop environment.	9	20
4	**Frameworks** Applications on Big Data Using Pig and Hive – Data processing operators in Pig – Hive services – HiveQL – Querying Data in Hive - fundamentals of HBase and ZooKeeper - IBM InfoSphereBigInsights and Streams.	7	20
5	**Predictive Analytics** Simple linear regression- Multiple linear regression- Interpretation of regression coefficients. Visualizations - Visual data analysis techniques- interaction techniques - Systems and applications.	7	20

Course Outcome:

Sr. No.	CO statement	Unit No
CO-1	Work with big data platform and explore the big data analytics techniques business applications	1
CO-2	Design efficient algorithms for mining the data from large volumes	2
CO-3	Analyze the HADOOP and Map Reduce technologies associated with big data analytics	3
CO-4	Explore on Big Data applications Using Pig and Hive.	4
CO-5	Understand the fundamentals of various big data analytics techniques.	5

List of Experiments/Tutorials:

1. Set up a pseudo-distributed, single-node Hadoop cluster backed by the Hadoop Distributed File System, running on Ubuntu Linux. After successful installation on one node, configuration of a multi-node Hadoop cluster (one master and multiple slaves).
2. MapReduce application for word counting on Hadoop cluster
3. Unstructured data into NoSQL data and do all operations such as NoSQL query with API.
4. K-means clustering using map reduce
5. Page Rank Computation
6. Mahout machine learning library to facilitate the knowledge build up in big data analysis.
7. Application of Recommendation Systems using Hadoop/mahout libraries

Major Equipment:
XMLSpy, RSS Feed, RSS Reader

Books Recommended:

- Bill Franks, ―Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics‖, Wiley and SAS Business Series, 2012.
- David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", 2013.
- Michael Berthold, David J. Hand, ―Intelligent Data Analysis‖, Springer, Second Edition, 2007.
- Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
- P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
- Richard Cotton, "Learning R – A Step-by-step Function Guide to Data Analysis, , O‘Reilly Media, 2013.

List of Open Source Software/learning website:

http://in.reuters.com/tools/rss

http://www.altova.com/xmlspy.html

https://www.w3.org/RDF/