DKTES Textile and Engineering Institute, Ichalkaranji
Third Year B. Tech. (Semester – VI)
CSL704: Big Data Analytics

Teaching Scheme:	Credits	Evaluation Scheme:
Lectures : 03 Hrs/Week	03	SE-I: 25 Marks
Tutorials : 00 Hrs/Week		SE-II: 25 Marks
Practicals: 00 Hrs/Week		SEE: 50 Marks

Course Outcomes:

On completion of the course, student will be able to–

- Explain the fundamental concepts of big data and its analytics
- Explain how to Analyze the big data using Hadoop and intelligent techniques
- Explain NoSQL big data management
- Choose the suitable secure models for building competitive business decisions

Course Contents

Unit I	**Importance of Big Data**	**05 Hours**

Classification of Digital Data, Characteristics of Data, Evolution of Big Data, Definition of Big Data, Challenges with Big Data, What is Big Data? Other Characteristics of Data Which are not Definitional Traits of Big Data, Why Big Data? Traditional Business Intelligence (BI) versus Big Data, A Typical Data Warehouse Environment, A Typical Hadoop Environment, What is New Today? What is Changing in the Realms of Big Data? What is Big Data Analytics?, What Big Data Analytics isn't, Classification of Analysis, Challenges that prevent business from capitalizing Big Data, Top challenges facing Big Data, Why is Big Data analytics important?

Unit II	**Hadoop Architecture**	**06 Hours**

Hadoop ecosystem, Design of Hadoop distributed file system (HDFS), Data Flow-Anatomy of a File Read, Anatomy of a File Write, Coherency Model, Parallel Copying with distcp, Keeping an HDFS Cluster Balanced, MapReduce, Classic Map-reduce, Analyzing data with Hadoop, Anatomy of a MapReduce Job Run, Failures,

Unit III	**YARN and Hadoop I/O**	**06 Hours**

Yet Another Resource Negotiator (YARN), Anatomy of a YARN Application Run, YARN Compared to MapReduce, Scheduling in YARN, Data Integrity - Data Integrity in HDFS, LocalFileSystem, ChecksumFileSystem, Compression – Codecs, Compression and Input Splits, Using Compression in MapReduce, Serialization - The Writable Interface, Writable Classes, Serialization Frameworks, File-Based Data Structures – SequenceFile, MapFile

Unit IV	**NoSQL Management**	**06 Hours**

Why NoSQL?, Impedance mismatch, Emergence of NoSQL, Aggregate data models, Key-value and document data models, Column-family stores, Graph databases, Schema less databases, Distribution models - sharding, Master-slave replication, Peer-peer replication, Consistency - Update Consistency, Read Consistency, Relaxing Consistency, The CAP Theorem, Relaxing Durability

Unit V	**Analytics Framework**	**05 Hours**

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

Unit VI	**Securing Ecosystem**	**05 Hours**

Why do we need to secure Hadoop, Challenges for securing the Hadoop ecosystem, Key security considerations, Reference architecture for Big Data security, What is Kerberos, How Kerberos works, Hadoop Kerberos security implementation, Configuring Hadoop with Kerberos authentication, Securing ecosystem components –Hive, Oozie, Flume, Pig,

Text Books:

1. Seema Acharya, Subhasini Chellappan, "Big Data Analytics", Wiley.

2. Tom White, “Hadoop: The Definitive Guide” (O’Reilly Media)
3. P. J. Sadalage, M. Flower, “NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence” (Addison-Wesley)
4. Sudeesh Narayanan, “Securing Hadoop” (O’Reilly Media)

References Books:

1. Michael Mineli, Michele Chambers, Ambiga Dhiraj, “Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses” (Wiley Publication)
2. Chris Eaton,Dirk derooset al., “Understanding Big data”, McGraw Hill.
3. G James, D. Witten, T Hastie, R. Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer.
4. Douglas Eadline, “Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem”, Pearson Education.
5. E. Capriolo, D. Wampler, J. Rutherglen, “Programming Hive”, O’ Reilly.
6. Lars George, “HBase: The Definitive Guide”, O’ Reilly.

   Alan Gates, "Programming Pig”, O’ Reilly

Useful Links:

1. Analytics Vidhya (http://www.analyticsvidhya.com/) ...
2. Dataversity (http://www.dataversity.net/) ...
3. R Bloggers (http://www.r-bloggers.com/) ...
4. SmartData Collective (http://www.smartdatacollective.com/) ...
5. Data Science Central (http://www.datasciencecentral.com/) ...
6. Planet Big Data (http://planetbigdata.com/)