Data Analytics

1.1 Course Number: CS457

1.2 Contact Hours 40 Credits: 3-0-2 [11]

1.3 Semester-offered:

1.4 Prerequisite: Should have knowledge of one Programming Language (Java preferably),

Practice of SQL (queries and sub queries), exposure to Linux Environment.

1.5 Syllabus Committee Member:

2. **Objective:**

- Understand the Big Data Platform and its Use cases
- Provide an overview of Apache Hadoop
- Provide HDFS Concepts and Interfacing with HDFS
- Understand Map Reduce Jobs
- Provide hands on Hadoop EcoSystem
- Apply analytics on Structured, Unstructured Data.
- Exposure to Data Analytics with R.

3. **Course Content:**

Unit-wise distribution of content and number of lectures

Unit	Topics	Sub-topic	Lectures
1	Introduction to Big Data and Hadoop	Types of Digital Data, Introduction to Big Data, Big Data Analytics, History of Hadoop, Apache Hadoop, Analysing Data with Unix tools, Analysing Data with Hadoop, Hadoop Streaming, Hadoop Ecosystem, IBM Big Data Strategy, Introduction to Infosphere BigInsights and Big Sheets.	8
2	HDFS (Hadoop Distributed File System)	The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro and File Based Data structures.	8
3	Map Reduce	Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features.	8

4	Hadoop EcoSystem	Pig : Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators. Hive : Hive Shell, Hive Services, Hive Metastore, Comparison with Traditional Databases, HiveQL, Tables, Querying Data and User Defined Functions. Hbase : HBasics, Concepts, Clients, Example, Hbase Versus RDBMS. Big SQL : Introduction	8
5	Data Analytics with R	Machine Learning : Introduction, Supervised Learning, Unsupervised Learning, Collaborative Filtering. Big Data Analytics with BigR.	8
	Total		40

4. Readings

4.1 Textbook:

- Tom White “ Hadoop: The Definitive Guide” Third Edition, O’reily Media, 2012.
- Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.

4.2 Reference books:

- Michael Berthold, David J. Hand, "Intelligent Data Analysis”, Springer, 2007. ● Jay Liebowitz, “Big Data and Business Analytics” Auerbach Publications, CRC press (2013)
- Tom Plunkett, Mark Hornick, “Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R
- Enterprise and Oracle R Connector for Hadoop”, McGraw-Hill/Osborne Media (2013), Oracle press.
- Anand Rajaraman and Jeffrey David Ulman, “Mining of Massive Datasets”, Cambridge University Press, 2012.
- Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley & sons, 2012.
- Glen J. Myat, “Making Sense of Data”, John Wiley & Sons, 2007
- Pete Warden, “Big Data Glossary”, O’Reily, 2011.
- Michael Mineli, Michele Chambers, Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley Publications, 2013.
- ArvindSathi, “BigDataAnalytics: Disruptive Technologies for Changing the Game”, MC Press, 2012
- Paul Zikopoulos ,Dirk DeRoos , Krishnan Parasuraman , Thomas Deutsch , James Giles , David Corigan ,
- "Harness the Power of Big Data The IBM Big Data Platform ", Tata McGraw Hill

Publications, 2012.

5 Outcome of the Course: The students will be able to:

- Identify Big Data and its Business Implications.
- List the components of Hadoop and Hadoop Eco-System
- Access and Process Data on Distributed File System
- Manage Job Execution in Hadoop Environment
- Develop Big Data Solutions using Hadoop Eco System
- Analyze Infosphere BigInsights Big Data Recommendations.
- Apply Machine Learning Techniques using R.