COURSE CODE	COURSE TITLE	L	T	P	C
1152CS210	BIG DATA ANALYTICS AND TOOLS	2	0	2	3

Course Category: Program Elective

A. Preamble : This course covers foundational techniques and tools required for data science and big data analytics. The course focuses on concepts, principles, and techniques applicable to any technology environment and industry and establishes a baseline that can be enhanced by further formal training and additional real-world experience.

B. Prerequisite Courses:

Sl. No	Course Code	Course Name
1	1151CS107	Database Management Systems

C. Related Courses:

Sl. No	Course Code	Course Name
1	1152CS118	Distributed and Parallel Computing

D. Course Educational Objectives :

Learners are exposed to

- To explore the fundamental concepts of big data analytics.
- To learn to analyze the big data using intelligent techniques.
- To understand the various search methods and visualization techniques.
- To learn to use various techniques for mining data stream.

E. Course Outcomes :

Upon the successful completion of the course, students will be able to:

CO No's	Course Outcomes	Knowledge Level
CO1	Differentiate traditional data processing with Big Data Analytics.	K2
CO2	Explain the technology landscape behind the Big Data Analytics using Hadoop and NoSQL	K2
CO3	Solve distributed computing challenges with the help of Hadoop and MongoDB.	K3
CO4	Perform CRUD operations using Cassandra and Hive	K3
CO5	Differentiate between Pig and Hive in terms of processing and to design JasperReports using Jaspersoft studio using data from NoSQL databases.	K3

F. Correlation of COs with POs :

COs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	M		H												
CO2	M			M											
CO3	M	M	H	M	M										
CO4	M	M	M		M						M				
CO5	M	H	M		M						M				

H- High; M-Medium; L-Low

G. Course Content:

UNIT I Introduction to Digital Data and Big Data 6

Types of Digital Data - Structured Data - Semi-Structured Data - Unstructured Data. Characteristics of Data- Evolution of Big Data- Definition of Big Data - Challenges of Big Data- Other Characteristics of Data - Traits of Big Data- Traditional Business Intelligence (BI) versus Big Data- Typical Data Warehouse Environment- Typical Hadoop Environment- Realms of Big Data.

UNIT II Introduction to Big Data Analytics and Technology landscape 6

Big Data Analytics - Hype around Big Data Analytics- Classification of Analytics and Challenges on Big Data- Data Science- Data Scientist- -Terminologies Used in Big Data Environments -Basically Available Soft State Eventual Consistency (BASE)- Analytics Tools.

NoSQL (Not Only SQL)- Hadoop- Features and Advantages of Hadoop - Overview of Hadoop Ecosystems, Hadoop Distributions, Hadoop versus SQL, Integrated Hadoop Systems Offered by Leading Market Vendors - Cloud based Hadoop solutions.

UNIT III Introduction to Hadoop and MongoDB 6

Introducing Hadoop - RDBMS versus Hadoop- Distributed Computing Challenges - History of Hadoop - Hadoop Overview – Hadoop Distributors- Hadoop Distributed File System - HDFS - Processing Data with Hadoop - Managing Resources and Application with Hadoop YARN- Interacting with Hadoop Ecosystem. Introduction to MongoDB – JSON - Terms used in RDBMS and MongoDB - Data Types in MongoDB – MongoDB Query Language.

UNIT IV Introduction to Cassandra and Hive 6

Apache Cassandra- An Introduction- Features of Cassandra- CQL Data Types- CQLSH- Keyspaces- CRUD- Collections- Using a Counter -Time To Live (TTL) - Alter Commands -

Import and Export- Querying System Tables- Practice Examples.
Introduction to Hive - Hive Architecture - Hive Data Types - Hive File Format- Hive Query Language- RCFILE Implementation –SerDe - UDF.

UNIT V Introduction to Pig and Jasper Report **6**

Apache Pig - The Anatomy of Pig - Pig on Hadoop - Pig Philosophy - Use Case for Pig- ETL Processing - Pig Latin Overview - Data Types in Pig - HDFS Commands- Relational Operators- Complex Data Type - Piggy Bank- UDF (User Defined Function)- Parameter Substitution- Diagnostic Operator- Pig versus Hive- Hive Vs Pig.
Jasper Report using Jasper Soft - Introduction to Jasper Reports - Connecting to MongoDB NoSQL database- Connecting to Cassandra NoSQL Databases.

Lab Experiment: **L-15**

1. Installation of Mongo DB with CRUD Operation in MongoDB
2. Mongo DB Query Language
3. Mongo DB with Java Connectivity
4. Installation of Cassandra
5. CRUD Operation in Cassandra
6. Cassandra with Java Connectivity
7. Pig and Hive Query Language
8. Jasper Report -Connecting to Cassandra NoSQL Databases
9. JasperReports - Connecting to MongoDB NoSQL database
10. Single Node Hadoop Installation

TOTAL: 60

H. Learning Resources

i. Text Books

1. Seema Acharya and Subhashini C: Big Data and Analytics, First Edition, Wiley India Pvt. Ltd, 2015.
2. Judith Hurwitz, Alan Nugent,Fern Halper, Marcia Kaufman : Big data for dummies – Judith Hurwitz, Alan Nugent,Fern Halper, Marcia Kaufman, Wiley India Pvt. Ltd, April 2013.

ii. Reference Books:

1. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.
2. Tom White “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 2012.
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “UnderstandingBig Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Publishing, 2012
4. Big Data: A Revolution That Will Transform How We Live, Work, and Think by Viktor Mayer-Schoenberger& Kenneth Cukier
5. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems

iii. Web References

1. http://tdan.com/matching-unstructured-data-and-structured-data/5009
2. https://www.mongodb.com/
3. http://cassandra.apache.org/
4. https://hadoop.apache.org/docs/r2.8.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html