COURSE CODE	COURSE TITLE	L	T	P	C
1152CS121	BIG DATA AND ANALYTICS	3	0	0	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	1151CS108	Operating System
2	1152CS119	Python Programming
3	1151CS117	Java Programming

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 (Based on revised Bloom's Taxonomy)
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	PSO 2	PSO 3
CO1	M														
CO2	M	M			M										
CO3	M	M			M						M				L
CO4	M	M			M						L				
CO5	M	H			H										

H- High; M-Medium; L-Low

G. Course Content:

UNIT 1Introduction to Digital Data and Big Data 7

Types of Digital Data:

Classification of Digital Data- Structured Data: Sources of Structured Data, Ease of Working with Structured Data- Semi-Structured Data: Sources of Semi-Structured Data- Unstructured Data: Issues with Unstructured Data, How to Deal with Unstructured Data.

Introduction to Big Data:

Characteristics of Data- Evolution of Big Data- Definition of Big Data: Volume, Velocity, Variety - Challenges of Big Data- What is Big Data?- Other Characteristics of Data Which are Not Definitional Traits of Big Data- Why Big Data?- Are We Just an Information Consumer or Do We Also Produce Information- Traditional Business Intelligence (BI) versus Big Data- A Typical Data Warehouse Environment- A Typical Hadoop Environment- What is Changing in the Realms of Big Data?- What is New Today?: Coexistence of Big Data and Data Warehouse.

UNIT 2 Introduction to Big Data Analytics and Technology landscape 8

Introduction to Big Data Analytics:

Where do we Begin?- What is Big Data Analytics?- What Big Data Analytics isn't?- Why this Sudden Hype around Big Data Analytics?- Classification of Analytics, Greatest Challenges that Prevent Businesses from Capitalizing on Big Data- Top Challenges Facing Big Data- Why is Big Data Analytics Important?- What Kind of Technologies are we Looking Toward to Help Meet the Challenges Posed by Big Data?- Data Science- Data Scientist- Terminologies Used in Big Data Environment: In Memory Analytics, In Database Processing, Symmetric Multiprocessor System, Massively Parallel Processing, Difference between Parallel versus Distributed Systems, Shared Nothing Architecture, Consistency, Availability, Partition Tolerance (CAP): Theorem Explained, Basically Available Soft State Eventual Consistency (BASE)- Top Analytics Tools.

The big data technology landscape:

NoSQL: Where is it used? What is it? Types of NoSQL Databases, Why NoSQL? Advantages of NoSQL, what we miss with NoSQL? NoSQL Vendors, SQL versus NoSQL, NewSQL, Comparison of SQL, NoSQL and NewSQL. Hadoop: Features of Hadoop, Key Advantages of Hadoop, Versions of Hadoop: Hadoop 1.0- Hadoop 2.0, Overview of Hadoop Ecosystems, Hadoop Distributions, Hadoop versus SQL, Integrated Hadoop Systems Offered by Leading Market Vendors, Cloud based Hadoop solutions.

UNIT 3 Introduction to Hadoop and MongoDB **10**

Introduction to Hadoop:

Introducing Hadoop: Data- The Treasure Trove- Why Hadoop? - Why not RDBMS? - RDBMS versus Hadoop- Distributed Computing Challenges: Hardware Failure, How to Process this Gigantic Store of Data? - A Brief History of Hadoop: The Origin of the Name Hadoop- Hadoop Overview: Key Aspects of Hadoop- Hadoop Component- Hadoop Conceptual Layer- High Level Architecture of Hadoop. Business Value of Hadoop: Clickstream Data - Hadoop Distributors- Hadoop Distributed File System: HDFS Daemons, Anatomy of File Read, Anatomy of File Write, Replica Placement Strategy, Working with HDFS commands, Special Features of HDFS- Processing Data with Hadoop: MapReduce Daemons, how does MapReduce work? MapReduce Example- Managing Resources and Application with Hadoop YARN: Limitations of Hadoop 1.0 Architecture, HDFS Limitation, Hadoop 2: HDFS, Hadoop 2 YARN: Taking Hadoop Beyond Batch- Hadoop Ecosystem: Pig, Hive, Sqoop, HBase.

Introduction to MongoDB:

What is MongoDB?- Why MongoDB? : Using JSON, Creating or Generating a Unique Key, Support for Dynamic Queries, Storing Binary Data, Replication, Sharding, Updating Information In-Place - Terms used in RDBMS and MongoDB - Data Types in MongoDB – CRUD(Create, Read, Update and Delete): Insert(), Update(), Save(), Remove(), find() – Arrays- MapReduce Functions- Aggregation- Java Scripting- Cursor- Index- MongoImport- MongoExport- Automatic generation of unique numbers for the "_id" field.

UNIT 4 Introduction to Cassandra and Hive **10**

Introduction to Cassandra:

Apache Cassandra : An Introduction- Features of Cassandra: Peer-to-Peer Network, Gossip and Failure Detection, Partitioner, Replication Factor, Anti-Entropy and Read Repair, Writes in Cassandra, Hinted Handoffs, Tunable Consistency: Read Consistency and Write Consistency- CQL Data Types- CQLSH- Key spaces- CRUD: Insert, Update, Delete, Select - Collections: Set, List, Map- Using a Counter -Time To Live (TTL)- Alter: Alter Table to Change the Data Type of a Column, Alter Table to Delete a Column, Drop a Table, Drop a Database- Import and Export: Export to CSV, Import from CSV, Import from STDIN, Export to STDOUT -System Tables- Practice Examples.

Introduction to Hive:

What is Hive?: History of Hive and Recent Releases of Hive, Hive Features, Hive Integration and Work Flow, Hive Data Unit - Hive Architecture - Hive Data Types: Primitive Data Types, Collection Data Types - Hive File Format: Text File, Sequential File, RCFile (Record Columnar File)- Hive Query Language: DDL (Data Definition Language) Statements, DML (Data Manipulation Language) Statements, Starting Hive Shell, Database, Tables, Partitions, Buckets, Views, Sub Query, Joins, Aggregation, Group BY and Having. RCFILE Implementation, SERDE, UDF.

UNIT 5 Introduction to Pig and Jasper Report **L-10**

Introduction to Pig:

What is Pig?: Key Features of Pig - The Anatomy of Pig - Pig on Hadoop - Pig Philosophy - Use Case for Pig: ETL Processing - Pig Latin Overview: Pig Latin: Statements, Pig Latin: Keywords, Pig Latin: Identifiers, Pig Latin: Comments, Pig Latin: Case Sensitivity- Data Types in Pig: Simple Data Types, Complex Data Types- Running Pig: Interactive Mode, Batch Mode- Execution Modes of Pig: Local Mode, Map Reduce Mode- HDFS Commands- Relational Operators- Eval Function- Complex Data Type: Tuple, Map - Piggy Bank- UDF (User Defined Function)- Parameter Substitution- Diagnostic Operator- Word Count Example- When to use Pig?- When NOT to use Pig?- Pig at Yahoo - Pig versus Hive- Hive Vs Pig.

Jasper Report using Jasper Soft:

Introduction to JasperReports, Jaspersoft Studio: JasperReports, Jaspersoft Studio - Connecting to MongoDB NoSQL database: Syntax of Few MongoDB Query Language, Elements and Attributes, Creating Variables, Creating Report Parameters- Connecting to Cassandra NoSQL Databases.

TOTAL: 45 periods

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.
3. Tom White: Hadoop: The Definitive Guide, O'Reilly Media 4th Edition, March 2015.
4. Chuck Lam: Hadoop in action, Manning Publications Co, 2011
5. Dirk Deroos, Paul C. Zikopoulos, Roman B. Melnyk,BruceBrown: Hadoop for dummies, Wiley publications, 2014.

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, “Understanding Big 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. www.iannauniversity.com/.../it2024-user-interface-design-u...
2. www.cramster.com › ... › software design › resource › lecture note
3. www.aw-bc.com/DTUI3/lecnotes.doc.
4. https://www.cosc.brocku.ca/~bockusd/3p94/webui1.pdf