Semester: VI						
BIG DATA TECHNOLOGIES **Category: Professional Core Course** **(Theory & Practice)**						
Course Code	**:**	**AI362IA**		**CIE**	**:**	**100 +50Marks**
Credits: L:T:P	**:**	**3:0:1**		**SEE**	**:**	**100 + 50Marks**
Total Hours	**:**	**45L+30P**		**SEE Duration**	**:**	**3.00+ 3.00 Hours**

The Hadoop Distributed File system
The Design of HDFS - HDFS Concepts – Blocks, Name nodes and Data nodes, HDFS Federation, HDFS High Availability
Data Flow – Anatomy of a File Read, Anatomy of a File Write

Unit – II **09 Hrs**

Map Reduce – Distributed Processing Framework- A Weather Dataset – Data format, Analysing the data with Unix Tools, Analyzing the Data with Hadoop – Java MapReduce, Scaling Out
Working of Map Reduce - Anatomy of a Map Reduce Job Run, Failures, Shuffle and Sort, Task Execution

Unit –III **09 Hrs**

Hive - Configuring Hive, Hive Services ,The Metastore
Comparison with Traditional Databases -Schema on Read Versus Schema on Write, Updates, Transactions, and Indexes ,SQL-on-Hadoop Alternatives
HiveQL - Data Types, Operators and Functions
Tables -Managed Tables and External Tables, Partitions and Buckets, Storage Formats, Importing Data, Altering Tables, Dropping Tables,
Querying Data -Sorting and Aggregating, Map Reduce Scripts, Joins, Subqueries,Views

Unit –IV **09 Hrs**

Flume - Installing Flume, **Transactions and Reliability** -Batching ,**The HDFS Sink** -Partitioning and Interceptors File Formats
Fan Out-Delivery Guarantees, Replicating and Multiplexing Selectors
Distribution: Agent Tiers-,Delivery Guarantees,
Sink Groups - Integrating Flume with Applications, Component Catalog

Unit –V **09 Hrs**

Spark Applications- Jobs, Stages, and Tasks, A Scala Standalone Application,
Resilient Distributed Datasets - Creation, Transformations and Actions, Persistence, Serialization
Shared Variables -Broadcast Variables, Accumulators
Anatomy of a Spark Job Run - Job Submission, DAG Construction, Task Scheduling, Task Execution

Lab Component	
Expt. No	**Programs**
1.	**Map Reduce Program on Counting** a) Write a Java Program using Mapper and Reducer function to find the number of records in the give dataset b) Submit the job to cluster c) Track the job information
2.	**Map Reduce Program using Temperature Dataset** 1. Write a Java program for finding Maximum recorded temperature by the year from Weather Dataset 2. Submit the job to cluster 3. Find the status of the Job and terminate it
3.	**Programs on Pig Script Using movie lens data** a) List all the movies and the number of ratings b) List all the users who have rated the same movie and find the number of ratings

	c) List all the Users who have rated the movies (Users who have rated at least one movie) d) Find the count of the Movie which has the ratings more than 3 e) Find the max, min, average ratings for all the movie
4.	**Program on Advanced Concepts in Pig** a) Group by Year and dump the result in a bag b) Write a pig script to find the maximum temperature c) Write a pig Script to find the average temperature of a state for 3 years and store the result in HDFS
5.	**Extract facts using Hive on movie lens data** a) Write a query to select only those records which correspond to starting, browsing, completing, or purchasing movies. Use a CASE statement to transform the RECOMMENDED column into integers where 'Y' is 1 and 'N' is 0. Also, ensure GENREID is not null. Only include the first 25 rows. b) Write a query to select the customer ID, movie ID, recommended state and most recent rating for each movie.

PART - B

Group of two students belongs to same batch are required to implement a problem statement which makes use of streaming data using Apache Spark.

Examples: Identifying Credit Card Fraud, Identifying prospective customers on a commerce website, real-time stock trades, up-to-the minute inventory management, fake-news detection, etc.

Course Outcomes: After completing the course, the students will be able to:-

CO1	Understand and apply the different building blocks of Big Data Technologies to a given problem
CO2	Articulate the programming aspect of Big Data Technologies to obtain solution to the problem through lifelong learning
CO3	Exhibit effective communication to represent the analytical aspects of Big Data Technologies for obtaining solution to the problems
CO4	Demonstrate solutions for societal and environmental concern problems using modern engineering tools through writing effective reports
CO5	Appraise the knowledge of Big Data Technologies as an Individual /as a team member to manage multidisciplinary projects

Reference Books

1.	Hadoop – The Definitive Guide; Storage and Analysis at Internet scale,Tom White ,4th Edition, 2015, O'Reilly, Shroff Publishers & Distributers Pvt. Ltd., ISBN – 978-93-5213-067-2
2.	DT Editorial Services,Big Data – Black Book,Dreamtech Press, 1st Edition – 2015, ISBN - 978-93-511-9-757-7
3.	Hadoop for Dummies, Dirk deRoos, Paul C. Zikopoulos, Roman B. Melnyk, Bruce Brown, Rafael Coss, 2014, John Wiley & Sons, Inc., ISBN: 978-1-118-60755-8 (pbk); ISBN 978-1-118-65220-6 (ebk); ISBN 978-1-118-70503-2 (ebk)
4.	Big Data Principles and best practices of scalable real-time data systems ,Nathan Marz and James Warren, 1st Edition, 2015, ISBN 9781617290343

RUBRICFOR THE CONTINUOUS INTERNAL EVALUATION		
#	**COMPONENTS**	**MARKS**
1.	**QUIZZES:** Quizzes will be conducted in online/offline mode. **TWO QUIZZES** will be conducted & Each Quiz will be evaluated for 10 Marks. Each quiz is evaluated for 10 marks adding up to 20 MARKS	**20**
2.	**TESTS:** Students will be evaluated in test, descriptive questions with different complexity levels (Revised Bloom's Taxonomy Levels: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating). **TWO tests will be conducted**. Each test will be evaluated for **50Marks**, adding upto 100 Marks. **FINAL TEST MARKS WILL BE REDUCED TO 40 MARKS.**	**40**
3.	**EXPERIENTIAL LEARNING:** Students will be evaluated for their creativity and practical implementation of the problem. Case study based teaching learning (10), Program specific requirements (10), Video based seminar/presentation/demonstration (10) Designing & Modeling (10) **Phase 2 will be done in the exhibition mode (Demo/Prototype/any outcome). ADDING UPTO 40 MARKS**.	**40**
4.	**LAB:** Conduction of laboratory exercises, lab report, observation, and analysis (20 Marks), lab test (10 Marks) and Innovative Experiment/ Concept Design and Implementation (20 Marks) adding up to 50 Marks. **THE FINAL MARKS WILL BE 50 MARKS**	**50**
	MAXIMUM MARKS FOR THE CIE(THEORY+LAB)	**150**

RUBRIC FOR SEMESTER END EXAMINATION (THEORY)		
Q.NO.	**CONTENTS**	**MARKS**
	PART A	
1	Objective type of questions covering entire syllabus	20
	PART B (Maximum of THREE Sub-divisions only)	
2	Unit 1 : (Compulsory)	16
3 & 4	Unit 2 : Question 3 or 4	16
5 & 6	Unit 3 : Question 5 or 6	16
7 & 8	Unit 4 : Question 7 or 8	16
9 & 10	Unit 5: Question 9 or 10	16
	TOTAL	**100**

RUBRIC FOR SEMESTER END EXAMINATION (LAB)		
Q.NO.	**CONTENTS**	**MARKS**
1	Write Up	10
2	Conduction of the Experiments	20
3	Viva	20
	TOTAL	**50**