Course Code	UDS21403J	Course Name	WORKING WITH BIG DATA	Course Category	C	Professional Core Course	L	T	P	C
							4	0	2	5

Pre-requisite Courses	Nil	Co-requisite Courses	Nil	Progressive Courses	Nil
Course Offering Department	Computer Applications		Data Book / Codes/Standards	Nil	

Course Learning Rationale (CLR):	The purpose of learning this course is to,	Learning	Program Learning Outcomes (PLO)

		1	2	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
CLR-1 :	To provide the participants with the comprehensive knowledge of different types of big data types like the structured, unstructured, semi- structured and streaming datasets.																		
CLR-2 :	To familiarize the participants with the Hadoop and Apache spark the two most popular big data processing frameworks available in the market.																		
CLR-3 :	To understand the Hadoop Ecosystem or a suite which provides various services to solve the big data problems																		
CLR-4 :	To introduce the participants to DataFrames in Apache Spark for large scale Data science applications.																		
CLR-5 :	To to introduce the participants to build real-time streaming data pipelines and real-time streaming applications with Apache Kafka.																		
CLR-6 :	Bring the users to an alignment, applies their learning to a real-world business problem, and then performs research, design, development, and delivers an end-to-end Big Data solution for a given industry problem. The students will be working either in a group or individually.]																		

Course Learning Outcomes (CLO):	At the end of this course, learners will be able to:	Level of Thinking (Bloom)	Expected Proficiency (%)	Expected Attainment (%)	Fundamental Knowledge	Application of Concepts	Link with Related Disciplines	Procedural Knowledge	Skills in Specialization	Ability to Utilize Knowledge	Skills in Modeling	Analyze, Interpret Data	Investigative Skills	Problem Solving Skills	Communication Skills	Analytical Skills	ICT Skills	Professional Behavior	Life Long Learning
CLO-1 :	To design and develop natural language processing solution artifacts and ultimately demonstrate an "end-to-end" machine learning solution for a given problem statement either in a group or individually.	2	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H
CLO-2 :	hands-on skills, knowledge and expertise in IoT communication protocols that are modes of communication to ensure optimum security of the data being exchanged between IoT connected devices	3	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H
CLO-3 :	publish (write) and subscribe to (read) streams of events, including continuous import/export of your data from other systems.	3	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H
CLO-4 :	efficiently work with Apache Kafka for process streaming data in real-time, and Publish and subscribe to streams of records	3	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H
CLO-5 :	utilize the power of spark and python in a nutshell and process data in a distributed environment	3	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H
CLO-6 :	Have a fundamental understanding of all the big data types, tools and techniques that are involved to process data	3	85	80	H	H	H	H	H	H	H	H	H	H	M	H	H	H	H

Note: All our curriculum, study materials, assignments, quizzes, lab works, and learning resources are personalized and dynamically generated using machine learning models based on the learner's learning ability. Users can review our learning curriculum only through our intelligent learning management platform (iLMSP), and our learning resources and lab infrastructures are available only in the digital form on our cloud infrastructures.

Duration (hour)		18	18	18	18	18
S-1	SLO-1	**Unit 1: Introduction to Big Data**	Apache Hadoop overview	Apache Kafka Streams	DataFrames in Spark Overview, Features of DataFrames in Spark , Why do we need Spark DataFrames, Sources for Spark DataFrames	NoSQL Databases Overview, Evolution of NoSQL, What makes NoSQL different
	SLO-2	Big Data Tools Overview	Business Benefits of Apache Hadoop	Apache Kafka Stream processing	Creation Spark DataFrames from JSON, Creation Spark DataFrames from existing RDD's, Creation Spark DataFrames from existing csv files, Spark DataFrame Operations	Business Benefits and Challenges of NoSQL, NoSQL vs Relational Databases
S-2	SLO-1	Hadoop	Need of Apache Hadoop	**Unit 5: Map Reduce, its Working and Developing a Map Reduce Application**	select(), withColumn() Transformation, filter() Transformation, orderBy(), sort(), sortWithinPartitions() Transformation	No SQL Data Store Types, No SQL Database management systems
	SLO-2	Apache Strom	Components of Hadoop	Map Reduce overview	distinct(), dropDuplicates() Transformation, join () Transformation, groupBy () Transformation	**Unit 14: Working with IIoT Technologies, Communication Protocols and Data Services**
S-3	SLO-1	MongoDB	Processing Layer (MapReduce)	How does MapReduce Work?	**Unit 9: Introduction to Apache Kafka**	IIoT Communication Protocols overview
	SLO-2	Cloudera	Storage Layer (HDFS)	Business benefits of MapReduce	Apache Kafka overview	IIoT Wireless Communication Protocols overview
S-4	SLO-1	Big Data Technologies Overview	Hadoop YARN	Business Challenges of MapReduce	Event Streaming, Uses of Event Streaming, Apache Kafka as event Streaming platform, Working of Apache Kafka	IIoT Communication Protocols overview
	SLO-2	Data Management	Apache Spark overview	MapReduce Architecture	Apache Kafka overview	IIoT Wireless Communication Protocols overview
S-5 & S-6	SLO-1 SLO-2	Lab 1 :	Lab 4 :	Lab 7:	Lab 10 :	Lab 13:
S-7	SLO-1	Data Mining	Business Benefits of Apache Spark	MapReduce Example	Event Streaming, Uses of Event Streaming, Apache Kafka as event Streaming platform, Working of Apache Kafka	Business Benefits and Challenges of IIoT Communication Protocols

	SLO-2	In-Memory Analytics	Need of Apache Spark	Implementation of MapReduce	Event, Producers, Consumer, Topic, Partition, Messaging System	Client/Server, pub/sub, Request/Response
S-8	SLO-1	Predictive Analytics	Components of Apache Spark	**Unit 6: Big Data HDFS Ecosystem, Tools and Technologies**	Broker, Kafka API's	RESTful Interface, MQTT, AMQP, OPC UA
	SLO-2	Text Mining	Spark Core Engine	Overview of Hadoop Ecosystem	**Unit 10: Data Streaming Setup and Configuration**	**Unit 15: Hands On Lab Usecase Implementation (Health -3)**
S-9	SLO-1	Big Data Analytics	Spark SQL	Components of Hadoop Ecosystem ✓ HDFS ✓ YARN ✓ MapReduce ✓ Spark ✓ Pig ✓ Hive ✓ Hbase ✓ Mahout ✓ Zookeeper ✓ Oozie	Introduction to Kafka Event Streaming, Understanding Architecture & Working of Kafka Event Streaming	Hospital readmission
	SLO-2	Text Analytics	Spark Streaming	**Unit 7: Introduction to PySpark**	Steps to Set Up Kafka Event Streaming, Set Up Kafka Environment, Create a Kafka Topic to Store Kafka Events, Write Kafka Events into the Topic	Problem statement
S-10	SLO-1	Information extraction	MLib	Spark Overview	Read Kafka Events, Import/ Export Streams of Events Using Kafka Connect, Process Kafka Events Using Kafka Streams,Terminate Kafka Environment	Problem type
	SLO-2	Text Summarization	GraphX	PySpark Overview	**Unit 11: Data Event Ingestion Setup and Configuration**	Data engineering
S-11 & S-12	SLO-1 SLO-2	Lab 2 :	Lab 5 :	Lab 8:	Lab 11:	Lab 14:
S-13	SLO-1	Question Answering	**Unit 4: Introduction to Stream Concepts**	Business Benefits and Challenges of PySpark	Introduction to Kafka Event Ingestion, Understanding Architecture & Working of Kafka Event Ingestion	Data pipeline
	SLO-2	**Unit 2: Role of Big Data for Data Engineering - Deep Dive**	Data Stream Overview	Components of PySpark	Steps to Set Up Kafka Event Ingestion, Set Up Kafka Environment	Model selection

S-14	SLO-1	Working with Semi-structured Data	Types of Data Stream ✓ Transactional Data Streams ✓ Measurement Data Streams	SparkSession Overview	Load Sample, Build a data cube, Examine the ingestion spec	Model engineering
	SLO-2	Working with Unstructured Data	Characteristics of Data Streams	SparkContext Overview	**Unit 12: Data and System Interoperability**	Model outcome, analysis
S-15	SLO-1	Working with Images	Examples of Data Streams	SparkConf Overview	Confluent Platform and Apache Kafka Compatibility, Using Confluent Platform system Service Unit Files	Model optimization
	SLO-2	Working with audio	Business Benefits of Data Streams	PySpark RDD, MLib, Serializers	Control Center, Apache Kafka, Kafka Connect	Model pipeline
S-16	SLO-1	Working with video	Business Challenges of Data Streams	**Unit 8: Data Processing, Transformations with Spark DataFrames**	Confluent REST Proxy, ksqlDB (ksql), Schema-Registry (schema-registry),ZooKeeper (zookeeper)	Data visualization
	SLO-2	**Unit 3: Big Data Hadoop and Apache Spark Framework**	Applications of Data Streams	DataFrames in Spark Overview	**Unit 13: Introduction to NoSQL Databases**	User interface
S-17 & S-18	SLO-1 SLO-2	Lab 3: Apache Hadoop overview	Lab 6:	Lab 9:	Lab 12:	Lab 15:

Learning Resources	1. Michael Berthold, David J. Hand, (2007), "Intelligent Data Analysis", Springer 2. Tom White (2012), " Hadoop:The Definitive Guide" Third Edition, O'reillyMedia 3. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge Press, 2012.

Learning Assessment

	Bloom's Level of Thinking	Continuous Learning Assessment (50% weightage)								Final Examination (50% weightage)	
		CLA – 1 (10%)		CLA – 2 (10%)		CLA – 3 (20%)		CLA – 4 (10%) #			
		Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice
Level 1	Remember Understand	20%	15%	20%	15%	20%	15%	20%	15%	20%	15%
Level 2	Apply Analyze	20%	20%	20%	20%	20%	20%	20%	20%	20%	20%
Level 3	Evaluate Create	10%	15%	10%	15%	10%	15%	10%	15%	10%	15%
	Total	100 %		100 %		100 %		100 %		100 %	

CLA – 4 can be from any combination of these: Assignments, Seminars, Tech Talks, Mini-Projects, Case-Studies, Self-Study, MOOCs, Certifications, Conf. Paper etc.,

Course Designers		
Experts from Industry	**Experts from Higher Technical Institutions**	**Internal Experts**
Mr.Jothi, Periyasamy , Chief AI Architect DeepSphere.AI, CA, USA	Dr.S.Gopinathan, Associate Professor, University of Madras, Chennai	Mrs.M.R.Sudha,SRMIST