Course Code	PAD21D01T	Course Name	DATA ENGINEERING AND GOVERNANCE	Course Category	D	Discipline Specific Elective	L	T	P	C
							4	0	0	4

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:*
CLR-1 :	*To learn the concepts of Big data*
CLR-2 :	*To impart in-depth knowledge of data lakes*
CLR-3 :	*Understand the principles of Data warehouse*
CLR-4 :	*Basic knowledge of lake on AWS*
CLR-5 :	*Basic knowledge of distributed computing using Spark*
CLR-6 :	*Design principles of* Resilient distributed datasets

Course Learning Outcomes (CLO):	*At the end of this course, learners will be able to:*	Learning			Program Learning Outcomes (PLO)														
		1	2	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		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 :	*Have a thorough Understanding of Big data*	*3*	*80*	*70*	*H*	*H*	*M*	-	-	-	-	-	*H*	*H*	-	-	*M*	*H*	*H*
CLO-2 :	*Understand the concepts of Data warehouse*	*3*	*85*	*75*	*H*	*H*	*H*	*H*	*H*	-	*M*	-	*H*	*H*	-	-	*M*	*H*	*H*
CLO-3 :	*Real time applications of* Data lake design principles	*3*	*75*	*70*	*H*	*H*	*M*	*H*	*H*	-	*M*	-	*H*	*H*	-	-	*M*	*H*	*H*
CLO-4 :	*Deployment knowledge of AWS*	*3*	*85*	*80*	*H*	*H*	*H*	-	-	-	-	-	*H*	*M*	-	-	*M*	*H*	*H*
CLO-5 :	*Design and implementation knowledge of* Spark data frames	*3*	*85*	*75*	*H*	*M*	*M*	*M*	*M*	*M*	*M*	-	*H*	*H*	-	*M*	*M*	*H*	*H*
CLO-6 :	*Real time application of* accumulator param	*3*	*80*	*70*	*H*	*H*	*M*	-	-	-	-	-	*H*	*H*	-	-	*M*	*H*	*H*

Duration (hour)		12	12	12	12	12
S-1	SLO-1	DataArchitecture: Data and Data lifecycle databases and it's types	BigData :Introduction to big data	DataLakeArchitecture: Data silos	Data LakeonAWS: Data lakes and Data warehouses	Distributed Computing usingSpark: PySpark and SQL basics introduction to spark and Hadoop

	SLO-2	SQL vs NoSQL , creating ERD (entity relationship diagram) ,implementing SQL with AWS	building systems to scale with data	Data lakes	Data lake selection criteria	Resilient distributed datasets(RDD) , Spark data frames
S-2	SLO-1	implementing NoSQL with AWS , create , NoSQL DB with python	building systems to scale with data	Data lakes	Data lake and data	Spark architecture
	SLO-2	create SQL DB with python , big data		characteristics of Data lakes	Democratization	working with RDD
S-3	SLO-1	reading data from csv	a quick overview of Hadoop	Data lake architecture	Data lake design principles	creating data frames for RDD – SQL context
	SLO-2	overview of the four vs the importance of volume			AWS Data lake architecture	map() function of RDD
S-4	SLO-1	the importance of variety , the importance of velocity	Map-Reduce overview	Data warehouse	Implement AWS data store	access content of data frame
	SLO-2	the importance of veracity		Data Lakes	Data lake for on-premise and multi-cloud	data frame in spark and Pandas , performance improvements in Spark
S-5	SLO-1	the relationship between the four VS	Map-Reduce overview	Data streams	data processing frameworks for data lake	broadcast variables and accumulators – loading data into a data frame
	SLO-2	Variety and Data structure			real-time big data architectures	Sampling the contents of a data frame
S-6	SLO-1	Validity and Volatility	map phase	Data streams	Data lake reference architecture	grouping and aggregations – visualizing data in a data frame
	SLO-2	finding balance in the four VS use cases			data ingestion and file formats	trimming and cleaning data
S-7	SLO-1	extracting value from the four VS Data driven organizations	map phase of Map-Reduce	migrate data to AWS	ingestion using Sqoop	user-defined functions and Data frames – combining filters –
	SLO-2				Data processing strategies	aggregations – and sorting – using broadcastvariables
S-8	SLO-1	decision making	Shuffle Phase	migrate data to AWS	deriving value from data lakes	using accumulators – exporting
	SLO-2	distributed systems			data life cycle – and glacier	data frame Contents – Custom accumulators
S-9	SLO-1	batch vs in- memory processing	Shuffle Phase	data lakes on AWS	create role for AWS glue service	Join operations – the Spark catalyst optimizer
	SLO-2	tools for data management			upload data to explore the glue web console	Introduction To Spark SQL – Preparing Data For Analysis – Running SQL Queries

S-10	SLO-1	understanding ETL	reducephase	data lakes on AWS	manually create glue table	Inferred And Explicit Schemas – Windowing In Spark
	SLO-2	ETL with Talend open studio			query data lake using amazon Athena	applying – Window Functions – PySpark Basics – sparkconf
S-11	SLO-1	ETL pipeline in Python , AI and machine learning	reducephase	Working with data lakes	configure and run glue crawlers	Spark context – Spark files – RDD – Storage level – Broadcast
	SLO-2	data modelling			access data in crawled tables	accumulator – accumulatorparam – marshalserializer – Pickle serializer – Status tracker
S-12	SLO-1	data partitioning/engineering/ reporting	Difference between Shuffle and Reduce Phase	Data Lakes on AWS	Crawl CSV files, merge data	sparkjobinfo – sparkstageinfo – profiler – basic profiler
	SLO-2				Same schema file manipulation	evaluatorpyspark.ml.tuning Module – PySpark SQL functions – PySpark SQL data Types

Learning Resources	**Text Book:** 1.Data Architecture: A Primer for the Data Scientist, 2nd Edition, By W.H.Inmon, Daniel Linstedt and Mary Levins, April 2019 2.Data Architecture, By Charles Tupper, May 2011	**Reference Book:** 1.Concise Guide to Databases by Peter Lake and Paul Crowther, Springer, 2013 2.The Enterprise Big, Data Lake, By Alex Gorelik, March 2019 3.Apache Spark with Python - Big Data with PySpark and Spark, By Pedro Magalhães Bernardo, Tao W and James Lee, May 2018

Learning Assessment

Level	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	40%	-	40%	-	40%	-	40%	-	40%	-
Level 2	Apply Analyze	30%	-	30%	-	30%	-	30%	-	30%	-
Level 3	Evaluate Create	30%	-	30%	-	30%	-	30%	-	30%	-
	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