A course on Big Data Analytics with Apache Spark in Python

GITAA and GYAN DATA are companies incubated under IIT Madras Research Park. Their founders are Chair Professors in Chemical Engineering Dept. IIT Madras

Course Outline **(Duration 10 weeks / 35 hrs)**

Week	Module	No. of hours
1.	**Introduction** • Introduction to Big Data • Characteristics of Big Data • Challenges with Big Data • Big Data Frameworks • Framework for solving Data Science Problems • Typology of Data Science problems	3 hours 45 mins (1 hour 15 mins /day)
2.	• Installing and Configuring Python, Hadoop, Spark and Jupyter • Hands on- Basics of Python using Jupyter	3 hours 45 mins (1 hour 15 mins /day)
3.	**Distributed Computing** • What and Why of Distributed Systems • Distributed File System • Distributed Programming Model • Parallel Processing explained with WordCount • Concept of Cloud Computing • Big Data and Cloud Computing – Benefits	3 hours 45 mins (1 hour 15 mins /day)
4.	**Hadoop and MapReduce** • Introduction to Hadoop • How MapReduce works • Parallelism in MapReduce • Example: K means Clustering – Sequential and with MapReduce • When does MapReduce work and Why? Comparison among Algorithms • Implementation in Python – Regular and Spark Version of KMeans	3 hours 45 mins (1 hour 15 mins /day)

Course Outline **Big Data Analytics**

Week	Module	No. of hours

5.	**Apache Spark**
	
	• Introduction to Apache Spark,
	• Spark ecosystem and architecture
	• Spark lifecycle
	• Spark API overview
	  ○ Structured Spark types
	  ○ API execution flow
	  ○ What happens when a Spark Session is initiated - Architecture?
	• Spark cluster managers
	  ○ Comparison to other tools
	  ○ Components
	  ○ Program flow
	• Resilient distributed dataset
	  ○ Basics
	  ○ RDD as abstract data type
	  ○ Transformations and actions
	  ○ Caching and checkpointing
		3 hours 45 mins
		(1 hour 15 mins /day)

6.	**Getting started with Spark**
	
	• Understanding spark environment with spark shell and user interface
	• RDD
	• Spark SQL
	  ○ Overview
	  ○ Uses
	  ○ Spark SQL in dataframe and dataset
	  ○ Spark SQL data description language
	  ○ Spark SQL data manipulation language
	• Hands-on session- Spark SQL and functions
		3 hours 45 mins
		(1 hour 15 mins /day)

7.	**Spark DataFrame**
	
	• Spark dataframe and dataframe functions
	  ○ Schema, columns, rows
	  ○ Dataframe operations
	• Working with data types and functions
	  ○ Standard data type (bools, numbers, strings etc)
	  ○ Complex type (structs, arrays etc)
		3 hours 45 mins
		(1 hour 15 mins /day)

	○ Aggregations, grouping, windowing ○ Joins • Hands-on session- Spark dataframes and illustration of data types and functions • Distributed shared variables ○ Broadcast variables ○ Accumulators • Data sources	
8.	• Spark streaming overview • Spark ML pipeline • Case study using PySpark covering ○ Starting Spark session ○ Basic spark operations ○ Reading data ○ Exploratory data analysis ○ Pre-processing data ○ ML algorithms ○ Measuring performance	3 hours 45 mins (1 hour 15 mins /day)
9.	• Case Study in AWS	3 hours 45 mins (1 hour 15 mins /day)
10.	**Course review for Final exam**	1 hour and 15 mins