Topic	Importance	Sub-Topic	Hours
Java Fundamentals (minimal Java knowledge that is required for Hadoop)	Medium	What is Java? How to install, set classpath and path, Hello world program, read from console, write to console, read and write files, play with collections and memory	4
		Exposure to OOPs in Java: Interface, Class, Abstract Class, Inheritance concepts	
		How to use Java for arithmentic, algorithmic and interactive requirements	
		How to create jar and run it	
		How to use external libraries?	
		Threads, Parallel processing vs Concurrent processing - understand what Java can and cannot in handling big volume of data	
		Assignments: 1) Setup Eclipse and run Hello world program 2) Setup Maven with Eclipse, use commons-io for file operations 3) Write a program to demonstrate your understanding of Inheritance 4) Use IVehicle interface to define the contact of a Vehicle, and create multiple Vehicle implementations	
	V.V.High		
Hadoop - Introduction		Concurrent Processsing vs Parallel Processing vs Distributed Processing	4
		What is Map-Reduce? Map-Reduce Framework and its components	
		Commodity Hardware Evolution, Moores law	
		Hadoop Architecture	
		Hadoop Ecosystem	
		Hadoop Distributions	
		Hadoop Evolution	
		Use cases of Hadoop	
		What is HDFS, concepts and how to work with it?	
		Is Java Mandatory to work with Hadoop? Alternatives	
		Hadoop configuration files, shell and hands-on	
		Hadoop Administration	
		Kaggle datasets overview	
		Assignments: 1) How to start and stop Hadoop? 2) How do I setup Hadoop in my local machine?	

Topic	Importance	Sub-Topic	Hours
HDFS, Scoop and Ooozie	Medium	Deep-dive into HDFS - Architecture - Redundancy - Integrity - Fault torelance - Security	8
		Introduction to Scoop - Setup Scoop - How to move bulk data from local to HDFS and vice versa	
		Introduction to Oozie - Architecture - Usage	
		Assignments: 1) Use HDFS CLI commands 2) Copy large files from local file system to HDFS and otherwise using Scoop 3) Create a workflow in Oozie to move data from local to HDFS, trigger a Split, Select some data, move to local file system as one big JSON 4) Monitor the progress of the copy, redundancy, integrity, failures … 5) How to add security?	
Map Reduce	High	Why we need MapReduce?	8
		MapReduce classic example - word count	
		Big volume data: Split, Combine and Partition concepts	
		Using Text, XML and JSON formats in MapReduce	
		What is YARN and how is it supporing MapReduce? - Architecture - Execution workflow - View tasks in the workflow	
		Assignments: (large dataset to be provided - from Kaggle) 1) Split datasets by criteria 2) Combine datasets by criteria 3) Produce aggregate of the datasets 4) Study a YARN workflow and show one of the above problems end to end executed by YARN with HDFS exchange	

Topic	Importance	Sub-Topic	Hours
Pig	Low	Pig - Architecture - As a non-Java programmer how can I use Hadoop using Pig? - Pig Latin scripting - How to deploy Pig Latin scripts?	2
		Assignments: (large dataset to be provided - from Kaggle) 1) Split datasets by criteria 2) Combine datasets by criteria 3) Produce aggregate of the datasets 4) Study a YARN workflow and show one of the above problems end to end executed by YARN with HDFS exchange for a Pig Latin program	
Hive	V.V.High	Typical Data warehousing vs Bigdata warehousing	12
		Hive vs Pig vs MapReduce	
		Why Hive is better than Pig?	
		Metastore and Data warehouse in Hive	
		Data modelling - available Data Types mapped to Java - ANSI SQL	
		ANSI-Joins Introduction - cartesan product - different types of joins	
		Partitions and Bucketing	
		Managed vs External Tables in Warehouse	
		UDF	
		Transactional data processing in Hive - Commit and Rolbacks	
		Schemas and Evolution of Schema	
		HiveQL, Indexing and Views	
		Thrift Server setup and architecture	
		Assignments: 1) Locate Hive Datawarehouse location, change it another location 2) Split datasets by criteria 3) Combine datasets by criteria 4) Produce aggregate of the datasets 5) Study a YARN workflow and show one of the above problems end to end executed by YARN with Hive queries	
Zoo Keeper	Medium	What is Zookeeper, co-ordination, APIs, consistency	2
		Assignments: 1) Comeup with the understanding and use cases of Zookeeper	

Topic	Importance	Sub-Topic	Hours
Hbase	Medium	Denormaliziation, Columnar Databases, Hbase Introduction, Architecture and Components	6
		Hbase CLI	
		Hbase vs Hive vs RDBMS	
		Hbase datamodel	
		Zookeeper co-ordination	
		Hbase - CRUD operations	
		Assignments: 1) CRUD operations in Hbase 2) Bulk Loading of data	
Spark	V.V.High	How Spark complements Hadoop	24
		Spark Ecosystem, Components, Clusters, Nodes, Jobs, Tasks	
		Scala Primer	
		Python Primer	
		PySpark	
		Spark Context, RDD, Dataset, Transformations and Actions - Split - Map - Reduce - Combine	
		Zippelin Notebooks Introduction	
		Connect to HDFS, Hive and Hbase	
		Spark Shell, Sheduler, Jobs, Tasks	
		Delta Tables - ACID Transactions - Data warehouse - Bulk operations - Single Inserts, Partitions - Time Travel feature - Partitions, Bloom Filter	
		Assignments: 1) Setup Apache Spark with Zippelin and Jupyter 2) Spark shell 3) Submit Spark job using Java, Scala, Python, SQL 4) Analyze large dataset and arrive at 10 data points using Spark	
		Provision Databricks and Connect to Azure Data Lake	

Topic	Importance	Sub-Topic	Hours
Azure Databricks	High	Create Delta Tables	16
		Perform Transformations in Databricks Notebook	
		Databricks CLI and REST APIs	
		Assignments: 1) Setup Standlone Databricks cluster 2) Install pytest library 3) Perform the same exercise done in Hive using Delta Tables	
Kafka	V.V.High	Kafka architecture, installation	4
		Message Producer and Consumer	
		Streams handling	
		Assignments: 1) Setup Kafka Infrastructure 2) Demonstrate Kafka near real time streaming	
Case Study	V.V.High	Based on BFSI industry another case study will be given to the participants to work on	6
			96