Hadoop course content

COURSE DETAILS	
1. In-detail explanation on the concepts of HDFS & MapReduce frameworks 2. What is Hadoop 2.X Architecture & How to set up Hadoop Cluster 3. How to write complex MapReduce Programs 4. In-detail explanation on how to load data using tools like Sqoop & Flume 5. How to perform data analysis using tools like PIG, HIVE & YARN 6. How to implement & integrate HBASE & MapReduce 7. How to execute Advanced Usage and Indexing 8. How to schedule jobs using Oozie 9. What are the best practices for overall Hadoop development 10. RTAs on Data Analytics 11. What is Spark & brief about its ecosystem & how to work on RDD Using Spark	**Programming languages**: Java & Scala **Frame works**: Hadoop Distributed File System (HDFS) & MapReduce, spark **Loading Tools**: Sqoop & Flume **Analytical Tools**: Pig, Hive and YARN **Scheduling Tools**: Oozie

CURRICULUM for HADOOP 2.X

S No	Concepts	Syllabus Objectives	Topics	RTAs
1	Understanding Big Data and Hadoop	The syllabus for this lecture would brief about: 1. Big Data 2. Big Data problems & solutions, their limitations 3. HADOOP's solutions that handle Big Data issue 4. Common Hadoop Ecosystem and its Architecture 5. Introduction to HDFS 6. What is a file and how to write & read 7. Brief on MapReduce	1. Big Data, Limitations and Solutions of existing Data Analytics Architecture, 2. Hadoop, 3. Hadoop Features, 4. Hadoop Ecosystem, 5. Hadoop 2.x core components, 6. Hadoop Storage: HDFS, 7. Hadoop Processing 8. MapReduce Framework, 9. Hadoop Different Distributions.	

		Framework and its working style.		
2	Hadoop requirements	The syllabus for this lecture would brief about: Hadoop Requirements	1. Machine learning algorithms(field of study that gives computer the ability to learn without being explicitly programmed) R, PYTHON Mlib, Weka 2. Large scale distributed systems, MPP columnar stores(v Machine learning algorithms(field of study that gives computer the ability to learn without being explicitly programmed) R, PYTHON Mlib, Weka 3. large scale distributed systems, MPP columnar stores (vertica), Graph and NOSQL 4. UNIX/LINUX, Shell Scripts 5. java program frameworks in Agile/scrum methodology 6. SPLUNK (BI tool), TALEND (it's an open studio for big data) 7. Agile software development practices 8. cloud based testing 9. Experience with writing test strategy/plan for IT projects and test status reporting 10. Tableau, Datameer, Azure 11. Serialization such as JSON(JavaScript 12. Object Notation) and BSON(Binary 13. Script Object Notation) 14. Apache Mahout 15. Sentiment analysis 16. Spark streaming and SparkQL 17. Data pipelines, Data API's 18. Oracle/DBA2/SQL 19. Basic ML tools(Machine learning) (weka/R/RapidMiner/SAS/SPSS) 20. Perl 21. Knowledge on Data cleaning and transformations	

			22. JSON transformations 23. RStudio and R Progamming language 24. Informatica ETL products 25. Javascript, HTML, CSS 26. Mesos, Swarm 27. TeraData 28. ertica), Graph and NOSQL 29. UNIX/LINUX, Shell Scripts 30. java program frameworks in 31. Agile/scrum methodology 32. Splunk(BI tool), Talend(it's an open 33. studio for big data) 34. Agile software development practices 35. cloud based testing 36. Experience with writing test strategy/plan for IT projects and test status reporting 37. Tableau, Datameer, Azure 38. Serialization such as JSON(JavaScript Object Notation) and BSON(Binary Script Object Notation) 39. Apache Mahout 40. Sentiment analysis 41. Spark streaming and SparkQ	
3	Hadoop Architecture and HDFS	The syllabus for this lecture would brief about: 1. What is Hadoop Cluster Architecture 2. What are the important Configuring files in a Hadoop Cluster 3. What are the various Data loading techniques 4. What are Single node and Multi nodes and their setups	1. Hadoop 2.x Cluster Architecture Federation and High Availability, 2. A Typical Production Hadoop Cluster, 3. Hadoop Cluster Modes, 4. Common Hadoop Shell Commands, 5. Hadoop 2.x Configuration Files, 6. Single node cluster and Multi node cluster set up Hadoop Administration.	
4	Hadoop MapReduce Framewor	The syllabus for this lecture would brief about: 1. In-depth analysis on Hadoop MapReduce Framework	1. MapReduce Use Cases, 2. Traditional way Vs MapReduce way, 3. Why MapReduce, 4. Hadoop 2.x MapReduce Architecture, 5. Hadoop 2.x MapReduce Components,	

	k	2. How MapReduce works on data stored in HDFS. 3. What are Splits, Combiner & Partitioner. 4. How to work on MapReduce using different data sets	6. YARN MR Application Execution Flow, 7. YARN Workflow, 8. Anatomy of MapReduce Program, 9. Demo on MapReduce. 10. Input Splits, 11. Relation between Input Splits and HDFS Blocks, 12. MapReduce Combiner & Partitioner, 13. Demo on de-identifying Health Care Data set, 14. Demo on Weather Data set.	
5	Advanced MapReduce	The syllabus for this lecture would brief about: 1. Advanced concepts in MapReduce 2. What are Counters, Distributed Cache, MRUNIR, Reduce Join, Custom Input format & Sequence Input Format 3. What is XML Parsing	1. Counters, 2. Distributed Cache, 3. MRunit, 4. Reduce Join, 5. Custom Input Format, 6. Sequence Input Format, 7. Xml file Parsing using MapReduce.	
6	Pig	The syllabus for this lecture would brief about: 1. What is PIG & types of use, demo case 2. How to couple PIG with MapREduce 3. What are PIG Latin Scripting 4. What are PIG running Modes PIG UDF, Pig Streaming, Testing PIG Scripts.	1. About Pig, 2. MapReduce Vs Pig, 3. Pig Use Cases, 4. Programming Structure in Pig, 5. Pig Running Modes, 6. Pig components, 7. Pig Execution, 8. Pig Latin Program, 9. Data Models in Pig, 10. Pig Data Types, 11. Shell and Utility Commands, 12. Pig Latin Relational Operators, 13. File Loaders, 14. Group Operator, 15. COGROUP Operator, 16. Joins and COGROUP, 17. Union, 18. Diagnostic Operators, 19. Specialized joins in Pig,	

			20. Built In Functions (Eval Function, Load and Store Functions, Math function, String Function, Date Function, Pig UDF, Piggybank), 21. Parameter Substitution (PIG macros and Pig Parameter 22. substitution), 23. Pig Streaming, 24. Testing Pig scripts with Punit, 25. Aviation use case in PIG, 26. Pig Demo on Healthcare Data set.	
7	Hive	The syllabus for this lecture would brief about: 1. What are HIVE concepts 2. What are HIVE data types 3. What are Loading & Querying in HIVE, 4. How to run HIVE scripts 5. What are Hive UDF .	1. Hive Background, 2. Hive Use Case, 3. About Hive, 4. Hive Vs Pig, 5. Hive Architecture and Components, 6. Metastore in Hive, 7. Limitations of Hive, 8. Comparison with Traditional Database, 9. Hive Data Types and Data Models, 10. Partitions and Buckets, 11. Hive Tables(Managed Tables and External Tables), 12. Importing Data, 13. Querying Data, 14. Managing Outputs, 15. Hive Script, 16. Hive UDF, 17. Retail use case in Hive, 18. Hive Demo on Healthcare Data set.	
8	Advanced Hive and HBase	The syllabus for this lecture would brief about: 1. What are Advanced HIVE concepts 2. What are UDF, Dynamic Partitioning, HIVE indexes & Views 3. What are Optimizations in HIVE 4. In-depth analysis on HBase,	1. Hive QL: Joining Tables, 2. Dynamic Partitioning, 3. Custom Map/Reduce Scripts, 4. Hive Indexes and views 5. Hive query optimizers, 6. Hive : Thrift Server, 7. User Defined Functions, 8. HBase: 9. Introduction to NoSQL 10. Databases and HBase,	

		its Architecture, components and its running modes	11. HBase v/s RDBMS, 12. HBase Components, 13. HBase Architecture, 14. Run Modes & Configuration, 15. HBase Cluster Deployment.	
9	Advanced HBase	The syllabus for this lecture would brief about: 1. What are Advanced HBase Concepts 2. How to perform bulk loading 3. What are filters 4. What is Zookeeper and how it helps in Cluster monitoring. 5. Why HBase utilizes Zookeeper	1. HBase Data Model, 2. HBase Shell, 3. HBase Client API, 4. Data Loading Techniques, 5. ZooKeeper 6. Data Model, 7. Zookeeper Service, 8. Zookeeper, 9. Demos on Bulk Loading, 10. Getting and Inserting Data, 11. Filters in HBase.	
10	Processing Distributed Data with Apache Spark	The syllabus for this lecture would brief about: 1. What is Spark Ecosystem 2. What is Scala and its utility in Spark 3. What is SparkContext 4. How to work on RDD in Spark 5. How to run a Spark Cluster 6. Comparison of MapReduce vs Spark	1. What is Apache Spark, 2. Spark Ecosystem, 3. Spark Components, 4. History of Spark 5. Spark Versions/Releases, 6. Spark a Polyglot, 7. What is Scala?, 8. Why Scala?, 9. SparkContext, 10. RDD.	
11	Oozie	The syllabus for this lecture would brief about: 1. How multiple Hadoop ecosystem components work 2. How they should be implemented to solve Big Data Issues	1. Flume and Sqoop Demo, 2. Oozie, 3. Oozie Components, 4. Oozie Workflow, 5. Scheduling with Oozie, 6. Demo on Oozie Workflow, 7. Oozie Co-ordinator, 8. Oozie Commands, 9. Oozie Web Console, 10. Oozie for MapReduce, 11. PIG, Hive, and Sqoop, 12. Combine flow of MR, PIG, Hive in	

			Oozie	
12	Hadoop Project	The syllabus for this lecture would brief about: 1. What are Flume & Scoop 2. How to utilize Apche Oozie Workflow Scheduler for Hadoop Jobs 3. What is Hadoop Tableau Integration	1. Hadoop Integration with Tableau	