Big Data Analytics using Hadoop Training

The Course

Through instructor-led discussion and interactive hands-on exercises participants will navigate the Hadoop ecosystem

The Rulepaper Promise

Our training methodologies promises to give the students hands on art enterprise skills to delve deeper into the technologies from a practical and enterprise point of view. Extreme Hands-on-Lab with a self doable on the fly practical based training approaches makes transformation of the student from a no vice to a capable experienced cloud computing engineer.

The Eligibility

Passionate Technology Enthusiasts with a minimal knowledge on IT and Operating Systems.

Good to have basic knowledge on Windows, OS & IT Infrastructure.

The Instructor

Enterprise Architect with huge experience on Private and Public Cloud Technologies. The trainers are advisors and members of larger Cloud Computing Forums and seasoned integrators of IT Cloud Computing technologies with more than 12+ years in global large enterprise giants.

Course Contents

Module 1

Hadoop Fundamentals

- What is Big Data and role of Hadoop in Big Data?
- The Motivation for Hadoop and use cases.
- Hadoop 2.0 Overview
- Distributed Data Processing: YARN
- Hadoop vs RDBMS
- Hadoop Ecosystem
- Data Processing and Analysis: Pig, Hive, Spark
- Data Integration: Sqoop
- Streaming Analysis: Flume
- Workflow: Oozie

Module 2

Data Storage: HDFS

- HDFS components (Blocks, Name Node, Data Node)
- HDFS High Availability
- Important HDFS commands
- Anatomy of File Read and Write

Module 3

Input Data into HDS

- Using Hadoop client
- Web HDFS
- Using Sqoop (data transfer between RDBMS and Hadoop)
- Flume (extract streaming data)

Module 4

MapReduce – Analysing Data with Hadoop

- Understanding Map and Reduce concepts
- WordCount MapReduce Program

Module 5

Introduction to Apache Pig

- What Is Apache Pig? its features and use cases
- Interacting with Pig – Pig Latin and Grunt shell
- Running Pig – Local Mode, MapReduce Mode

Module 6

Basic Data Analysis with Pig

- Pig Latin syntax and data types
- Defining and viewing the schemas
- Loading and storing data.
- Grouping, filtering and sorting data

Module 7

Advanced Data analysis using Pig

- Using operators – FOREACH, NESTED FOREACH, CASE, FLATTEN, PARALLEL
- Frequently used built in functions.
- Joining data sets – performing inner joins, outer joins, right outer joins, left outer joins, replicated joins, COGROUP
- Pig User Defined function. An UDF example. How to invoke a UDF?
- Pig scripts and parameter substitution
- Advance data analysis using Pig
- Tips for optimizing performance of Pig jobs

Module 8

Introduction to Hive

- What is Hive? Hive Query Language (HQL) versus SQL
- Hive Architecture
- Hive QL syntax and data types
- Invoking Hive, Hive Shell, submitting Hive queries

Module 9

Hive Data Management

- Creating Hive databases and Tables.
- - Managed Tables and External tables
- Different ways of loading data into Hive table
- Simplifying queries using Views
- Storing query results to a file

Module 10

Hive Data Storage

- Hive partitions, buckets and skewed tables
- Hive File Formats – SerDe, ORC, sequential
- Sorting Data – ORDER BY and SORT BY

Module 11

Hive Data Analysis

- Hive Joins – Inner joins and Outer Join
- Commonly used Hive Built in Functions
- Hive user defined function.
- Using Aggregation and Windowing. PARTITION BY clause
- Analytical functions – RANK, DENSE RANK

Module 12

Hive Performance Optimization

- Hive CBO, computing column and table statistics
- Tips for Hive Performance Optimization

Module 13

Hive metadata integration with Pig

- About Hcatalog
- Hcatalog in the Hadoop ecosystem
- Using HCatloader to load data into Pig relation from Hive table
- Using HCatstorer to store data from Pig into Hive table.
- Lab : using HCatalog with Pig.

Module 14

Oozie for Scheduling jobs

- Oozie components – Actions, Fork, Join Nodes, Workflow, Coordinator
- Submit a Oozie Workflow
- Lab: create a oozie workflow to run Pig and Hive job.

Module 15

Introduction to Apache Spark

- What is Apache Spark? Spark Origin
- Spark Ecosystem
- Spark use cases
- Spark versus MapReduce
- What is Spark context?
- Understanding RDD.
- Create an RDD
- Spark Operations - Transformations and Actions
- Examples of Actions and Transformations.
- Spark WordCount program using Python
- Lab : getting started with Spark

Module 16

Spark SQL and DataFrames

- Spark SQL overview
- DataFrames overview
- SQLContext and HiveContext
- Performing Spark SQL queries
- Performing DataFrame operations
- Lab: Spark SQL and DataFrame exercise to perform data analysis.

The Duration

Duration of the Course is 40 hours.

The Lab Requirements

Students must bring their own laptops with basic configuration

The cost of the Training

Please send an email or contact us at enquiry@rulepaper.com to know more about the cost and next batch schedules.