Big Data - Hadoop

Description

Data scientists build information platforms to provide deep insight and answer previously unimaginable questions. Hadoop are transforming how data scientists work by allowing interactive and iterative data analysis at scale. Learn how Hadoop enable data scientists to help companies reduce costs, increase profits, improve products, retain customers, and identify new opportunities. This course helps participants understand what data scientists do, the problems they solve, and the tools and techniques they use. Through in-class simulations, participants apply data science methods to real-world challenges in different industries and, ultimately, prepare for data scientist roles in the field.

Expectations and Goals

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, and develop concrete skills such as:

- How to identify potential business use cases where data science can provide impactful results?
- How to obtain, clean and combine disparate data sources to create a coherent picture for analysis?
- What statistical methods to leverage for data exploration that will provide critical insight into your data?
- Where and when to leverage Hadoop streaming and Apache Flume for data science pipelines? What machine learning technique to use for a particular data science project?

Prerequisites

- Object Oriented Programming in Java, Exception in Java
- Knowledge of SQL Command
- Basic Command in Linux

Course Schedule

Module	Topic
Module 1	**Introduction to Big Data** Features of Hadoop Components in Hadoop Concept of Hadoop Ecosystem Introduction to HDFS HDFS Practical
Module 2	**Collection Framework in Java (List, Map, Iterator)** String Tokenizer, File Handling, String Handling Concept of Map Reduce Map Reduce Practical
Module 3	**Introduction to Hive** Why we need Hive Architecture of Hive Hive Data Types Hive Complex Datatypes Managed Tables External Tables Running Hive Queries Perform Joining Handling JSON Data

	Handling XML Data Partitioned Table Hive UDF Scripting in Hive Performance Tuning in Hive Case Study in Hive based on Dataset
Module 4	**Introduction to Pig** Why we need Pig Technology Architecture of Pig Pig Data Types Different Modes in Pig Running Pig Command Tuple,Bag,Map Pig UDF Script in Pig Case Study in Pig Based on Dataset
Module 5	**Introduction to SQOOP** Importing and Exporting the RDBMS to HDFS Import data from RDBMS to Hive Export Data from Hive to RDBMS Assignment
Module 6	**Introduction to Flume** Introduction Source,Sink,Flume Agents Fetching Twitter Data into Solr Configuration to create twitter data into HDFS Use HiveSerde to Analyze the data
Module 7	Sentiment Analysis based on Twitter Data
Module 8	**Concept of Impala** Running the Queries on Impala Compare Impala with Hive Concept of HUE Access The Hadoop Component by means of HUE
Module 9	**Introduction to NOSQL Database** Compare between NOSQL and RDBMS Introduction to HBASE Why we need HBASE HBASE Commands
Module 10	**Installation of Tableau** Communicate Tableau with impala Plotting the Graph
Module 11	Project Work and Documentation