Hadoop + R Training + Machine Learning

Hadoop

1. HADOOP ECOSYSTEM & CLUSTER

Available version Hadoop 1.x & 2
Available Distributions of Hadoop (Cloudera, Hortonworks)
Hadoop Projects & Components
Architecture of Hadoop & Planning for cluster
The Hadoop Distributed File System (HDFS)
Cluster Daemons & Its Functions.
Name Node
Secondary Node
Data Nodes
Application Master and Task Tracker
YARN Responsibilities
Deployment of Hadoop Cluster

2. CLOUDERA SANDBOX OR QUICK START

Installation of cloudera quick start
Difference in sandbox and distributed environment
Overview of apache HUE

3. MAP-REDUCE, MAP-REDUCE STEAMING (IN JAVA)

All Map-Reduce API Concepts
Architecture of Map-Reduce
Writing Map-Reduce Drivers, Mappers, and Reducers in Java
Speeding Up Hadoop Development by Using Eclipse
Differences between the Old and New Map-Reduce APIs
Writing Mappers and Reducers with the Streaming API
Different question raised for Map-Reduce

4. HADOOP SHELL AND COMMANDS

Hadoop Developer commands using shell
Map-Reduce job deployment
Oozie workflow design
Installing Hadoop(Cloudera)
Different Components Jobs design.

5. HCATALOG OR METASTORE TABLES

Introduction of apache Hcatalog
Creating tables using Hcatalog
Bulk uploads using MetaStore Tables
Play with semi-structured data
Integration of Hcatalog with Hive
Hive SQL query analysis

6. HIVE

Problems with No-SQL Database
Introduction & Installation Hive
Hive Schema and Data Storage
Data Types & Introduction to SQL
Hive-SQL: DML & DDL
Hive-SQL: Views & Indexes
Explain and use the various Hive file formats
Use Hive to run SQL-like queries to perform data analysis
Use Hive to join data sets using a variety of techniques, including Mapside joins and Sort-Merge-Bucket joins
Integration to HBase & Cassandra
Sentiment Analysis and N-Grams
Hive Thrift Service

7. APACHE DRILL – REPLACEMENT OF MAP-REDUCE

Installation of Drill
Query data using apache drill
Query data from Hadoop/HDFS file system
Drill & Hbase integration
Drill & Hive integration & Replacement

8. FLUME

Installation of Flume
Ingesting Data from External Sources with Flume
Configuration for flume
REST Interfaces
Best Practices for Importing Data

9. SQOOP

Installation of Sqoop
Ingesting Data from External (RDBMS) Sources with Sqoop
Ingesting Data from/to Relational Databases with Sqoop
Integration of Sqoop and Hbase
Integration of Sqoop and Hive
Best Practices for Importing Data

10. CONCLUSION & FAQS

Note:
Every Topic has practical session
Hadoop uses different components which discussed in required sessions
Hue
Cloudera Manager
Zookeeper
Ooozie
Etc

PREREQUISITES

This course is best suited to developers and engineers who have some or little bit programming experience. Knowledge of Java is not mandatory, Any programming language can be used with Hadoop and is required to complete the hands-on exercises.

R Training

1: FUNDAMENTALS OF 'R'

First steps with R
Discover the data types & variable in R
Installing R on personal machines. retrieving R packages.
Basics of R, R-Studio, R Markdown.
Data types, variable assignment

2: VECTORS

Analyze gambling behaviour using vectors. Create, name and select elements from vectors.
Comparing Vectors
Selection from Vectors
Sorting of Vectors

3: MATRICES

Work with matrices in R
Computations with matrices
Demonstrate your knowledge by analyzing the any data figures
Comparing Matrices
Selection from Matrices
Sorting of Matrices

4: FACTORS

R stores categorical data in factors
Learn how to create subset and compare categorical data.
Comparing Factors
Selection from Factors
Sorting of Factors

5: DATA FRAMES

Learn how to create data frames
Data sets and structure
Selection from data frames
Sorting of data frames

6: LISTS

Learn how to create list
list and data structure
Selection and Sorting from/of list

7: MODULE

If/else statements.
For/while loops.
Functions
Apply() family over data
Utilities like with(), grepl(), sub() to specify environment

8:MODULE

Writing Functions in R
A quick refresher for functions
Functional programming
Advanced inputs and outputs
Robust functions

9: IMPORTING DATA INTO R

Importing data from flat files
Importing data from Excel
Importing data from Databases
Importing data from the web

ADVANCE R DATA VISUALIZATION

1: MODULE

The Grammar of Graphics
Lines and Syntax
Transformations
Interactivity and Layers
Customizing Axes, Legends.

2: MODULE

Data Visualization with ggplot2
Introduction & Data
Aesthetics
Geometries
qplot and wrap-up
Statistics
Coordinates and Facets
Themes

DATA VISUALIZATION - BEST PRACTICES & CASE STUDY

Statistical Modelling

Intro to Statistics with R:
Histograms and Distributions
Scales of Measurement
Measures of Central Tendency
Measures of Variability

Machine Learning

INTRODUCTION OF MACHINE LEARNING WITH R

Motivation of Machine Learning?
Use Cases of Machine Learning
Future Scope of Machine Learning
Real World Domain using ML
Types of Machine Learning
Different tools/framework available for ML
Limitation of Machine Learning

BASICS OF MACHINE LEARNING

Understanding supervised & unsupervised learning
Understanding bases associated with any machine learning algorithm
Better understnding with SPAM OR HAM detection
Ways of reducing bias and increasing generalisation
Difference between Rule base & ML based approach

SUPERVISED LEARNING - CLASSIFICATION PROBLEMS

Uses of classification in spam detection
How shapes(Line, Square, Cube) works in classification problems
An N-Dimensional hypercube

- Naive Bayes - Probabilistic Classifier
- K-Nearest Neighbours - Non-Probabilistic Classifier
- Support Vector Machines (SVM)

NAIVE BAYES - PROBABILISTIC CLASSIFIER

What is probability distribution
Random variable? & Types of Random variable?
Standard Deviation?
Bayes Theorem - Probabilistic Theorem?
What is conditional probability?
Use Naive Bayes with R.
Linear Regression with Multiple Variables

K-NEAREST NEIGHBOURS - NON-PROBABILISTIC CLASSIFIER

What is KNN classifier?
Calculate K in KNN?
Kind of problem instance in KNN
Difference between Naive Bayes & KNN
Definition of distance
Data reduction & Dimensionality Reduction?
Feature extraction in KNN

SUPPORT VECTOR MACHINES (SVM)

Learn the simple intuition behind Support Vector Machines.
Implement an SVM classifier in SKLearn/scikit-learn
Choose the right kernel for your SVM
Learn about RBF and Linear Kernels

REGRESSION

Logistic Regression
Linear Regression
Regularization

UNSUPERVISED LEARNING - CLUSTERING

What is unsupervised Learning?
Uses of clustering in facebook

- K-Means Clustering
- Hierarchical clustering
- Density-Based clustering
- Distribution-Based clustering