Hadoop and Hive Tutorial (3 Days)

This course would be a mix of theory and hands-on session. The primary focus would be to let participants get started with Hadoop and Hive. This can happen when they do more hands-on. But this can not justify less theoretical understandings.

For more information about the systems you are referred to the corresponding documentation pages:

- Apache Hadoop 2.6: http://hadoop.apache.org/docs/r2.6.0/
- Apache Hive 1.2: https://hive.apache.org/

1. Course Pre-requisite

1.1 Infrastructure

All the participants should have Oracle VirtualBox installed (https://www.virtualbox.org/). A Virtual Machine (VM) is provided with all the software already installed and configured. Within the machine the source code and jars are also provided, along with the used datasets. As a first step you have to download the VM and open it with VirtualBox. The VM holds an Ubuntu 14.04 LTS operative system with the essential packages needed for the execution of the examples. The VM is configured with 4GB RAM by default, but you can modify the corresponding settings to fit best with your host system.
There would be a hands on session on how to set up HDFS and YARN, But a VM with all set up will get you started easily. (installing everything on own is recommended)
The VM would be shared at the end of Day 1.

1.2 Language

Course Language would be Python and although it does not matter much, since python has a very easy syntax people who are not familiar with python will not face any issue. For Hive, participants should have knowledge of SQL and RDBMS. Since Hive is simple SQL on HDFS.

1.3 Datasets

1.3.1 NASA web logs

This dataset contains 205MB of uncompressed logs extracted from the HTTP requests submitted to the NASA Kennedy Space Center web server in July 1995.More information about the dataset can be found in http://ita.ee.lbl.gov/html/contrib/ NASA-HTTP.html

1.3.2 MovieLens dataset

The second dataset corresponds to the MovieLens 1M dataset, which contains 1 million ratings from 6000 users on 4000 movies. The description of the dataset is provided as part of the zip file and is also available at http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

The dataset is divided into three main files:

- Users: Each line contains information about a user: its identifier, gender, age, occupation and zip-code.

- Movies: Each line contains information about a movie: identifier, title (as it appears in IMDB, including release year) and a list of genres.

- Ratings: Each line contains a user rating with the following information: user identifier, movie identifier, rating (5-star scale) and timestamp.

2. Course Outline

2.1 Day-1

First day would be on Big Data Introduction and space of Hadoop and Hive in Big Data ecosystem. I will also cover basics of HDFS and Distributed Architecture. We will go through the architecture of HDFS (Hadoop Distributed Architecture). The topics which would be covered on day 1 are :

Introduction to Big Data
Big Data Definition
Challenges with traditional system

What is Big Data?
Attributes of Big Data
Types of Data
Facts about Big Data
Sources of Big Data
Big Data Use Cases
What is Hadoop?
Overview of Hadoop Ecosystem
History of Hadoop
HDFS Architecture
Need for a Distributed File System (DFS)
The Hadoop Distributed Filesystem
Features of HDFS Streaming data access
Features of HDFS – Commodity Hardware
Features of HDFS – Moving Computation to the Data
HDFS Concepts - Blocks
HDFS Components – NameNode and DataNode
HDFS Components – NameNode Operations
HDFS Components – Secondary NameNode
Features of HDFS – Data Replication
HDFS Architecture – Hadoop 1.0
HDFS- The anatomy of a File Write
HDFS- The anatomy of a File Read
Rack Awareness
Regular File System vs. HDFS
Apache Hadoop 2.0 and YARN
HDFS Federation
Hadoop 1.0 Vs Hadoop 2.0
HDFS Federation – Key Benefits

Modes of Hadoop Deployment
Hands-On (Installation and VM file distribution)
Hadoop Deployment in Pseudo Distributed Mode
Hadoop Configuration
JDK Installation
JPS and ssh installation
HDFS Basic Commands

2.2 Day-2

Day 2 would be more hands on session. We will go through the basics of Map -Reduce framework. How to write a map and reduce code and why it is required. What are its limitation. Once participants would got the basics of Map - Reduce.

2.2.1 Hadoop MapReduce

In the first part of the tutorial we are going to use Apache Hadoop MapReduce to process the information contained in the NASA logs datasets.

MapReduce programs basically consist in two user defined functions called respectively map and reduce. These functions are applied to records of key-value pairs and are executed in parallel. They also produce key-value pairs, the only restriction being that output key-value pairs of the map functions are of the same type as input key-value pairs of the reduce function.

Map Reduce Basics
Traditional Way of Processing Data
Hadoop Overview
Introduction to MapReduce
MapReduce - Analogy
MapReduce - Example word Count
Stages of MapReduce
MapReduce Daemons - JobTracker

MapReduce Daemons - TaskTracker
MapReduce – Web UI
MapReduce Principles
MapReduce and HDFS
MapReduce (Hands On) Nasa Web Log
The MapReduce in detail
The Map Side
The Reduce Side
MapReduce Types
Input Formats MapReduce
Output Formats MapReduce
Datatypes in MapReduce
Joins in MapReduce
Map Side Joins
Reduce Side Joins

2.3 Day-3

2.3.1 Hive

Hive is a data warehouse solution to be used on top of Hadoop. It allows to access the files in HDFS the same way as MapReduce and query them using an SQL-like query language, called HiveQL. The queries are transformed into MapReduce jobs and executed in the Hadoop framework.

The metadata containing the information about the databases and schemas is stored in a relational database called the metastore. You can plug any relational database providing JDBC. In this tutorial we are using the simpler embedded derby database provided with Hive.

In this part of the tutorial we are going to use Apache Hive to execute the same queries as in the previous example and observe that it is much easier to write them in the new framework. Additionally we are going to use the MovieLens dataset to execute other type of queries

MapReduce (Hands On) MovieLens Datset
Revisit SQL
Tables and Joins
Hive
Definition of Hive?
Features of Hive
Components of Hive
Architecture of Hive
Where to use Hive
Limitations of Hive
Hive Use Cases - Facebook
Hive Vs Traditional Databases
Hive Vs Pig
Why choose Hive over Pig?
Hive Example
Hive (Hands On) Nasa Log Datasets
Hive Deployment
Tables in Hive
Datatypes in Hive
Partitions and Buckets in Hive
Hive Examples MovieLens Dataset Analysis