Big Data Training

Sriram Ragavendran

Course Outline:

Module 1: Why Big Data and Hadoop? Need to learn.
Module 2: Introduction to HDFS.
Module 3: Introduction to Map Reduce.
Module 4: Dissecting a MapReduce Program.
Module 5: Introduction to Pig.
Module 6: Introduction to Hive.
Module 7: Introduction to Sqoop.
Module 8: Introduction to Oozie.
Module 9: Introduction to HBase.
Module 10: What Next?

Assumptions/Requirements:

- I will be using Cloudera's Virtual Machine to conduct the training. The trainees would have to install this on their machines so they can follow along and try the exercises at their end. This would require at least 4GB of RAM to run.

- You would have to know a bit of programming (object oriented) to understand the Map Reduce programs. You don't need to know Java in particular – familiarity with any language should help understand and follow the class.

- Understanding Pig / Hive/ Sqoop / HBase requires that you are good with RDBMS concepts and SQL Language. If you are completely new, you may struggle to follow the class. It is recommended that you gain some knowledge in this area before attending this course.

- The course will help you understand how to develop programs for the Hadoop eco system. This will go very light on the installation aspects. For instance, we won't discuss how to setup a multi node cluster as part of this course.

Big Data Training

Sriram Ragavendran

Details of Content Covered:

Module 1: Why Big Data and Hadoop? (Conceptual)

- A look at some of the recent happenings around us, in the fields of transportation, education, healthcare, commerce and sports to understand the role Big Data plays in our lives.
- History of Hadoop and its role in the Big Data Space.

Module 1a: Setting up of Cloudera VM: (Hands-on)

- Installing Virtual Box.
- Firing up an instance of Cloudera VM.
- Prepping up the VM instance for the course.

Module 2: Introduction to HDFS (Concepts)

- Role of HDFS in the Big Data world.
- Comparison of Local FS and HDFS.
- Discussion on Name Node and Data Node.
- Anatomy of File Read / Write in HDFS.
- Failures – NameNode / Data Node
- HDFS Federation – What / Why

Module 2a: Unix FS Commands and HDFS Commands (Demo)

- Working with files and folders (Local FS)
    - Creating / Removing / Copying / Moving
- Working with HDFS Commands (from the Unix Shell)
- Using Hue to work with HDFS.

Module 3: Introduction to Map Reduce (Concepts)

- Map Reduce explained with a simple real-life example
- Map Reduce introduced with the WordCount program (Demo)
- YARN – Introduction
    - Components, Failures, Scheduling

Big Data Training

Sriram Ragavendran

Module 4: Understanding a Map Reduce Program (Concepts + Demo)

- Discussions will cover
  - Tool / Tool Runner Classes
  - Key- Value Pairs
  - Input / Output File Formats
  - Serialization
  - Combines / Partitioners
  - Shuffle and Sort
  - Counters

Module 5: Pig (Demo Intensive)

- Need / History of Pig
- Discussion on Pig Latin
  - LOAD/DESCRIBE/ILLUSTRATE/DUMP commands
  - FILTERS / FOREACH / GENERATE/ GROUP commands
  - UNION / SPLIT
  - Writing UDFs

Module 6: Hive (Demo Intensive)

- Need / History of Hive
- HiveQL Vs SQL (RDBMS) – a very brief discussion
- Deep Dive into HiveQL
  - Managed / External Tables
  - Buckets / partitions
  - CTAS / Multiple Inserts
  - Different ways of interacting with Hive (shell / hue)
  - Writing UDFs and using them in Hive

Module 7: SQOOP (Demo Intensive)

- Why SQOOP?
- Importing Data into Hadoop using SQOOP
  - Import into HDFS
  - Direct import into Hive
- Exporting Data from Hadoop using SQOOP

Big Data Training

Sriram Ragavendran

Module 8: OOZIE (Demo Intensive)

- What purpose does Oozie serve?
- Discussion on Workflow.xml and Job.Properties files for various tools
- Discussion on Action / Control nodes
- Demo involving multiple operations using Hue Editor

Module 9: HBase

- Need for HBase
- HBase Vs SQL (RDBMS)
- Concepts – Regions / Zoo Keeper etc
- Operations
- Demo

Module 10: What Next?

- Brief Introduction to Data Analytics using R
- Introduction to Apache Spark
- Pointers to resources for further learning.