Apache Hadoop

Functional Area: Parallel Horizontally Scalable Data Processing:Batch

Overview:

Apache Hadoop is an open source software framework basically used for storage and processing of large-scale data sets on clusters.

The framework is composed of 2 main modules HDFS which is a distributed file system that stores data and Hadoop MapReduce a programming module for large scale data processing

Single-node Hadoop installation: Hadoop runs properly on 64 bit machine

Software Requirements: Java Development Kit  
 Linux OS or Ubuntu 14.04 on Virtual Box  
 Windows- Cygwin

Installation Tutorial

1. Install Java from Oracle
2. Download a stable release of Hadoop from <http://www.apache.org/dyn/closer.cgi/hadoop/common/>
3. Uncompress the archive:  
    tar xvzf hadoop-2.4.1-bin.tar.gz
4. Edit the file conf/hadoop-env.sh and add :  
    export JAVA\_HOME=<JDK DIRECTORY>
5. Export/set the Hadoop path in ~/.bashrc  
    export HADOPP\_INSTALL=/path/to your installation  
    PATH=$PATH:$HADOOP\_INSTALL/bin:$HADOOP\_INSTALL/sbin
6. Generate ssh key and authorize the keys  
    ssh-keygen –t rsa –P ‘ ’  
    cat ~/.ssh/id\_rsa.pub > ~/.ssh/authorized\_keys
7. If ssh is not installed  
   Install ssh server and then generate the ssh key  
    sudo –apt get install openssh-server

Check if hadoop is properly installed  
$hadoop version

Configuration for pseudo-Distributed mode:

Conf/core-site.xml:

<Configuration>  
 <property>  
 <name>fs.default.name</name>  
 <value>hdfs://localhost:9000</value>  
 </property>  
<configuration>

Conf/hdfs-site.xml:

<Configuration>  
 <property>  
 <name>dfs.replication</name>  
 <value>1</value>  
 </property>  
<configuration>

Conf/mapred-site.xml:

<Configuration>  
 <property>  
 <name>mapred.job.tracker</name>  
 <value>localhost:9001s</value>  
 </property>  
<configuration>

Execution :

1. Format a new distributed file system:  
    $bin/hadoop namenode –format
2. Start the hadoop daemons:  
    $bin/start-all.sh
3. Copy the input files into the distributed system:  
    $bin/hadoop fs –put conf input
4. Run some examples  
    $bin/hadoop jar hadoop-examples-\*.jar grep input output ‘dfs[a-z.]+’
5. View output files   
    $bin/hadoop fs –cat output/\*
6. Stop the daemons  
    $bin/stop-all.sh