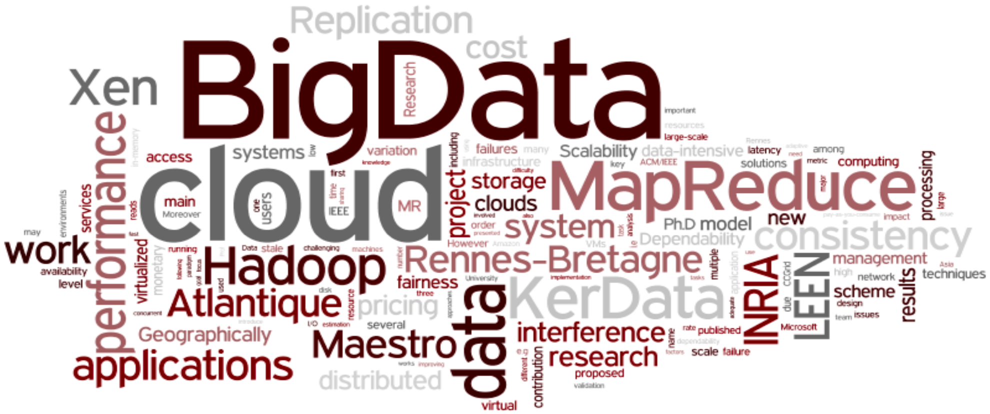
Hadoop
TP 1



Shadi Ibrahim

Inria, Rennes - Bretagne Atlantique Research Center

Getting started with Hadoop

Prerequisites
Basic Configuration
Starting Hadoop
Verifying cluster operation

Prerequisites

- Hadoop requires a working Java installation.

- Hadoop requires SSH access to manage its nodes

Basic Configuration

The **hadoop/conf** directory contains some configuration files for Hadoop.

hadoop-env.sh - This file contains some environment variable settings used by Hadoop. You can use these to affect some aspects of Hadoop daemon behavior, such as where log files are stored, the maximum amount of heap used etc. The only variable you should need to change in this file is JAVA_HOME, which specifies the path to the Java xx installation used by Hadoop.

hadoop-env.sh

```
# Set Hadoop-specific environment variables here.
# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.  Required.
#export JAVA_HOME=/System/Library/Frameworks/JavaVM.framework/Versions/1.6.0/
# export JAVA_HOME=/usr/lib/j2sdk1.6-sun
#export JAVA_HOME=/Library/Java/Home
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64/

# Extra Java CLASSPATH elements.  Optional.
# export HADOOP_CLASSPATH=x

# The maximum amount of heap to use, in MB. Default is 1000.
export HADOOP_HEAPSIZE=2000
```

hadoop-env.sh

```
# Where log files are stored.  $HADOOP_HOME/logs by default.
# export HADOOP_LOG_DIR=${HADOOP_HOME}/logs

# File naming remote slave hosts.  $HADOOP_HOME/conf/slaves by default.
# export HADOOP_SLAVES=${HADOOP_HOME}/conf/slaves

# host:path where hadoop code should be rsync'd from.  Unset by default.
# export HADOOP_MASTER=master:/home/$USER/src/hadoop
```

Basic Configuration

The **hadoop/conf** directory contains some configuration files for Hadoop.

masters - This file lists the host where the Hadoop master daemons (namenode and jobtracker) will run. By default this contains the single entry **localhost**

slaves - This file lists the hosts, one per line, where the Hadoop slave daemons (datanodes and tasktrackers) will run. By default this contains the single entry **localhost**

Basic Configuration

The **hadoop/conf** directory contains some configuration files for Hadoop.

core-site.xml - This file contains site specific settings for the Hadoop HDFS namenode daemon.

core-site.xml

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

Basic Configuration

The **hadoop/conf** directory contains some configuration files for Hadoop.

hdfs-site.xml - This file contains site specific settings for the Hadoop HDFS daemons.

hdfs-site.xml

```
<property>
   <name>dfs.replication</name>
     <value>1</value>
</property>
```

Basic Configuration

The **hadoop/conf** directory contains some configuration files for Hadoop.

mapred-site.xml - This file contains site specific settings for the Hadoop MapReduce daemons and jobs.

mapred-site.xml

```
<property>
    <name>mapred.job.tracker</name>
      <value>localhost:9001</value>
    </property>
<property>
```

Starting Hadoop

start-dfs.sh - Starts the Hadoop DFS daemons, the namenode and datanodes. Use this before start-mapred.sh

stop-dfs.sh - Stops the Hadoop DFS daemons.

start-mapred.sh - Starts the Hadoop Map/Reduce daemons, the jobtracker and tasktrackers.

stop-mapred.sh - Stops the Hadoop Map/Reduce daemons.

Staring Hadoop

start-all.sh - Starts all Hadoop daemons, the namenode, datanodes, the jobtracker and tasktrackers. Deprecated; use start-dfs.sh then start-mapred.sh

stop-all.sh - Stops all Hadoop daemons. Deprecated; use stop-mapred.sh then stop-dfs.sh

Starting Hadoop

Formatting the HDFS filesystem via the NameNode

The first step to starting up your Hadoop installation is formatting the Hadoop filesystem which is implemented on top of the local filesystem of your cluster.

It simply initializes the directory specified by the dfs.name.dir variable), run the command

```
bin/hadoop namenode -format
```

Verifying cluster operation

Access any of the GUI links:

http://localhost:50070/ – web UI of the NameNode daemon

http://localhost:50030/ – web UI of the JobTracker daemon

http://localhost:50060/ – web UI of the TaskTracker daemon

Thank you!

KerData

Shadi Ibrahim

shadi.ibrahim@inria.fr