Hadoop

By Dinesh Amatya

Hadoop

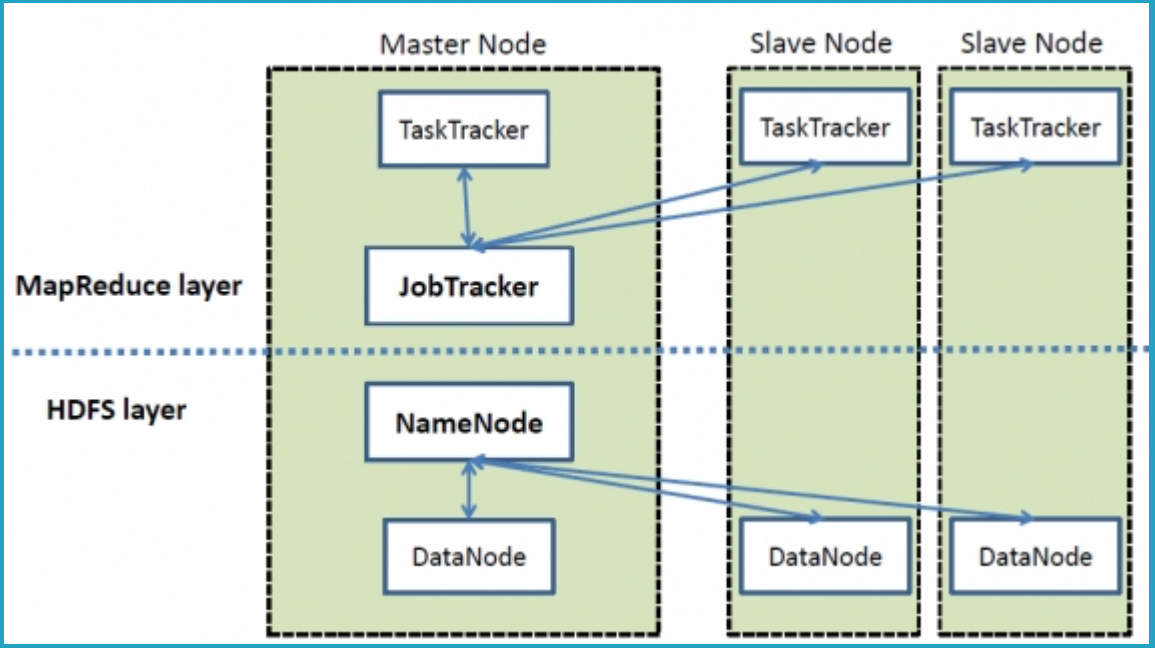
- The exponential growth of data first presented challenges to cutting-edge businesses such as Google, Yahoo, Amazon, and Microsoft
- Google publicize GFS, MapReduce
- Doug Cutting led the charge to develop an open source version of this MapReduce system called Hadoop
- Yahoo supported
-

Hadoop

- Hadoop is an open source framework for writing and running distributed applications that process large amounts of data
  - Hdfs - distributed storage
  - Mapreduce – distributed computation
- transfers code instead of data
- data replication
  -

Building blocks of Hadoop

- NameNode
- DataNode
- JobTracker
- TaskTracker
- Secondary NameNode



Setting up SSH for a Hadoop cluster

Define a common account

Verify SSH installation

[hadoop-user@master]$ which ssh

/usr/bin/ssh

[hadoop-user@master]$ which sshd

/usr/bin/sshd

[hadoop-user@master]$ which ssh-keygen

/usr/bin/ssh-keygen

Sudo apt-get install openssh-server
or
sudo dpkg -i openssh.deb

Setting up SSH for a Hadoop cluster

Generate SSH key pair

[hadoop-user@master]$ ssh-keygen -t rsa

Generating public/private rsa key pair.

Enter file in which to save the key (/home/hadoop-user/.ssh/id_rsa):

Enter passphrase (empty for no passphrase):

Enter same passphrase again:

Your identification has been saved in /home/hadoop-user/.ssh/id_rsa.

Your public key has been saved in /home/hadoop-user/.ssh/id_rsa.pub.

Setting up SSH for a Hadoop cluster

Distribute public key and validate logins

[hadoop-user@master]$ scp ~/.ssh/id_rsa.pub hadoop-user@target:~/master_key

[hadoop-user@target]$mkdir ~/.ssh

[hadoop-user@target]$chmod 700 ~/.ssh

[hadoop-user@target]$mv ~/master_key ~/.ssh/authorized_keys

[hadoop-user@target]$chmod 600 ~/.ssh/authorized_keys

[locally :: cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys]

[hadoop-user@master]$ ssh target

Last login: Sun Jan 4 15:32:49 2009 from master

Running Hadoop

[hadoop-user@master]$gedit **.bashrc**

export JAVA_HOME = /opt/jdk1.7.0

export PATH = $PATH:$JAVA_HOME/bin

Running Hadoop

[hadoop-user@master]$ cd $HADOOP_HOME/conf

hadoop-env.sh

export JAVA_HOME=/usr/share/jdk

Running Hadoop

core-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
        <name>fs.default.name</name>
        <value>hdfs://localhost:9000</value>
</property>
<property>
        <name>hadoop.tmp.dir</name>
        <value>/opt/hadoop_tmp</value>
</property>
</configuration>
```

Running Hadoop

mapred-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
    <name>mapred.job.tracker</name>
    <value>localhost:9001</value>
</property>
</configuration>
```

Running Hadoop

hdfs-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
    <name>dfs.replication</name>
    <value>3</value>
</property>
</configuration>
```

Running Hadoop

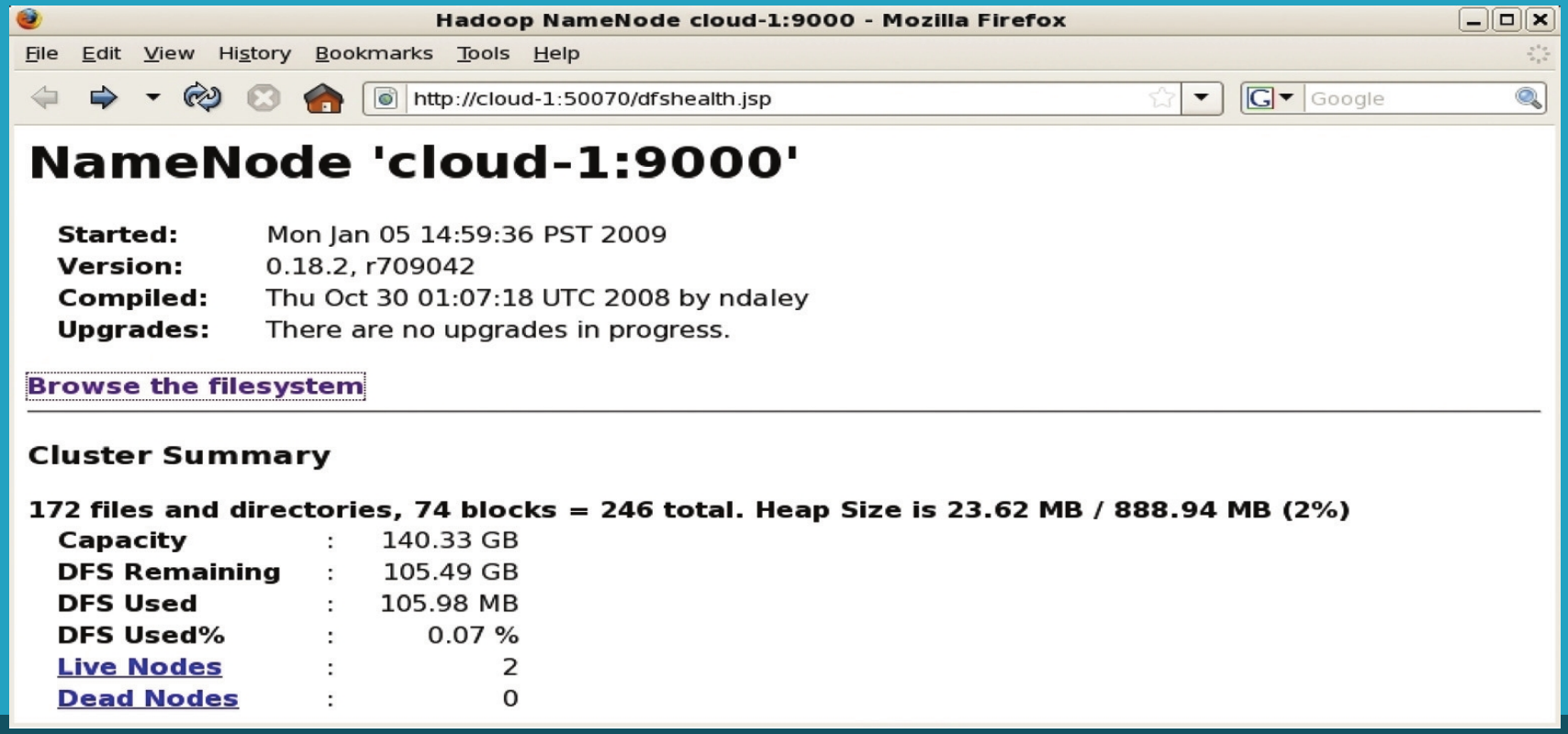
```
[hadoop-user@master]$ cat masters
localhost
[hadoop-user@master]$ cat slaves
localhost

[hadoop-user@master]$ bin/hadoop namenode -format
[hadoop-user@master]$ bin/start-all.sh
```

Running Hadoop

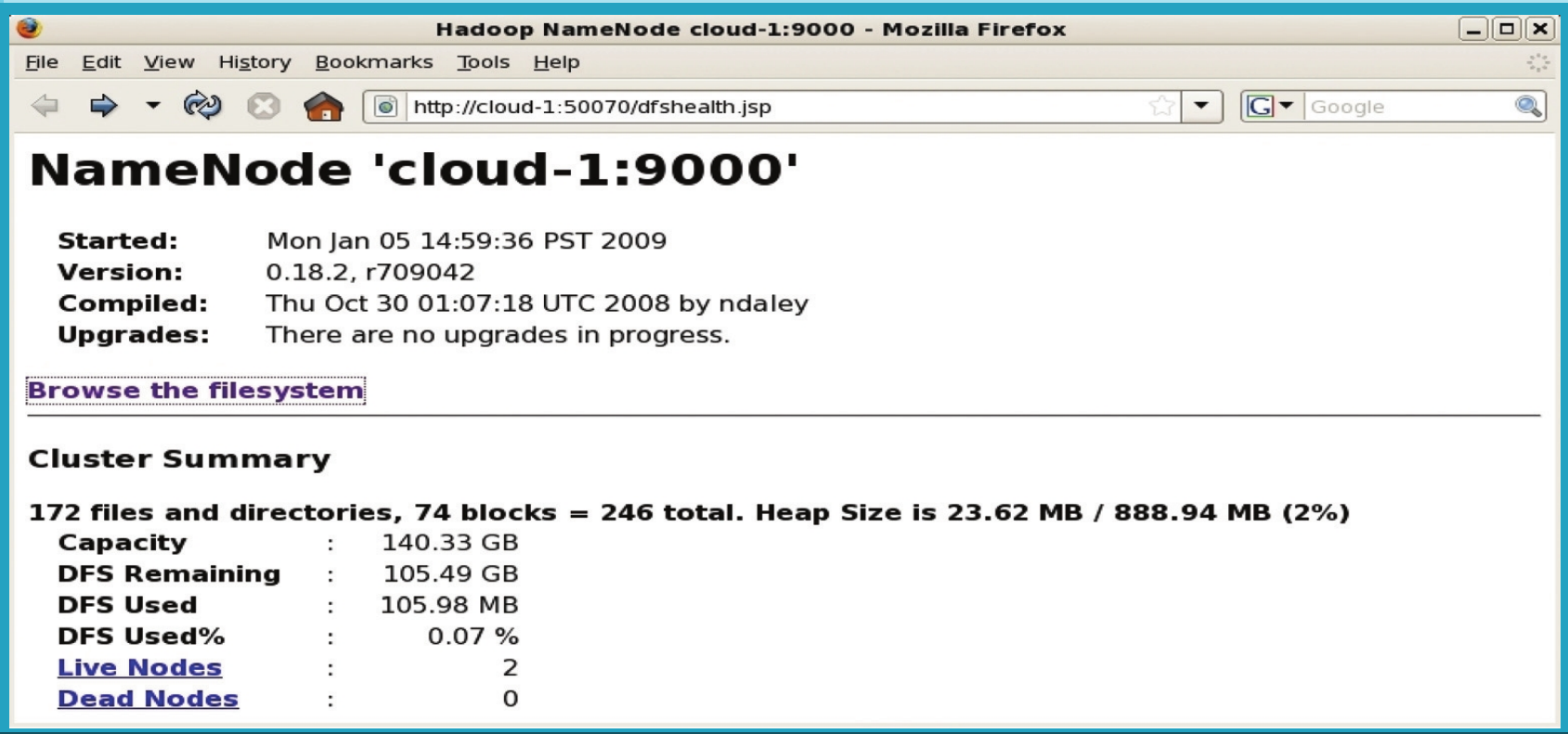
In file .bashrc

```
export HADOOP_HOME=/opt/programs/hadoop-0.20.2-cdh3u6
export PATH=$PATH:$HADOOP_HOME/bin
```

Web-based cluster UI



Web-based cluster UI



Working with files in HDFS

Basic file commands

hadoop fs -cmd <args>

hadoop fs –ls /

hadoop fs –mkdir /user/chuck

hadoop fs -put example.txt .

hadoop fs -put example.txt /user/chuck

hadoop fs -get example.txt .

Working with files in HDFS

```
hadoop fs -cat example.txt | head

hadoop fs –rm example.txt
hadoop fs –rmr /user/hdfs/dir1

hadoop fs -chmod 777 -R example.txt

hadoop fs -chown hdfs:hadoop example.txt
```

Working with files in HDFS

```
hadoop copyFromLocal  example.txt .
hadoop copyToLocal example.txt .

hadoop fs -getmerge  files/   mergedFile.txt

hadoop fs -cp /user/hadoop/file1 /user/hadoop/file2
hadoop fs -mv /user/hadoop/file1 /user/hadoop/file2
hadoop fs -du /user/hadoop/file1
```

References

- http://opensource.com/life/14/8/intro-apache-hadoop-big-data
- Hadoop In Action
- Hadoop : The definitive guide