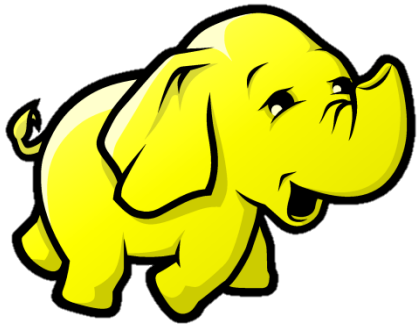
Diseño de Sistemas Distribuidos

Máster en Ciencia y Tecnología Informática
Curso 2018-2019

Sistemas escalables en entornos distribuidos. Introducción a Hadoop

Alejandro Calderón Mateos & Óscar Pérez Alonso
acaldero@inf.uc3m.es oscar@lab.inf.uc3m.es

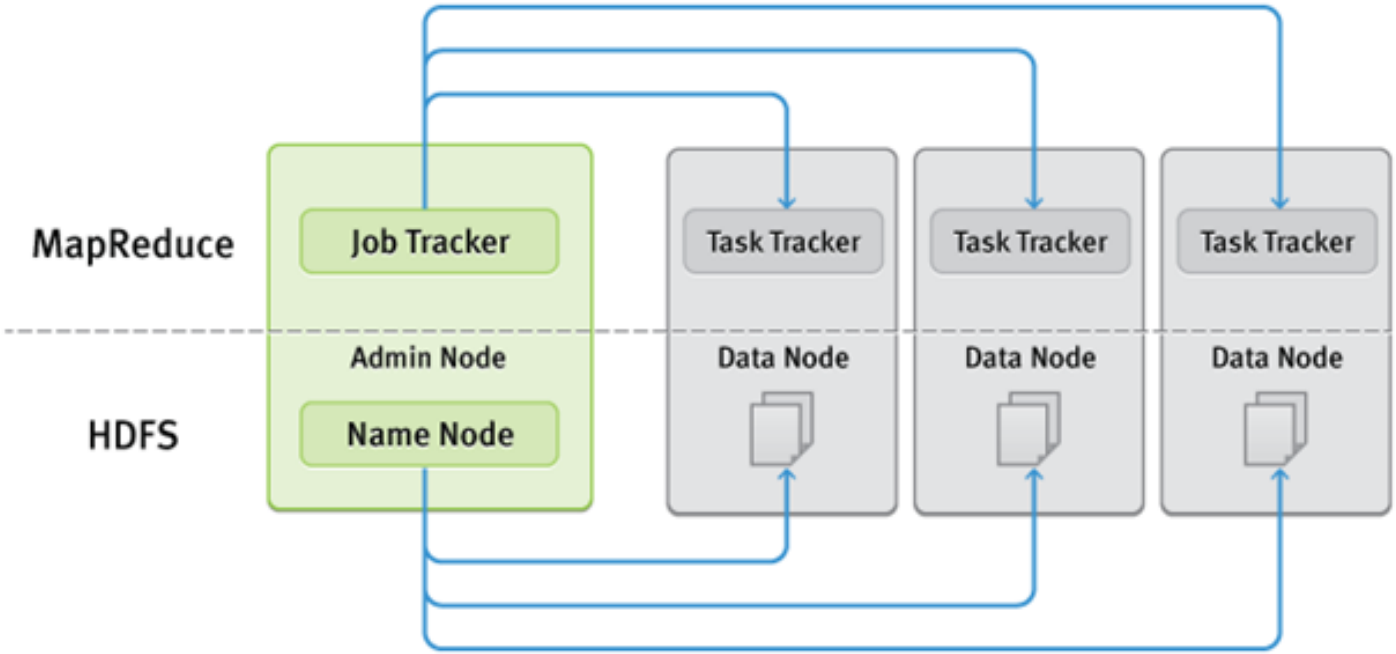
Contenidos



- **Introducción**
- *Hand-on*
- *Benchmarking*

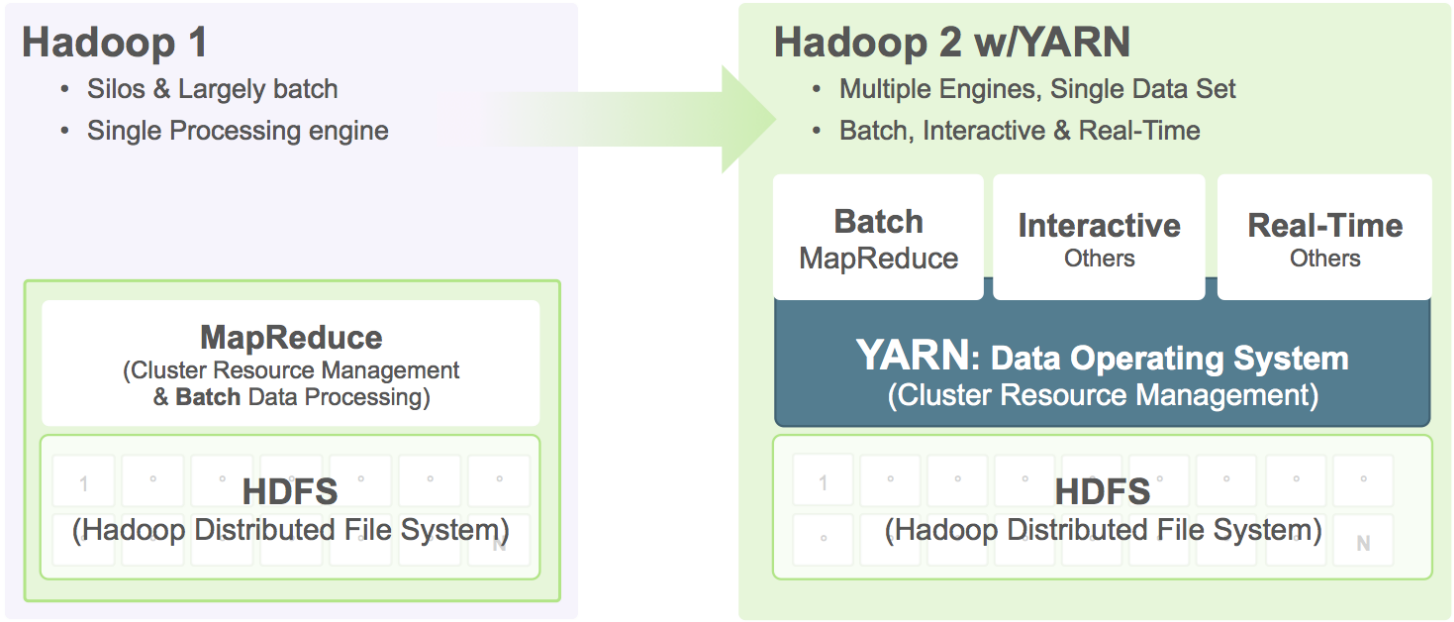
http://www.siliconweek.es/wp-content/uploads/2013/08/BigData-datos-guardar-almacenamiento-fichero-archivo.jpg
http://datameer2.datameer.com/blog/wp-content/uploads/2012/06/Hadoop-Ecosystem-Infographic-21.png

Arquitectura



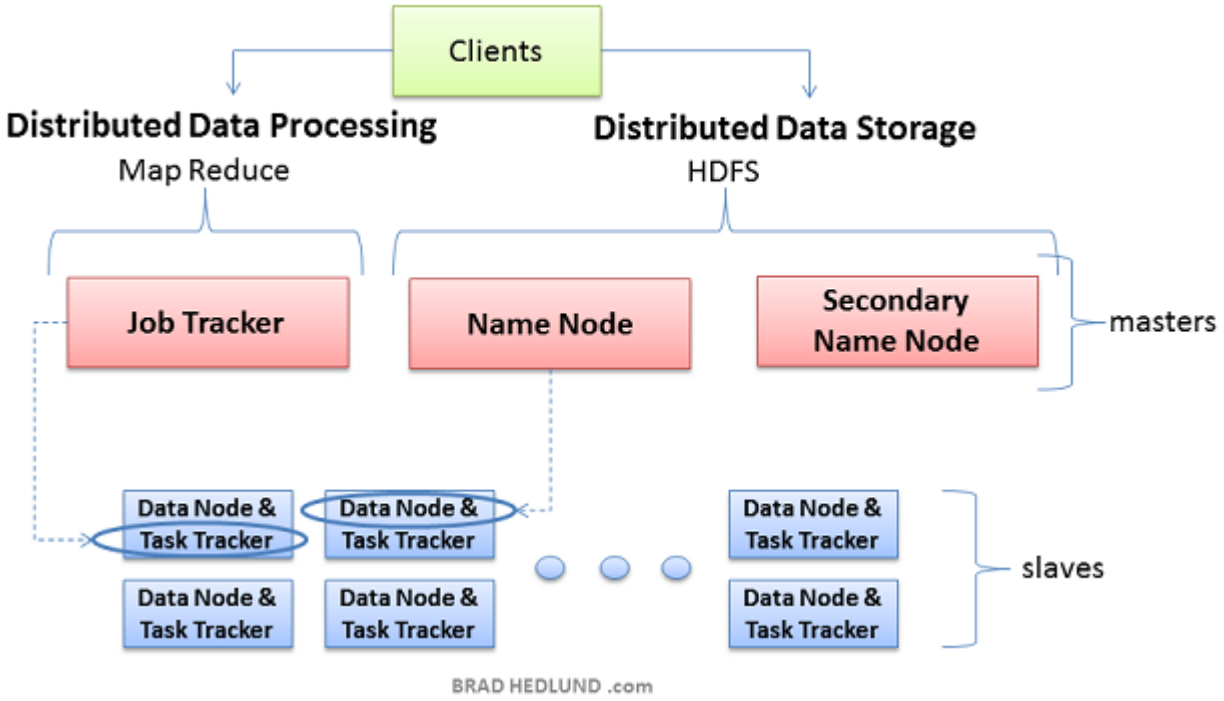
http://www.monitis.com/blog/2013/12/19/big-data-and-hadoop-whats-it-all-about/

Arquitectura



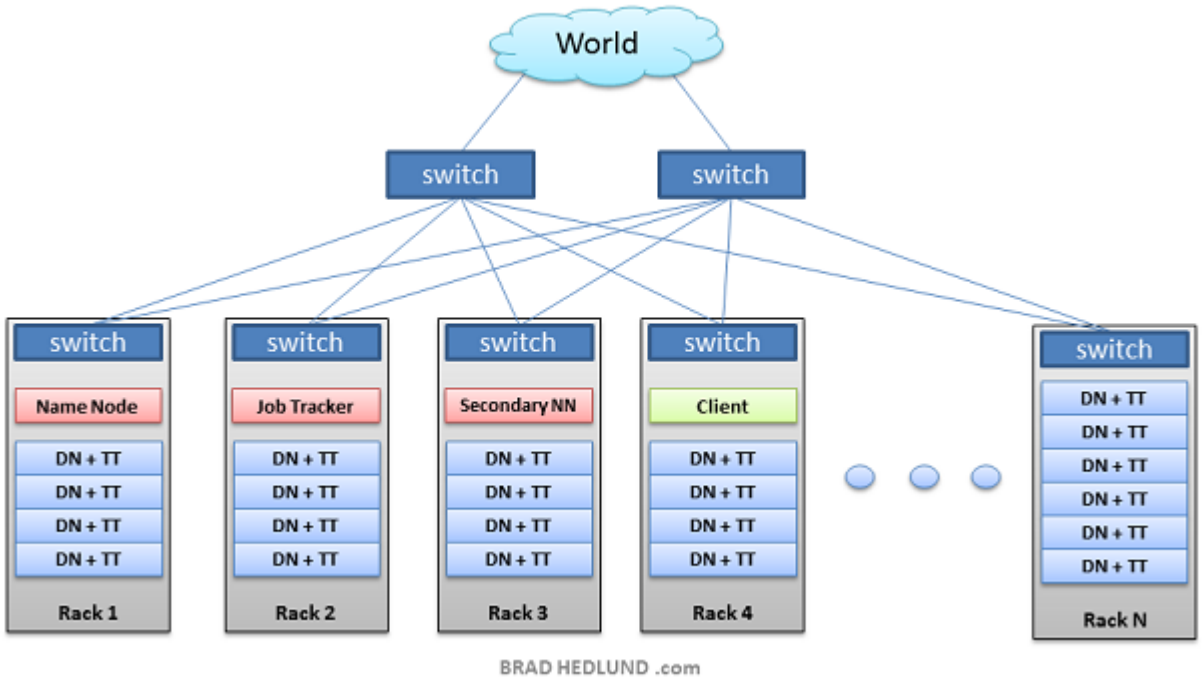
http://www.sachinpbuzz.com/2014/01/big-data-overview-of-apache-hadoop.html

Despliegue



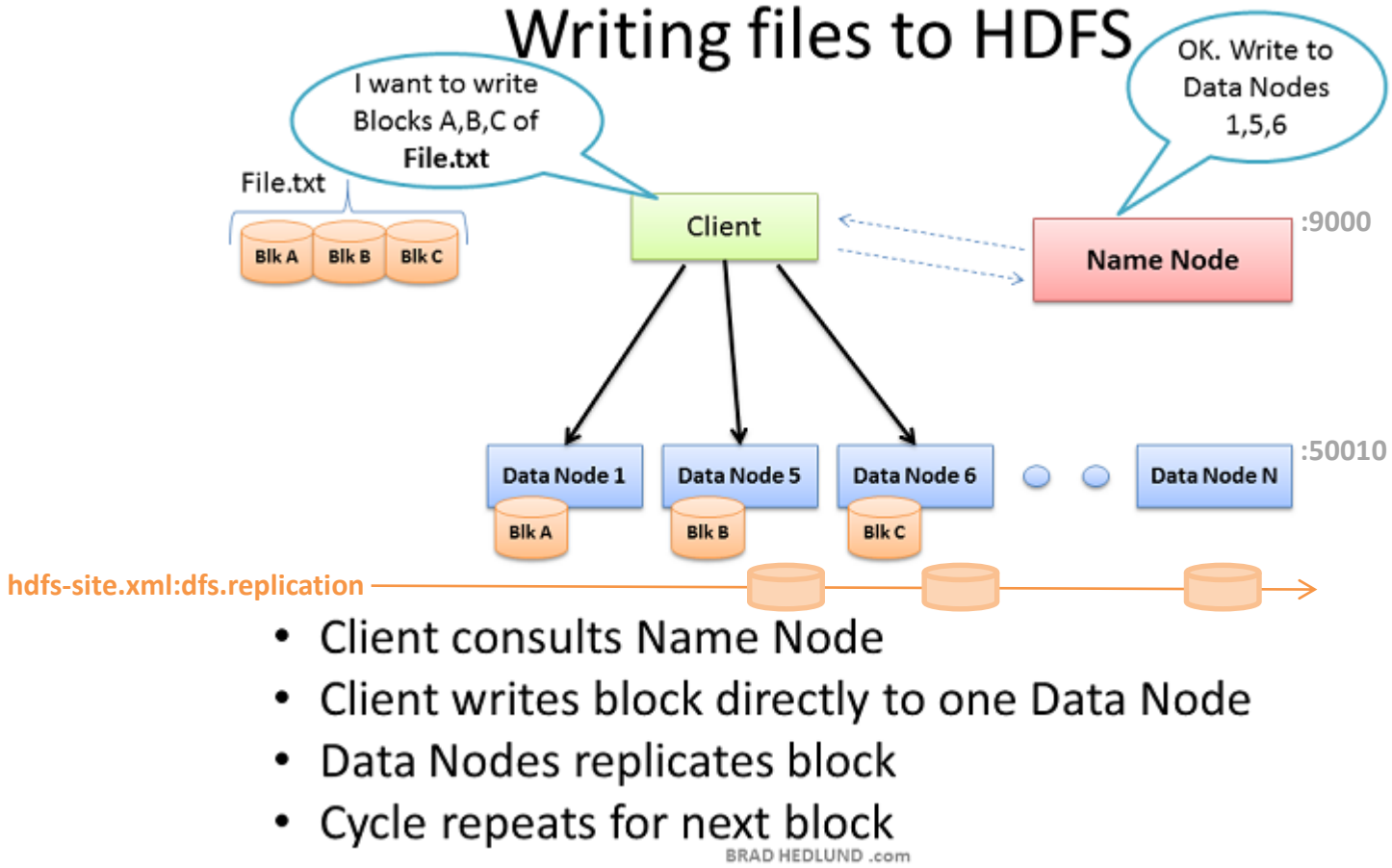
http://blog.csdn.net/suifeng3051/article/details/17288047

Despliegue



http://blog.csdn.net/suifeng3051/article/details/17288047

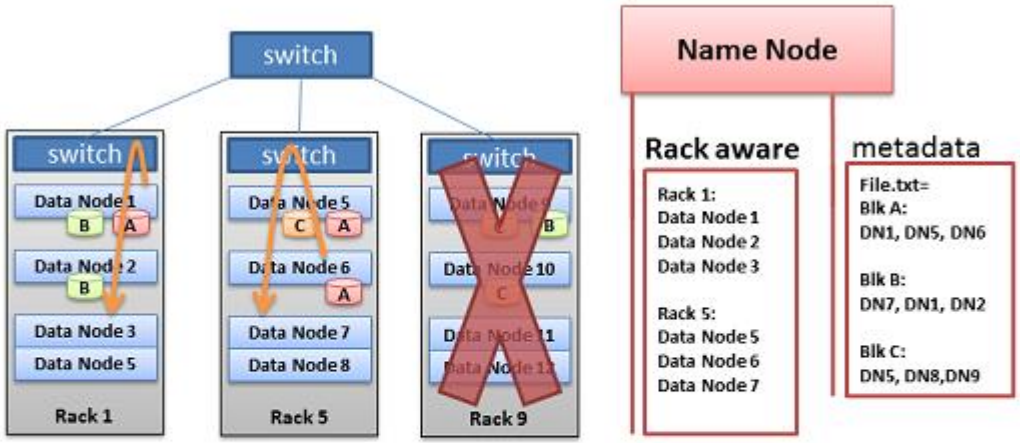
Despliegue



http://blog.csdn.net/suifeng3051/article/details/17288047

Despliegue

Hadoop Rack Awareness – Why?



- Never loose all data if entire rack fails
- Keep bulky flows in-rack when possible
- Assumption that in-rack is higher bandwidth, lower latency

BRAD HEDLUND .com

http://blog.csdn.net/suifeng3051/article/details/17288047

Despliegue

Typical Workflow

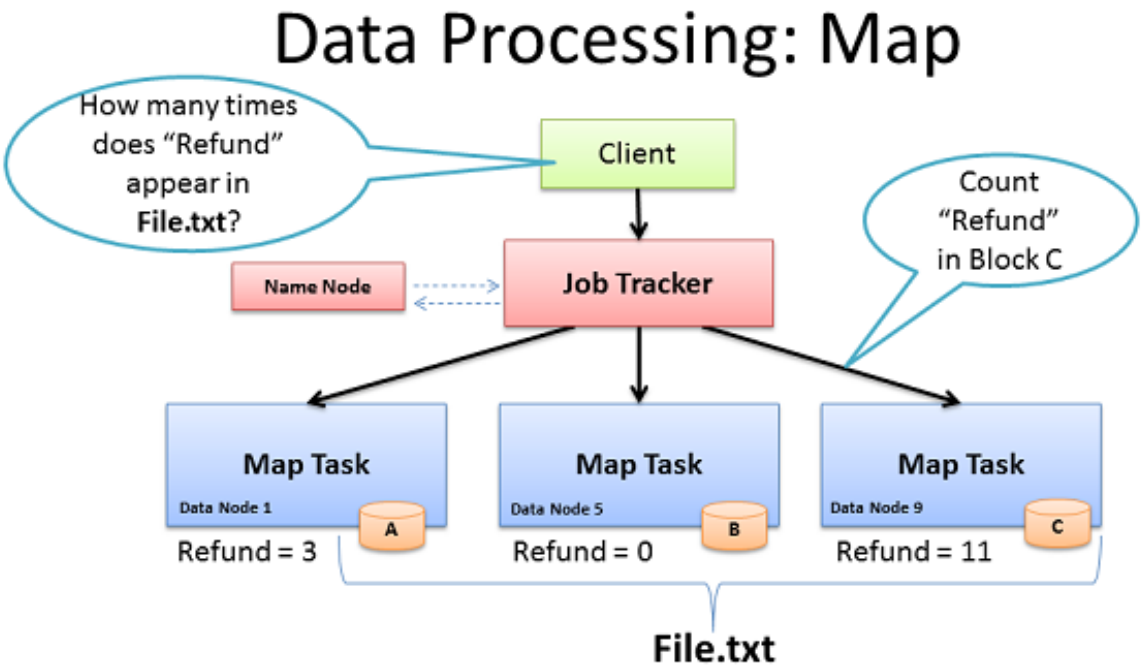
- Load data into the cluster (HDFS writes)
- Analyze the data (Map Reduce)
- Store results in the cluster (HDFS writes)
- Read the results from the cluster (HDFS reads)

Sample Scenario:

How many times did our customers type the word "**Refund**" into emails sent to customer service?

http://blog.csdn.net/suifeng3051/article/details/17288047

Despliegue



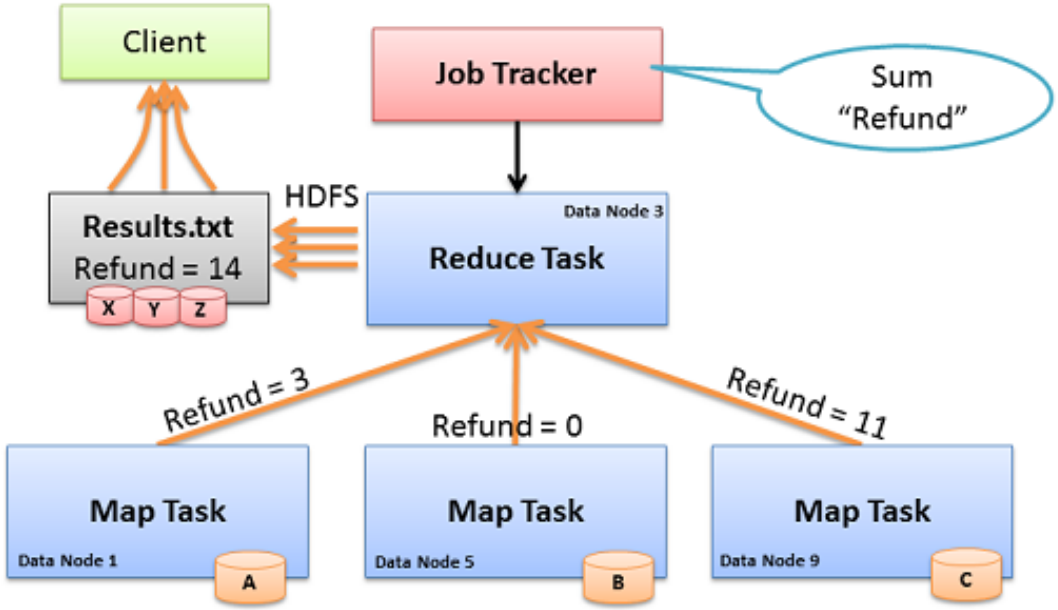
- **Map:** "Run this computation on your local data"
- Job Tracker delivers Java code to Nodes with local data

BRAD HEDLUND .com

http://blog.csdn.net/suifeng3051/article/details/17288047

Despliegue

Data Processing: Reduce

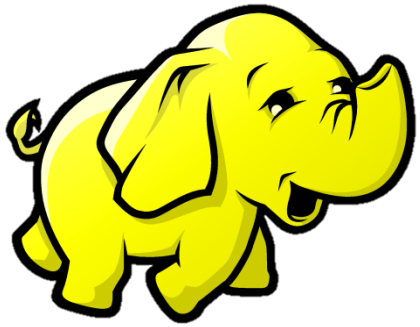


- **Reduce:** "Run this computation across Map results"
- Map Tasks send output data to Reducer over the network
- Reduce Task data output written to and read from HDFS

BRAD HEDLUND .com

http://blog.csdn.net/suifeng3051/article/details/17288047

Contenidos



- Introducción
- ***Hand-on***
- *Benchmarking*

http://www.siliconweek.es/wp-content/uploads/2013/08/BigData-datos-guardar-almacenamiento-fichero-archivo.jpg
http://datameer2.datameer.com/blog/wp-content/uploads/2012/06/Hadoop-Ecosystem-Infographic-21.png

Hadoop: solo un nodo

```
alejandro@h1:~$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.
```

```
alejandro@h1:~$ sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
    Full Name []:
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n]
```

http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php

Hadoop: solo un nodo

Prerequisitos Instalación Uso básico

```
alejandro@h1:~$ sudo apt-get install ssh rsync
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  rsync ssh
…
```

```
alejandro@h1:~$ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libx11-doc
  libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-7-jdk
…
```

http://wiki.apache.org/hadoop/HadoopJavaVersions

Hadoop: solo un nodo

```
alejandro@h1:~$ su hduser
Password:
```

```
hduser@h1:/home/alejandro$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
…
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
…
The key's randomart image is:
+--[ RSA 2048]----+
|         =+B+o.  |
|        ..B.o+.  |
…
```

Hadoop: solo un nodo

Prerequisitos Instalación Uso básico

```
hduser@h1:/home/alejandro$ cat $HOME/.ssh/id_rsa.pub >>
    $HOME/.ssh/authorized_keys
```

```
hduser@h1:/home/alejandro$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is eb:51:89:99:49:42:6a:6e:78:5d:79:6c:69:2a:8c:45.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known
    hosts.
Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-36-generic x86_64)
…
```

```
hduser@h1:~$ exit
logout
```

```
hduser@h1:/home/alejandro$ exit
exit
```

Hadoop: solo un nodo

Prerequisitos **Instalación** Uso básico

```
alejandro@h1:~$ wget http://apache.rediris.es/hadoop/common/current/hadoop-2.5.2.tar.gz
…
2014-09-26 21:57:25 (1,12 MB/s) - ‘hadoop-2.5.2.tar.gz’ saved [138656756/138656756]
```

```
alejandro@h1:~$ tar zxf hadoop-2.5.2.tar.gz
```

```
alejandro@h1:~$ ls -las hadoop-2.5.2
total 60
 4 drwxr-xr-x  9 alejandro alejandro  4096 jun 21 08:38 .
 4 drwxr-xr-x 16 alejandro alejandro  4096 sep 27 21:58 ..
 4 drwxr-xr-x  2 alejandro alejandro  4096 jun 21 08:05 bin
 4 drwxr-xr-x  3 alejandro alejandro  4096 jun 21 08:05 etc
 4 drwxr-xr-x  2 alejandro alejandro  4096 jun 21 08:05 include
 4 drwxr-xr-x  3 alejandro alejandro  4096 jun 21 08:05 lib
 4 drwxr-xr-x  2 alejandro alejandro  4096 jun 21 08:05 libexec
16 -rw-r--r--  1 alejandro alejandro 15458 jun 21 08:38 LICENSE.txt
 4 -rw-r--r--  1 alejandro alejandro   101 jun 21 08:38 NOTICE.txt
 4 -rw-r--r--  1 alejandro alejandro  1366 jun 21 08:38 README.txt
 4 drwxr-xr-x  2 alejandro alejandro  4096 jun 21 08:05 sbin
 4 drwxr-xr-x  4 alejandro alejandro  4096 jun 21 08:05 share
```

Hadoop: solo un nodo

Prerequisitos **Instalación** Uso básico

```
alejandro@h1:~$ sudo mv hadoop-2.5.2 /usr/local/hadoop
```

```
alejandro@h1:~$ sudo chown -R hduser:hadoop /usr/local/hadoop/
```

```
alejandro@h1:~$ cd /usr/local/hadoop/
```

```
alejandro@h1:/usr/local/hadoop$ ls -las
total 60
 4 drwxr-xr-x  9 hduser hadoop  4096 jun 21 08:38 .
 4 drwxr-xr-x 11 root   root    4096 sep 27 22:02 ..
 4 drwxr-xr-x  2 hduser hadoop  4096 jun 21 08:05 bin
 4 drwxr-xr-x  3 hduser hadoop  4096 jun 21 08:05 etc
 4 drwxr-xr-x  2 hduser hadoop  4096 jun 21 08:05 include
 4 drwxr-xr-x  3 hduser hadoop  4096 jun 21 08:05 lib
 4 drwxr-xr-x  2 hduser hadoop  4096 jun 21 08:05 libexec
16 -rw-r--r--  1 hduser hadoop 15458 jun 21 08:38 LICENSE.txt
 4 -rw-r--r--  1 hduser hadoop   101 jun 21 08:38 NOTICE.txt
 4 -rw-r--r--  1 hduser hadoop  1366 jun 21 08:38 README.txt
 4 drwxr-xr-x  2 hduser hadoop  4096 jun 21 08:05 sbin
 4 drwxr-xr-x  4 hduser hadoop  4096 jun 21 08:05 share
```

Hadoop: solo un nodo

- Configurar variables de entorno:
  - Encontrar los componentes de Hadoop
    - ~/.bashrc
  - Encontrar en Hadoop a JAVA_HOME
    - /usr/local/hadoop/etc/hadoop/hadoop-env.sh
- Configurar los componentes de Hadoop:
  - Configurar hadoop.tmp.dir y fs.default.name
    - /usr/local/hadoop/etc/hadoop/core-site.xml
  - Configurar qué framework usar para mapreduce
    - /usr/local/hadoop/etc/hadoop/mapred-site.xml
  - Configuración de los directorios para namenode y datanode
    - /usr/local/hadoop/etc/hadoop/hdfs-site.xml

Hadoop: solo un nodo

```
alejandro@h1:/usr/local/hadoop$ su hduser
Password:
```

```
hduser@h1:/usr/local/hadoop$ update-alternatives --config java
There is only one alternative in link group java (providing /usr/bin/java):
    /usr/lib/jvm/java-7-openjdk-amd64/jre/bin/java
Nothing to configure.
```

```
hduser@h1:/usr/local/hadoop$ cat >> ~/.bashrc
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib“
```

Hadoop: solo un nodo

```
hduser@h1:/usr/local/hadoop$ grep JAVA_HOME /usr/local/hadoop/etc/hadoop/hadoop-env.sh
…
export JAVA_HOME=${JAVA_HOME}
```

```
hduser@h1:/usr/local/hadoop$ cat >> /usr/local/hadoop/etc/hadoop/hadoop-env.sh
export JAVA_HOME="/usr/lib/jvm/java-7-openjdk-amd64"
```

Hadoop: solo un nodo

- Configurar variables de entorno:
  - Encontrar los componentes de Hadoop
    - ~/.bashrc
  - Encontrar en Hadoop a JAVA_HOME
    - /usr/local/hadoop/etc/hadoop/hadoop-env.sh
- Configurar los componentes de Hadoop:
  - Configurar hadoop.tmp.dir y fs.default.name
    - /usr/local/hadoop/etc/hadoop/core-site.xml
  - Configurar qué framework usar para mapreduce
    - /usr/local/hadoop/etc/hadoop/mapred-site.xml
  - Configuración de los directorios para namenode y datanode
    - /usr/local/hadoop/etc/hadoop/hdfs-site.xml

Hadoop: solo un nodo

Prerequisitos **Instalación** Uso básico

```
alejandro@h1:~$ sudo mkdir -p /hadoop/tmp ;
                sudo chown hduser:hadoop /hadoop/tmp/
```

```
hduser@h1:/usr/local/hadoop$ cat > /usr/local/hadoop/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
 <property>
  <name>hadoop.tmp.dir</name>
  <value>/hadoop/tmp</value>
  <description>A base for other temporary directories.</description>
 </property>
 <property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
  <description>The name of the default file system.</description>
 </property>
</configuration>
```

Hadoop: solo un nodo

Prerequisitos **Instalación** Uso básico

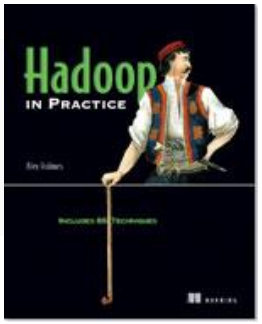
```
hduser@h1:/usr/local/hadoop$ cat > /usr/local/hadoop/etc/hadoop/mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
 <property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>The host and port that the MapReduce job tracker runs at.
  </description>
 </property>
</configuration>
```

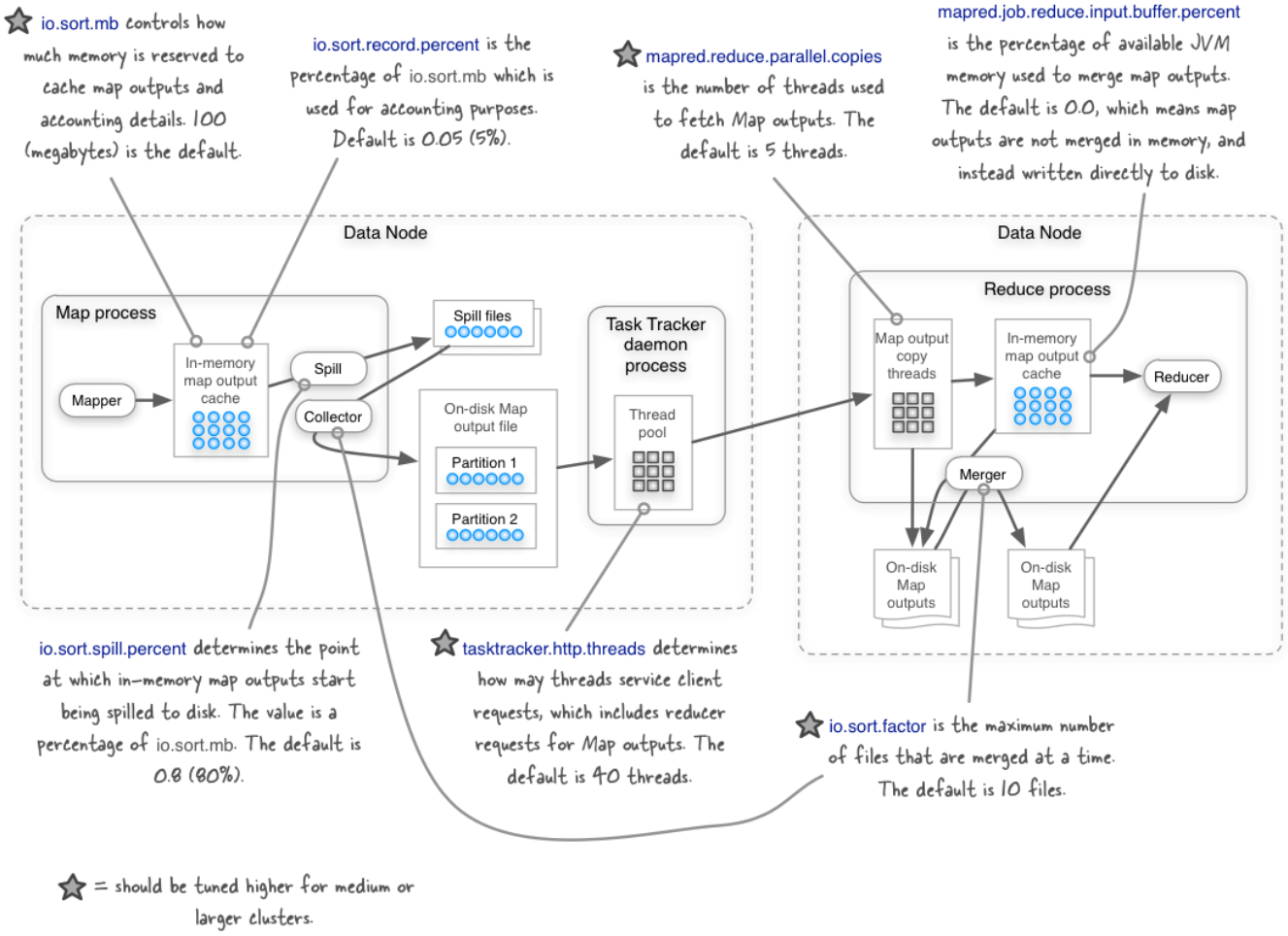
Hadoop: solo un nodo

```
alejandro@h1:~$ sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode ;
                sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode ;
                sudo chown -R hduser:hadoop /usr/local/hadoop_store
```

```
hduser@h1:/usr/local/hadoop$ cat > /usr/local/hadoop/etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
 <property><name>dfs.replication</name>
  <value>1</value>
 </property>
 <property><name>dfs.namenode.name.dir</name>
   <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
 </property>
 <property><name>dfs.datanode.data.dir</name>
   <value>file:/usr/local/hadoop_store/hdfs/datanode</value>
 </property>
</configuration>
```

Posible configuración adicional…

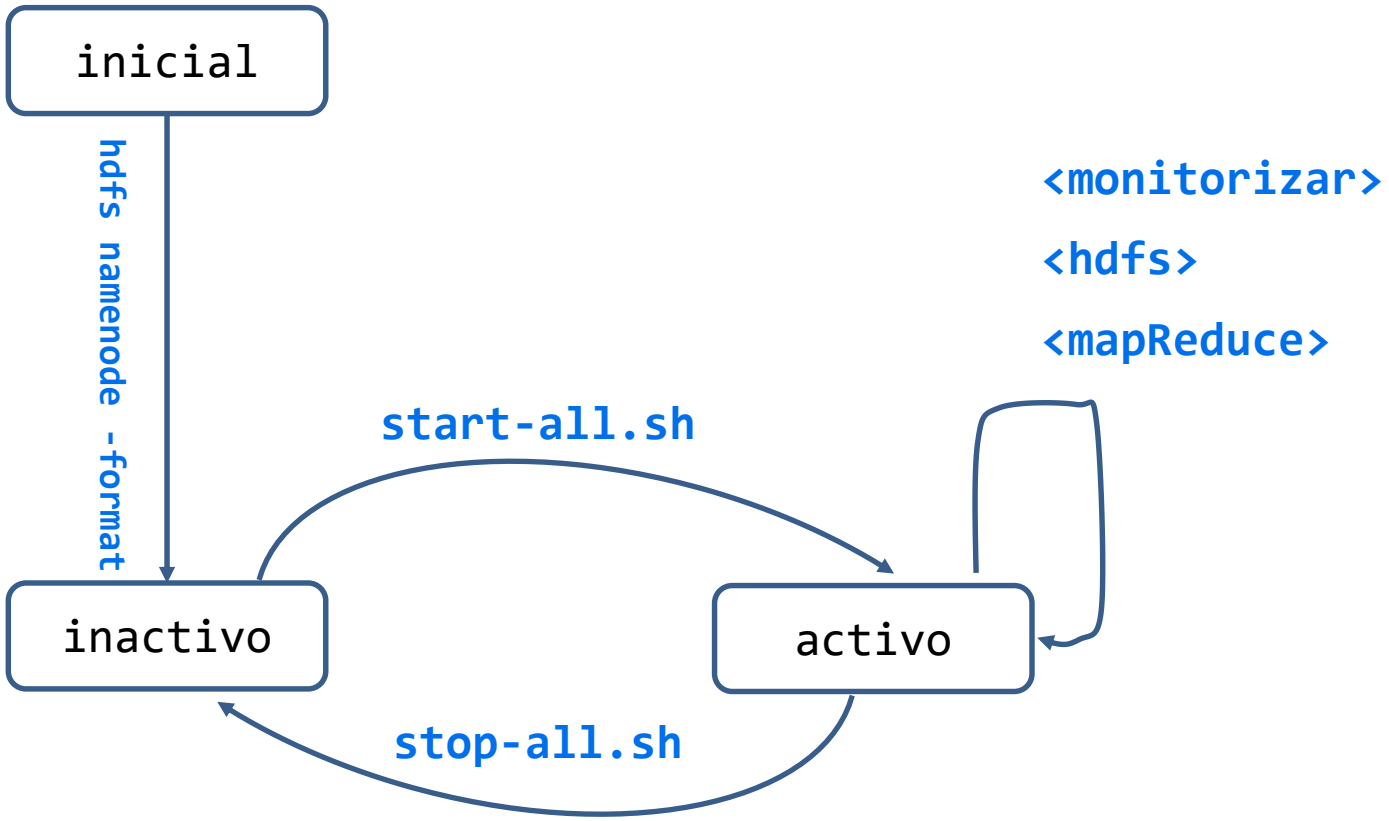




http://grepalex.com/2012/11/26/hadoop-shuffle-configurables/

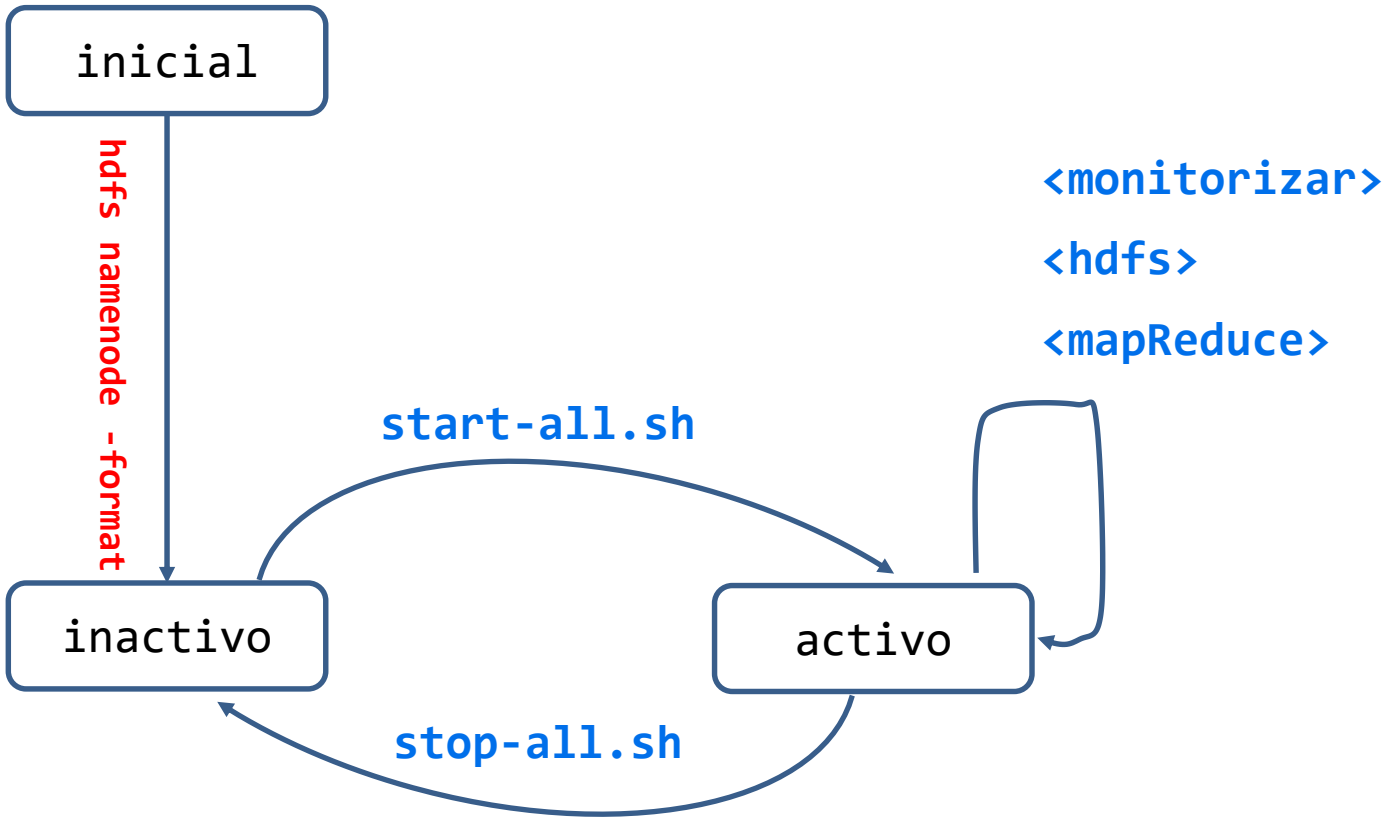
Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



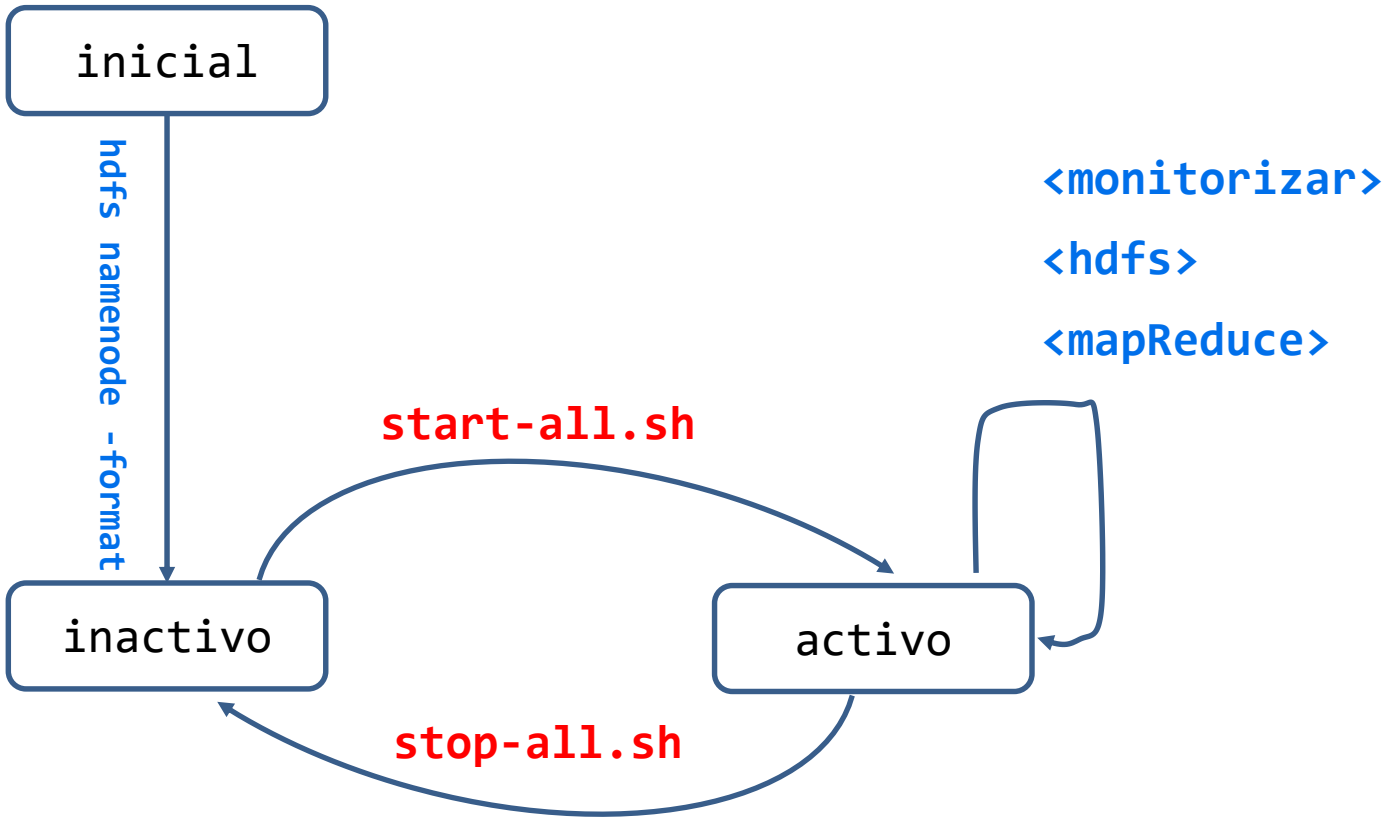
Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ hdfs namenode -format
14/09/25 23:02:59 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = h1/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.5.2
…
14/09/27 23:07:07 INFO blockmanagement.BlockManager: encryptDataTransfer = false
14/09/27 23:07:07 INFO namenode.FSNamesystem: fsOwner                    = hduser (auth:SIMPLE)
…
14/09/25 23:03:04 INFO util.ExitUtil: Exiting with status 0
14/09/25 23:03:04 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at h1/127.0.1.1
************************************************************/
```

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
14/09/28 13:31:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-h1.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-h1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-h1.out
14/09/28 13:32:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-h1.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-h1.out
```

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ jps
28026 ResourceManager
28147 NodeManager
27877 SecondaryNameNode
27564 NameNode
28448 Jps
27683 DataNode
```

```
hduser@h1:~$ nmap localhost
…
PORT     STATE SERVICE
22/tcp   open  ssh
631/tcp  open  ipp
8031/tcp open  unknown
8042/tcp open  fs-agent
8088/tcp open  radan-http
```

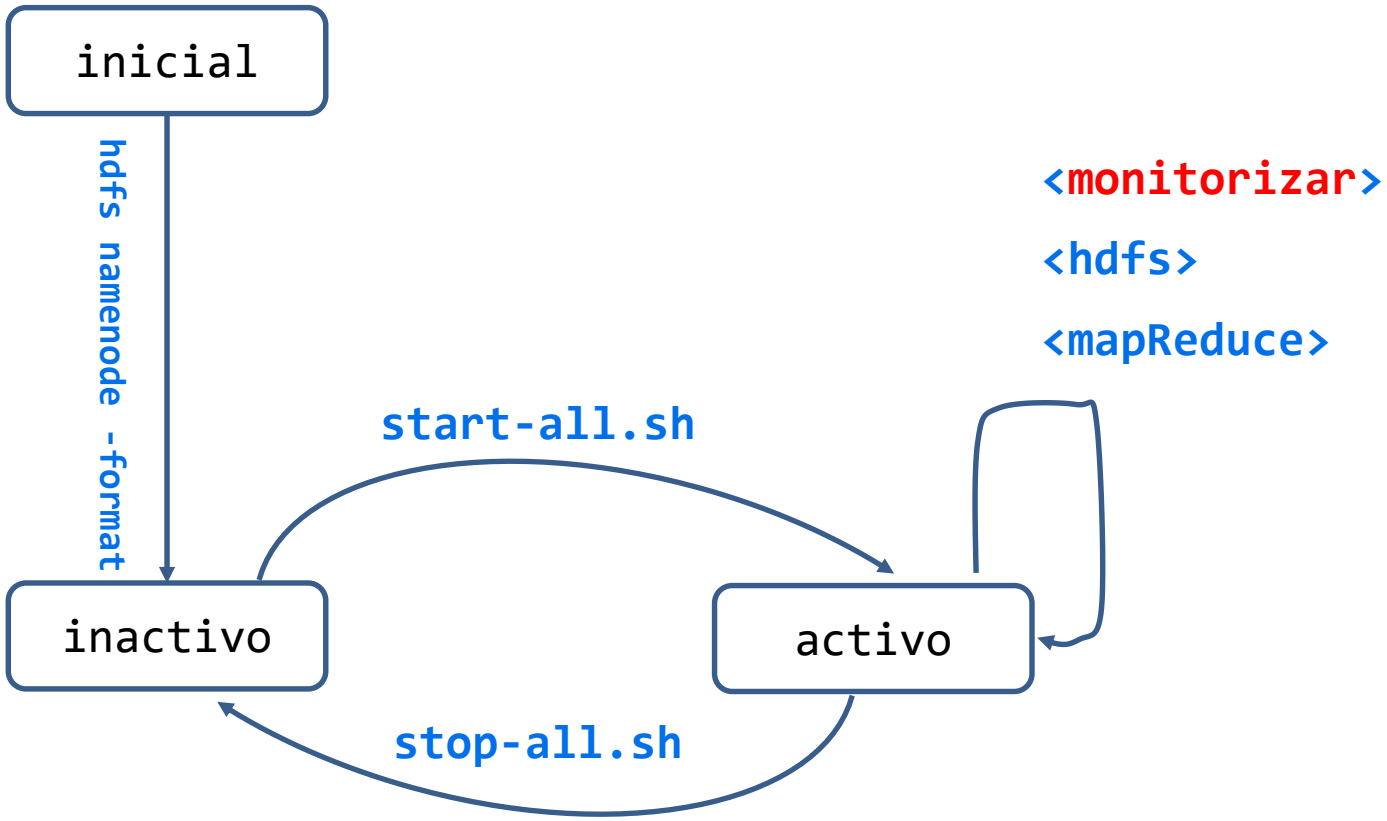
Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
14/09/28 13:33:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
14/09/28 13:33:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
```

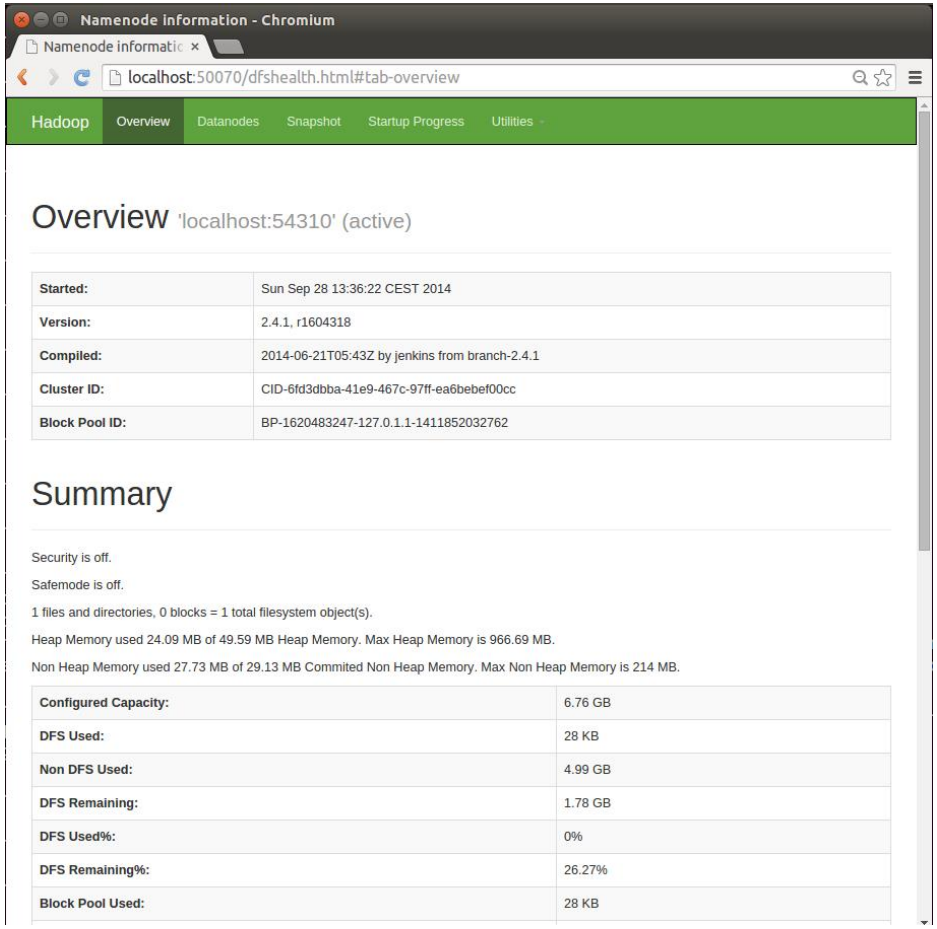
Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



Hadoop: solo un nodo

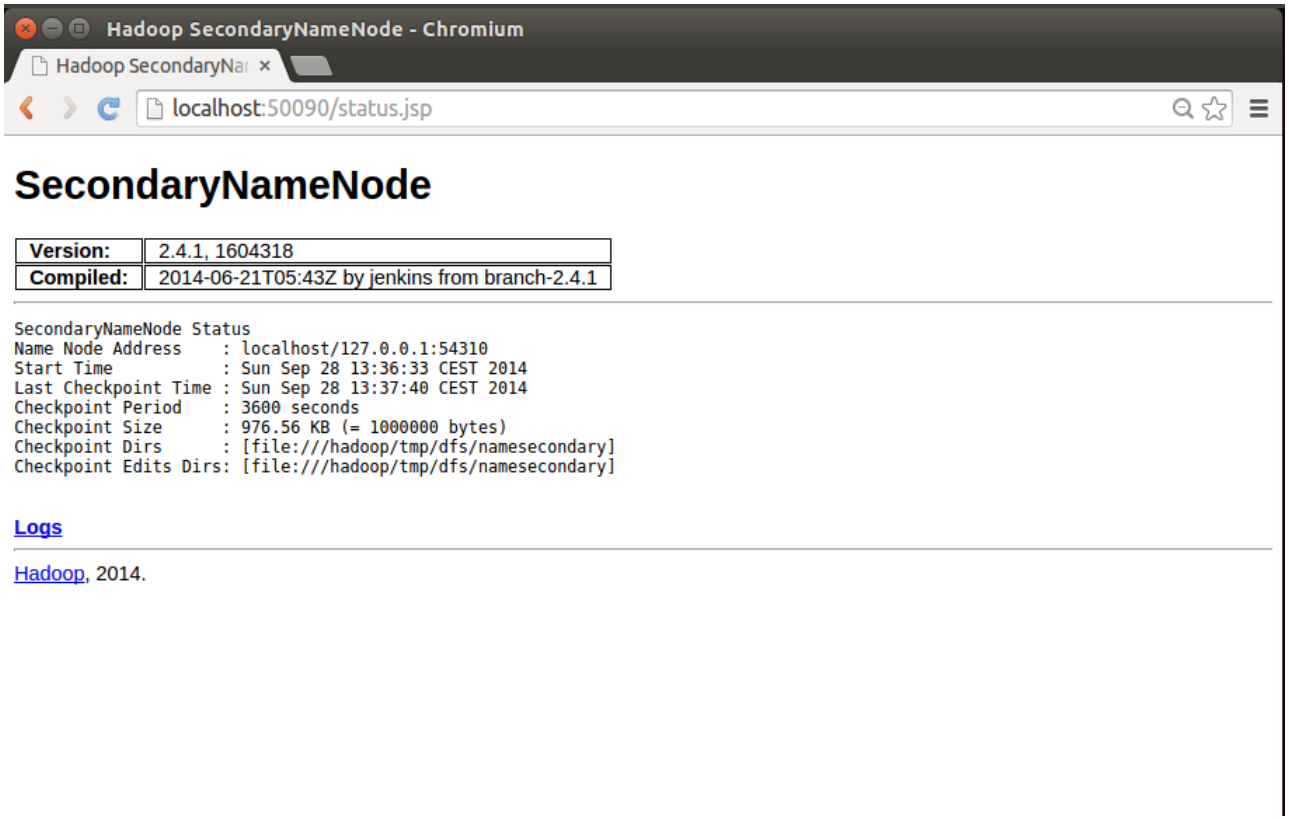
Prerequisitos Instalación **Uso básico**



- NameNode: http://localhost:50070/

Hadoop: solo un nodo

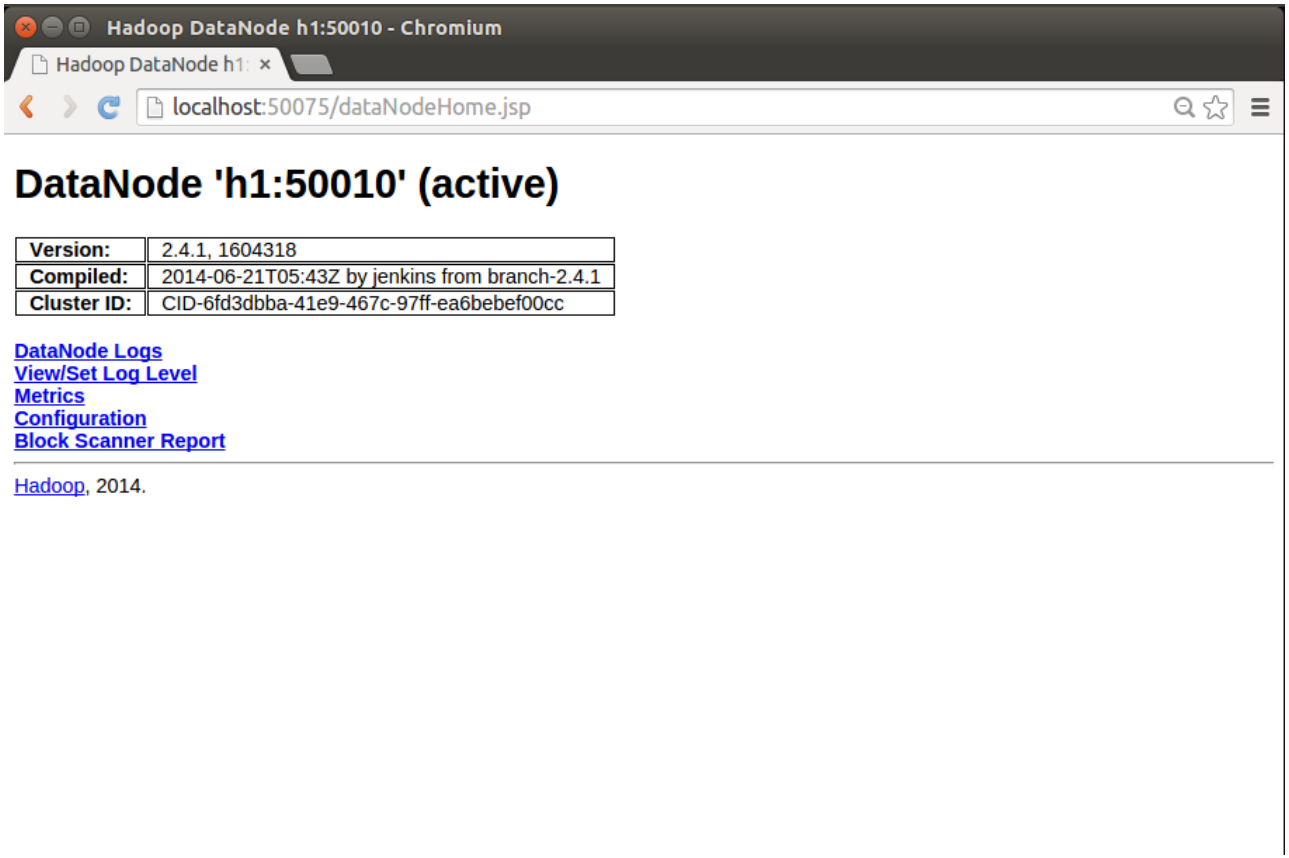
Prerequisitos Instalación **Uso básico**



- SecondaryNameNode: http://localhost:50090/

Hadoop: solo un nodo

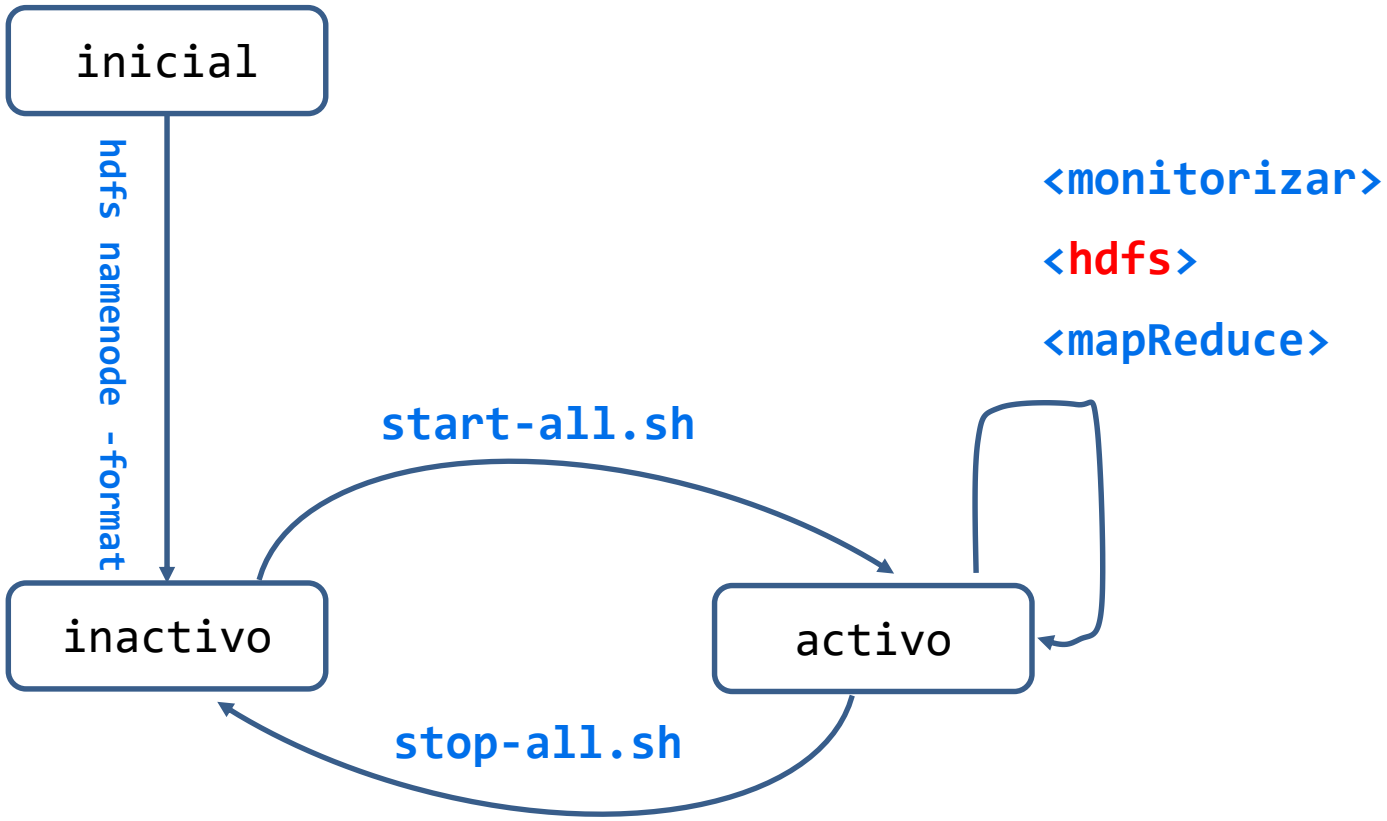
Prerequisitos Instalación **Uso básico**



- DataNode: http://localhost:50075/

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



Hadoop: solo un nodo


```
: crear un directorio
hduser@h1:~$ hadoop fs -mkdir -p /user/hduser

: copiar un fichero de local a hadoop
hduser@h1:~$ echo "hdfs test" > hdfsTest.txt
hduser@h1:~$ hadoop fs -copyFromLocal hdfsTest.txt hdfsTest.txt

: ver contenido de un directorio
hduser@h1:~$ hadoop fs -ls

: ver contenido de un archivo
hduser@h1:~$ hadoop fs -cat /user/hduser/hdfsTest.txt

: copiar un fichero de hadoop a local
hduser@h1:~$ hadoop fs -copyToLocal /user/hduser/hdfsTest.txt hdfsTest2.txt

: borrar un fichero
hduser@h1:~$ hadoop fs -rm hdfsTest.txt
```

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ wget http://www.gutenberg.org/files/2000/old/2donq10.txt
…
2014-10-04 12:53:30 (1,10 MB/s) - ‘2donq10.txt’ saved [2143292/2143292]
```

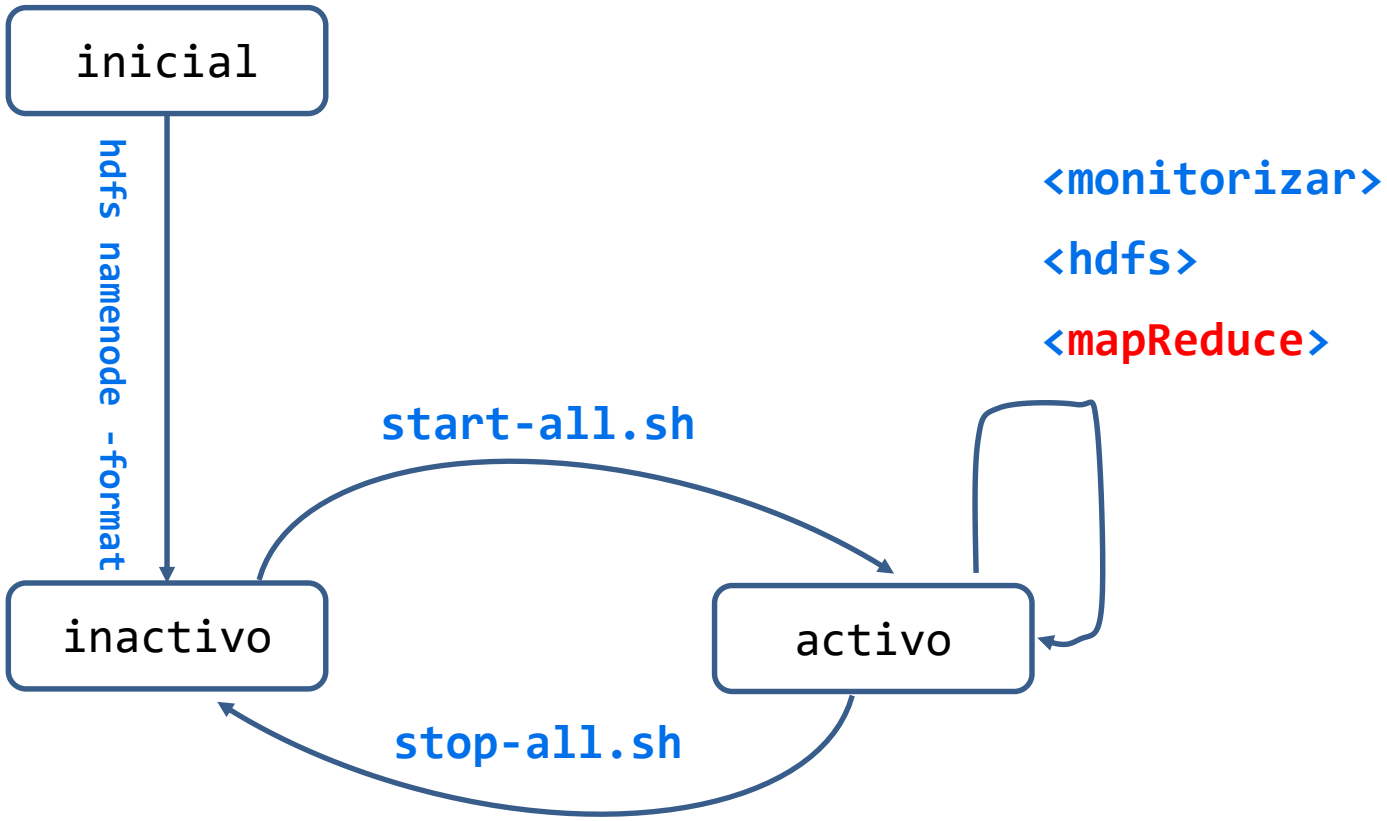
```
hduser@h1:~$ dos2unix -n 2donq10.txt dq.txt
dos2unix: converting file 2donq10.txt to file dq.txt in Unix format ...
```

```
hduser@h1:~$ hadoop fs –copyFromLocal -f dq.txt /user/hduser/dq.txt
```

```
hduser@h1:~$ hadoop fs -ls /user/hduser
…
Found 1 items
-rw-r--r--   3 hduser supergroup    2143292 2014-10-04 13:09 /user/hduser/dq.txt
```

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**



Hadoop: solo un nodo

Nativo

Java

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-
    mapreduce-examples-2.5.2.jar pi 2 5
Number of Maps  = 2
Samples per Map = 5
…
Job Finished in 11.536 seconds
Estimated value of Pi is 3.60000000000000000000
```

http://www.bogotobogo.com/Hadoop/BigData_hadoop_Running_MapReduce_Job.php

Hadoop: solo un nodo

1

```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
```

http://wiki.apache.org/hadoop/WordCount

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

2

```
public class WordCount {

 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map (LongWritable key, Text value, Context context)
    throws IOException, InterruptedException
    {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
 }
```

http://wiki.apache.org/hadoop/WordCount

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

3

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

   public void reduce (Text key, Iterable<IntWritable> values, Context context)
   throws IOException, InterruptedException
   {
       int sum = 0;
       for (IntWritable val : values) {
           sum += val.get();
       }
       context.write(key, new IntWritable(sum));
   }
}
```

http://wiki.apache.org/hadoop/WordCount

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

4

```
  public static void main (String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    job.waitForCompletion(true);
  }
} // class WordCount
```

http://wiki.apache.org/hadoop/WordCount

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:/usr/local/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-
    mapreduce-examples-*.jar wordcount /user/hduser/dq.txt
    /user/hduser/counterj
14/10/04 16:33:36 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
14/10/04 16:33:37 INFO input.FileInputFormat: Total input paths to process : 1
14/10/04 16:33:37 INFO mapreduce.JobSubmitter: number of splits:1
14/10/04 16:33:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local835374884_0001
…
     File Input Format Counters
             Bytes Read=2106143
     File Output Format Counters
             Bytes Written=454722
```

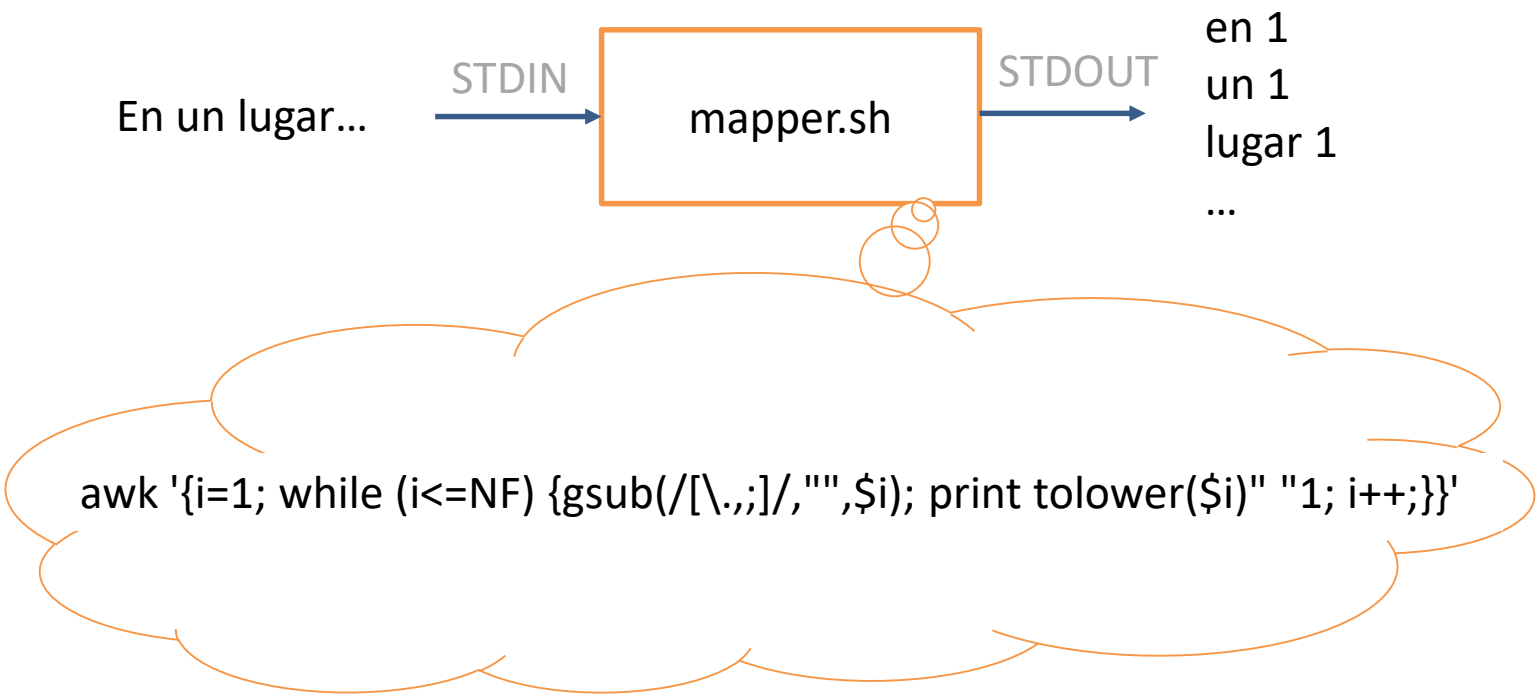
```
hduser@h1:/usr/local/hadoop$ hadoop fs -cat /user/hduser/counterj/* | sort
    -n -k 2 -r|head -5
…
que  19429
de   17986
y    15887
la   10199
a    9502
```

http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

Hadoop: solo un nodo

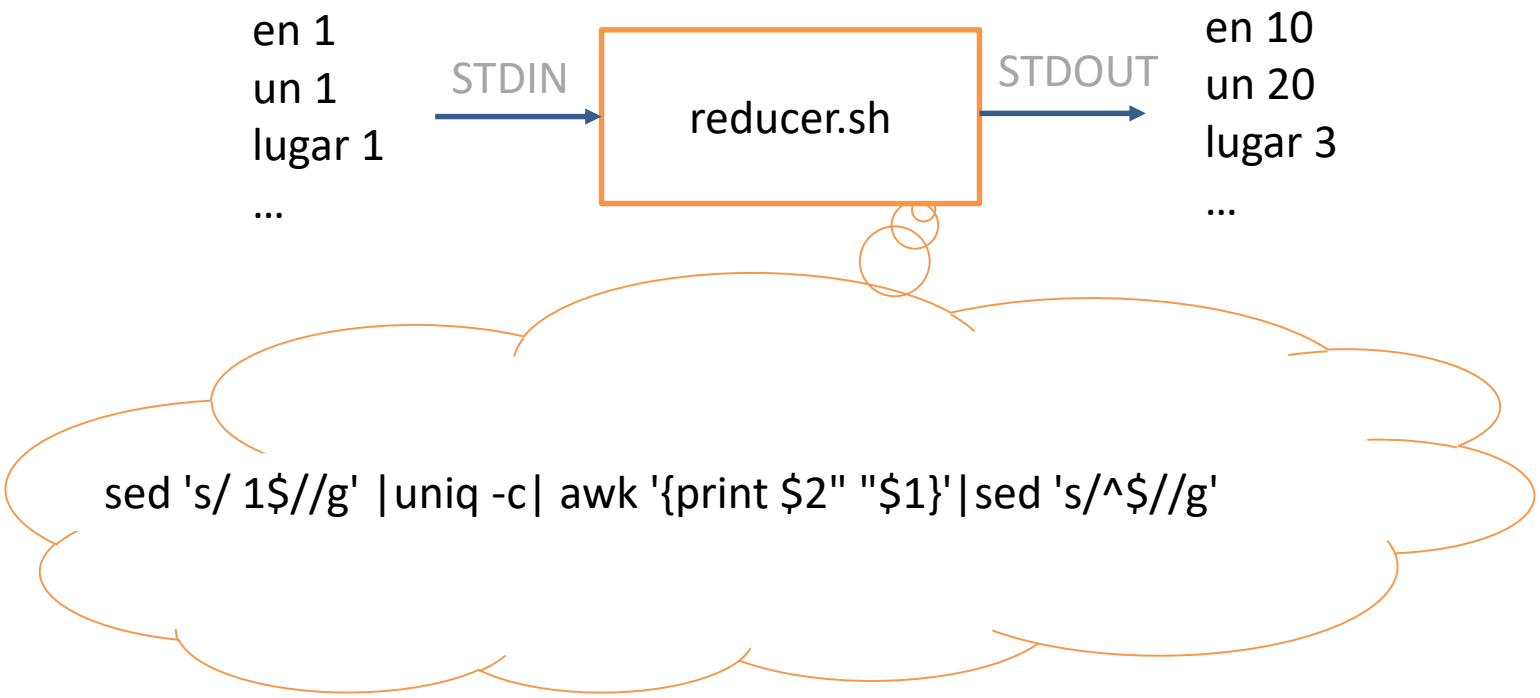
Nativo	Encapsulado
Java	Perl, Python, …

Hadoop Streaming API



http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

Hadoop Streaming API



http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

Hadoop: solo un nodo

Prerequisitos Instalación **Uso básico**

```
hduser@h1:~$ echo “uno uno dos dos tres” | ./mapper.sh | more
…
```

```
hduser@h1:~$ echo “uno uno dos dos tres” | ./mapper.sh|sort | more
…
```

```
hduser@h1:~$ echo “uno uno dos dos tres” | ./mapper.sh|sort|./reducer.sh |more
…
```

Hadoop: solo un nodo

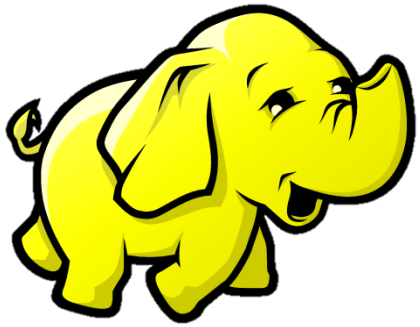
Prerequisitos Instalación **Uso básico**

```
hduser@h1:/usr/local/hadoop$ hadoop jar share/hadoop/tools/lib/hadoop-
    streaming-2.5.2.jar -file ./mapper.sh -mapper ./mapper.sh
    -file ./reducer.sh -reducer ./reducer.sh
    -input /user/hduser/ -output /user/hduser/counter
packageJobJar: [./mapper.sh, ./reducer.sh] [] /tmp/streamjob724842872862965882.jar tmpDir=null
14/10/04 15:48:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use
    dfs.metrics.session-id
…
    File Input Format Counters
            Bytes Read=2106143
    File Output Format Counters
            Bytes Written=320124
14/10/04 15:48:46 INFO streaming.StreamJob: Output directory: /user/hduser/counter
```

```
hduser@h1:/usr/local/hadoop$ hadoop fs -cat /user/hduser/counter/part-
    00000|sort -n -k 2 -r|head -5
…
que 20545
de 18154
y 18053
la 10338
a 9779
```

http://hadoop.apache.org/docs/r1.1.2/streaming.html#Hadoop+Streaming

Contenidos

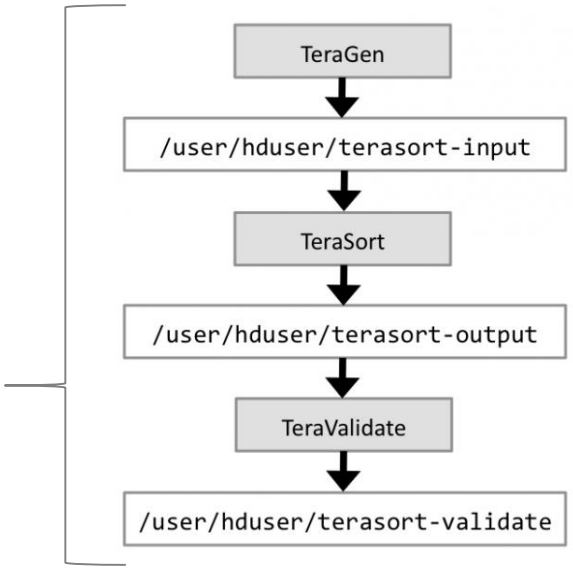


- Introducción
- *Hand-on*
- ***Benchmarking***

http://www.siliconweek.es/wp-content/uploads/2013/08/BigData-datos-guardar-almacenamiento-fichero-archivo.jpg
http://datameer2.datameer.com/blog/wp-content/uploads/2012/06/Hadoop-Ecosystem-Infographic-21.png

Benchmarking

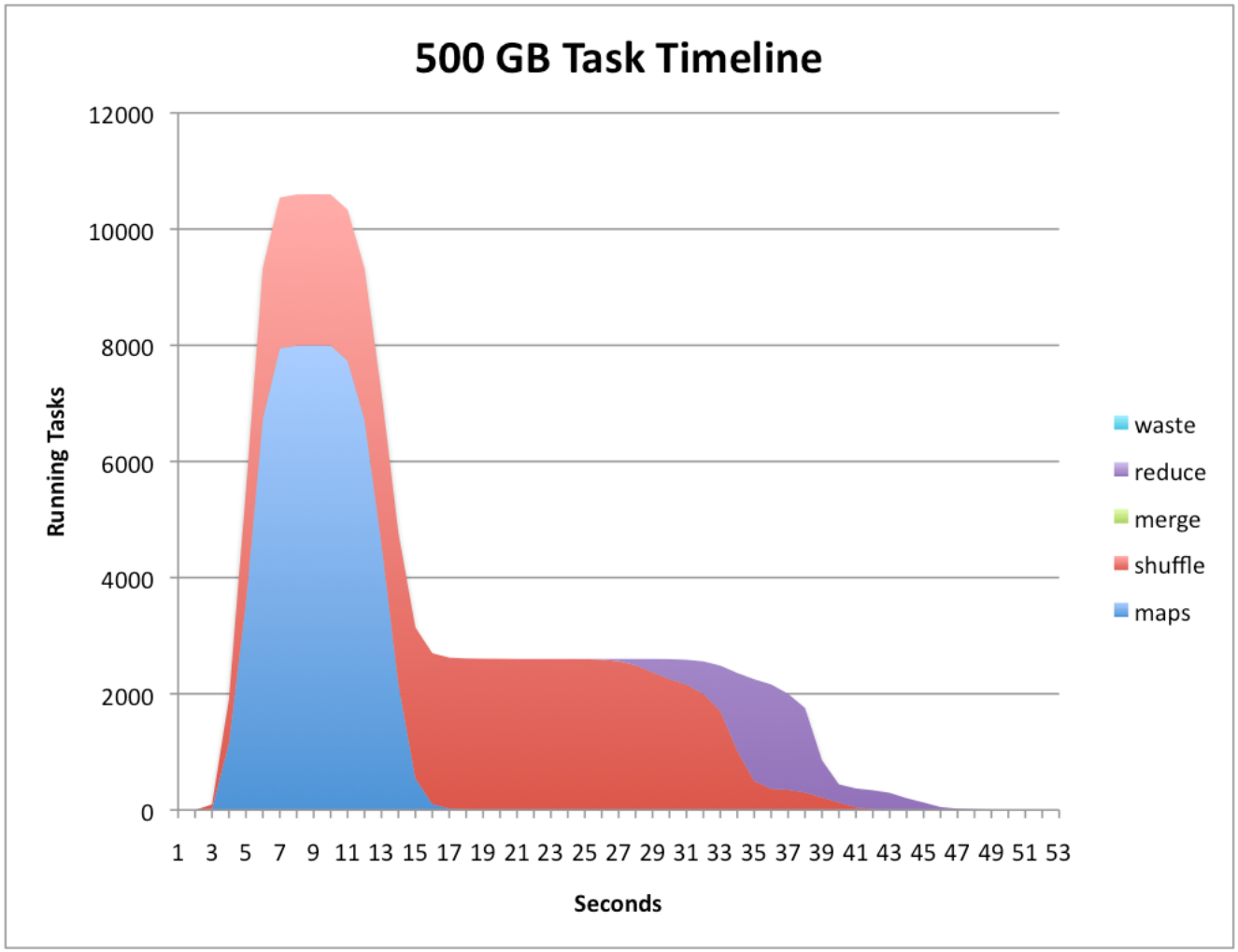
- TestDFSIO
- TeraSort benchmark suite
  - Yahoo! 2009: 1 PB de datos en 16 horas



- NameNode benchmark (nnbench)
- MapReduce benchmark (mrbench)

http://m.oschina.net/blog/74201

TeraSort (2009, 500GB)



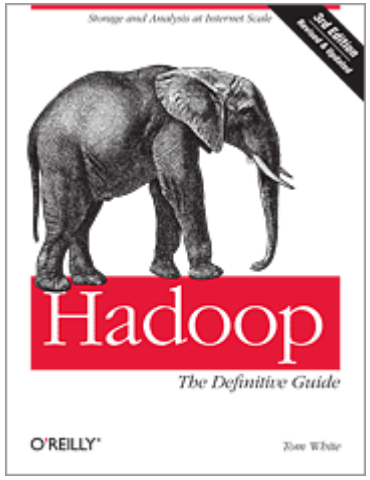
https://developer.yahoo.com/blogs/hadoop/hadoop-sorts-petabyte-16-25-hours-terabyte-62-422.html

Bibliografía: tutoriales

- Página Web oficial:
  - http://hadoop.apache.org/
- Introducción a cómo funciona Hadoop:
  - http://blog.csdn.net/suifeng3051/article/details/17288047
- Tutorial de cómo instalar y usar Hadoop:
  - http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php
  - http://www.bogotobogo.com/Hadoop/BigData_hadoop_Running_MapReduce_Job.php

Bibliografía: libro

- Hadoop: The Definitive Guide, 3rd Edition:
  - http://shop.oreilly.com/product/0636920021773.do
  - https://github.com/tomwhite/hadoop-book/



Bibliografía: TFG

- Extracción de información social desde Twitter y análisis mediante Hadoop.
  - Autor: Cristian Caballero Montiel
  - Tutores: Daniel Higuero Alonso-Mardones y Juan Manuel Tirado Martín
  - http://e-archivo.uc3m.es/handle/10016/16784

- Adaptation, Deployment and Evaluation of a Railway Simulator in Cloud Environments
  - Autora: Silvina Caíno Lores
  - Tutor: Alberto García Fernández

Agradecimientos

- Por último pero no por ello menos importante, agradecer al personal del Laboratorio del Departamento de Informática todos los comentarios y sugerencias para esta presentación.

Diseño de Sistemas Distribuidos

Máster en Ciencia y Tecnología Informática
Curso 2018-2019

Sistemas escalables en entornos distribuidos. Introducción a Hadoop

Alejandro Calderón Mateos & Óscar Pérez Alonso
acaldero@inf.uc3m.es oscar@lab.inf.uc3m.es