

Master Executive di II Livello  
**BIG DATA ANALYSIS AND  
BUSINESS INTELLIGENCE**

*Nome Docente*

## Installation of the Hadoop Ecosystem

**Francesco Pugliese, PhD**

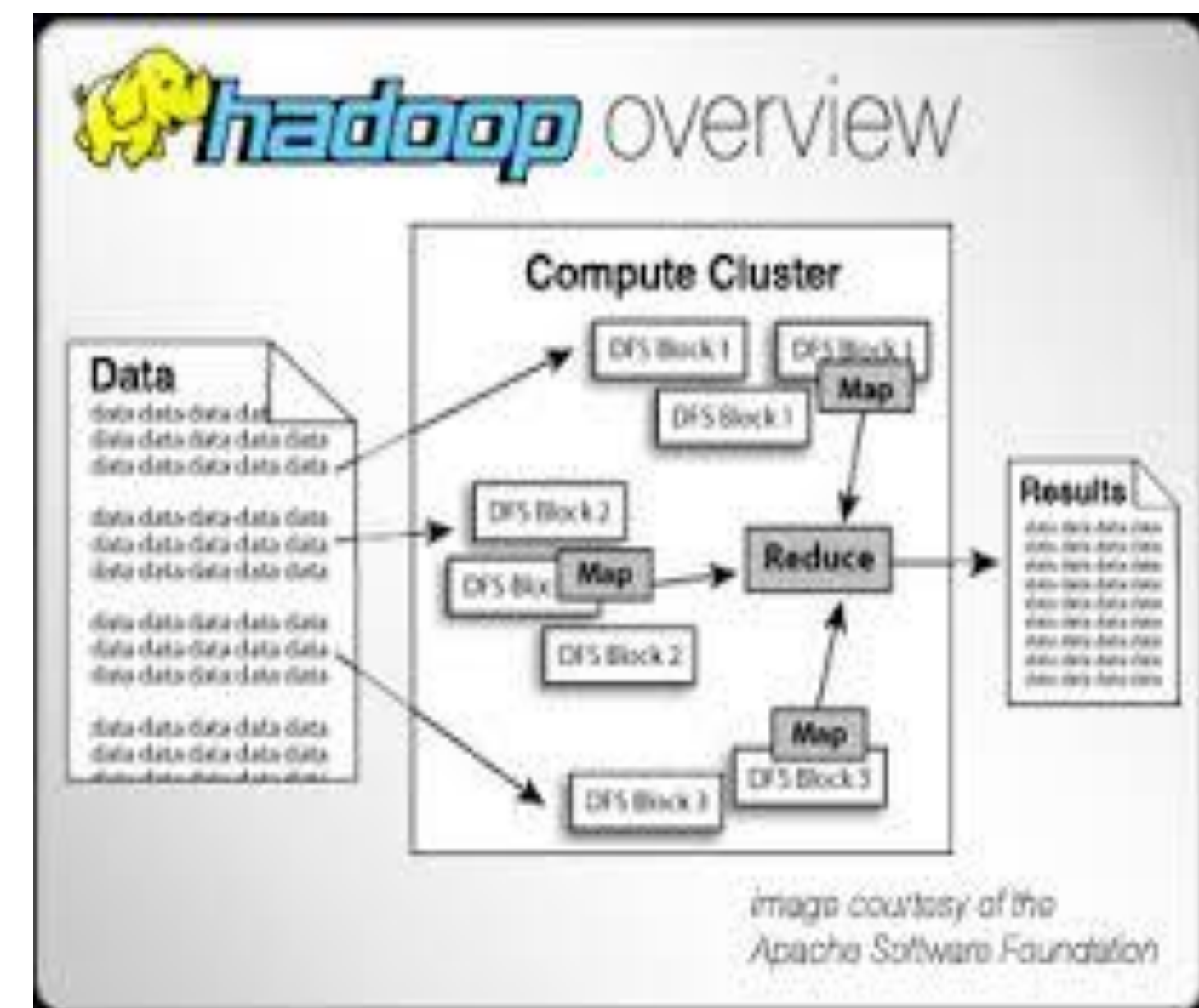
*Italian National Institute of Statistics, Division "Information and  
Application Architecture", Directorate for methodology and  
statistical design*

Email **Francesco Pugliese** : [francesco.pugliese@istat.it](mailto:francesco.pugliese@istat.it)



# Installation of a Single Node Cluster onto Linux Operating System

- There are two ways to install Hadoop, i.e. **Single node** and **Multi node**.
- **Single node cluster** means only one **DataNode** running and setting up all the **NameNode**, **DataNode**, **ResourceManager** and **NodeManager** on a single machine. This is used for studying and testing purposes.
- **Multi node cluster**, there are more than one **DataNode** running and each **DataNode** is running on different machines. This multi node mode is typically used in to analyze **Big Data**, because they require **Petabytes** of memory and **Teraflops** of computational power.



---

# Installation of a Single Node Cluster onto Linux Operating System

- In this lesson we will show how to install Hadoop in **Single Node** mode.

## Prerequisites:

*VIRTUAL BOX:* it is used for installing the operating system on it.

*OPERATING SYSTEM:* You can install Hadoop on Linux based operating systems. Ubuntu distribution is strongly recommended.

*JAVA:* You need to install the Java 8 64 bitpackage on your system.

*HADOOP:* From Hadoop version 2.7.3 to 3.2.0 package.



---

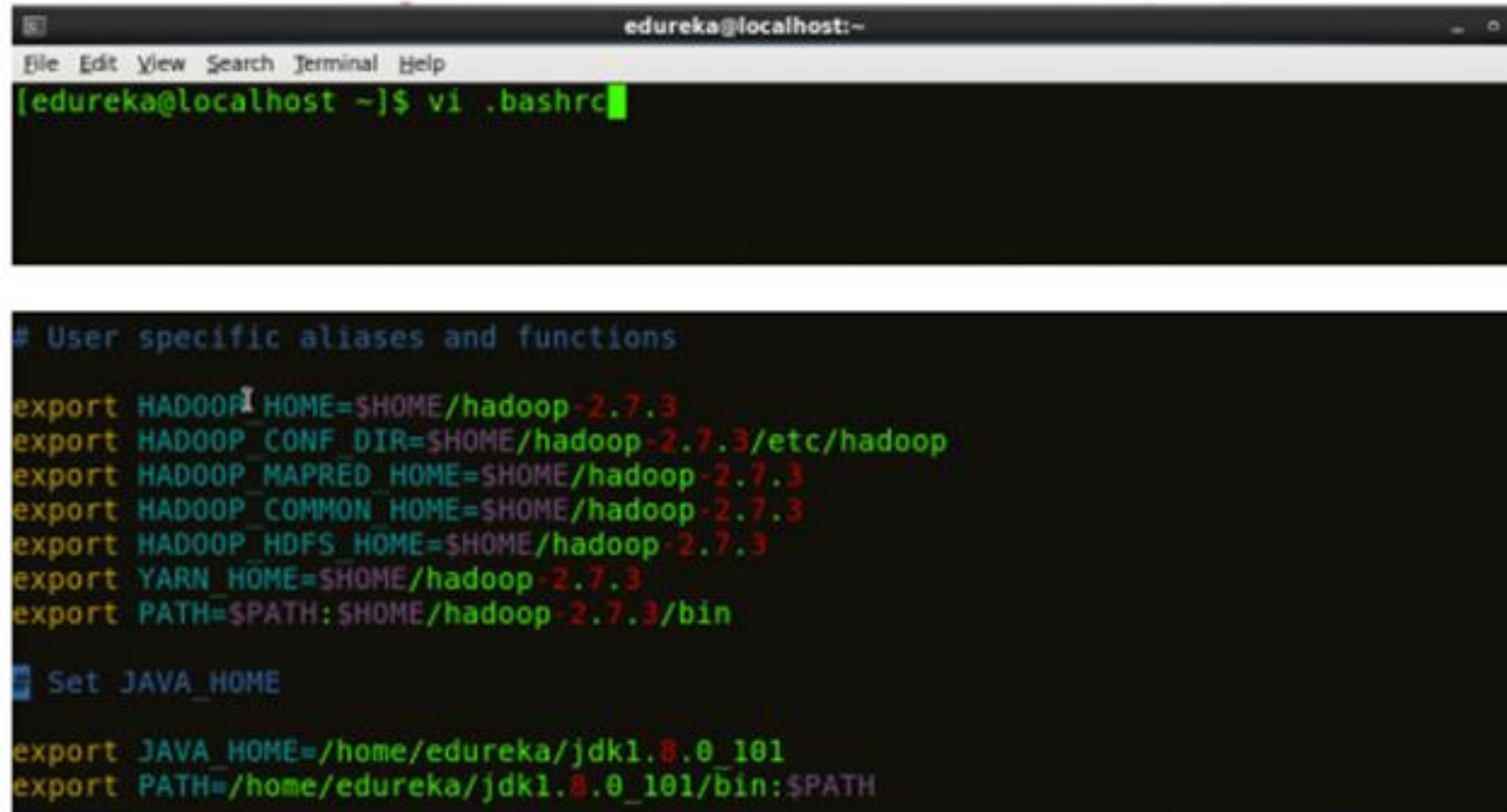
# Installation of a Single Node Cluster onto Linux Operating System

- 1) Download Java 64 bit from websites like this:  
<https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html> and extract the tar file with **tar -xvf file.tar**
  - 2) Download the Hadoop Package from  
<https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz>  
Use wget in Linux. Extract the Tar File
-



# Installation of a Single Node Cluster onto Linux Operating System

- 3) Add the Hadoop and Java Path into the *.bashrc*. Use *vi* or *gedit*.



```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ vi .bashrc  
  
# User specific aliases and functions  
  
export HADOOP_HOME=$HOME/hadoop-2.7.3  
export HADOOP_CONF_DIR=$HOME/hadoop-2.7.3/etc/hadoop  
export HADOOP_MAPRED_HOME=$HOME/hadoop-2.7.3  
export HADOOP_COMMON_HOME=$HOME/hadoop-2.7.3  
export HADOOP_HDFS_HOME=$HOME/hadoop-2.7.3  
export YARN_HOME=$HOME/hadoop-2.7.3  
export PATH=$PATH:$HOME/hadoop-2.7.3/bin  
  
# Set JAVA_HOME  
  
export JAVA_HOME=/home/edureka/jdk1.8.0_101  
export PATH=/home/edureka/jdk1.8.0_101/bin:$PATH
```

# Installation of a Single Node Cluster onto Linux Operating System

- 4) Check Java and Hadoop Version after installation with Hadoop version and java –version commands.

```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ java -version  
java version "1.8.0_101"  
Java(TM) SE Runtime Environment (build 1.8.0_101-b13)  
Java HotSpot(TM) 64-Bit Server VM (build 25.101-b13, mixed mode)  
[edureka@localhost ~]$
```

```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ hadoop version  
Hadoop 2.7.3  
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r baa91f7c6bc9cb92be5982de4719c1c8af91ccff  
Compiled by root on 2016-08-18T01:41Z  
Compiled with protoc 2.5.0  
From source with checksum 2e4ce5f957ea4db193bce3734ff29ff4  
This command was run using /home/edureka/hadoop-2.7.3/share/hadoop/common/hadoop-common-2.7.3.jar  
[edureka@localhost ~]$
```

---

# Installation of a Single Node Cluster onto Linux Operating System

- 5) Hadoop Configuration: **cd hadoop-ver/etc/Hadoop** and edit **core-site.xml**

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

---



---

# Installation of a Single Node Cluster onto Linux Operating System

- 6) Hadoop Configuration: edit **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.permission</name>
    <value>>false</value>
  </property>
</configuration>
```

---



---

# Installation of a Single Node Cluster onto Linux Operating System

- 7) Hadoop Configuration: edit `mapred-site.xml`

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

---



---

# Installation of a Single Node Cluster onto Linux Operating System

- 8) Hadoop Configuration: edit yarn-site.xml

```
<?xml version="1.0">
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

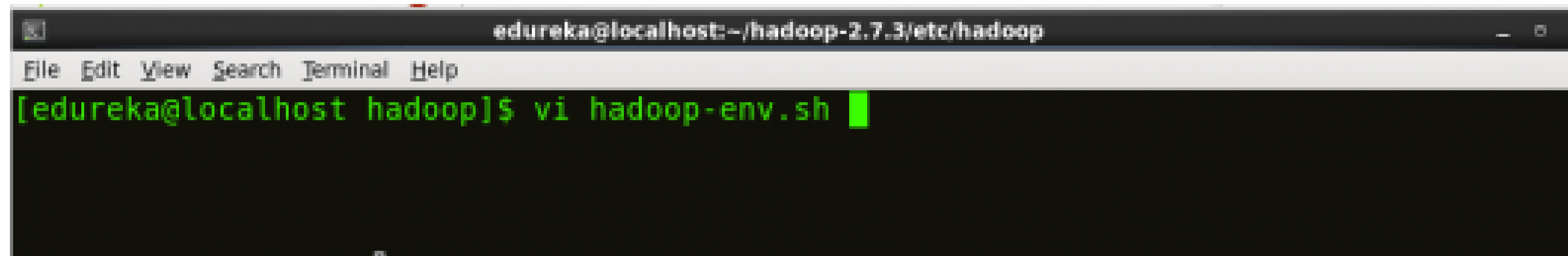
---



---

# Installation of a Single Node Cluster onto Linux Operating System

- 9) Hadoop Configuration: edit `hadoop-env.sh`

A terminal window with a dark background. The title bar shows the user 'edureka' at 'localhost' in the directory '/hadoop-2.7.3/etc/hadoop'. The menu bar includes 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The command prompt shows '[edureka@localhost hadoop]\$ vi hadoop-env.sh' with a green cursor at the end of the line.

```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ vi hadoop-env.sh
```

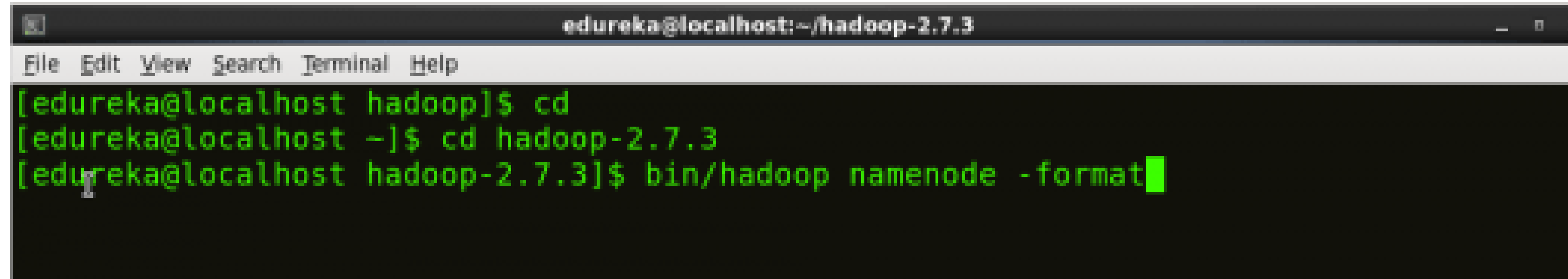
A terminal window showing the content of the 'hadoop-env.sh' file. It displays a comment and an export command for the Java Home path, with a green cursor at the end of the line.

```
# The java implementation to use.
export JAVA_HOME=/home/edureka/jdk1.8.0_101
```



# Installation of a Single Node Cluster onto Linux Operating System

- 10) Hadoop Configuration: format the NameNode with **bin/hadoop namenode -format**

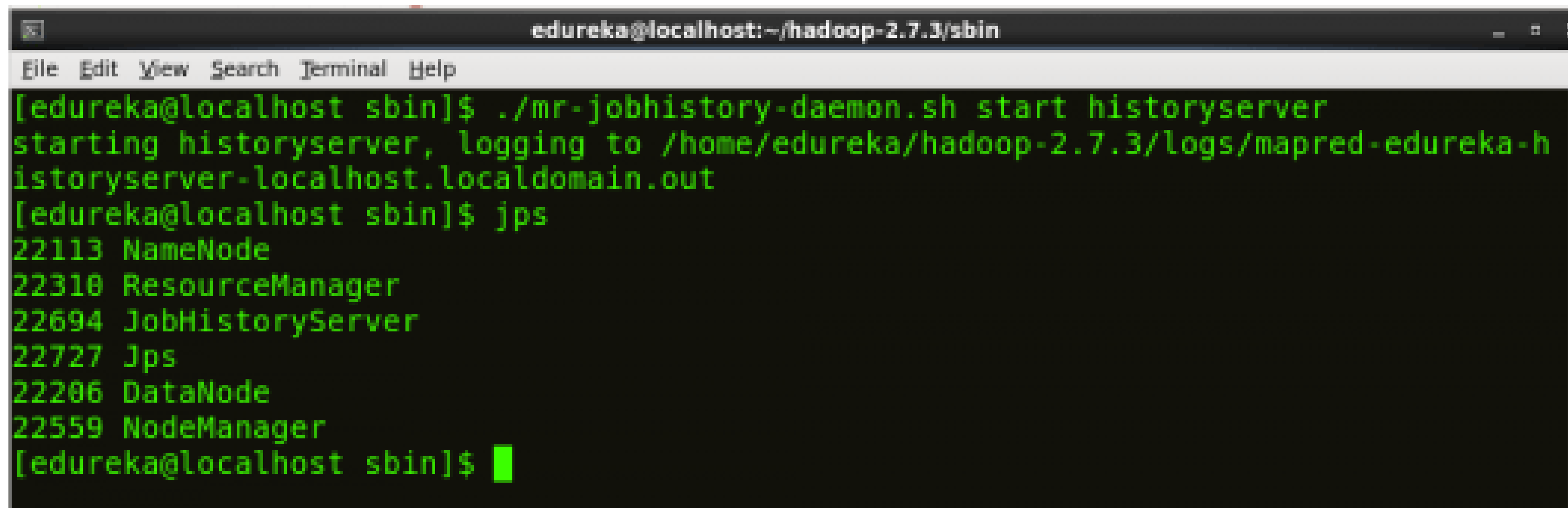
A terminal window titled 'edureka@localhost:~/hadoop-2.7.3' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows three commands being executed in green text on a black background: '[edureka@localhost hadoop]\$ cd', '[edureka@localhost ~]\$ cd hadoop-2.7.3', and '[edureka@localhost hadoop-2.7.3]\$ bin/hadoop namenode -format' with a green cursor at the end.

```
edureka@localhost:~/hadoop-2.7.3
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ cd
[edureka@localhost ~]$ cd hadoop-2.7.3
[edureka@localhost hadoop-2.7.3]$ bin/hadoop namenode -format
```



# Installation of a Single Node Cluster onto Linux Operating System

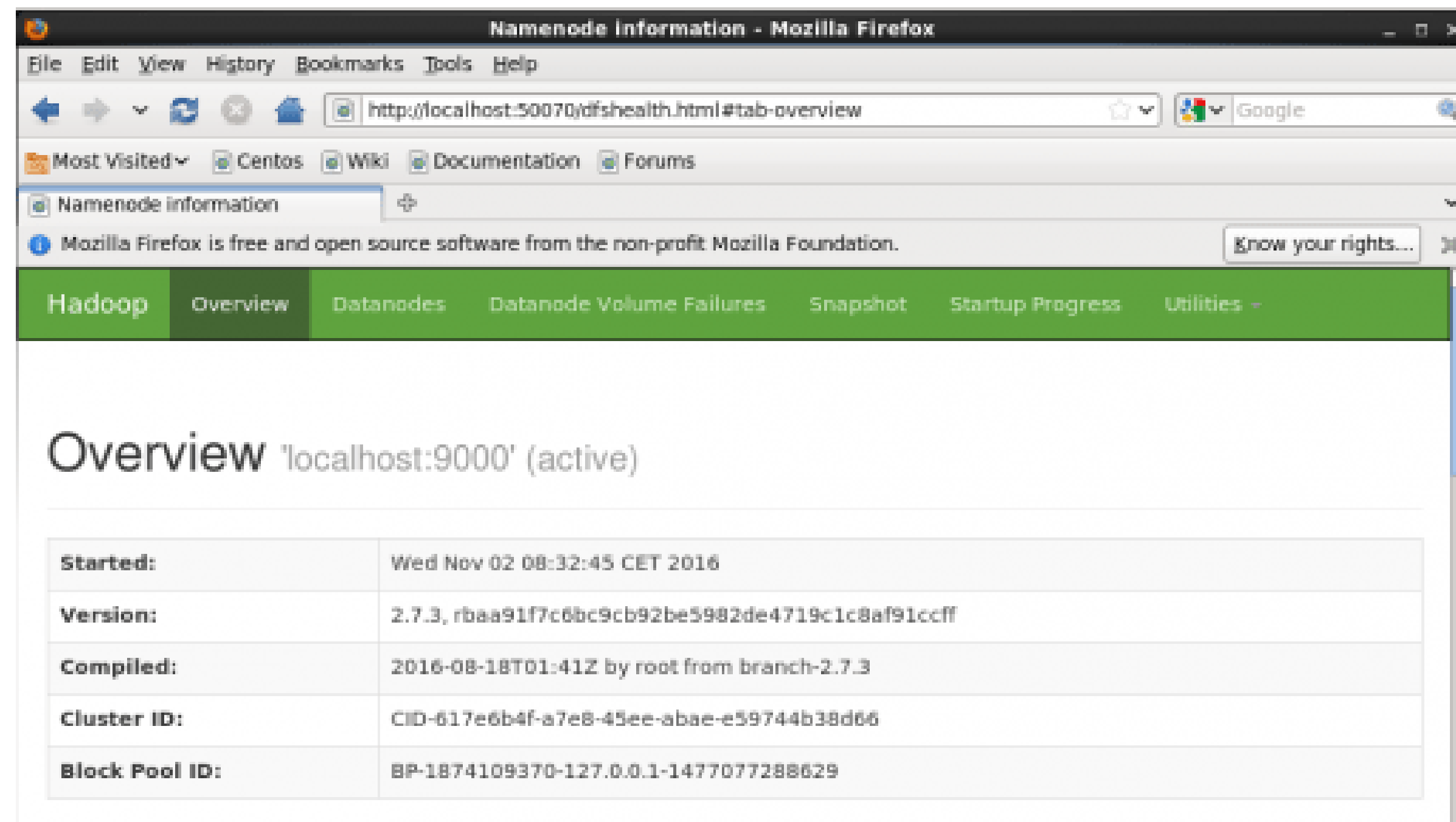
- 11) Start Hadoop services with **sbin/start-all.sh** and after check if all the services are running with **jps**.

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/sbin' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the execution of './mr-jobhistory-daemon.sh start historyserver', which logs to '/home/edureka/hadoop-2.7.3/logs/mapred-edureka-historyserver-localhost.localdomain.out'. Then, the 'jps' command is run, displaying a list of running processes: NameNode (PID 22113), ResourceManager (PID 22310), JobHistoryServer (PID 22694), Jps (PID 22727), DataNode (PID 22206), and NodeManager (PID 22559).

```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/edureka/hadoop-2.7.3/logs/mapred-edureka-h
istoryserver-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22310 ResourceManager
22694 JobHistoryServer
22727 Jps
22206 DataNode
22559 NodeManager
[edureka@localhost sbin]$
```

# Installation of a Single Node Cluster onto Linux Operating System

- 12) Open the NameNode interface at: **localhost:50070/dfshealth.html**





Master Executive di II Livello  
**BIG DATA ANALYSIS AND  
BUSINESS INTELLIGENCE**

*Nome Docente*

# Grazie

**Francesco Pugliese, PhD**

*Italian National Institute of Statistics, Division "Information and  
Application Architecture", Directorate for methodology and  
statistical design*

Email **Francesco Pugliese** : [francesco.pugliese@istat.it](mailto:francesco.pugliese@istat.it)

fondazione  
**INOIT**  
TORVERGATA