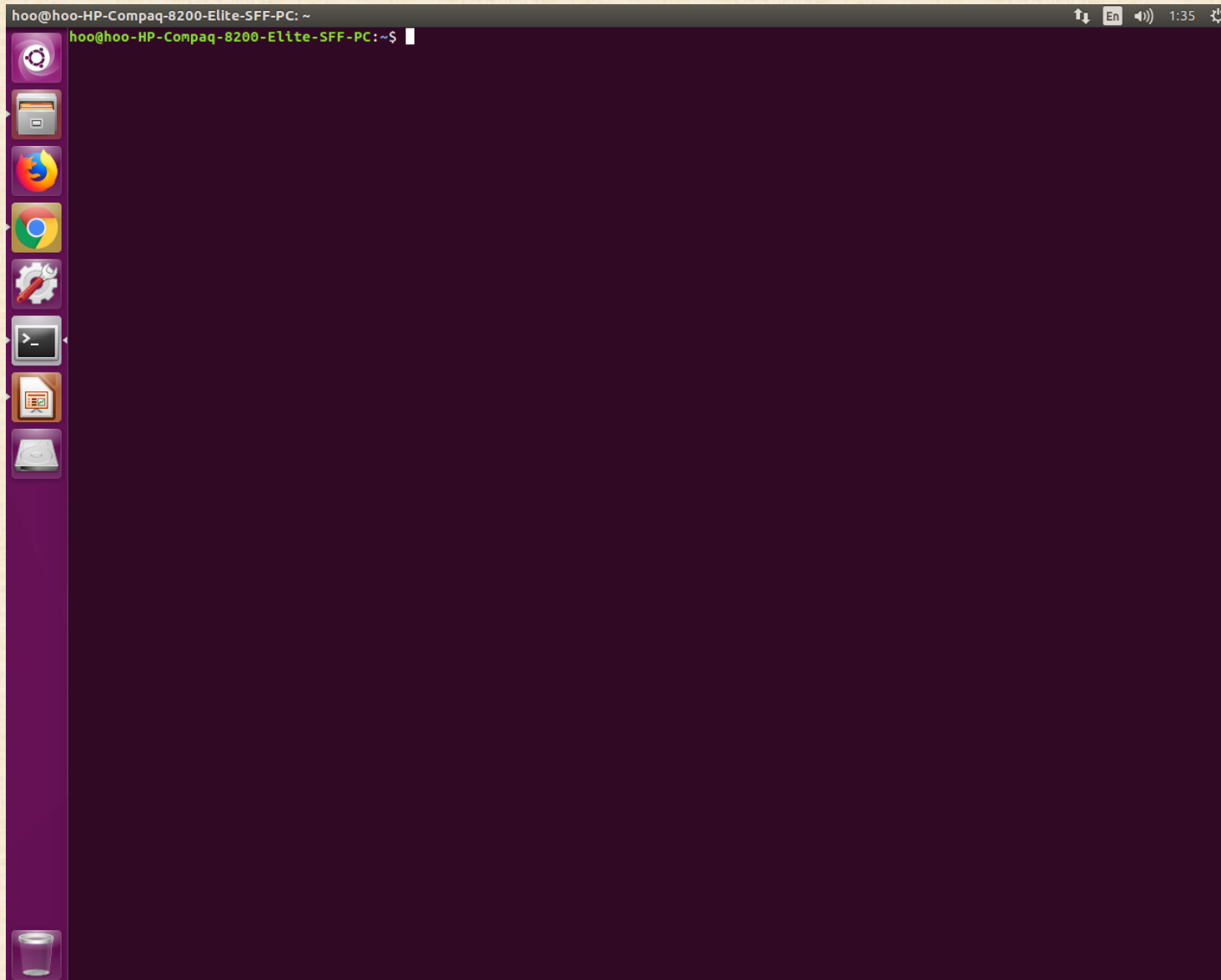




# Lab 1: Installing Hadoop to Linux System (Ubuntu)

# Start terminal





# Updating your system

```
sudo apt-get update  
sudo apt-get upgrade  
sudo apt-get install openssh-server
```

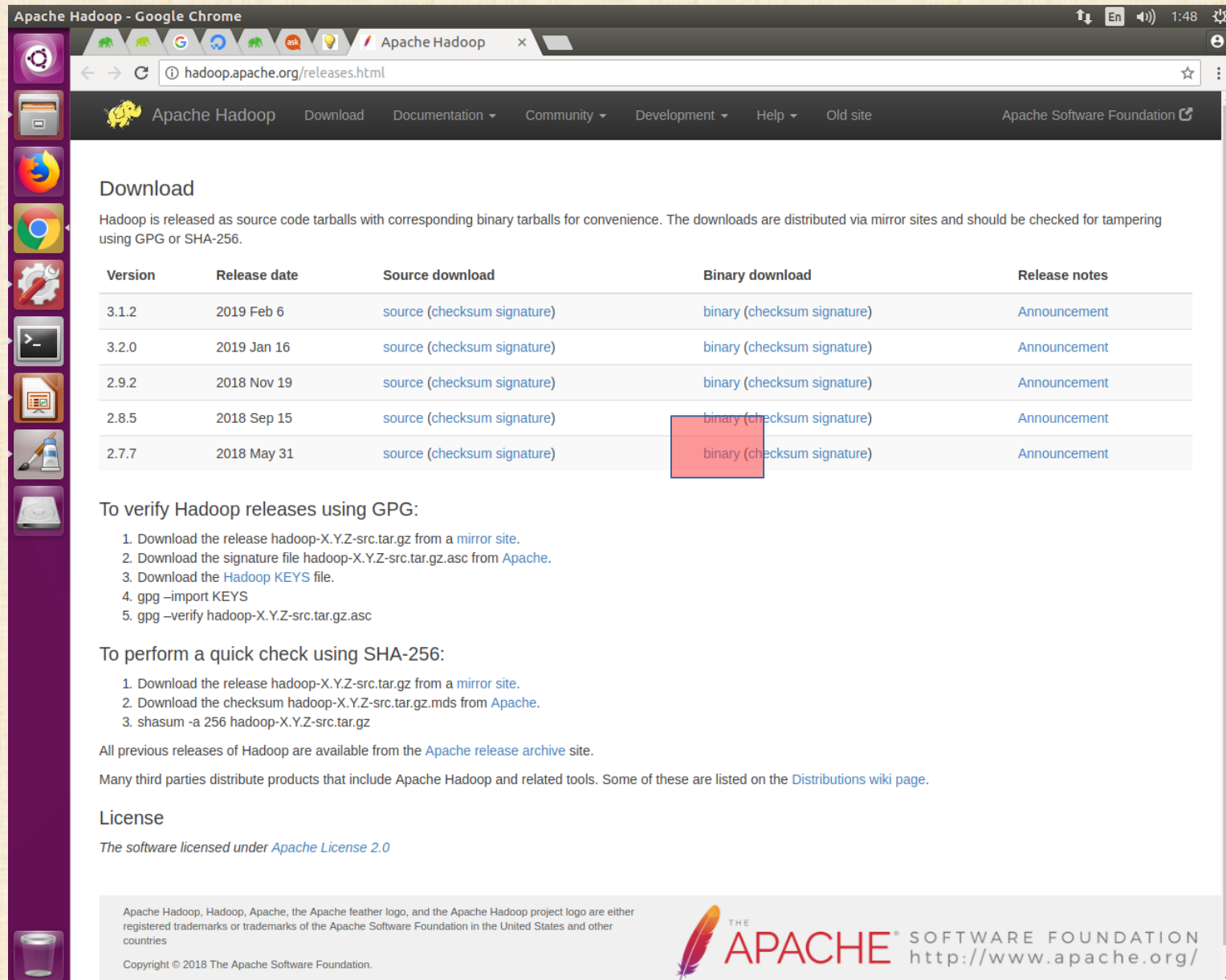


# Check Java installation

```
sudo apt-get install openjdk-8-jdk  
java -version
```

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$ java -version  
openjdk version "1.8.0_191"  
OpenJDK Runtime Environment (build 1.8.0_191-8u191-b12-2ubuntu0.16  
.04.1-b12)  
OpenJDK 64-Bit Server VM (build 25.191-b12, mixed mode)  
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$
```

# Look for Hadoop distributions



Apache Hadoop - Google Chrome

hadoop.apache.org/releases.html

Apache Hadoop Download Documentation Community Development Help Old site Apache Software Foundation

## Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

| Version | Release date | Source download                             | Binary download                             | Release notes                |
|---------|--------------|---------------------------------------------|---------------------------------------------|------------------------------|
| 3.1.2   | 2019 Feb 6   | <a href="#">source (checksum signature)</a> | <a href="#">binary (checksum signature)</a> | <a href="#">Announcement</a> |
| 3.2.0   | 2019 Jan 16  | <a href="#">source (checksum signature)</a> | <a href="#">binary (checksum signature)</a> | <a href="#">Announcement</a> |
| 2.9.2   | 2018 Nov 19  | <a href="#">source (checksum signature)</a> | <a href="#">binary (checksum signature)</a> | <a href="#">Announcement</a> |
| 2.8.5   | 2018 Sep 15  | <a href="#">source (checksum signature)</a> | <a href="#">binary (checksum signature)</a> | <a href="#">Announcement</a> |
| 2.7.7   | 2018 May 31  | <a href="#">source (checksum signature)</a> | <a href="#">binary (checksum signature)</a> | <a href="#">Announcement</a> |

### To verify Hadoop releases using GPG:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the signature file `hadoop-X.Y.Z-src.tar.gz.asc` from [Apache](#).
3. Download the [Hadoop KEYS](#) file.
4. `gpg --import KEYS`
5. `gpg --verify hadoop-X.Y.Z-src.tar.gz.asc`

### To perform a quick check using SHA-256:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the checksum `hadoop-X.Y.Z-src.tar.gz.md5` from [Apache](#).
3. `shasum -a 256 hadoop-X.Y.Z-src.tar.gz`

All previous releases of Hadoop are available from the [Apache release archive](#) site.

Many third parties distribute products that include Apache Hadoop and related tools. Some of these are listed on the [Distributions wiki page](#).

## License

The software licensed under [Apache License 2.0](#)

Apache Hadoop, Hadoop, Apache, the Apache feather logo, and the Apache Hadoop project logo are either registered trademarks or trademarks of the Apache Software Foundation in the United States and other countries

Copyright © 2018 The Apache Software Foundation.

THE APACHE SOFTWARE FOUNDATION  
<http://www.apache.org/>





# Download Hadoop 2.7.7

Apache Download Mirrors - Google Chrome

Secure | <https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

Home » Dyn About Projects People Get Involved Download Support Apache

Google Custom

[The Apache Way](#)

[Contribute](#)

[ASF Sponsors](#)

We suggest the following mirror site for your download:

<https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

Other mirror sites are suggested below.

It is essential that you [verify the integrity](#) of the downloaded file using the PGP signature ( `.asc` file) or a hash ( `.md5` or `.sha*` file).

Please only use the backup mirrors to download KEYS, PGP and MD5 sigs/hashes or if no other mirrors are working.

**HTTP**

<https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

<https://www-us.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

**BACKUP SITES**

Please only use the backup mirrors to download KEYS, PGP and MD5 sigs/hashes or if no other mirrors are working.

<https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

<https://www-us.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>

The full listing of mirror sites is also available.

**BECOMING A MIRROR**

The procedure for setting up new mirrors is described in [How to become a mirror](#).

**VERIFY THE INTEGRITY OF THE FILES**

It is essential that you verify the integrity of the downloaded file using the PGP signature ( `.asc` file) or a hash ( `.md5` or `.sha*` file). Please read [Verifying Apache](#)

# To download via terminal

```
wget https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz
```

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$ wget https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz
--2019-02-26 13:58:24-- https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz
Resolving www-eu.apache.org (www-eu.apache.org)... 2a01:4f9:2a:185f::2, 95.216.24.32
Connecting to www-eu.apache.org (www-eu.apache.org)|2a01:4f9:2a:185f::2|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 218720521 (209M) [application/x-gzip]
Saving to: 'hadoop-2.7.7.tar.gz'

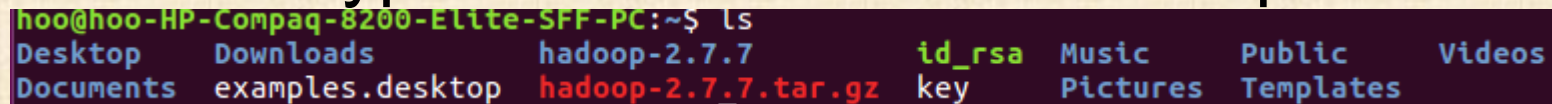
hadoop-2.7.7.tar.gz      100%[=====>] 208.59M  5.89MB/s   in 56s
2019-02-26 13:59:41 (3.73 MB/s) - 'hadoop-2.7.7.tar.gz' saved [218720521/218720521]
```



# Unzip the folder

```
tar -xzvf hadoop-2.7.7.tar.gz
```

- -x flag to extract, -z to uncompress, -v for verbose output, and -f to specify that we're extracting from a file
- You should notice a number of file extracted via the terminal
- You can type “ls” to check the uncompressed folder



A terminal window screenshot showing the output of the 'ls' command. The prompt is 'hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~\$'. The output lists several files and directories: Desktop, Downloads, hadoop-2.7.7, id\_rsa, Music, Public, Videos, Documents, examples.desktop, hadoop-2.7.7.tar.gz, key, Pictures, and Templates. The file 'hadoop-2.7.7' is highlighted in green, and 'hadoop-2.7.7.tar.gz' is highlighted in red.

```
sudo mv hadoop-2.7.7 /home/student/hadoop/
```

- Assuming your username is “student”



# Configuring Hadoop Java Home

```
readlink -f /usr/bin/java | sed "s:bin/java::"
```

- The path to Java, /usr/bin/java is a symlink to /etc/alternatives/java, which is in turn a symlink to default Java binary. We will use readlink with the -f flag to follow every symlink in every part of the path, recursively. Then, we'll use sed to trim bin/java from the output to give us the correct value for JAVA\_HOME.

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$ readlink -f /usr/bin/java | sed "s:bin/java::"  
/usr/lib/jvm/java-8-openjdk-amd64/jre/
```

```
sudo nano /home/{yourname}/hadoop/etc/hadoop/hadoop-env.sh
```

- Update JAVA\_HOME

```
# The java implementation to use.  
#export JAVA_HOME=${JAVA_HOME}  
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/
```



# Run Hadoop

/home/{yourname}/hadoop/bin/Hadoop

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:/home/wlhoo$ /home/wlhoo/hadoop/bin/hadoop
Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]
  CLASSNAME          run the class named CLASSNAME
or
  where COMMAND is one of:
  fs                  run a generic filesystem user client
  version             print the version
  jar <jar>           run a jar file
                      note: please use "yarn jar" to launch
                      YARN applications, not this command.
  checknative [-a|-h] check native hadoop and compression libraries availability
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath           prints the class path needed to get the
  credential           interact with credential providers
                      Hadoop jar and the required libraries
  daemonlog           get/set the log level for each daemon
  trace              view and modify Hadoop tracing settings

Most commands print help when invoked w/o parameters.
```

- Any other way?



# Test installation with sample MapReduce code

```
mkdir ~/input
```

```
cp /home/{yourname}/hadoop/etc/hadoop/*.xml ~/input
```

```
/home/{yourname}/hadoop/bin/hadoop jar  
/home/{yourname}/hadoop/share/hadoop/mapreduce/hadoop-  
mapreduce-examples-2.7.7.jar grep ~/input ~/grep_example  
'principal[.]*'
```

# Part of the generated output

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC: ~  
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$ /home/wlhoo/hadoop/bin/hadoop jar /home/wlhoo/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar grep ~/input ~/grep_example 'principal[.]*' 19/02/26 14:39:55 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id  
19/02/26 14:39:55 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=  
19/02/26 14:39:55 INFO input.FileInputFormat: Total input paths to process : 8  
19/02/26 14:39:55 INFO mapreduce.JobSubmitter: number of splits:8  
19/02/26 14:39:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local635374620_0001  
19/02/26 14:39:56 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
19/02/26 14:39:56 INFO mapreduce.Job: Running job: job_local635374620_0001  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: OutputCommitter set in Config null  
19/02/26 14:39:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: Waiting for map tasks  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: Starting task: attempt_local635374620_0001_m_000000_0  
19/02/26 14:39:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  
19/02/26 14:39:56 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  
19/02/26 14:39:56 INFO mapred.MapTask: Processing split: file:/home/hoo/input/hadoop-policy.xml:0+9683  
19/02/26 14:39:56 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)  
19/02/26 14:39:56 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
19/02/26 14:39:56 INFO mapred.MapTask: soft limit at 83886080  
19/02/26 14:39:56 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600  
19/02/26 14:39:56 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  
19/02/26 14:39:56 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer  
19/02/26 14:39:56 INFO mapred.LocalJobRunner:  
19/02/26 14:39:56 INFO mapred.MapTask: Starting flush of map output  
19/02/26 14:39:56 INFO mapred.Task: Task:attempt_local635374620_0001_m_000000_0 is done. And is in the process of committing  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: map  
19/02/26 14:39:56 INFO mapred.Task: Task 'attempt_local635374620_0001_m_000000_0' done.  
19/02/26 14:39:56 INFO mapred.Task: Final Counters for attempt_local635374620_0001_m_000000_0: Counters: 18  
File System Counters  
FILE: Number of bytes read=306708  
FILE: Number of bytes written=587865  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
Map-Reduce Framework  
Map input records=226  
Map output records=0  
Map output bytes=0  
Map output materialized bytes=6  
Input split bytes=103  
Combine input records=0  
Combine output records=0  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=6  
Total committed heap usage (bytes)=216006656  
File Input Format Counters  
Bytes Read=9683  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: Finishing task: attempt_local635374620_0001_m_000000_0  
19/02/26 14:39:56 INFO mapred.LocalJobRunner: Starting task: attempt_local635374620_0001_m_000001_0  
19/02/26 14:39:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  
19/02/26 14:39:56 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  
19/02/26 14:39:56 INFO mapred.MapTask: Processing split: file:/home/hoo/input/kms-site.xml:0+5540  
19/02/26 14:39:56 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)  
19/02/26 14:39:56 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
19/02/26 14:39:56 INFO mapred.MapTask: soft limit at 83886080
```



# Final outcome

```
cat ~/grep_example/*
```

```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:~$ cat ~/grep_example/*  
6      principal  
1      principal.
```

- What this means?