DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics

digitalent.kominfo.go.id

Hadoop

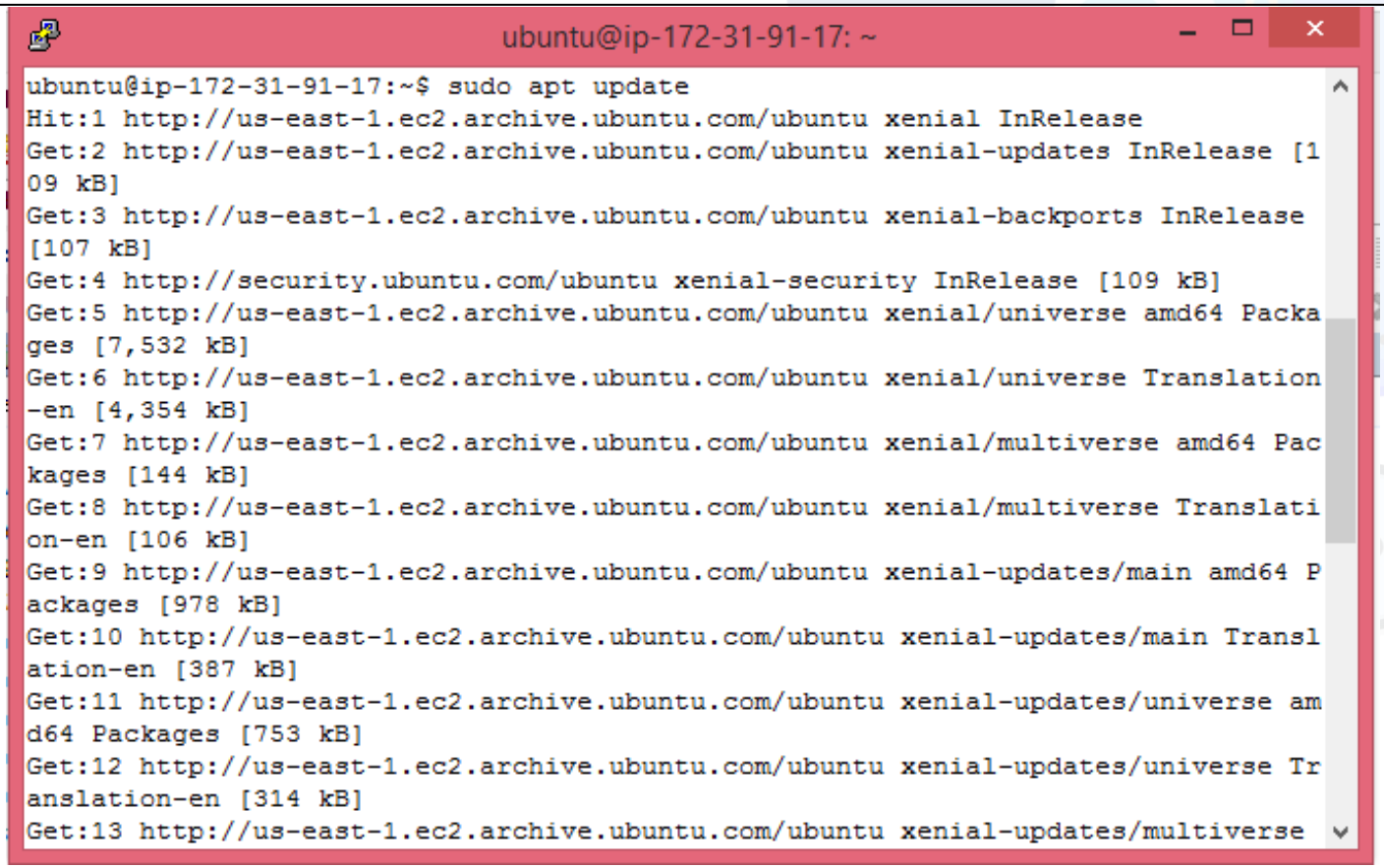
Instalasi Hadoop Single Node dan Case Study

Requirement

- EC2 Instance
- Akses ke EC2 Instance.
  - (SSH Client, Putty, WinSCP, dll)

Update apt Database

```
$ sudo apt update
```



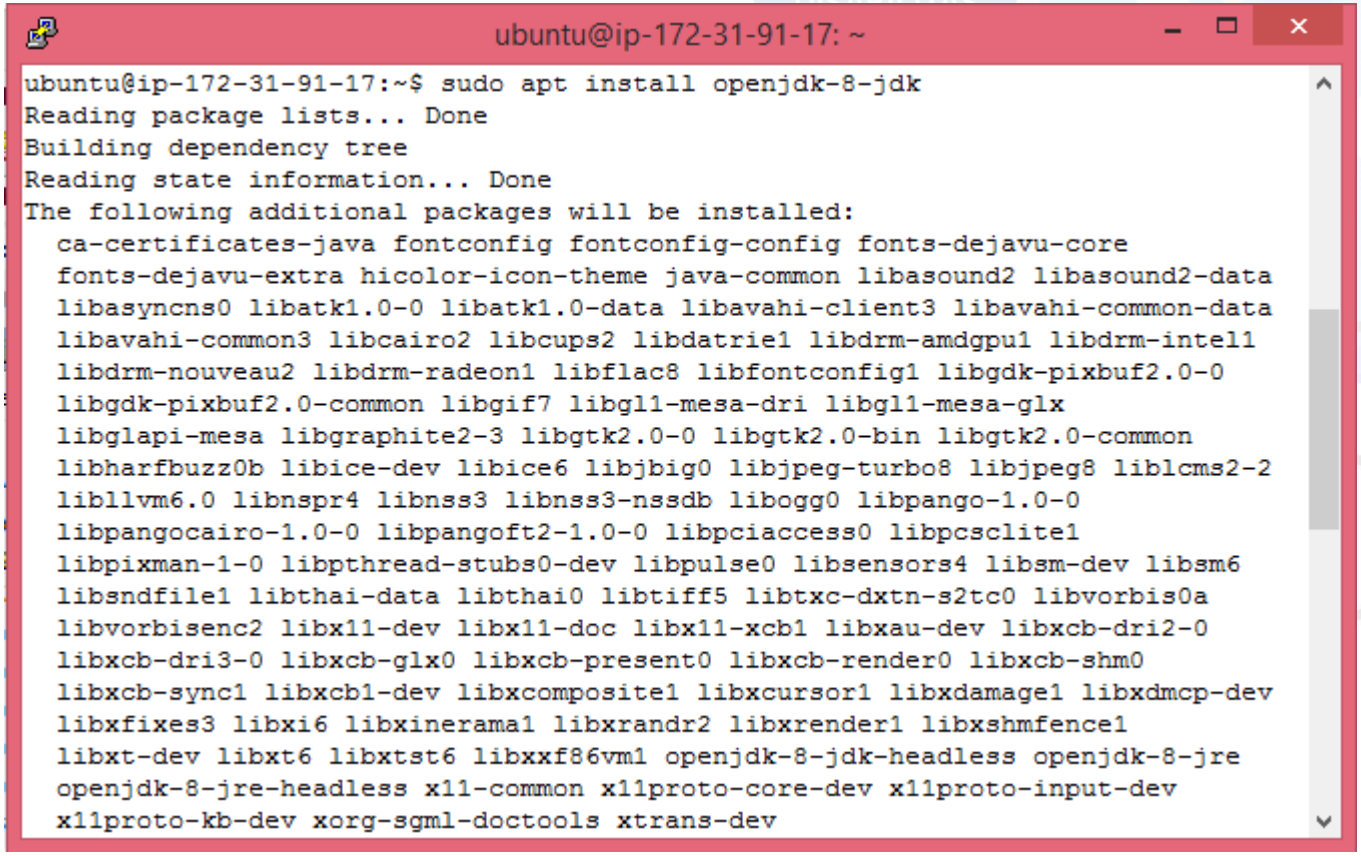
Instalasi Java

Big Data

Hadoop

Install OpenJDK-8

$ sudo apt install openjdk-8-jdk

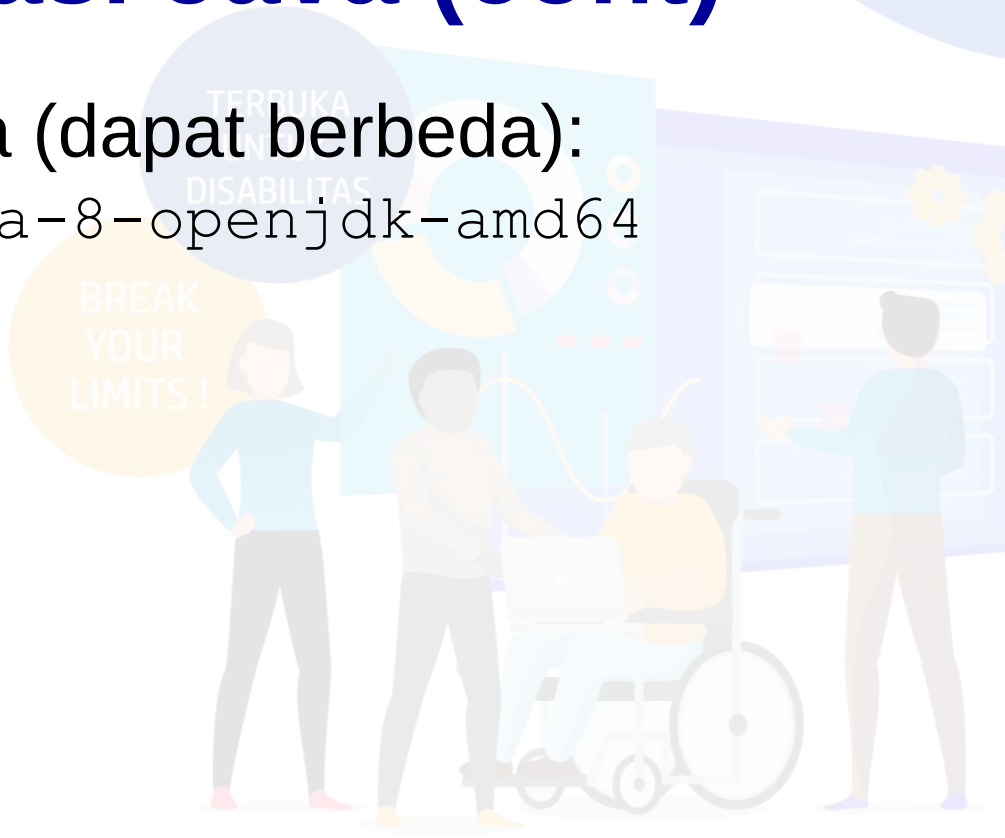


Direktori instalasi Java

```
$ update-alternatives --display java
```

Direktori instalasi Java (cont)

- Direktori instalasi java (dapat berbeda):
  - `/usr/lib/jvm/java-8-openjdk-amd64`



Variabel JAVA_HOME

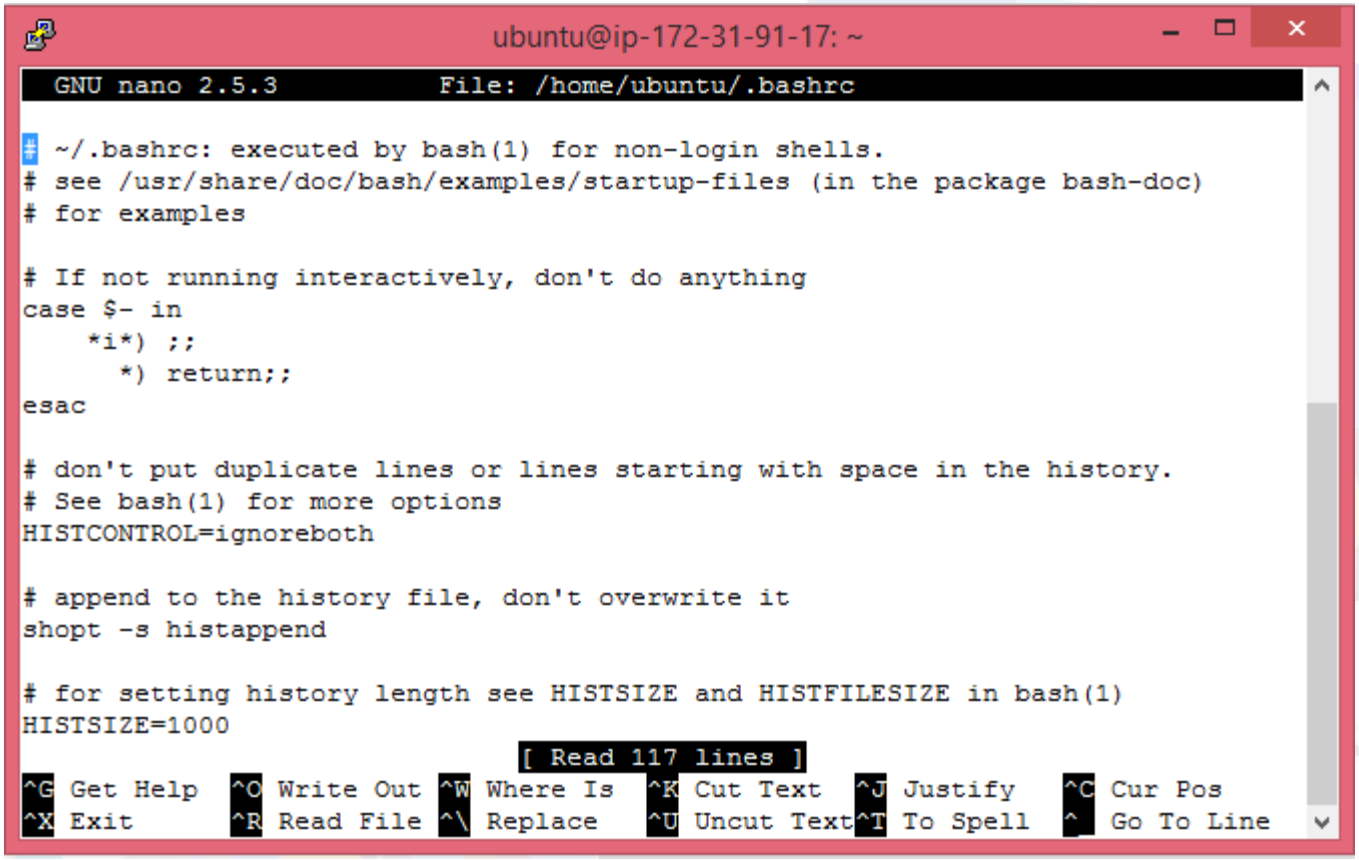
```
$ nano ~/.bashrc
```

Atau..

```
$ nano ~/.profile
```

Gunakan text editor lain apabila tersedia.

Variabel JAVA_HOME (cont)

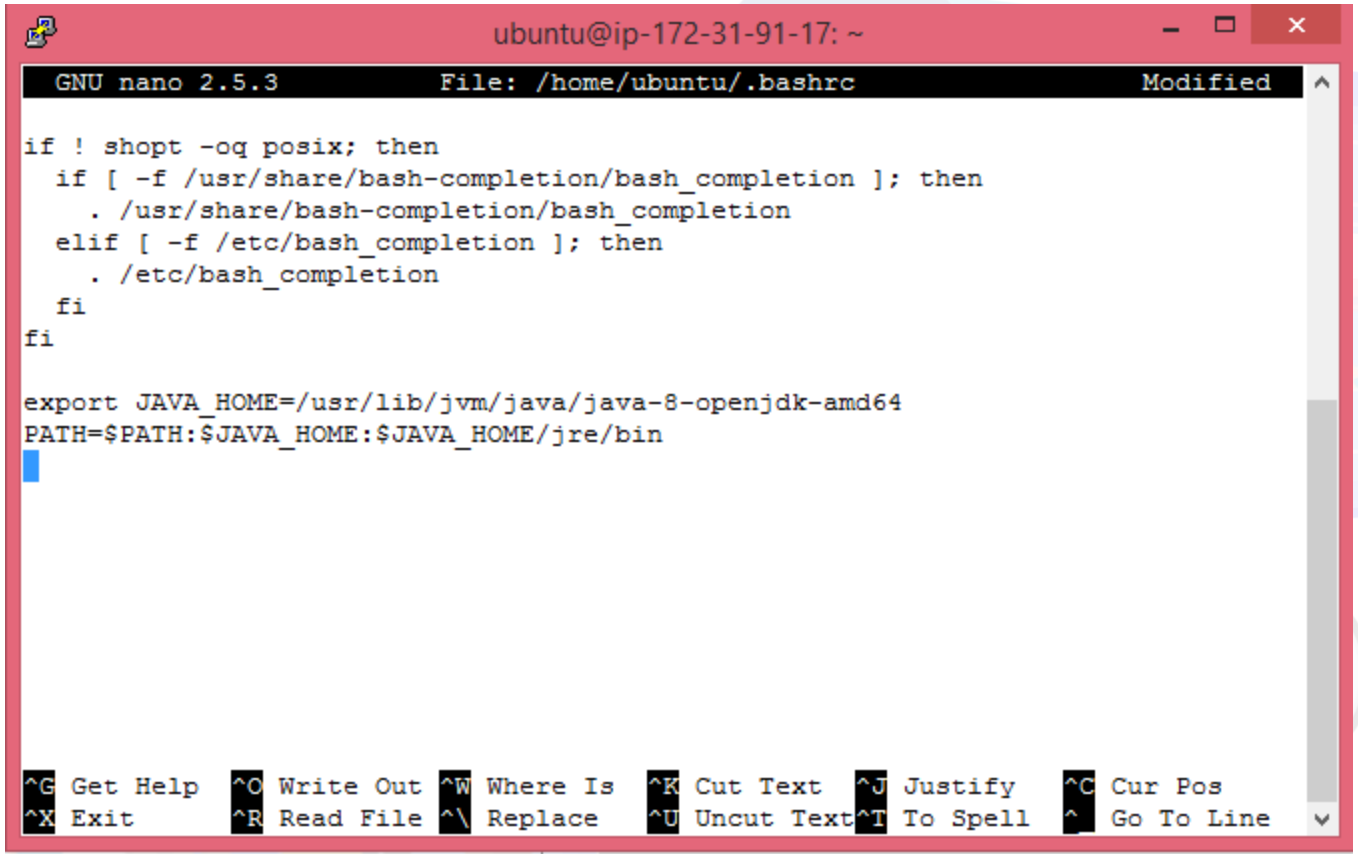


Variabel JAVA_HOME (cont)

Tambahkan string berikut (sesuaikan dengan direktori java anda):

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

PATH=$PATH:$JAVA_HOME:$JAVA_HOME/jre/bin
```

Variabel JAVA_HOME (cont)

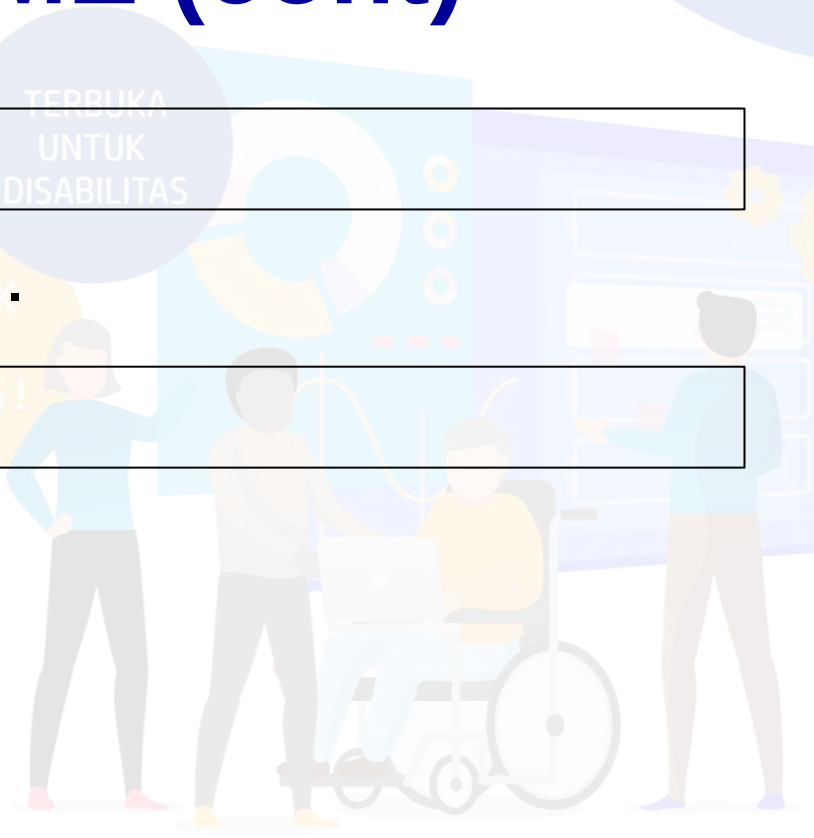


Variabel JAVA_HOME (cont)

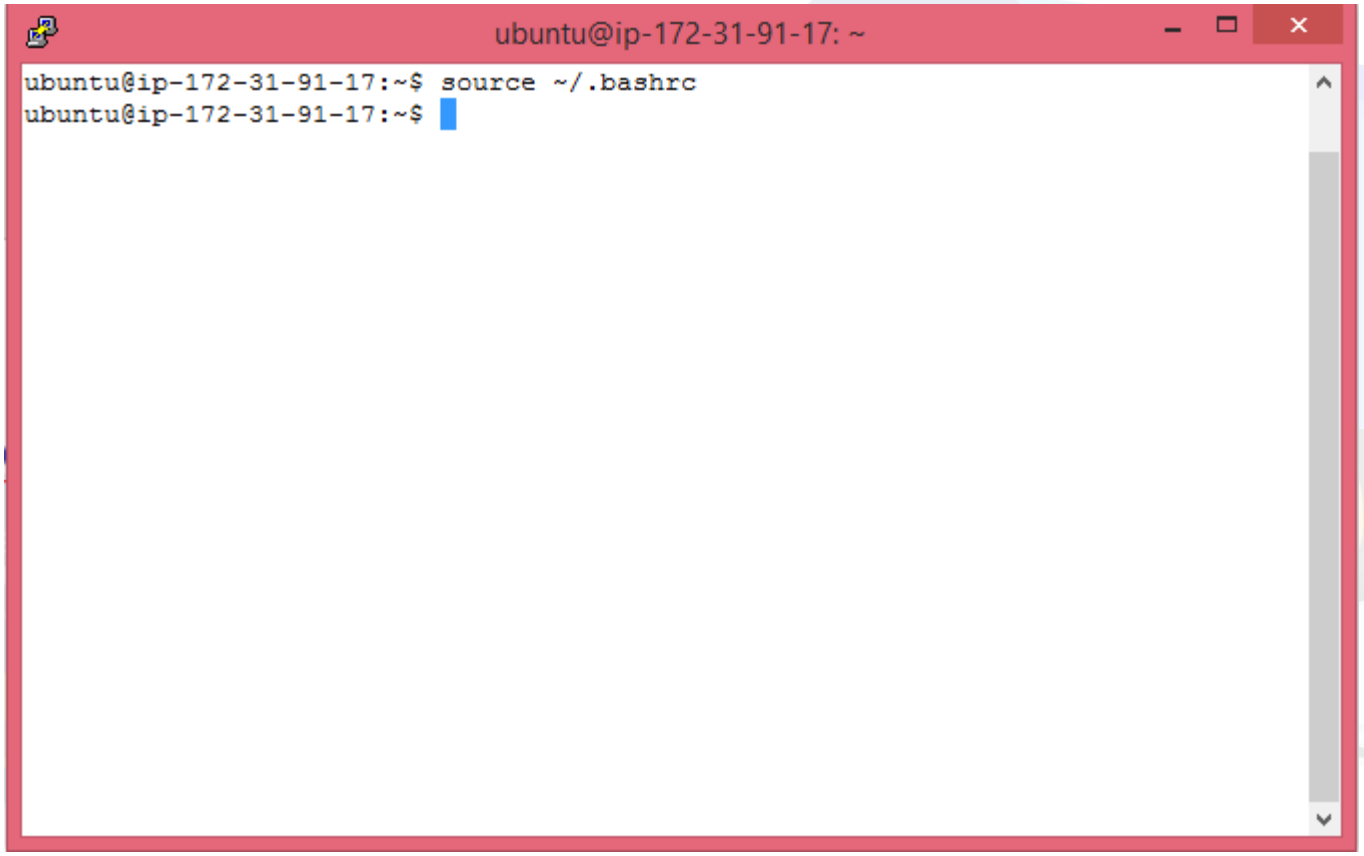
```
$ source ~/.bashrc
```

Atau..

```
$ source ~/.profile
```

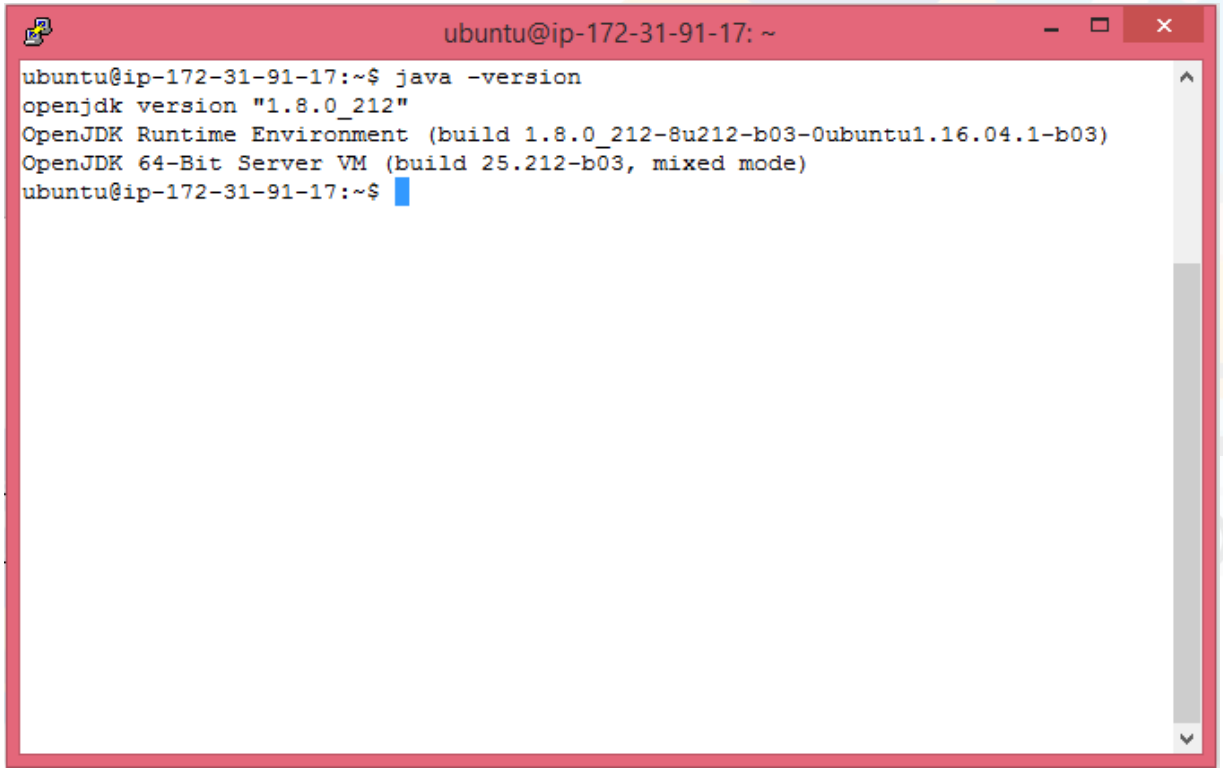


Variabel JAVA_HOME (cont)



Periksa instalasi Java

```
$ java -version
```



Instalasi Python (Anaconda)

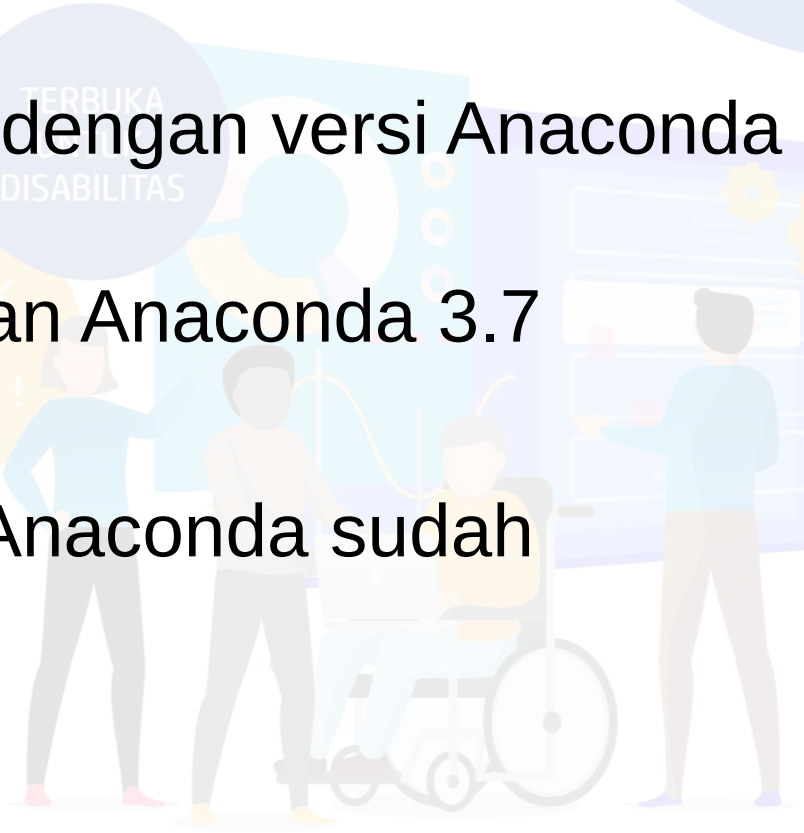
Big Data

Hadoop

Versi Anaconda

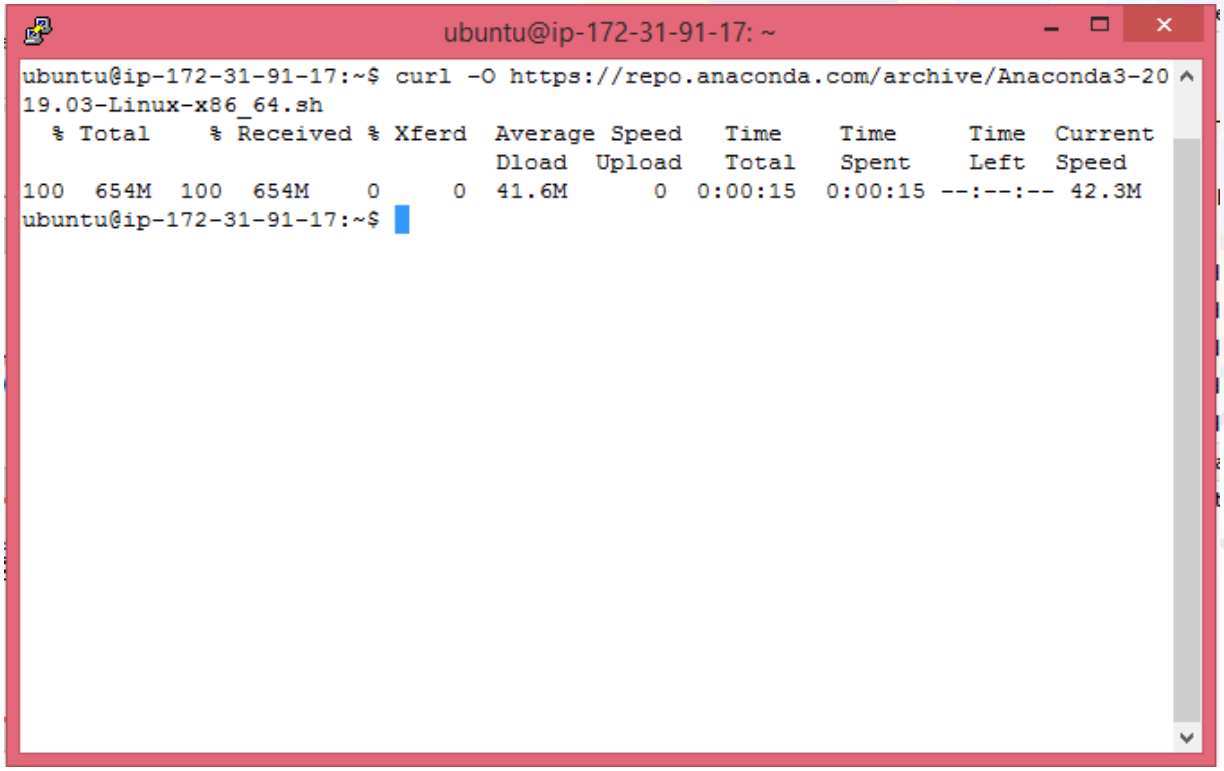
- Sesuaikan versi Anaconda dengan versi Anaconda yang anda pilih.
- Pada tutorial ini digunanakan Anaconda 3.7

- Lewati langkah ini apabila Anaconda sudah terinstall



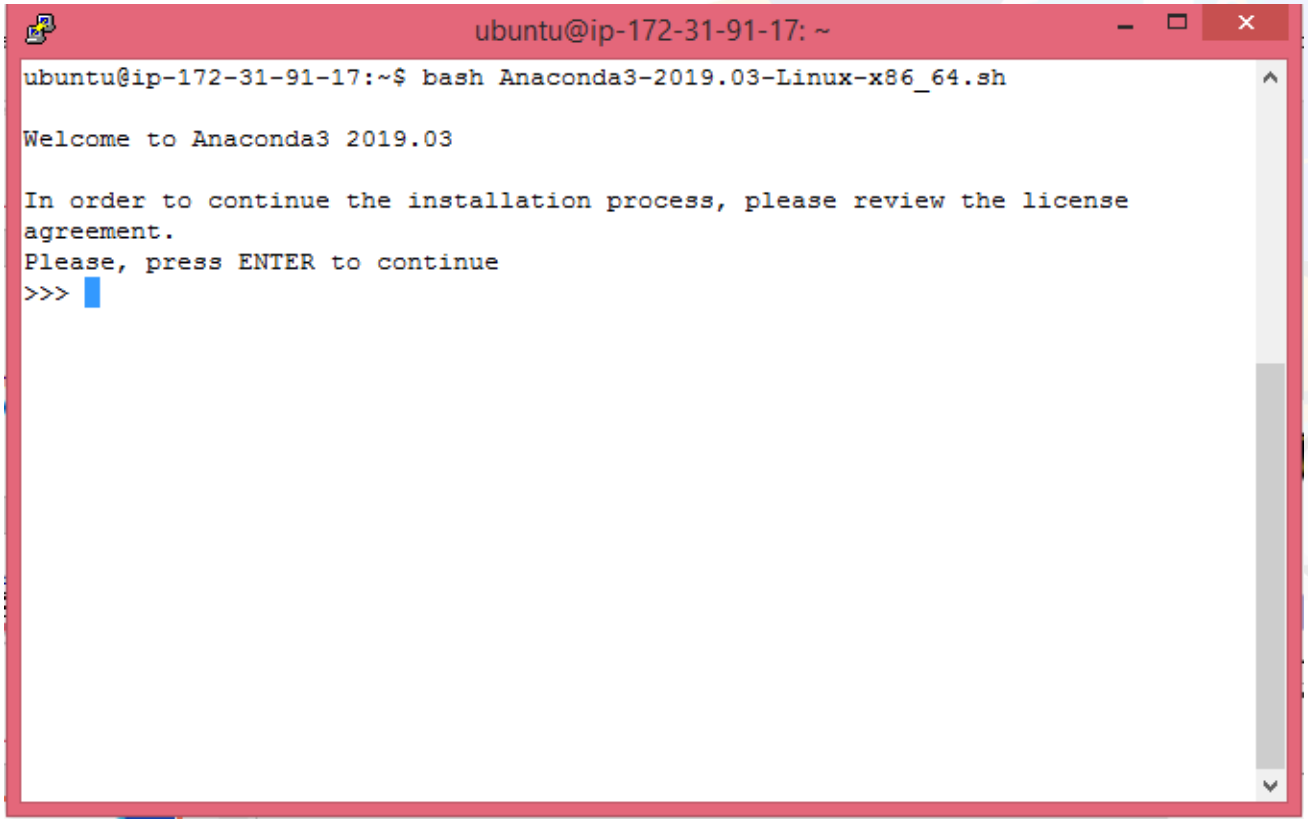
Download Anaconda

```
$ curl -o https://repo.anaconda.com/archive/Anaconda3-2019.03-Linux-x86_64.sh
```



Install Anaconda

```
$ bash Anaconda3-2019.03-Linux-x86_64.sh
```

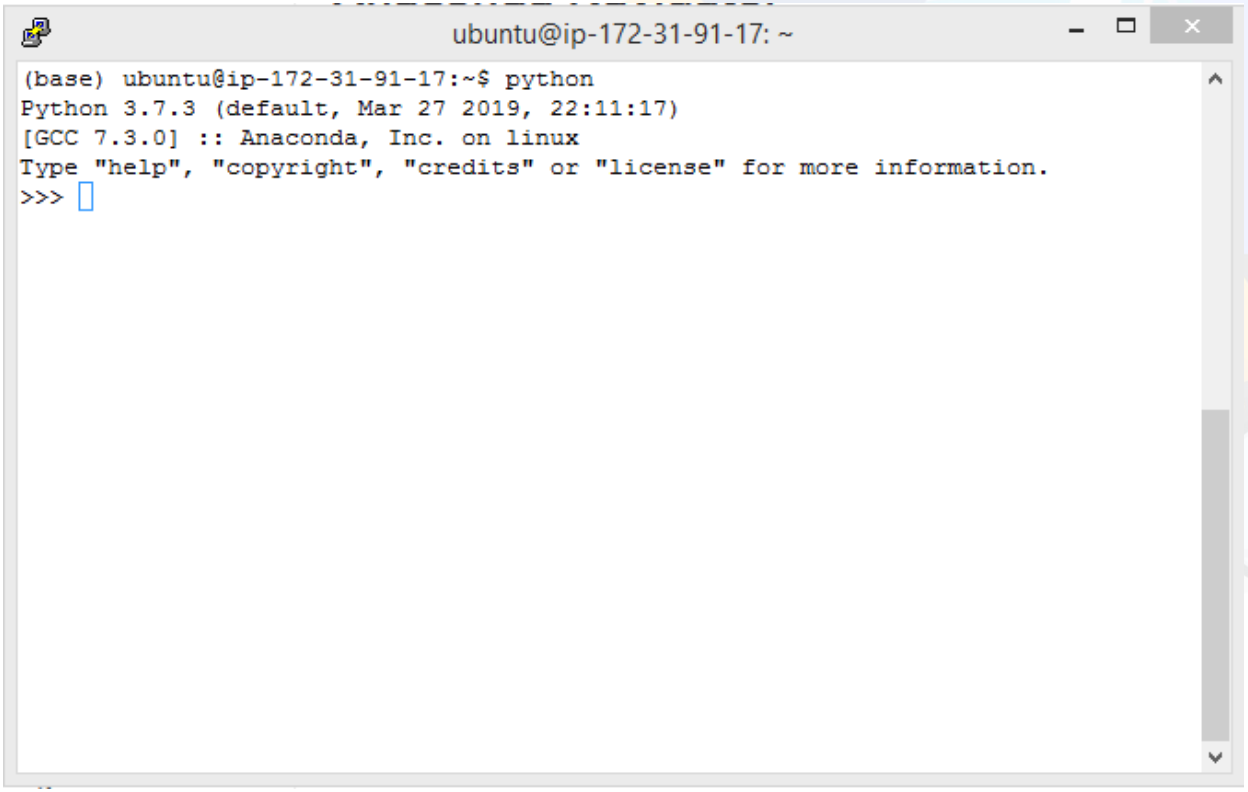


Instalasi Anaconda (cont)

- Sesuaikan konfigurasi instalasi Anaconda anda dengan preferensi anda.
- Pada tutorial ini, Anaconda dipasang pada direktori `/home/ubuntu/anaconda3`
- Jalankan `$conda init` pada direktori `anaconda3/bin` untuk menjalankan inisialisasi Anaconda apabila proses instalasi tidak menjalankan.
- Jalankan `$source ~/.bashrc` jika tidak dapat membuka Anaconda atau python

Python Console

```
$ python
```



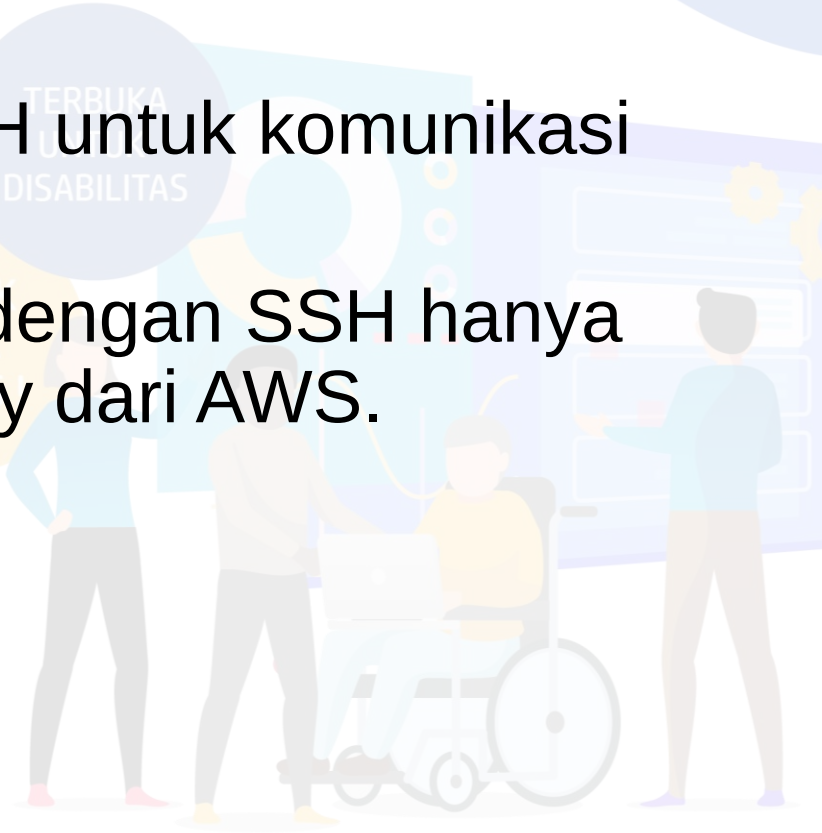
Konfigurasi SSH

Big Data

Hadoop

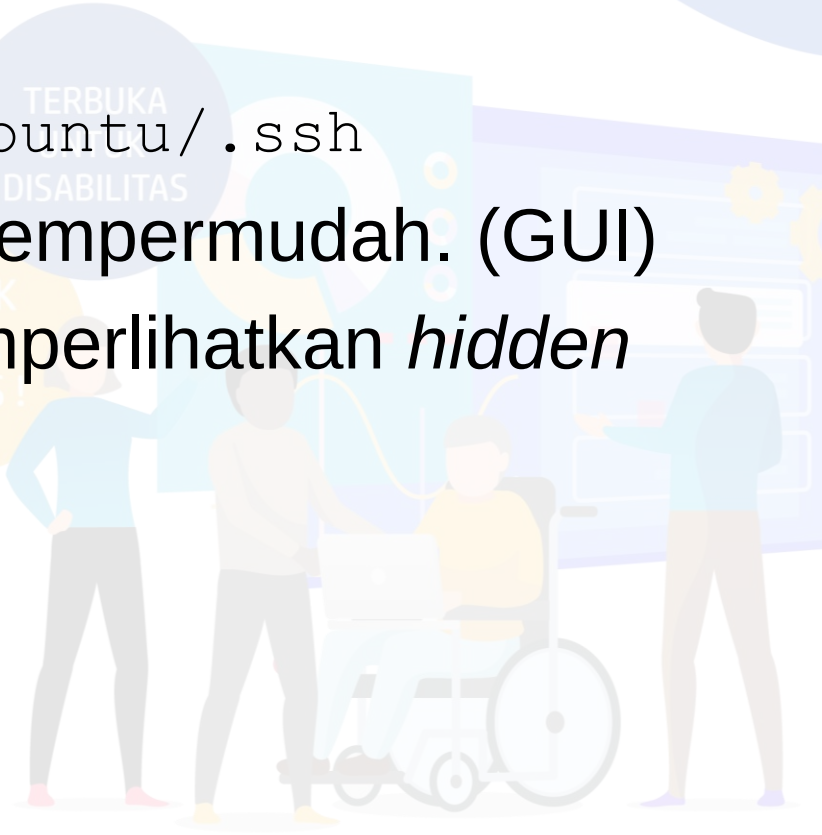
Why?

- Hadoop menggunakan SSH untuk komunikasi antar node.
- Mengakses Instance EC2 dengan SSH hanya dapat dilakukan dengan key dari AWS.

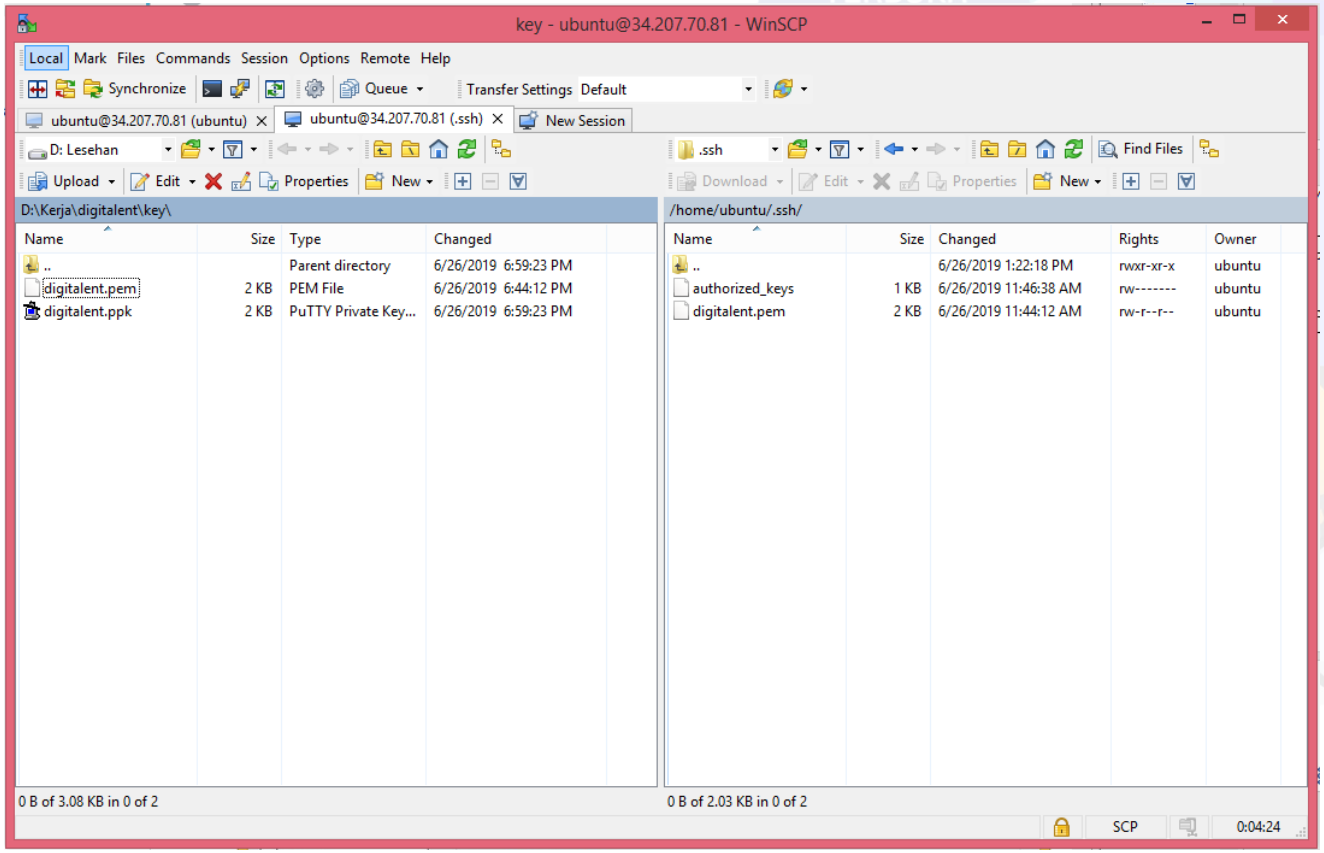


.pem Key

- Salin .pem key ke `/home/ubuntu/.ssh`
- Gunakan WinSCP untuk mempermudah. (GUI)
- Tekan ctrl+alt+h untuk memperlihatkan *hidden files.*

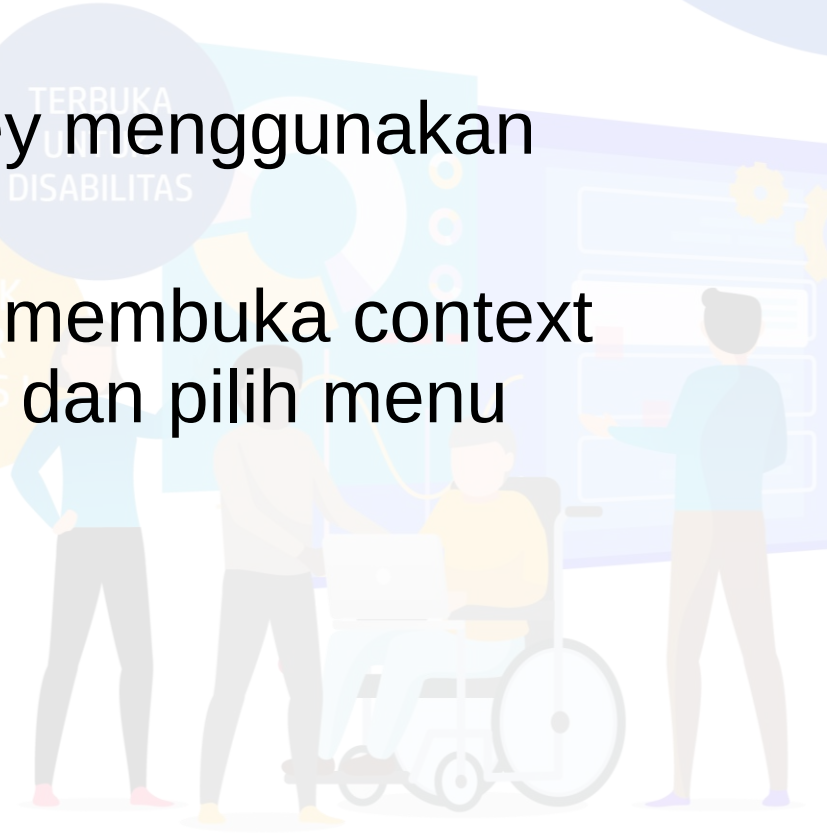


.pem Key (cont)

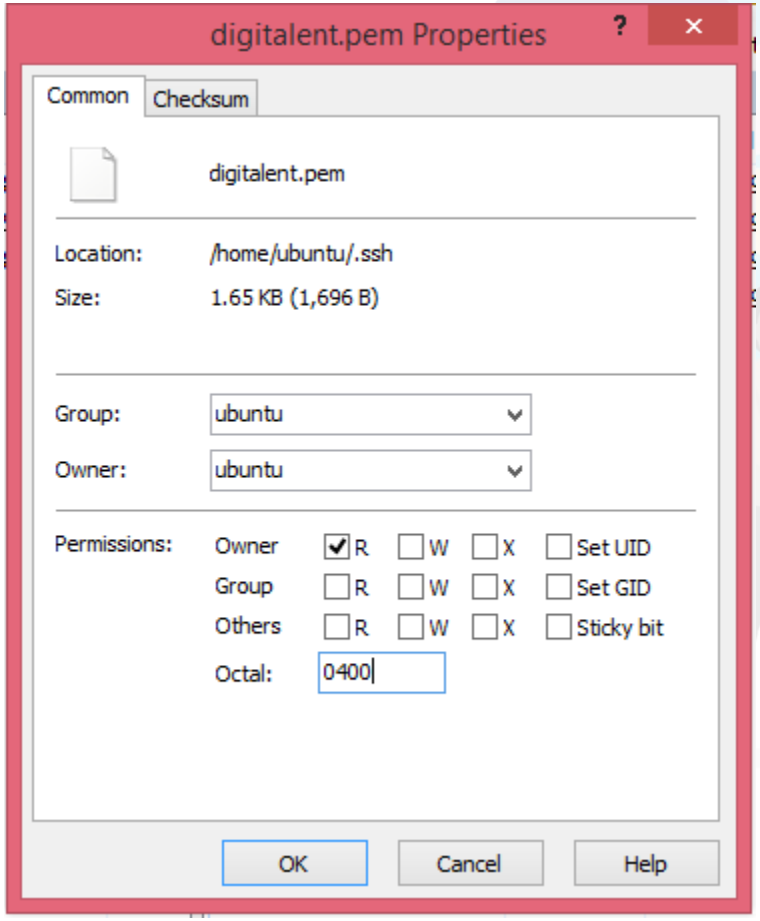

key - ubuntu@34.207.70.81 - WinSCP
Local Mark Files Commands Session Options Remote Help
Synchronize Queue Transfer Settings Default
ubuntu@34.207.70.81 (ubuntu)
ubuntu@34.207.70.81 (.ssh)
New Session
D: Lesehan
Upload Edit Properties New
D:\Kerja\digitalent\key\
Name Size Type Changed
.. Parent directory 6/26/2019 6:59:23 PM
digitalent.pem 2 KB PEM File 6/26/2019 6:44:12 PM
digitalent.ppk 2 KB PuTTY Private Key... 6/26/2019 6:59:23 PM
0 B of 3.08 KB in 0 of 2
.ssh
Download Edit Properties New
Find Files
/home/ubuntu/.ssh/
Name Size Changed Rights Owner
.. 6/26/2019 1:22:18 PM rwxr-xr-x ubuntu
authorized_keys 1 KB 6/26/2019 11:46:38 AM rw------- ubuntu
digitalent.pem 2 KB 6/26/2019 11:44:12 AM rw-r--r-- ubuntu
0 B of 2.03 KB in 0 of 2
SCP
0:04:24

.pem Key (cont)

- Rubah *Permission* .pem key menggunakan Octal 0400.
- Gunakan WinSCP dengan membuka context menu (klik kanan pada file) dan pilih menu properties.



.pem Key (cont)

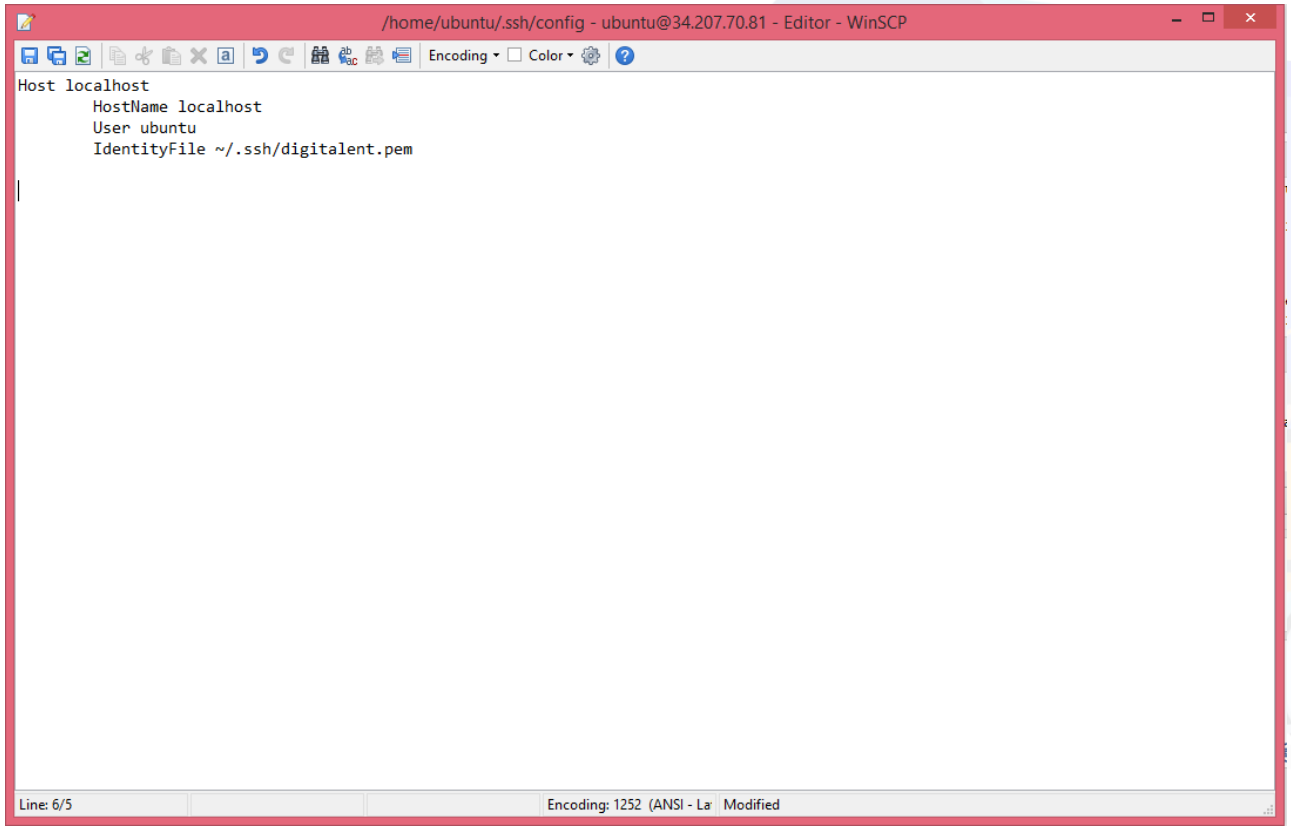

digitalent.pem Properties
Common
Checksum
digitalent.pem
Location: /home/ubuntu/.ssh
Size: 1.65 KB (1,696 B)
Group: ubuntu
Owner: ubuntu
Permissions: Owner R W X Set UID
Group R W X Set GID
Others R W X Sticky bit
Octal: 0400
OK
Cancel
Help

SSH Config

- Buat file pada direktori `/home/ubuntu/.ssh` dengan nama `config`
- Tambahkan string berikut ke file `config` dan ganti key dengan nama key anda

```
Host localhost
        HostName localhost
        User ubuntu
        IdentityFile ~/.ssh/digitalent.pem
```

SSH Config


/home/ubuntu/.ssh/config - ubuntu@34.207.70.81 - Editor - WinSCP
Encoding
Color
Host localhost
HostName localhost
User ubuntu
IdentityFile ~/.ssh/digitalent.pem
Line: 6/5
Encoding: 1252 (ANSI - La
Modified

SSH

- Buat koneksi SSH ke localhost untuk memeriksa.

```
$ ssh localhost
```

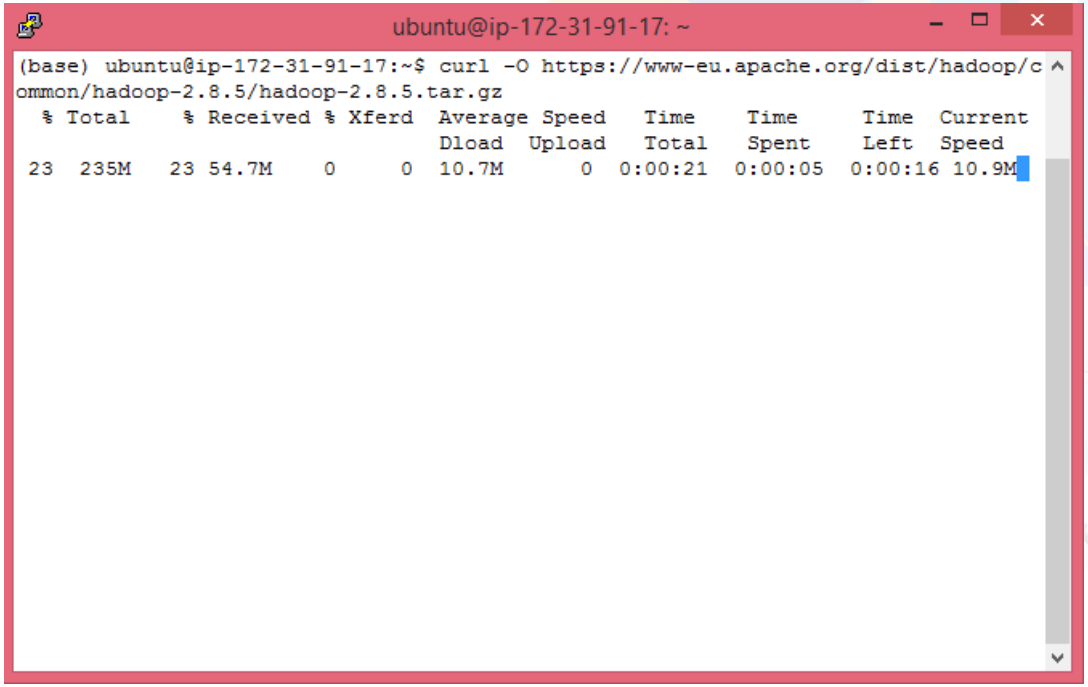
Instalasi Hadoop Single Node

Big Data

Hadoop

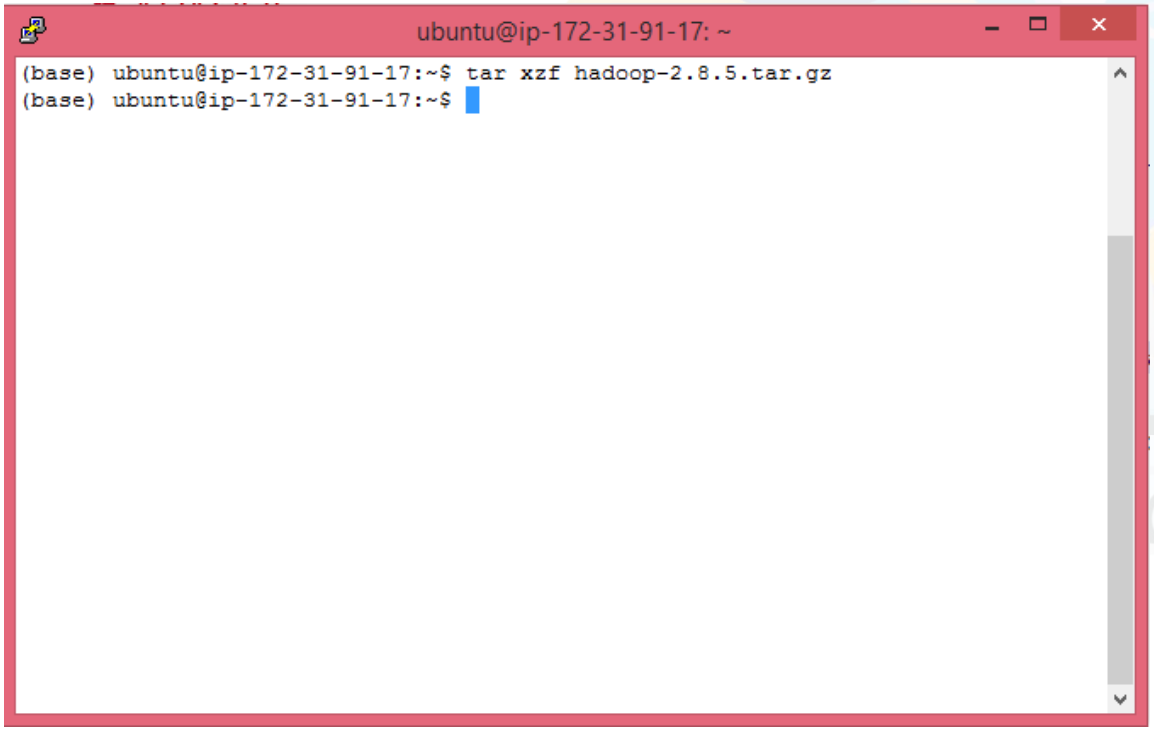
Download Hadoop

```
$ curl -O https://www-eu.apache.org/dist/hadoop/common/hadoop-2.8.5/hadoop-2.8.5.tar.gz
```



Extract Hadoop

```
$ tar xzf hadoop-2.8.5.tar.gz
```



Env Variabel Hadoop

Tambahkan string berikut ke `.bashrc` atau `.profile` (Gunakan WinSCP)

```
export HADOOP_HOME=/home/ubuntu/hadoop-2.8.5
export
LD_LIBRARY_PATH=$HADOOP_HOME/lib/native/:$LD_L
IBRARY_PATH
export
HADOOP_JAR=$HADOOP_HOME/share/hadoop/mapreduce
PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:
$HADOOP_JAR
```

Env Variabel Hadoop

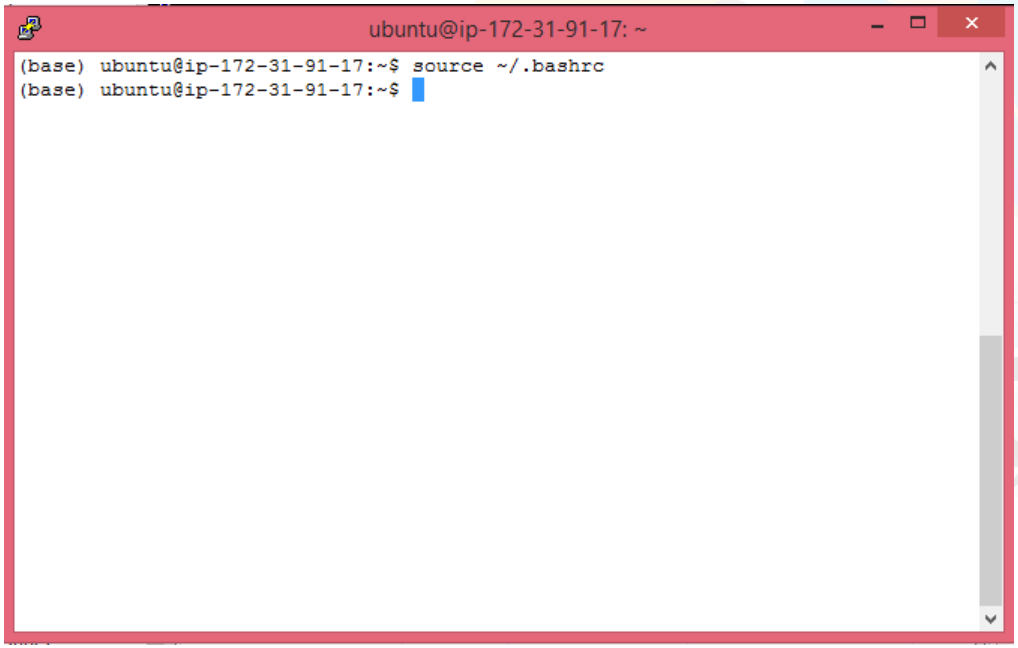
/home/ubuntu/.bashrc - ubuntu@34.207.70.81 - Editor - WinSCP

```
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export JAVA_HOME=/usr/lib/jvm/java/java-8-openjdk-amd64
PATH=$PATH:$JAVA_HOME:$JAVA_HOME/jre/bin


# >>> conda initialize >>>
# !! Contents within this block are managed by 'conda init' !!
__conda_setup="$('/home/ubuntu/anaconda3/bin/conda' 'shell.bash' 'hook' 2> /dev/null)"
if [ $? -eq 0 ]; then
    eval "$__conda_setup"
else
    if [ -f "/home/ubuntu/anaconda3/etc/profile.d/conda.sh" ]; then
        . "/home/ubuntu/anaconda3/etc/profile.d/conda.sh"
    else
        export PATH="/home/ubuntu/anaconda3/bin:$PATH"
    fi
fi
unset __conda_setup
# <<< conda initialize <<<

export HADOOP_HOME=/home/ubuntu/hadoop-2.8.5
export LD_LIBRARY_PATH=$HADOOP_HOME/lib/native/:$LD_LIBRARY_PATH
export HADOOP_JAR=$HADOOP_HOME/share/hadoop/mapreduce
PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HADOOP_JAR
```

Line: 142/143 | Column: 1 | Encoding: 1252 (ANSI - La | Saving...

Env Variabel Hadoop

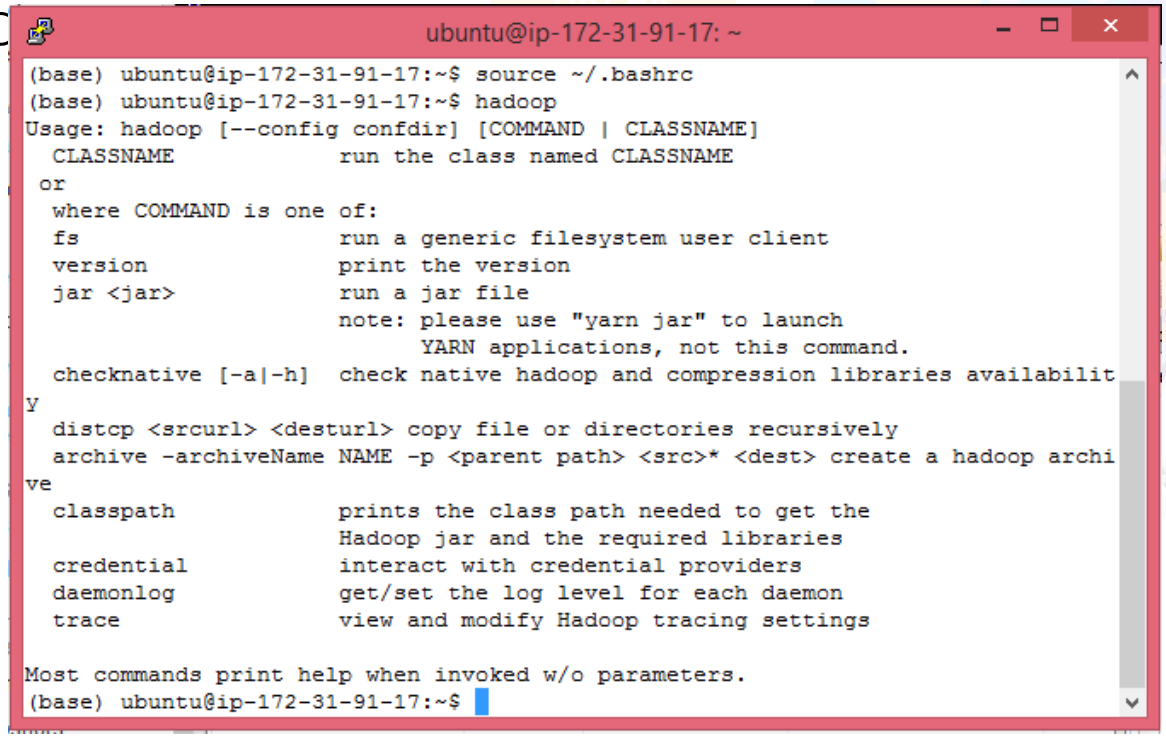
Jalanakan `$source ~/.bashrc` atau `source ~/.profile`



Env Variabel Hadoop

Coba instalasi hadoop dengan menjalankan perintah

```
$ hadoo
```



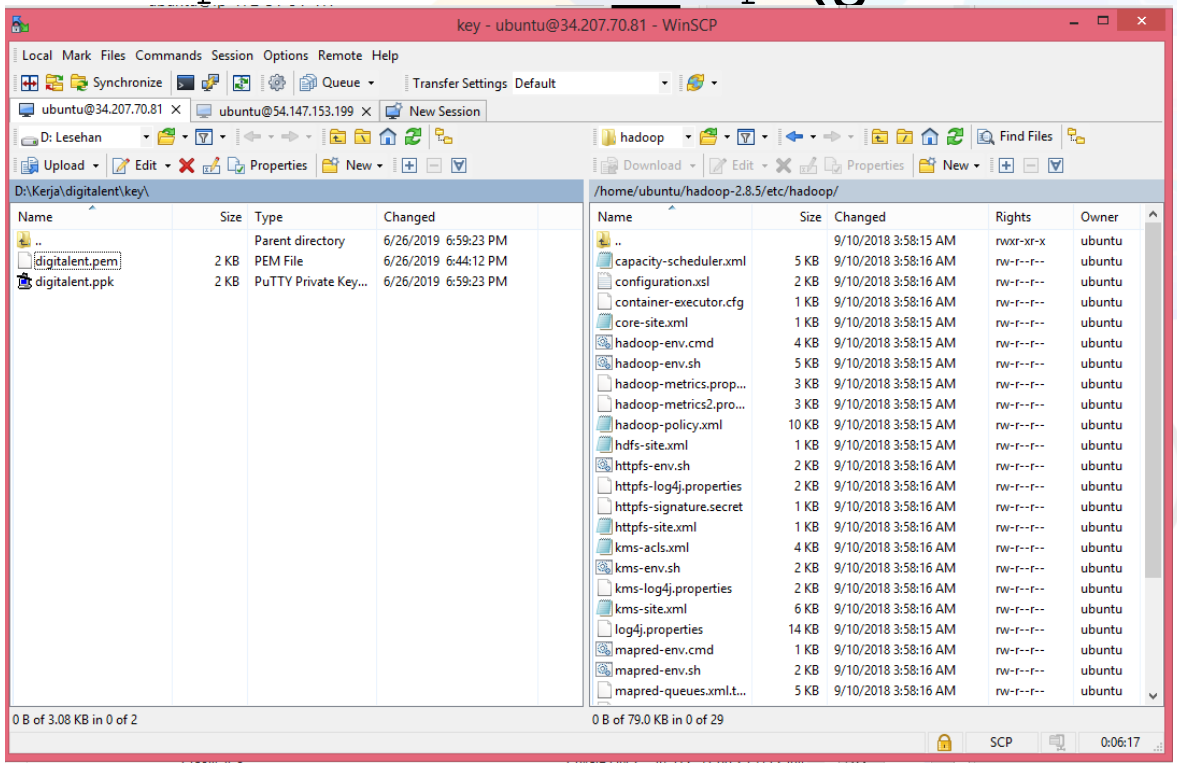
Konfigurasi Hadoop Single Node

Big Data

Hadoop

Konfigurasi Hadoop

Buka direktori konfigurasi hadoop pada `/home/ubuntu/hadoop-2.8.5/etc/hadoop` (gunakan WinSCP)

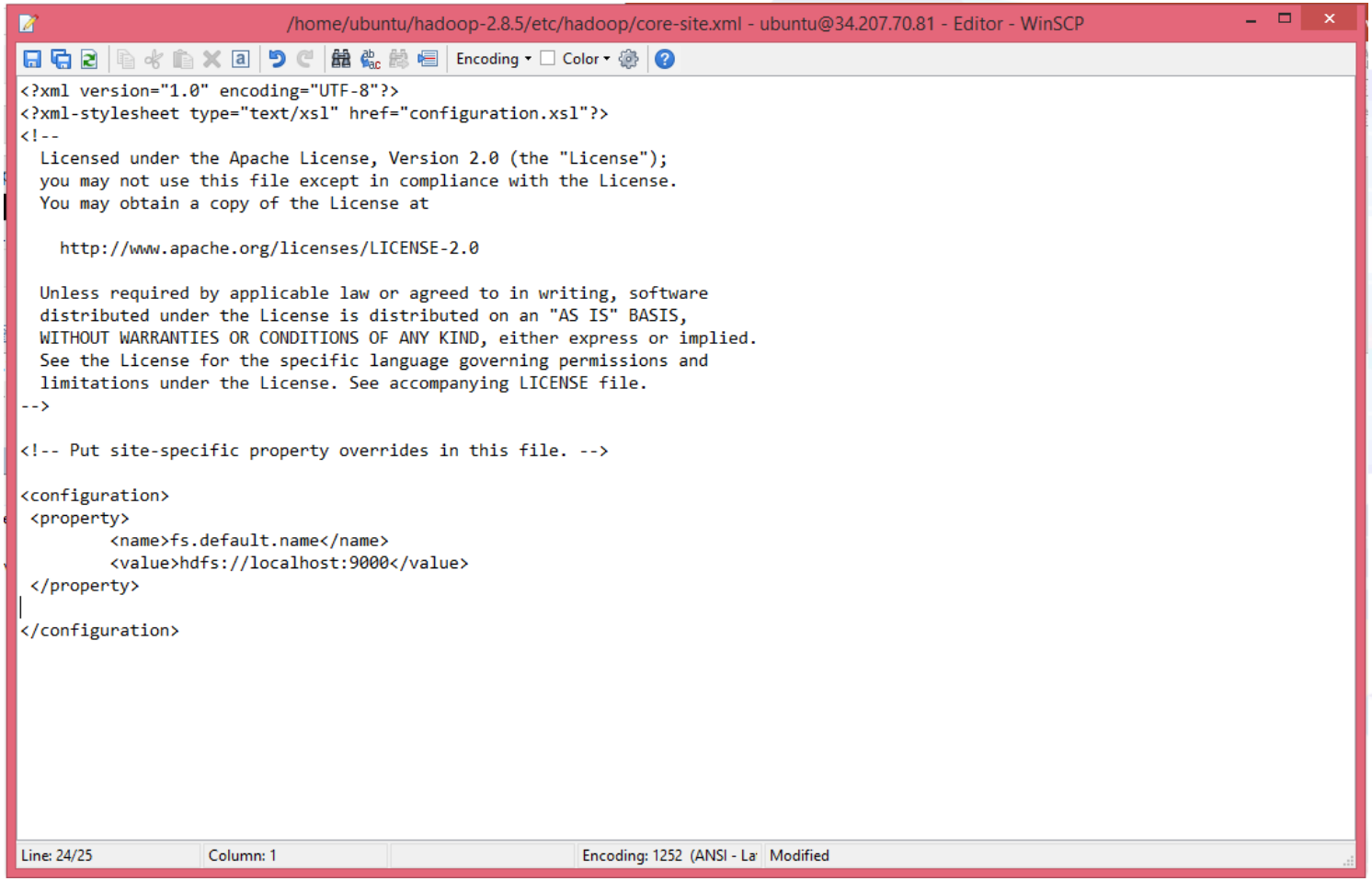


core-site.xml

Buka file `core-site.xml` dan tambahkan string berikut pada tag `<configuration>`

```
<property>
        <name>fs.default.name</name>
        <value>hdfs://localhost:8020</value>
</property>
```

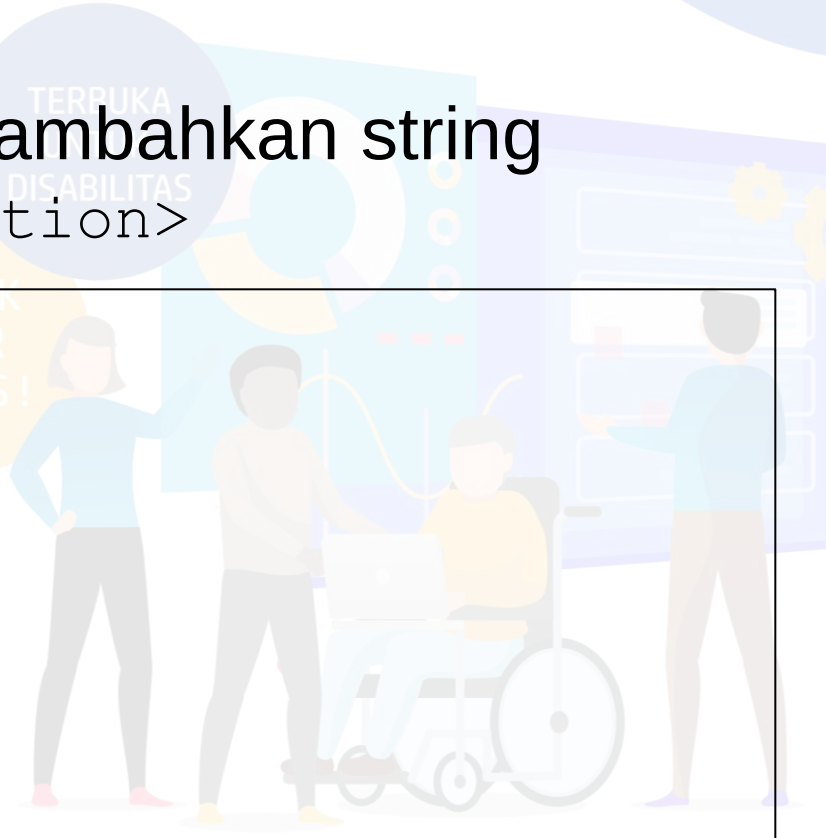
core-site.xml



core-site.xml

Buka file `hdfs-site.xml` dan tambahkan string berikut pada tag `<configuration>`

```
<property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/ubuntu/data/nameNode</value>
 </property>
 <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/ubuntu/data/dataNode</value>
 </property>
 <property>
        <name>dfs.replication</name>
        <value>2</value>
 </property>
```



hdfs-site.xml

/home/ubuntu/hadoop-2.8.5/etc/hadoop/hdfs-site.xml - ubuntu@34.207.70.81 - Editor - WinSCP

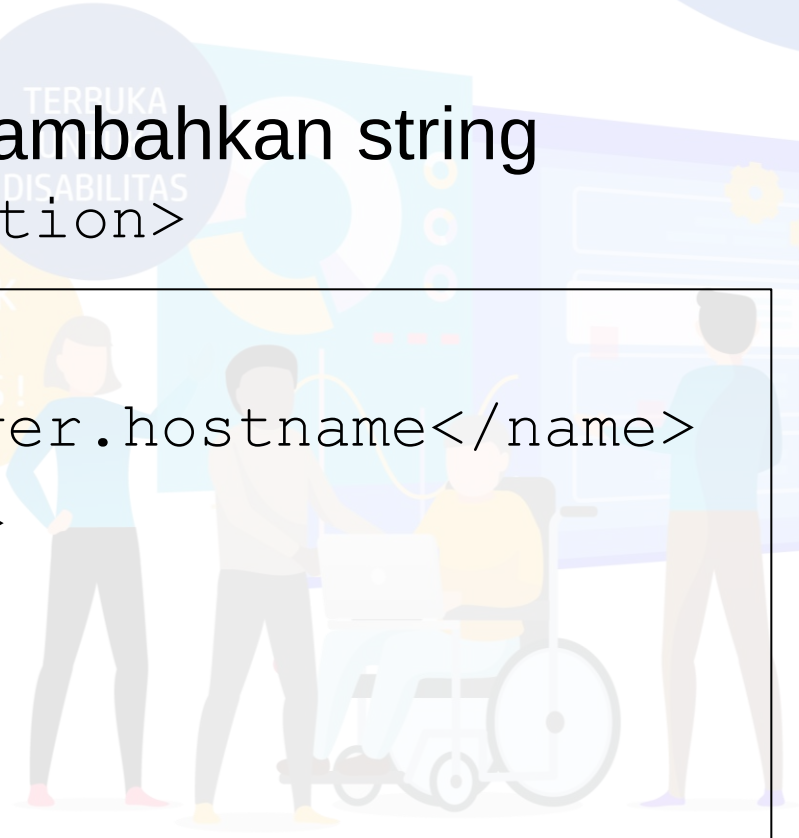
Encoding Color

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>/home/ubuntu/data/nameNode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/home/ubuntu/data/dataNode</value>
        </property>
        <property>
                <name>dfs.replication</name>
                <value>2</value>
        </property>

</configuration>
```

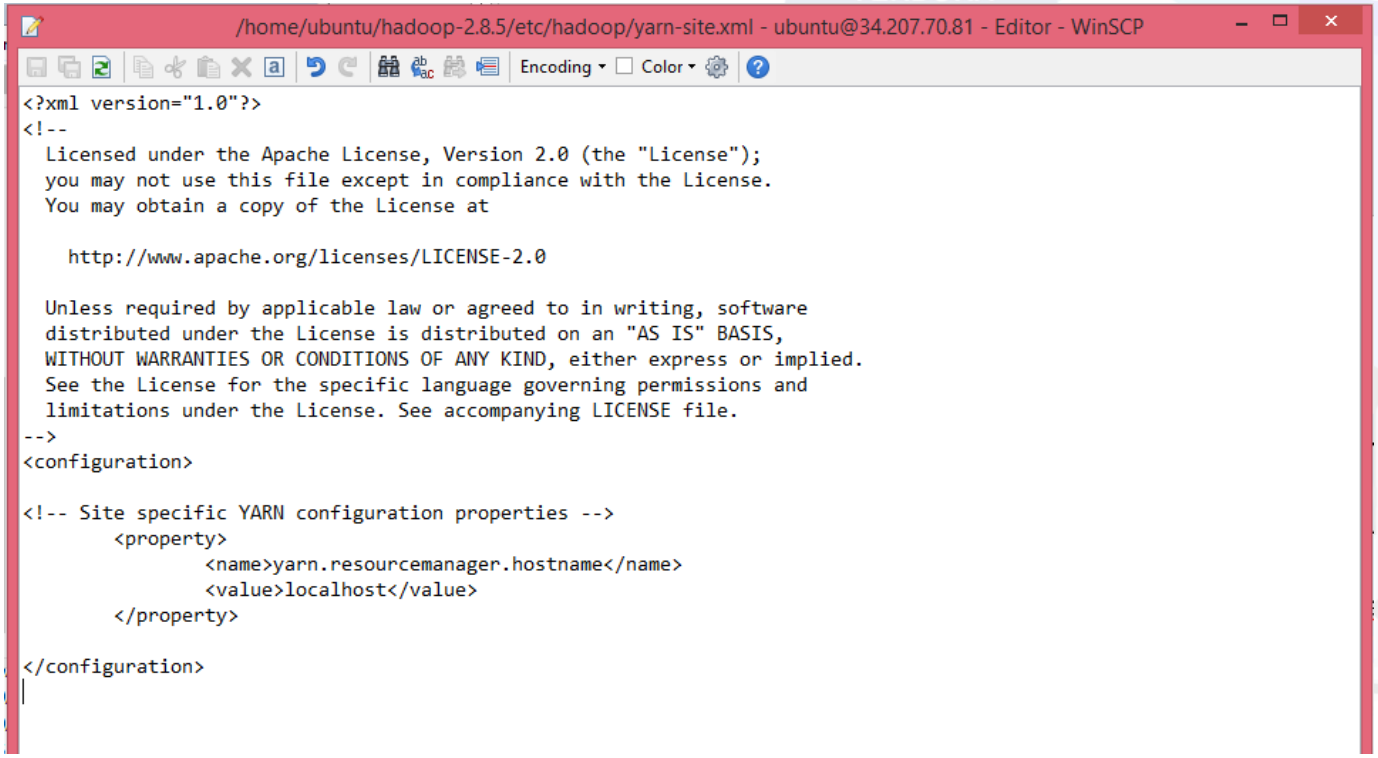
yarn-site.xml

Buka file `yarn-site.xml` dan tambahkan string berikut pada tag `<configuration>`

```
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>localhost</value>
</property>
```



yarn-site.xml




```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>localhost</value>
        </property>

</configuration>
```

mapred-site.xml

copy file `mapred-site.xml.template` ke `mapred-site.xml` dan tambahkan string berikut pada tag `<configuration>`

```
<property>
        <name>mapreduce.framework.name</name>
        <value>local</value>
</property>
```

mapred-site.xml

/home/ubuntu/hadoop-2.8.5/etc/hadoop/mapred-site.xml - ubuntu@54.147.153.199 - Editor - WinSCP

Encoding Color

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>local</value>
        </property>

</configuration>
```

mapred-site.xml

Apabila anda mengkonfigurasikan Hadoop multinode maka gunakan nilai yarn pada property `mapreduce.framework.name` sehingga YARN digunakan sebagai resource manager.

```
<property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
</property>
```

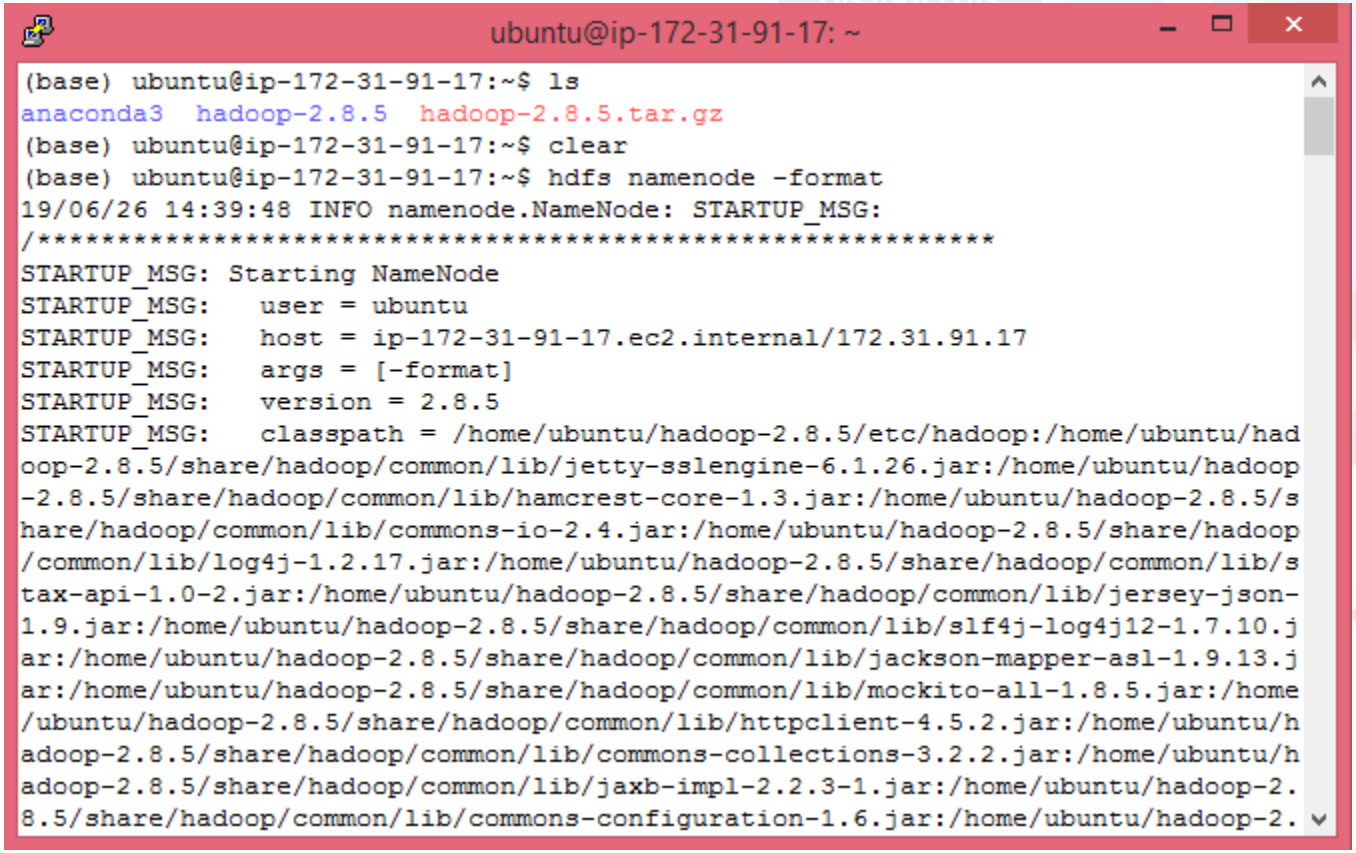
HDFS

Big Data

Hadoop

Format HDFS

```
$ hdfs namenode -format
```



Format HDFS

- Pastikan direktori data telah terbentuk pada /home/ubuntu

/home/ubuntu/

Name	Size	Changed	Rights	Owner
..		6/26/2019 11:46:38 AM	rwxr-xr-x	root
.cache		6/26/2019 12:07:16 PM	rwx------	ubuntu
.nano		6/26/2019 12:44:33 PM	rwxrwxr-x	ubuntu
.ssh		6/26/2019 1:32:20 PM	rwx------	ubuntu
anaconda3		6/26/2019 1:04:07 PM	rwxrwxr-x	ubuntu
data		6/26/2019 2:39:49 PM	rwxrwxr-x	ubuntu
hadoop-2.8.5		9/10/2018 3:58:16 AM	rwxr-xr-x	ubuntu
.bash_history	2 KB	6/26/2019 2:36:40 PM	rw-------	ubuntu
.bash_logout	1 KB	8/31/2015 11:27:45 PM	rw-r--r--	ubuntu
.bashrc	5 KB	6/26/2019 2:37:49 PM	rw-r--r--	ubuntu
.profile	1 KB	5/16/2017 12:49:38 PM	rw-r--r--	ubuntu
.python_history	0 KB	6/26/2019 1:12:07 PM	rw-------	ubuntu
.sudo_as_admin_succ...	0 KB	6/26/2019 12:22:55 PM	rw-r--r--	ubuntu
hadoop-2.8.5.tar.gz	240,766 KB	6/26/2019 2:00:16 PM	rw-rw-r--	ubuntu

HDFS : Start Service

- Jalankan service HDFS dengan menjalankan file `start-dfs.sh`

```
$ start-dfs.sh
```

HDFS : Membuat Direktori

- Buat direktori user anda dengan perintah berikut.

```
$ hdfs dfs –mkdir –p /user/ubuntu
```

Buat direktori bernama `input`

```
$ hdfs dfs –mkdir input
```

Melihat direktori

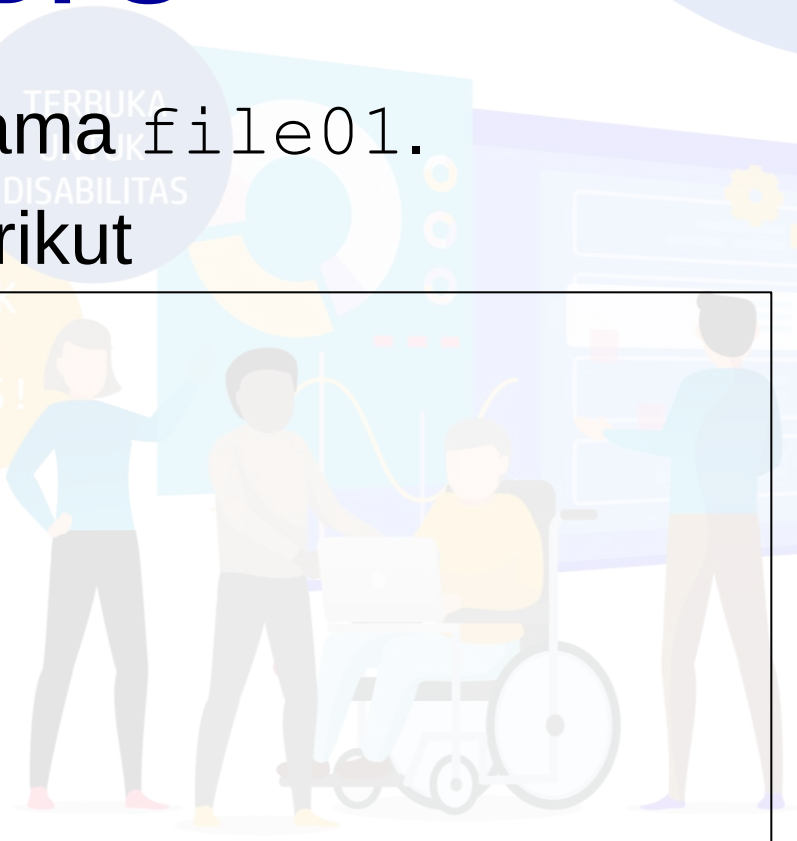
```
$ hdfs dfs –ls
```

HDFS : Membuat Direktori

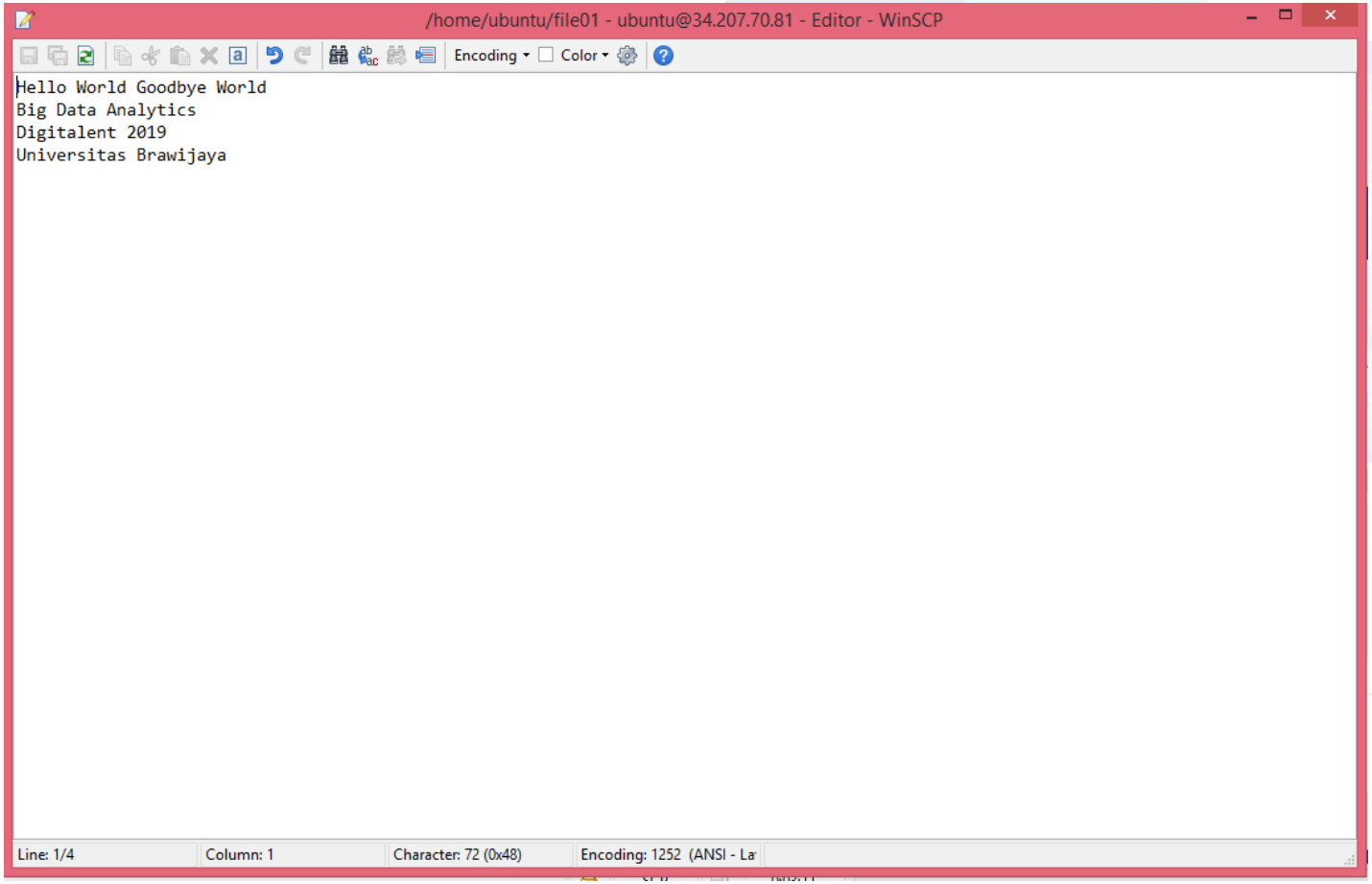
Mengunggah ke HDFS

- Buat sebuah file dengan nama `file01`.
- Isikan file dengan string berikut

```
Hello World Goodbye World
Big Data Analytics
Digitalent 2019
Universitas Brawijaya
```



Mengunggah ke HDFS



Mengunggah ke HDFS

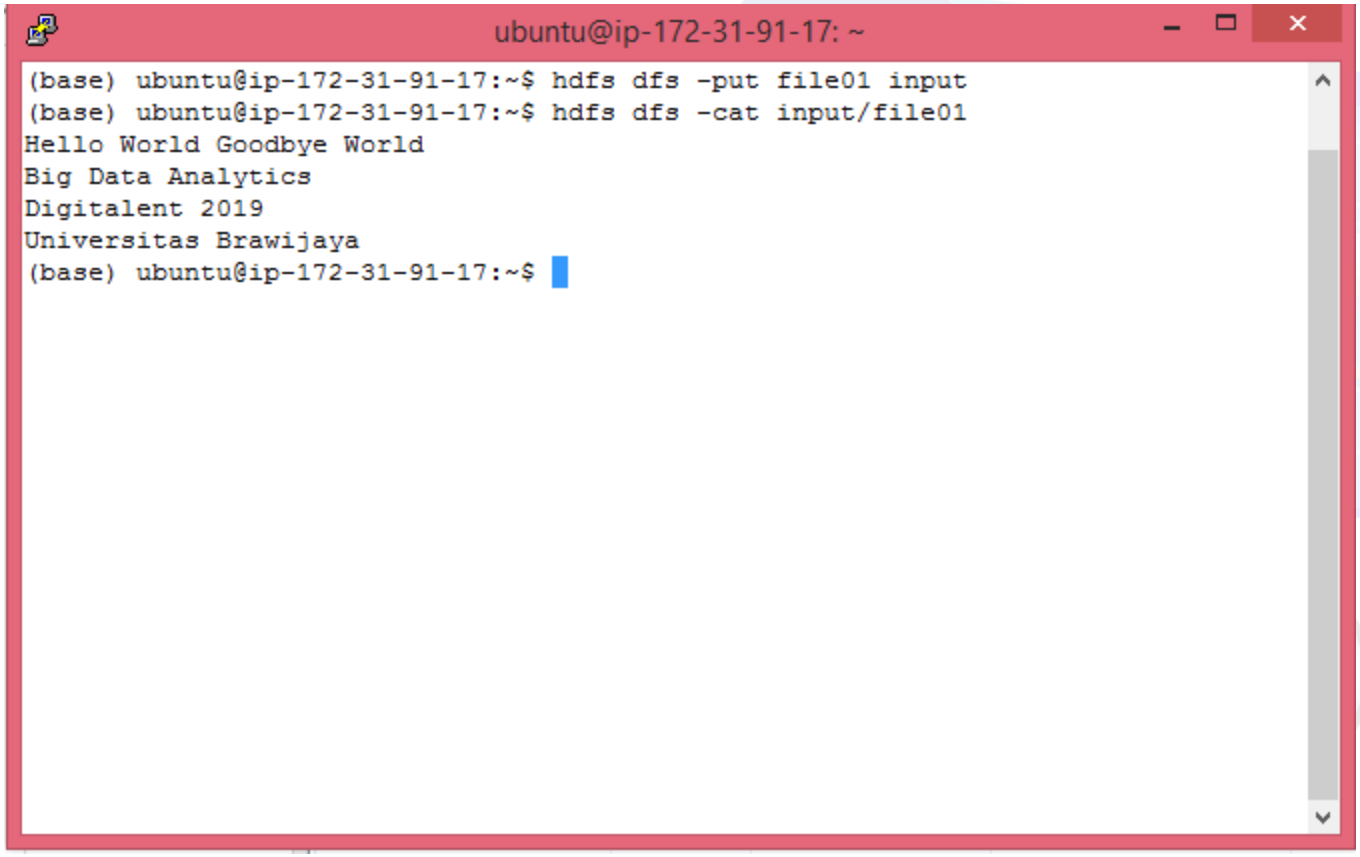
- Unggah file `file01` ke direktori input pada `hdfs`

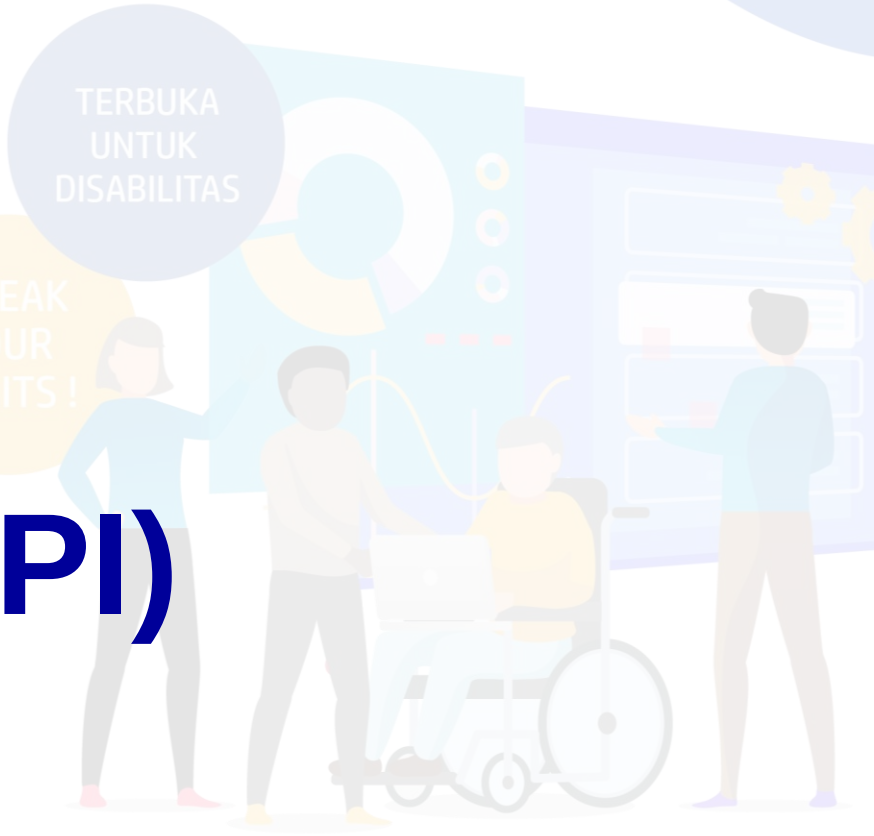
```
$ hdfs dfs -put file01 input
```

- Mecetak `file01` ke terminal

```
$ hdfs dfs -cat input/file01
```

Mengunggah ke HDFS





MapReduce (Streaming API)

Big Data

Hadoop

Python Wordcount

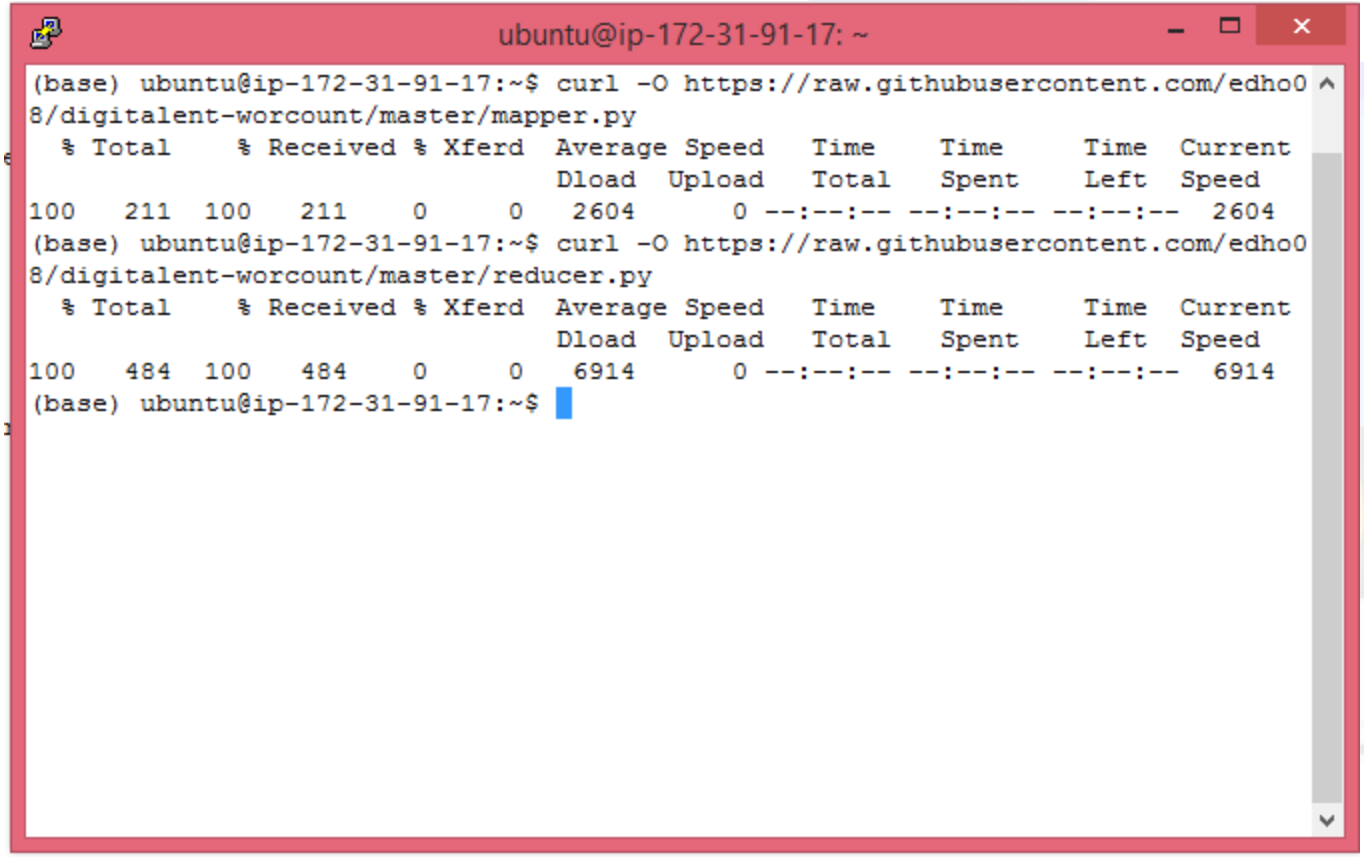
- Jalankan perintah berikut untuk mengunduh mapper program wordcount python menggunakan hadoop streaming api

```
$ curl -O
https://raw.githubusercontent.com/edho08/digitalent-worcount/master/mapper.py
```

- Jalankan perintah berikut untuk mengunduh reducer

```
$ curl -O
https://raw.githubusercontent.com/edho08/digitalent-worcount/master/reducer.py
```

Python Wordcount

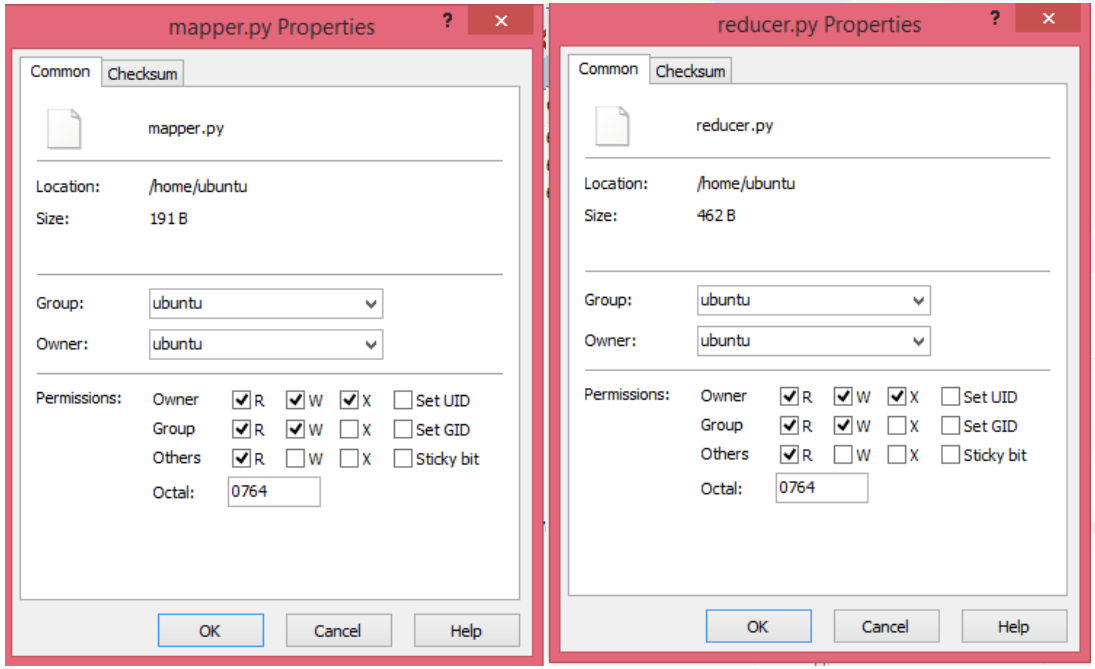


Python Wordcount

- Pastikan terdapat file `mapper.py` dan `reducer.py` dengan menggunakan perintah `$ ls`

Python Wordcount

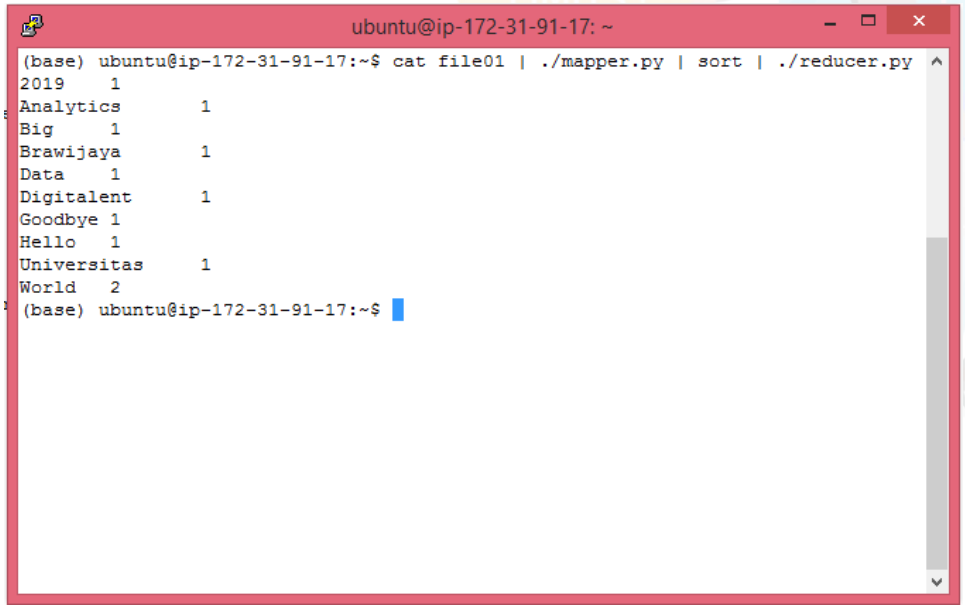
- Tambahkan ijin eksekusi pada kedua file. (gunakan WinSCP)



Python Wordcount

- Jalankan program secara lokal tanpa hadoop.

```
$ cat file01 | ./mapper.py | sort | ./reducer.py
```

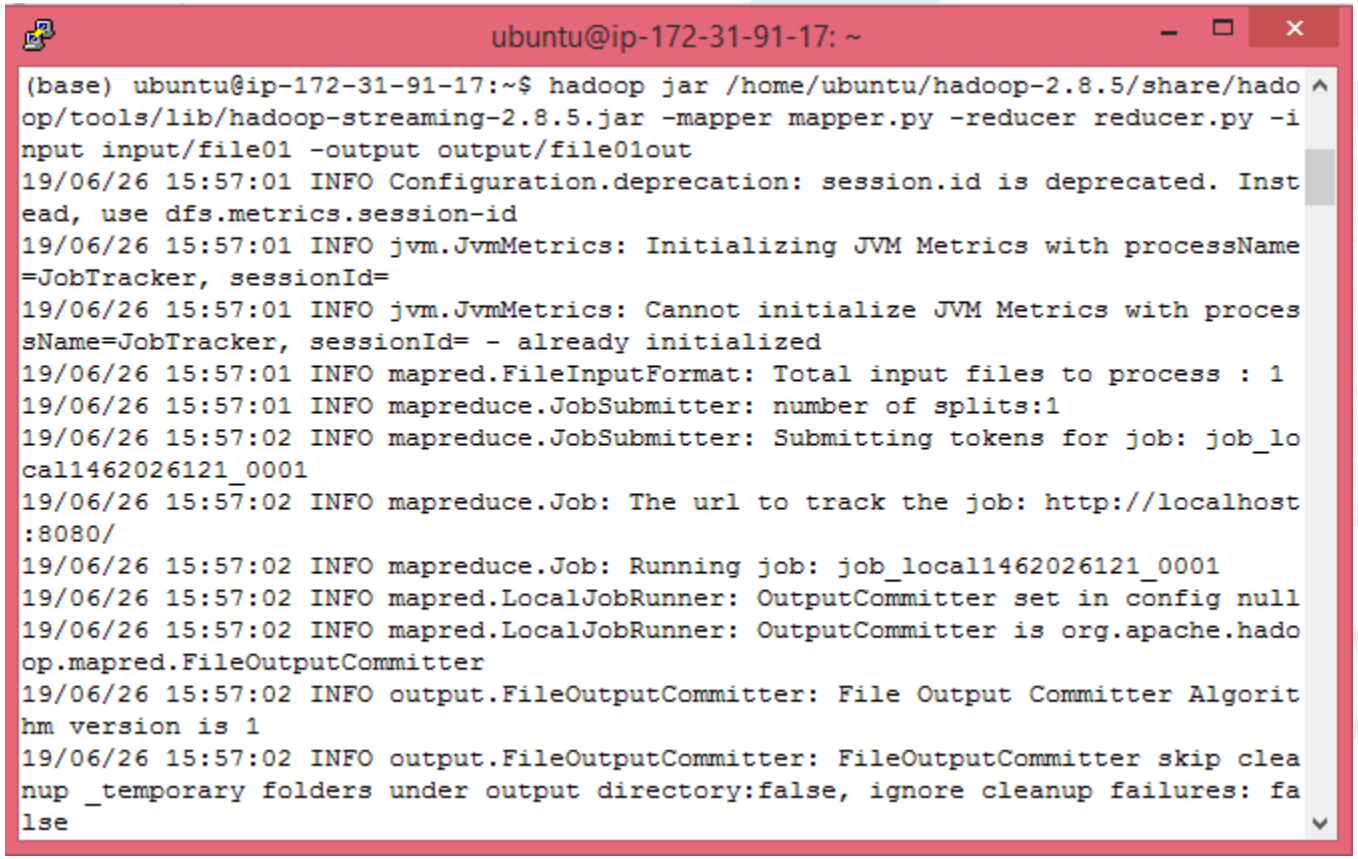


Python Wordcount

- Jalankan program dengan hadoop

```
$ hadoop jar /home/ubuntu/hadoop-
2.8.5/share/hadoop/tools/lib/hadoop-
streaming-2.8.5.jar \
-mapper mapper.py \
-reducer reducer.py \
-input input/file01 \
-output output/file01out
```

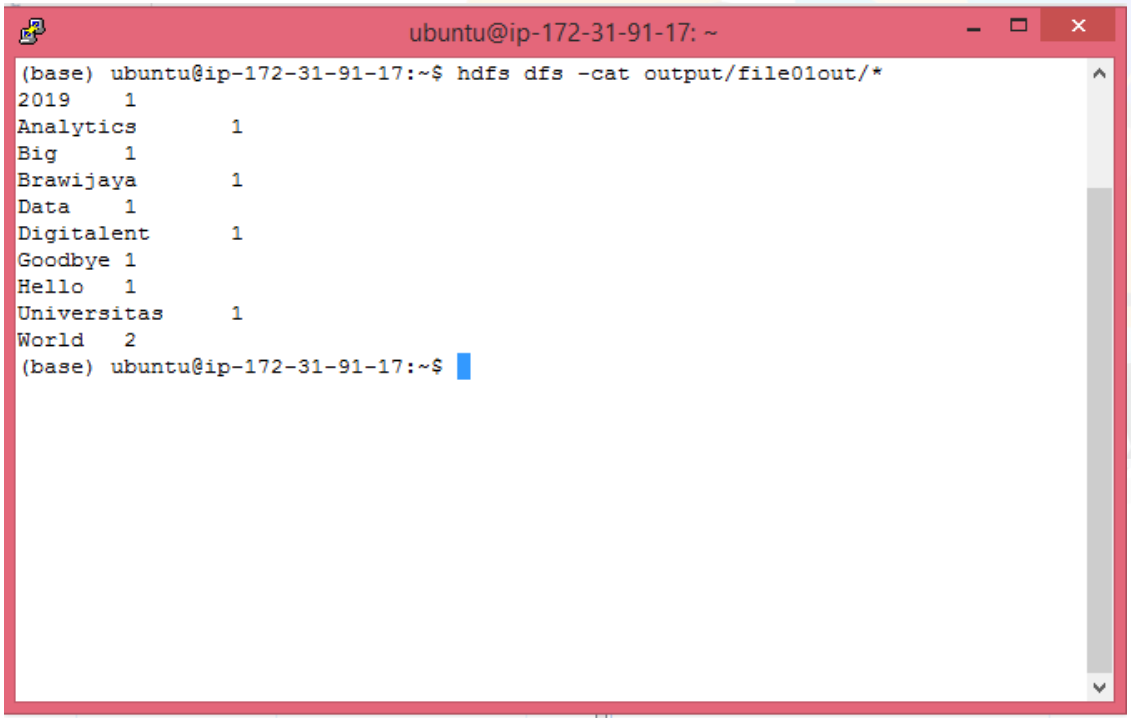
Python Wordcount



Python Wordcount

- Cetak file output

```
$ hdfs dfs –cat output/file01out/*
```



IKUTI KAMI

DIGITAL TALENT SCHOLARSHIP

digitalent.kominfo
digitalent.kominfo
DTS_kominfo
Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110