

10-10-2023

Hadoop 3.1 installation.

```
sudo apt update
```

```
sudo apt install openjdk-17-jdk -y
```

```
ssh-keygen
```

```
cat id_rsa.pub >> authorized_keys
```

Hadoop 3 setup

```
sudo apt update
```

```
sudo apt install openjdk-17-jdk -y
```

```
ssh-keygen
```

```
cat id_rsa.pub >> authorized_keys
```

```
wget
```

```
https://archive.apache.org/dist/hadoop/core/hadoop-3.3.1/hadoop-3.3.1.tar.g
```

```
z
```

```
ubuntu@ip-172-31-17-223:~$ ls
authorized_keys  hadoop-3.3.1.tar.gz
```

```
tar -xvzf hadoop-3.3.1.tar.gz
```

```
nano .bashrc
```

```
export HADOOP_PREFIX=/usr/local/hadoop/  
export PATH=$PATH:$HADOOP_PREFIX/bin  
export HADOOP_HOME=/usr/local/hadoop/  
export PATH=$PATH:HADOOP_HOME/sbin  
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64  
export PATH=$PATH:$JAVA_HOME
```

core-site.xml

```
<configuration>  
  <property>  
    <name>fs.defaultFS</name>  
    <value>hdfs://localhost:9000</value>  
  </property>  
</configuration>
```

hdfs-site.xml

```
<configuration>  
  <property>  
    <name>dfs.replication</name>  
    <value>1</value>  
  </property>  
</configuration>
```

hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64  
hdfs namenode -format
```

start-dfs.sh

```
authorized_keys  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-17-223:~$ sbin/start-dfs.sh
-bash: sbin/start-dfs.sh: No such file or directory
ubuntu@ip-172-31-17-223:~$ cd /usr/local/hadoop/
ubuntu@ip-172-31-17-223:/usr/local/hadoop$ sbin/start-dfs.sh
```

```
ubuntu@ip-172-31-17-223:~$ cd /usr/local/hadoop/share/hadoop/mapreduce/
ubuntu@ip-172-31-17-223:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar had
oop-mapreduce-examples-3.3.1.jar wordcount input output
```

download sample file - wget

<https://raw.githubusercontent.com/ErikSchierboom/sentencegenerator/master/samples/the-king-james-bible.txt> > sample.txt

hdfs dfs -mkdir -p "/user/ubuntu"

hdfs dfs -mkdir input

hdfs dfs -put sample.txt input

cd /usr/local/hadoop/share/hadoop/mapreduce/

```
ubuntu@ip-172-31-17-223:~$ cd /usr/local/hadoop/share/hadoop/mapreduce/
ubuntu@ip-172-31-17-223:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar had
oop-mapreduce-examples-3.3.1.jar wordcount input output
```

hdfs dfs -ls output

hdfs dfs -tail output/part-r-00000 | tail

```
at java.base/java.lang.reflect.Method.invoke(Method.java:568)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
ubuntu@ip-172-31-17-223:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -ls o
tput
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_P
REFIX.
Found 2 items
-rw-r--r--  3 ubuntu supergroup          0 2023-10-09 11:19 output/_SUCCESS
-rw-r--r--  3 ubuntu supergroup          0 2023-10-09 11:19 output/part-r-00000
ubuntu@ip-172-31-17-223:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -tail
output/part-r-00000 | tail
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_P
REFIX.
ubuntu@ip-172-31-17-223:/usr/local/hadoop/share/hadoop/mapreduce$
```

Setup jupyterlab on the instance

Sudo apt install python3 pip

```
ubuntu@ip-172-31-17-223: ~  
hadoop-metrics2.properties      mapred-site.xml  
hadoop-policy.xml              shellprofile.d  
hadoop-user-functions.sh.example  ssl-client.xml.example  
hdfs-site.xml                  ssl-server.xml.example  
httpfs-env.sh                  user_ec_policies.xml.template  
httpfs-log4j.properties        workers  
httpfs-signature.secret        yarn-env.cmd  
httpfs-site.xml                yarn-env.sh  
kms-acls.xml                    yarn-site.xml  
kms-env.sh  
ubuntu@ip-172-31-17-223:/usr/local/hadoop/etc/hadoop$ sudo nano hdfs-site.xml  
ubuntu@ip-172-31-17-223:/usr/local/hadoop/etc/hadoop$ cd  
ubuntu@ip-172-31-17-223:~$ sudo apt install python3-pip  
Reading package lists... Done  
Building dependency tree... Done  
Reading state information... Done  
The following additional packages will be installed:  
  build-essential bzip2 cpp cpp-11 dpkg-dev fakeroot g++ g++-11 gcc gcc-11  
  gcc-11-base gcc-12-base javascript-common libalgorithm-diff-perl  
  libalgorithm-diff-xs-perl libalgorithm-merge-perl libasan6 libatomic1  
  libc-dev-bin libc-devtools libc6 libc6-dev libcc1-0 libcrypt-dev  
  libdpkg-perl libexpat1-dev libfakeroot libfile-fcntllock-perl libgcc-11-dev  
  libgcc-s1 libgd3 libgomp1 libisl23 libitm1 libjs-jquery libjs-sphinxdoc  
  libjs-underscore liblsan0 libmpc3 libnsl-dev libpython3-dev libpython3.10
```

Pip 3 install jupyterlab

ubuntu@ip-172-31-17-223: ~

```
Scanning candidates...
Scanning linux images...

Running kernel seems to be up-to-date.

Restarting services...
systemctl restart acpid.service chrony.service cron.service irqbalance.service
multipathd.service packagekit.service polkit.service rsyslog.service serial-gett
y@ttyS0.service snapd.service ssh.service systemd-journald.service systemd-netwo
rkd.service systemd-resolved.service systemd-udev.service
Service restarts being deferred:
/etc/needrestart/restart.d/dbus.service
systemctl restart getty@tty1.service
systemctl restart networkd-dispatcher.service
systemctl restart systemd-logind.service
systemctl restart unattended-upgrades.service
systemctl restart user@1000.service

No containers need to be restarted.

No user sessions are running outdated binaries.

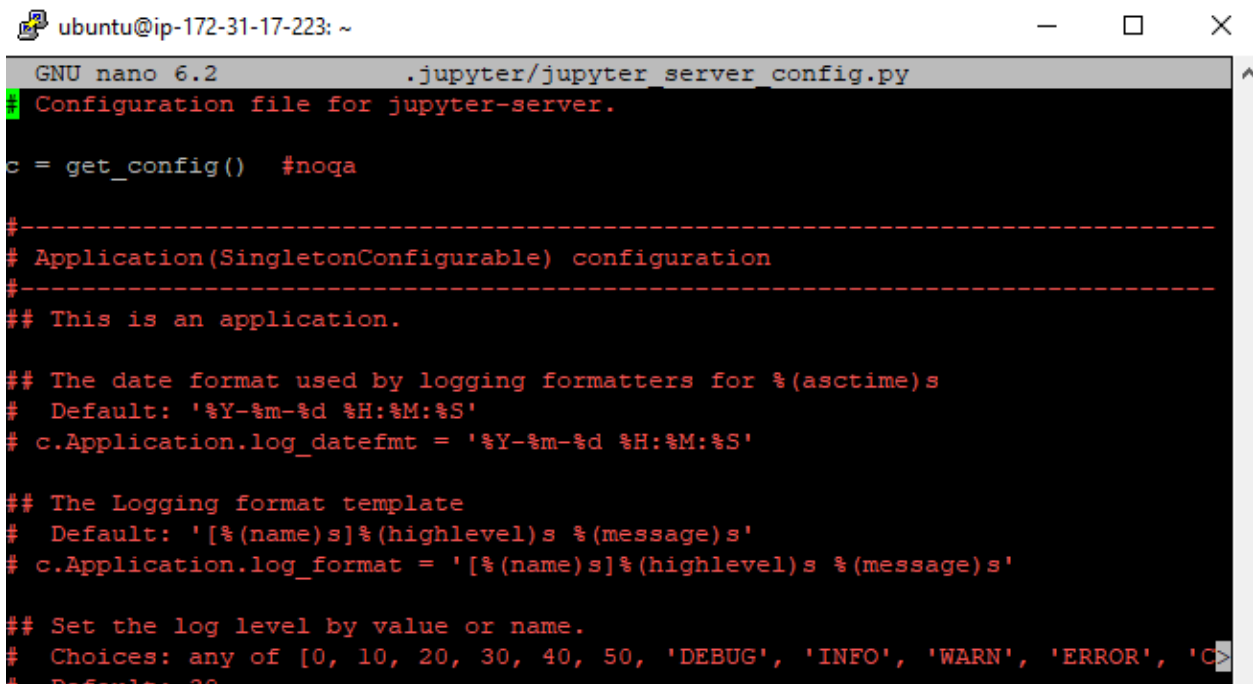
No VM guests are running outdated hypervisor (qemu) binaries on this host.
ubuntu@ip-172-31-17-223:~$ pip3 install jupyterlab
```

```
ubuntu@ip-172-31-17-223: ~  
WARNING: The scripts jlpm, jupyter-lab, jupyter-labextension and jupyter-labhub  
are installed in '/home/ubuntu/.local/bin' which is not on PATH.  
Consider adding this directory to PATH or, if you prefer to suppress this warn  
ing, use --no-warn-script-location.  
Successfully installed anyio-4.0.0 argon2-cffi-23.1.0 argon2-cffi-bindings-21.2.  
0 arrow-1.3.0 asttokens-2.4.0 async-lru-2.0.4 attrs-23.1.0 babel-2.13.0 backcall  
-0.2.0 beautifulsoup4-4.12.2 bleach-6.1.0 cffi-1.16.0 charset-normalizer-3.3.0 c  
omm-0.1.4 debugpy-1.8.0 decorator-5.1.1 defusedxml-0.7.1 exceptiongroup-1.1.3 ex  
ecuting-2.0.0 fastjsonschema-2.18.1 fqdn-1.5.1 ipykernel-6.25.2 ipython-8.16.1 i  
soduration-20.11.0 jedi-0.19.1 json5-0.9.14 jsonschema-4.19.1 jsonschema-specifi  
cations-2023.7.1 jupyter-client-8.3.1 jupyter-core-5.3.2 jupyter-events-0.7.0 ju  
pyter-lsp-2.2.0 jupyter-server-2.7.3 jupyter-server-terminals-0.4.4 jupyterlab-4  
.0.6 jupyterlab-pygments-0.2.2 jupyterlab-server-2.25.0 matplotlib-inline-0.1.6  
mistune-3.0.2 nbclient-0.8.0 nbconvert-7.9.2 nbformat-5.9.2 nest-asyncio-1.5.8 n  
otebook-shim-0.2.3 overrides-7.4.0 packaging-23.2 pandocfilters-1.5.0 parso-0.8.  
3 pickleshare-0.7.5 platformdirs-3.11.0 prometheus-client-0.17.1 prompt-toolkit-  
3.0.39 psutil-5.9.5 pure-eval-0.2.2 pycparser-2.21 pygments-2.16.1 python-dateut  
il-2.8.2 python-json-logger-2.0.7 pyzmq-25.1.1 referencing-0.30.2 requests-2.31.  
0 rfc3339-validator-0.1.4 rfc3986-validator-0.1.1 rpds-py-0.10.4 send2trash-1.8.  
2 sniffio-1.3.0 soupsieve-2.5 stack-data-0.6.3 terminado-0.17.1 tinycss2-1.2.1 t  
omli-2.0.1 tornado-6.3.3 traitlets-5.11.2 types-python-dateutil-2.8.19.14 typing  
-extensions-4.8.0 uri-template-1.3.0 wcwidth-0.2.8 webcolors-1.13 webencodings-0  
.5.1 websocket-client-1.6.4  
ubuntu@ip-172-31-17-223:~$
```

jupyter server --generate-config

```
ubuntu@ip-172-31-17-223:/usr/local/hadoop$ cd  
ubuntu@ip-172-31-17-223:~$ jupyter server --generate-config  
Writing default config to: /home/ubuntu/.jupyter/jupyter_server_config.py  
ubuntu@ip-172-31-17-223:~$
```

nano .jupyter/jupyter_server_config.py



```
ubuntu@ip-172-31-17-223: ~
GNU nano 6.2 .jupyter/jupyter_server_config.py
# Configuration file for jupyter-server.

c = get_config()  #noqa

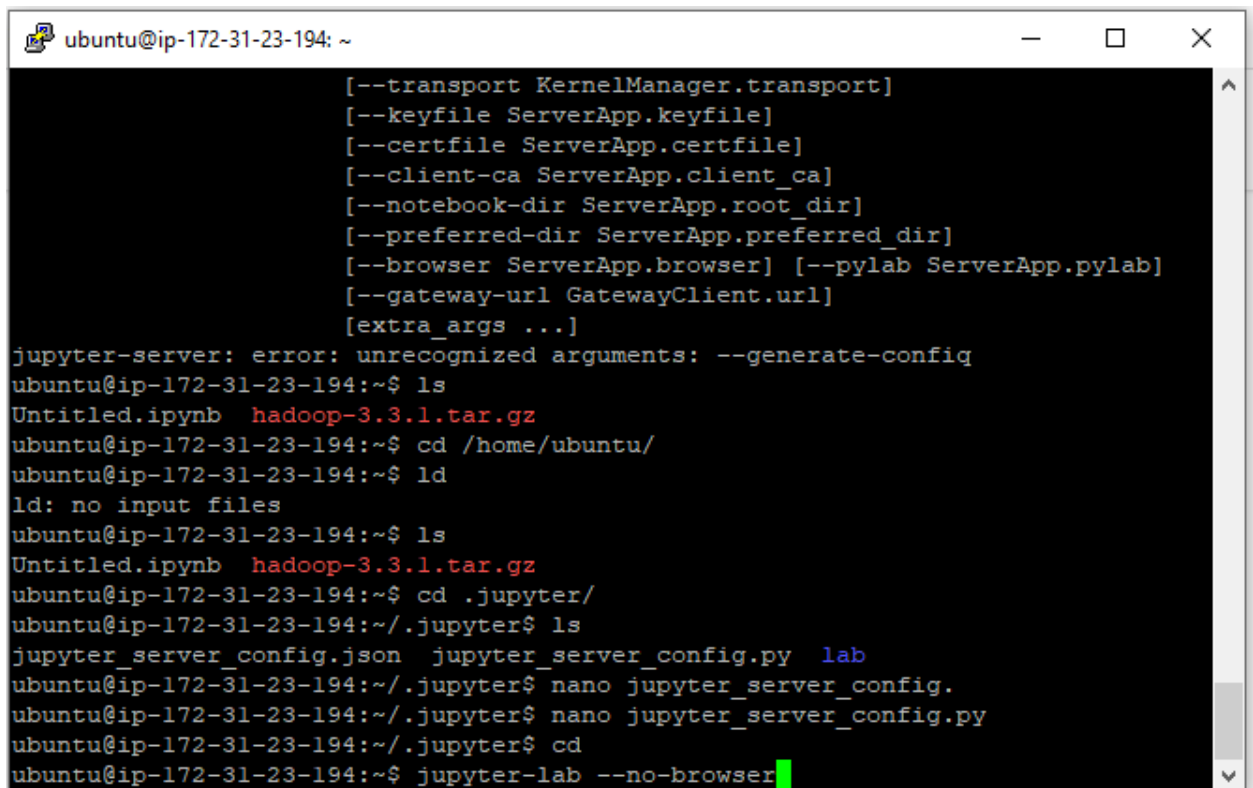
#-----
# Application(SingletonConfigurable) configuration
#-----
## This is an application.

## The date format used by logging formatters for %(asctime)s
# Default: '%Y-%m-%d %H:%M:%S'
# c.Application.log_datefmt = '%Y-%m-%d %H:%M:%S'

## The Logging format template
# Default: '[(name)s]%(highlevel)s %(message)s'
# c.Application.log_format = '[(name)s]%(highlevel)s %(message)s'

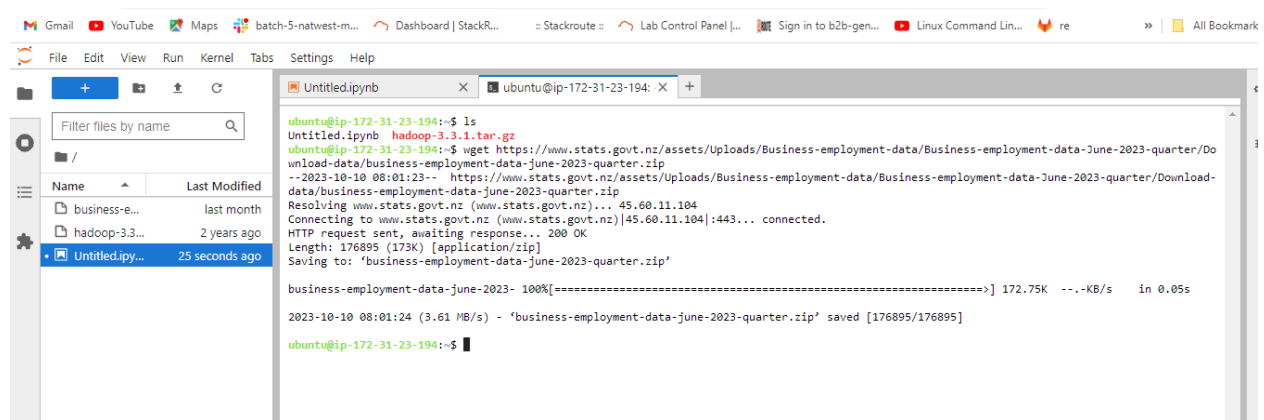
## Set the log level by value or name.
# Choices: any of [0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR', 'CRITICAL']
# Default: 30
```

jupyter-lab --no-browser



```
ubuntu@ip-172-31-23-194: ~
[--transport KernelManager.transport]
[--keyfile ServerApp.keyfile]
[--certfile ServerApp.certfile]
[--client-ca ServerApp.client_ca]
[--notebook-dir ServerApp.root_dir]
[--preferred-dir ServerApp.preferred_dir]
[--browser ServerApp.browser] [--pylab ServerApp.pylab]
[--gateway-url GatewayClient.url]
[extra_args ...]
jupyter-server: error: unrecognized arguments: --generate-config
ubuntu@ip-172-31-23-194:~$ ls
Untitled.ipynb  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-23-194:~$ cd /home/ubuntu/
ubuntu@ip-172-31-23-194:~$ ld
ld: no input files
ubuntu@ip-172-31-23-194:~$ ls
Untitled.ipynb  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-23-194:~$ cd .jupyter/
ubuntu@ip-172-31-23-194:~/.jupyter$ ls
jupyter_server_config.json  jupyter_server_config.py  lab
ubuntu@ip-172-31-23-194:~/.jupyter$ nano jupyter_server_config.
ubuntu@ip-172-31-23-194:~/.jupyter$ nano jupyter_server_config.py
ubuntu@ip-172-31-23-194:~/.jupyter$ cd
ubuntu@ip-172-31-23-194:~$ jupyter-lab --no-browser
```


Copy paste the token and download file on browser:



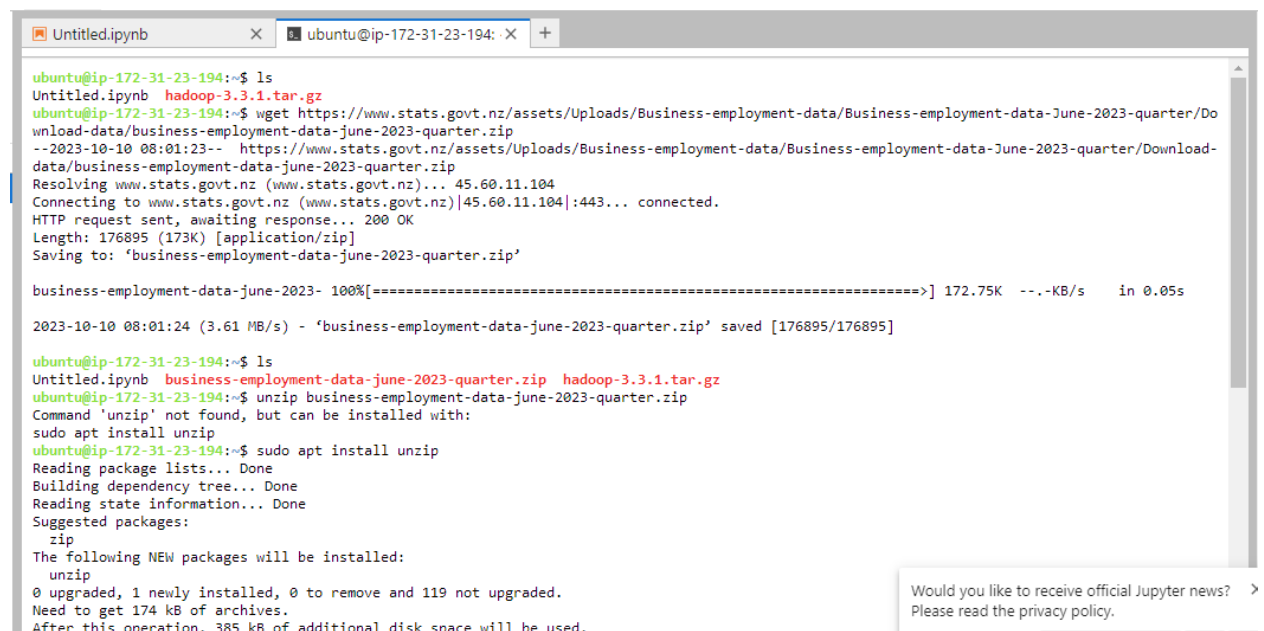
```
ubuntu@ip-172-31-23-194:~$ ls
Untitled.ipynb  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-23-194:~$ wget https://www.stats.govt.nz/assets/Uploads/Business-employment-data/Business-employment-data-June-2023-quarter/Download-data/business-employment-data-june-2023-quarter.zip
--2023-10-10 08:01:23-- https://www.stats.govt.nz/assets/Uploads/Business-employment-data/Business-employment-data-June-2023-quarter/Download-data/business-employment-data-june-2023-quarter.zip
Resolving www.stats.govt.nz (www.stats.govt.nz)... 45.60.11.104
Connecting to www.stats.govt.nz (www.stats.govt.nz)|45.60.11.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 176895 (173K) [application/zip]
Saving to: 'business-employment-data-june-2023-quarter.zip'

business-employment-data-june-2023- 100%[=====] 172.75K  --.-KB/s   in 0.05s

2023-10-10 08:01:24 (3.61 MB/s) - 'business-employment-data-june-2023-quarter.zip' saved [176895/176895]

ubuntu@ip-172-31-23-194:~$
```

Unzip the file



```
ubuntu@ip-172-31-23-194:~$ ls
Untitled.ipynb  business-employment-data-june-2023-quarter.zip  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-23-194:~$ unzip business-employment-data-june-2023-quarter.zip
Command 'unzip' not found, but can be installed with:
sudo apt install unzip
ubuntu@ip-172-31-23-194:~$ sudo apt install unzip
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Suggested packages:
  zip
The following NEW packages will be installed:
  unzip
0 upgraded, 1 newly installed, 0 to remove and 119 not upgraded.
Need to get 174 kB of archives.
After this operation, 385 kB of additional disk space will be used.
```

Read csv file

```
[14]: import pandas as pd
```

```
[15]: fh=pd.read_csv('business-employment-data-june-2023-quarter.zip')
print(fh.head(5))
```

	Series_reference	Period	Data_value	Suppressed	STATUS	UNITS	Magnitude	\
0	BDCQ.SEA1AA	2011.06	80078.0	NaN	F	Number	0	
1	BDCQ.SEA1AA	2011.09	78324.0	NaN	F	Number	0	
2	BDCQ.SEA1AA	2011.12	85850.0	NaN	F	Number	0	
3	BDCQ.SEA1AA	2012.03	90743.0	NaN	F	Number	0	
4	BDCQ.SEA1AA	2012.06	81780.0	NaN	F	Number	0	

Head value

The screenshot shows a Jupyter Notebook environment. On the left is a file explorer with a search bar and a list of files: 'business-e...', 'hadoop-3.3...', 'machine-re...', and 'Untitled.ipyn...' (3 minutes ago). The main area contains two code cells. The first cell imports pandas and reads a CSV file. The second cell sorts the data by 'Series_reference' in descending order and prints the first five rows. The output of the second cell shows a table with columns: Series_reference, Period, Data_value, Suppressed, STATUS, UNITS, and Magnitude. The data is sorted by Series_reference in descending order.

```
[16]: sfh=fh.sort_values(by=['Series_reference'], ascending=False)
print(sfh.head(5))
```

	Series_reference	Period	Data_value	Suppressed	STATUS	UNITS	Magnitude	\
21801	BDCQ.SEE3999A	2018.06	NaN	Y	C	Number		
21787	BDCQ.SEE3999A	2014.12	NaN	Y	C	Number		
21773	BDCQ.SEE3999A	2011.06	NaN	Y	C	Number		
21774	BDCQ.SEE3999A	2011.09	NaN	Y	C	Number		
21775	BDCQ.SEE3999A	2011.12	NaN	Y	C	Number		

Install pyspark

```
[*]: pip install pyspark
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting pyspark
  Downloading pyspark-3.5.0.tar.gz (316.9 MB)
    9 MB 31.7 MB/s eta 0:00:01
```

Hdfs storage

```
employment-data-jun-2023-quarter.csv /user/ubuntu/input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value
of HADOOP_PREFIX.
ubuntu@ip-172-31-23-194:~$ hdfs dfs -ls input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value
of HADOOP_PREFIX.
Found 2 items
-rw-r--r--    1 ubuntu supergroup    3400510 2023-10-10 08:48 input/m
achine-readable-business-employment-data-jun-2023-quarter.csv
-rw-r--r--    1 ubuntu supergroup      0 2023-10-10 06:59 input/s
ample.txt
ubuntu@ip-172-31-23-194:~$
```