

Name: Mohd Amirul Shafiq Bin Shafiee

Matric No: WQD180114

Milestone 3: Processing of data - Accessing hive data warehouse using Python

Part 1: Configure HiveServer2

- Add proxy user in coresite.xml in hadoop as follows:

```
<property>
  <name>hadoop.proxyuser.amirul.groups</name>
  <value>*</value>
</property>
<property>
  <name>hadoop.proxyuser.amirul.hosts</name>
  <value>*</value>
</property>
```

- Set the hiveserver2 at localhost in hive-site.xml as follows:

```
<property>
  <name>hive.server2.thrift.bind.host</name>
  <value>localhost</value>
</property>
<property>
  <name>hive.server2.thrift.port</name>
  <value>10000</value>
</property>
```

- `hadoop fs -chmod 777 /tmp`
- Run in terminal: `hiveserver2 &`
- Test connection to hiveserver2 via beeline:

```
beeline
beeline>!connect jdbc:hive2://localhost:10000
```

```
amirul Add Configuration...
Terminal: Local x Local (2) x Local (3) x Local (4) x Local (5) x +
amirul@amirul-VirtualBox:~$ beeline
bSLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/amirul/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/amirul/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Beeline version 3.1.2 by Apache Hive
beeline> !connect jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Enter username for jdbc:hive2://localhost:10000: amirul
Enter password for jdbc:hive2://localhost:10000: *****
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
```

- Open <http://localhost:10002>

HiveServer2

localhost:10002

HiveServer2

Active Sessions

User Name	IP Address	Operation Count	Active Time (s)	Idle Time (s)
amirul	127.0.0.1	0	132	131

Total number of sessions: 1

Open Queries

User Name	Query	Execution Engine	State	Opened Timestamp	Opened (s)	Latency (s)	Drilldown Link
-----------	-------	------------------	-------	------------------	------------	-------------	----------------

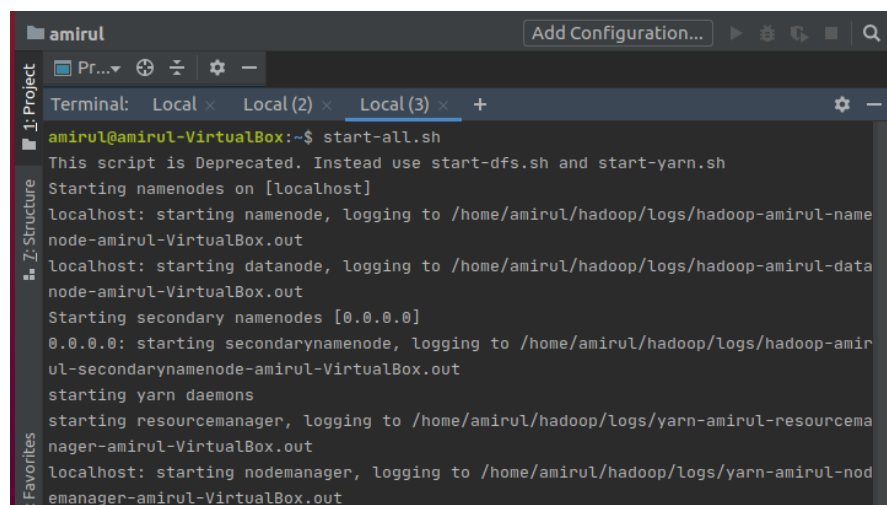
Part 2: Install PyHive and its components

- gcc: `sudo apt-get install gcc`
- Python developer package : `sudo apt-get python3.8-devel.x86_64`
- libsasl2-dev: `sudo apt-get install libsasl2-dev`
- sasl: `pip3 install sasl`
- thrift sasl : `pip3 install thrift_sasl`

Part 3: Access Hive through Python

At command prompt

- Start HDFS



```
amirul
Add Configuration...
Terminal: Local x Local (2) x Local (3) x +
amirul@amirul-VirtualBox:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/amirul/hadoop/logs/hadoop-amirul-name
node-amirul-VirtualBox.out
localhost: starting datanode, logging to /home/amirul/hadoop/logs/hadoop-amirul-data
node-amirul-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/amirul/hadoop/logs/hadoop-amir
ul-secondarynamenode-amirul-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /home/amirul/hadoop/logs/yarn-amirul-resourcem
anager-amirul-VirtualBox.out
localhost: starting nodemanager, logging to /home/amirul/hadoop/logs/yarn-amirul-nod
emanager-amirul-VirtualBox.out
```

- Initiate hiveserver2 &

```

amirul@amirul-VirtualBox:~$ hiveserver2 &
[1] 3893
amirul@amirul-VirtualBox:~$ 2020-06-18 21:53:05: Starting HiveServer2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/amirul/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/amirul/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = cb467462-49cc-4aa5-943e-71c75519ff08

```

At Jupyter Notebook

- Import pyhive
- Define hive connection
- Read data file (df2) and make sure it is correct

```

In [1]: import pandas as pd
        from pyhive import hive

In [2]: conn = hive.Connection(host="localhost", port=10000)

In [3]: df = pd.read_sql("SELECT * FROM default.df2", conn)

In [4]: df.head()

```

Out[4]:

	df2.quarter	df2.axiata	df2.digi	df2.maxis	df2.tm
0	2009-09-30	3380922	1238766	2156000	2101087
1	2009-12-31	3693781	1247612	2211000	2272544

Milestone3 - Jupyter Notebook Milestone 3_ - Jupyter Notebook

localhost:8888/notebooks/Milestone 3_

jupyter Milestone 3_ Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
In [5]: df.describe()
```

Out[5]:

	df2.axiata	df2.digi	df2.maxis	df2.tm
count	4.300000e+01	4.300000e+01	4.300000e+01	4.300000e+01
mean	4.982791e+06	1.586184e+06	2.214674e+06	2.694206e+06
std	8.641711e+05	1.388962e+05	9.578068e+04	3.345451e+05
min	3.380922e+06	1.238766e+06	2.065000e+06	2.101087e+06
25%	4.344400e+06	1.547072e+06	2.154000e+06	2.404248e+06
50%	4.747318e+06	1.618345e+06	2.214000e+06	2.778922e+06
75%	5.874018e+06	1.674696e+06	2.245000e+06	2.943220e+06
max	6.267007e+06	1.798623e+06	2.590000e+06	3.237032e+06