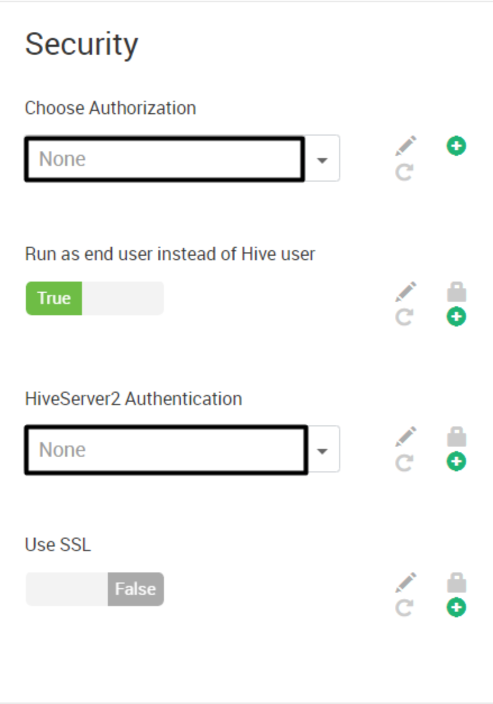
Tan Sia Hong & Tan Chang Jung

Milestone 3: Accessing Hive data warehouse using Python.

PART 1: Setting up the connection

1. Set the hive authorization to be PLAIN, which is ‘None’ in Hortonworks Sandbox.



Security

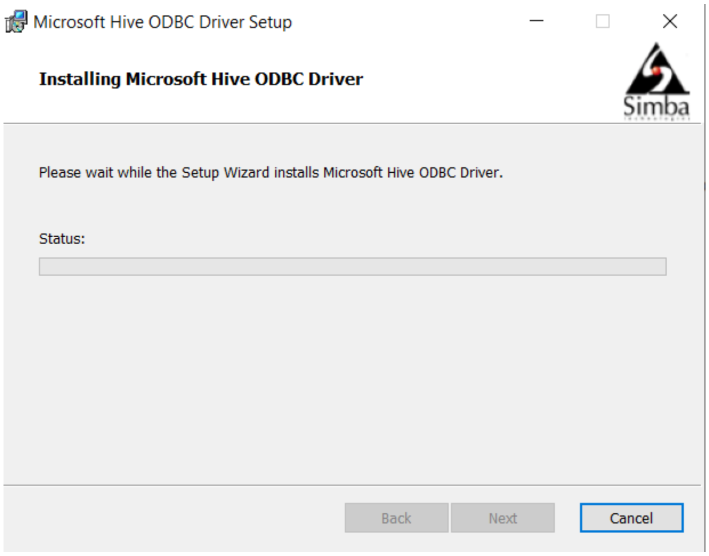
Choose Authorization: None

Run as end user instead of Hive user: True

HiveServer2 Authentication: None

Use SSL: False

2. Download and install the Hive ODBC Driver
   - Open a web browser and navigate to https://drivers.softpedia.com/get/Other-DRIVERS-TOOLS/MICROSOFT/Microsoft-Hive-ODBC-Driver-1100-64-bit.shtml
   - After download the driver start to install it.



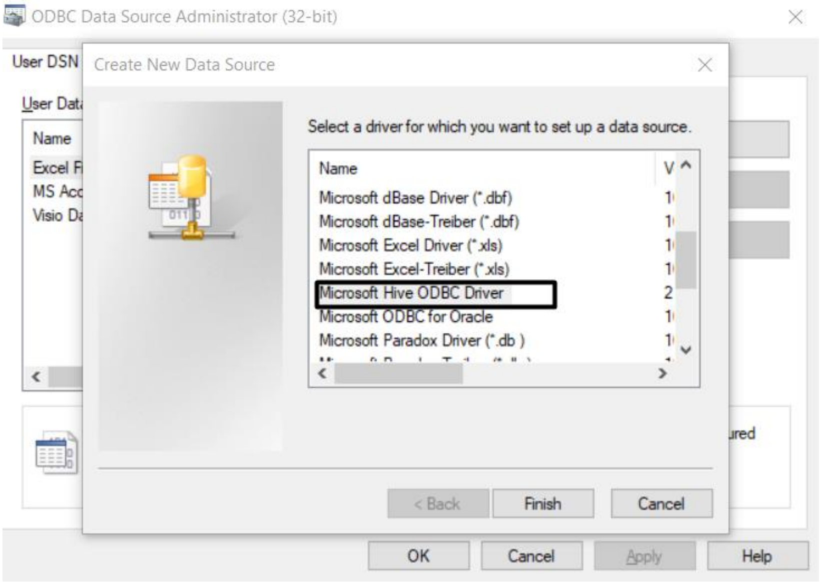
Microsoft Hive ODBC Driver Setup

Installing Microsoft Hive ODBC Driver

Please wait while the Setup Wizard installs Microsoft Hive ODBC Driver.

Status:

3. Configure the Hive ODBC driver.
   - Open ‘ODBC Data Source Administration’ in Windows.
   - Add the hive driver



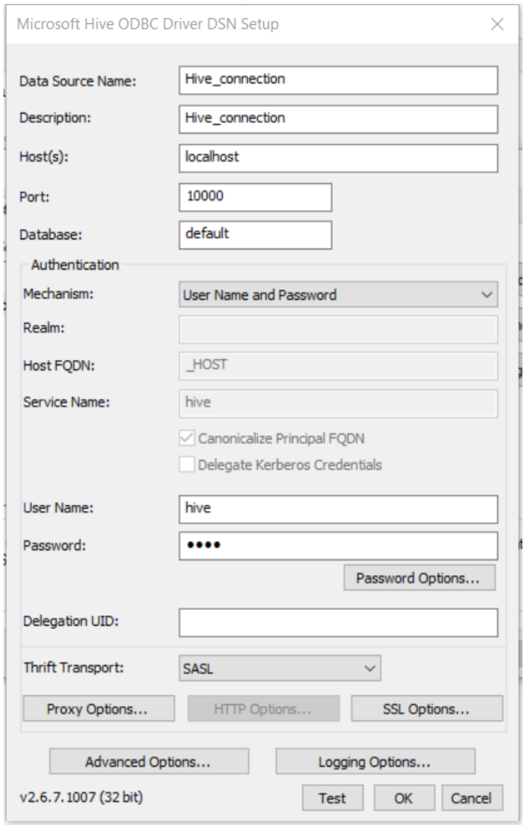
ODBC Data Source Administrator (32-bit)

Create New Data Source

Select a driver for which you want to set up a data source.

Name	V
Microsoft dBase Driver (*.dbf)	1
Microsoft dBase-Treiber (*.dbf)	1
Microsoft Excel Driver (*.xls)	1
Microsoft Excel-Treiber (*.xls)	1
Microsoft Hive ODBC Driver	2
Microsoft ODBC for Oracle	1
Microsoft Paradox Driver (*.db)	1

   - Go to system DSN and Setup the Hive ODBC connection as below:
     The default port for Hive is **10000.**
     The default credentials for connection is:
     User: hive   Password: hive



Microsoft Hive ODBC Driver DSN Setup

Data Source Name: Hive_connection

Description: Hive_connection

Host(s): localhost

Port: 10000

Database: default

Authentication

Mechanism: User Name and Password

Realm:

Host FQDN: _HOST

Service Name: hive

☑ Canonicalize Principal FQDN

□ Delegate Kerberos Credentials

User Name: hive

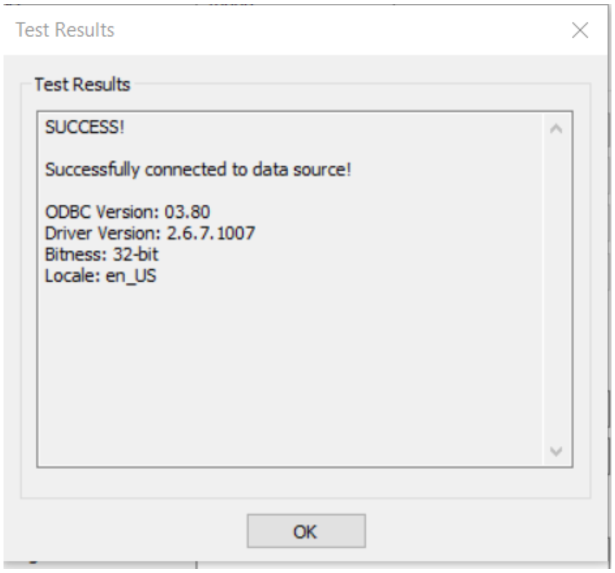
Password: ••••

Delegation UID:

Thrift Transport: SASL

v2.6.7.1007 (32 bit)

Then click 'Test' for testing the connection, make sure the connection is work.



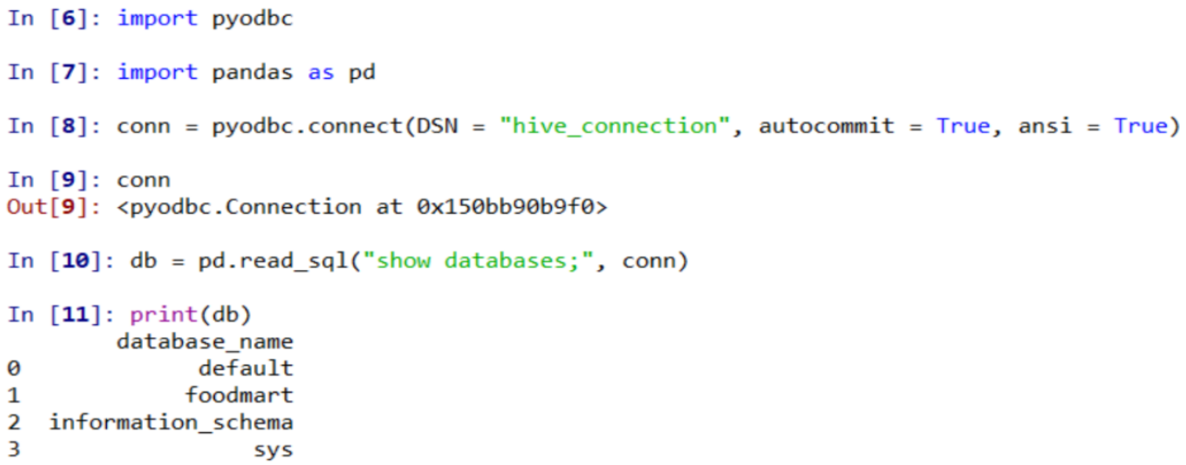
SUCCESS!

Successfully connected to data source!

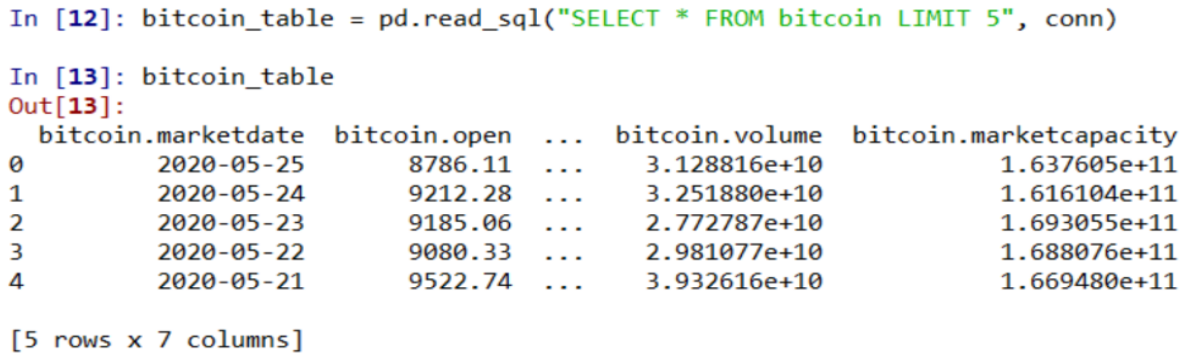
ODBC Version: 03.80
Driver Version: 2.6.7.1007
Bitness: 32-bit
Locale: en_US

PART 2: Run the python script to access data from hive data warehouse.

Before run the python script, install the necessary library 'pyodbc'.
In Python IDE, use library 'pyodbc' to connect Hive and access Hive database.



```
In [6]: import pyodbc

In [7]: import pandas as pd

In [8]: conn = pyodbc.connect(DSN = "hive_connection", autocommit = True, ansi = True)

In [9]: conn
Out[9]: <pyodbc.Connection at 0x150bb90b9f0>

In [10]: db = pd.read_sql("show databases;", conn)

In [11]: print(db)
        database_name
0             default
1           foodmart
2  information_schema
3                 sys
```



```
In [12]: bitcoin_table = pd.read_sql("SELECT * FROM bitcoin LIMIT 5", conn)

In [13]: bitcoin_table
Out[13]:
  bitcoin.marketdate  bitcoin.open  ...  bitcoin.volume  bitcoin.marketcapacity
0         2020-05-25       8786.11  ...    3.128816e+10            1.637605e+11
1         2020-05-24       9212.28  ...    3.251880e+10            1.616104e+11
2         2020-05-23       9185.06  ...    2.772787e+10            1.693055e+11
3         2020-05-22       9080.33  ...    2.981077e+10            1.688076e+11
4         2020-05-21       9522.74  ...    3.932616e+10            1.669480e+11

[5 rows x 7 columns]
```