Cloud Computing Programming Assignment 1 - Query on Hadoop and Spark Cluster

Dataset: https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data?select=1.+Data+description.txt

Data are stored in /home/admin/stock_data/
Sample data are stored in /home/admin/test_data/
Hadoop and Spark are installed in /usr/local/
All scripts are put in /home/admin/
All output log files are designed to put in /home/admin/spark_output/

Execution

Ref: https://sparkbyexamples.com/category/pyspark/

```
ssh admin@128.123.160.205
pwd: admin

# Locate at /home/admin/
# master node named as hadoop-master
# worker nodes named as hadoop-worker#
# You can access three worker nodes from master node with admin user
ssh admin@hadoop-worker1
ssh admin@hadoop-worker2
ssh admin@hadoop-worker3
```

a. Find the maximum number of transactions in a day for all companies in a user supplied given time window.

```
spark-submit spark_stock_max_trans_all_given_win.py -start <start_date> -end <end_date>
```

Note: Input format for the start and end dates of the time window is "mm/dd/yyyy".
Sample script:

```
spark-submit spark_stock_max_trans_all_given_win.py -start 01/01/2016 -end 12/31/2016 >sp

# >> see output log file at spark_output/output_stock_max_trans_all_given_win.log
# >> see result only at spark_output/result_stock_max_trans_all_given_win.txt
```

b. Maximum stock deviation is defined as (highest price for a day – lowest price for a day)/lowest price for a day. Find the stock that had the highest maximum stock deviation in a day among all stocks and what was the corresponding value for a given time window.

`spark-submit spark_stock_max_deviation_all_given_win.py -start <start_date> -end <end_date>`

Note: Input format for the start and end dates of the time window is "mm/dd/yyyy".
Sample script:

```
spark-submit spark_stock_max_deviation_all_given_win.py -start 01/01/2016 -end 12/31/2016

# >> see output log file at spark_output/output_stock_max_deviation_all_given_win.log
# >> see result only at spark_output/result_stock_max_deviation_all_given_win.txt
```

c. Find the maximum sell price in a day for a given company in the entire data set (No time window required).

`spark-submit spark_stock_max_sell_by_given_company.py -company <company_name>`

Note: Input company must presents in stock records.
Sample script:

```
spark-submit spark_stock_max_sell_by_given_company.py -company ADANIENT >spark_output/out

# >> see output log file at spark_output/output_stock_max_sell_given_company.log
# >> see result only at spark_output/result_stock_max_sell_given_company.txt
```

Configuration

Set up Firewall with ufw

- https://phoenixnap.com/kb/configure-firewall-with-ufw-on-ubuntu \
- https://www.cyberciti.biz/faq/ubuntu-22-04-lts-set-up-ufw-firewall-in-5-minutes/ \
- https://www.cyberciti.biz/faq/ufw-allow-incoming-ssh-connections-from-a-specific-ip-address-subnet-on-ubuntu-debian/

```
sudo apt install ufw
sudo ufw default deny incoming
sudo ufw default allow outgoing
sudo ufw allow ssh
sudo ufw enable
```

```
sudo ufw status verbose
sudo ufw status numbered
sudo ufw delete [#]
sudo ufw allow from <IP address>
sudo ufw allow from <IP address> to any port <port number>
```

Create user named admin and give sudo access

https://www.cyberciti.biz/faq/add-new-user-account-with-admin-access-on-linux/

```
sudo adduser admin
sudo usermod -aG sudo admin
su admin
```

Add ssh keys as authorized keys for the admin user

https://askubuntu.com/questions/46424/how-do-i-add-ssh-keys-to-authorized-keys-file
https://www.digitalocean.com/community/tutorials/how-to-set-up-ssh-keys-on-ubuntu-20-04

```
------------------both master and worker----------------
vi ~/.ssh/authorized_keys
# add ssh-rsa keys in authorized_keys file
ssh-copy-id admin@128.123.160.205
```

==========

Install and configure the Hadoop and Spark Frameworks

Setting Up Hadoop Multi-Node Cluster

https://computingforgeeks.com/how-to-install-apache-spark-on-ubuntu-debian/
https://tecadmin.net/how-to-install-apache-hadoop-on-ubuntu-22-04/

```
------------------both master and worker----------------
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz
tar xzf hadoop-3.3.5.tar.gz
mv hadoop-3.3.5 hadoop
```

https://blog.devgenius.io/setting-up-hadoop-multi-node-cluster-on-ubuntu-multiple-devices-637f539ce73a

```
ssh-keygen -t rsa -P “”
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh localhost
sudo apt install openjdk-11-jdk

------------------both master and worker----------------
vi hadoop/etc/hadoop/hadoop-env.sh
    export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
sudo mv hadoop /usr/local/hadoop
sudo vi /etc/environment
    JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64

rename /etc/hostname for hadoop-master and hadoop-worker nodes
repeat adduser, rename step
reboot

------------------both master and worker----------------
sudo usermod -aG admin admin
sudo chown admin:root -R /usr/local/
sudo chmod g+rwx -R /usr/local/
sudo adduser admin sudo

------------------both master and worker----------------
sudo vi ~/.bashrc
#Hadoop Related Options
export HADOOP_HOME="/usr/local/hadoop"
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

source ~/.bashrc
sudo vi /etc/hosts
192.168.202.205 hadoop-master
192.168.202.206 hadoop-worker1
192.168.202.207 hadoop-worker2
192.168.202.208 hadoop-worker3

-------------------------master-------------------------------
ssh-keygen -t rsa
ssh-copy-id -i ~/.ssh/id_rsa.pub admin@hadoop-master
```

```
ssh-copy-id -i ~/.ssh/id_rsa.pub admin@hadoop-worker1
ssh-copy-id -i ~/.ssh/id_rsa.pub admin@hadoop-worker2
ssh-copy-id -i ~/.ssh/id_rsa.pub admin@hadoop-worker3

sudo vi /usr/local/hadoop/etc/hadoop/workers, add
hadoop-worker1
hadoop-worker2
hadoop-worker3

------------------------worker-----------------------------
sudo vi /usr/local/hadoop/etc/hadoop/yarn-site.xml
<configuration>
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>hadoop-master</value>
    </property>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>

------------------both master and worker (vi on master then cp to worker)----------------
sudo vi /usr/local/hadoop/etc/hadoop/core-site.xml, add
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://hadoop-master:9000</value>
    </property>
</configuration>

sudo vi /usr/local/hadoop/etc/hadoop/hdfs-site.xml, add
<configuration>
    <property>
        <name>dfs.name.dir</name>
        <value>/usr/local/hadoop/data/nameNode</value>
    </property>
    <property>
        <name>dfs.data.dir</name>
        <value>/usr/local/hadoop/data/dataNode</value>
    </property>
    <property>
        <name>dfs.replication</name>
        <value>2</value>
    </property>
</configuration>

sudo vi /usr/local/hadoop/etc/hadoop/mapred-site.xml
<configuration>
```

```
    <property>
        <name>mapred.job.tracker</name>
        <value>hadoop-master:9001</value>
    </property>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>

-------------------------master--------------------------------
# make sure the admin user has all the permissions
sudo scp -r /usr/local/hadoop/* admin@hadoop-worker1:/usr/local/hadoop
sudo scp -r /usr/local/hadoop/* admin@hadoop-worker2:/usr/local/hadoop
sudo scp -r /usr/local/hadoop/* admin@hadoop-worker3:/usr/local/hadoop

hdfs namenode -format
start-dfs.sh
jps
start-yarn.sh
stop-all.sh
```

Setting Up Spark on Hadoop Cluster

https://medium.com/@jootorres_11979/how-to-install-and-set-up-an-apache-spark-cluster-on-hadoop-18-04-b4d70650ed42 | from step 10th https://dwbi.org/pages/192

```
------------------both master and worker----------------
sudo apt-get install scala
wget https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz
tar xzf spark-3.3.2-bin-hadoop3.tgz
sudo mv spark-3.3.2-bin-hadoop3 /usr/local/spark
sudo vi ~/.bashrc
    export SPARK_HOME=/usr/local/spark
    export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
source ~/.bashrc

sudo chown admin:root -R /usr/local/
sudo chmod g+rwx -R /usr/local/
sudo scp -r /usr/local/spark admin@hadoop-worker1:/usr/local
sudo scp -r /usr/local/spark admin@hadoop-worker2:/usr/local
sudo scp -r /usr/local/spark admin@hadoop-worker3:/usr/local

-------------------------master--------------------------------
cp /usr/local/spark/conf/spark-env.sh.template /usr/local/spark/conf/spark-env.sh
sudo vi /usr/local/spark/conf/spark-env.sh
    export SPARK_MASTER_HOST=192.168.202.205
```

```
    export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/

sudo vi /usr/local/spark/conf/slaves
    hadoop-worker1
    hadoop-worker2
    hadoop-worker3

cd /usr/local/spark
./sbin/start-all.sh
./sbin/stop-all.sh
```

==========

Retrieve data available using scp and use MD5sum for verification

```
scp user@server:src_path des_path
MD5sum file
sudo apt-get install zip unzip
unzip -l archive.zip
unzip archive.zip -d stock_data/
```

==========

Spark Job Submission (demo)

ref: https://www.youtube.com/watch?v=nDfO7N8n_Pg&ab_channel=AmpCode

```
hdfs dfs -mkdir spark_output
hdfs dfs -ls
hadoop fs -rm -r
hdfs dfs -copyFromLocal Spark_Hadoop.py Spark_Hadoop.py
hdfs dfs -copyFromLocal friends.csv friends.csv
spark-submit Spark_Hadoop.py.py
```

==========

Trouble Shooting

```
# Add extra space whenever available in system partition
# Increase space in /dev/mapper/ubuntu--vg-ubuntu--lv
sudo lvextend -l +100%FREE /dev/ubuntu-vg/ubuntu-lv
sudo resize2fs /dev/mapper/ubuntu--vg-ubuntu--lv
```

```
# hdfs dfs -mkdir -p hdfs://localhost:9000/user/admin/stock_data
# mkdir: `hdfs://localhost:9000/user/admin/stock_data': No such file or directory
https://stackoverflow.com/questions/40143528/hdfs-dfs-mkdir-no-such-file-or-directory/427

hadoop namenode -format

# Name node is in safe mode
sudo -u hdfs hdfs dfsadmin -safemode leave

>> hdfs-site.xml
<configuration>
    <property>
        <name>dfs.safemode.threshold.pct</name>
        <value>0</value>
    </property>
</configuration>
```

(additional) Enabling clipboard copy and paste for a single VM

https://kb.vmware.com/s/article/57122
https://www.altaro.com/vmware/clipboard-vsphere/#Enabling%C2%A0clipboard%C2%A0copy_and_paste%C2%A0globally

```
Step 1 – Power off the VM
Step 2 – Right-click on the VM and select Edit Settings
Step 3 – Select the VM Options tab and expand Advanced. Click on Edit Configuration next

Name:                                Value:
isolation.tools.copy.disable         FALSE
isolation.tools.paste.disable        FALSE
isolation.tools.setGUIOptions.enable  TRUE
```