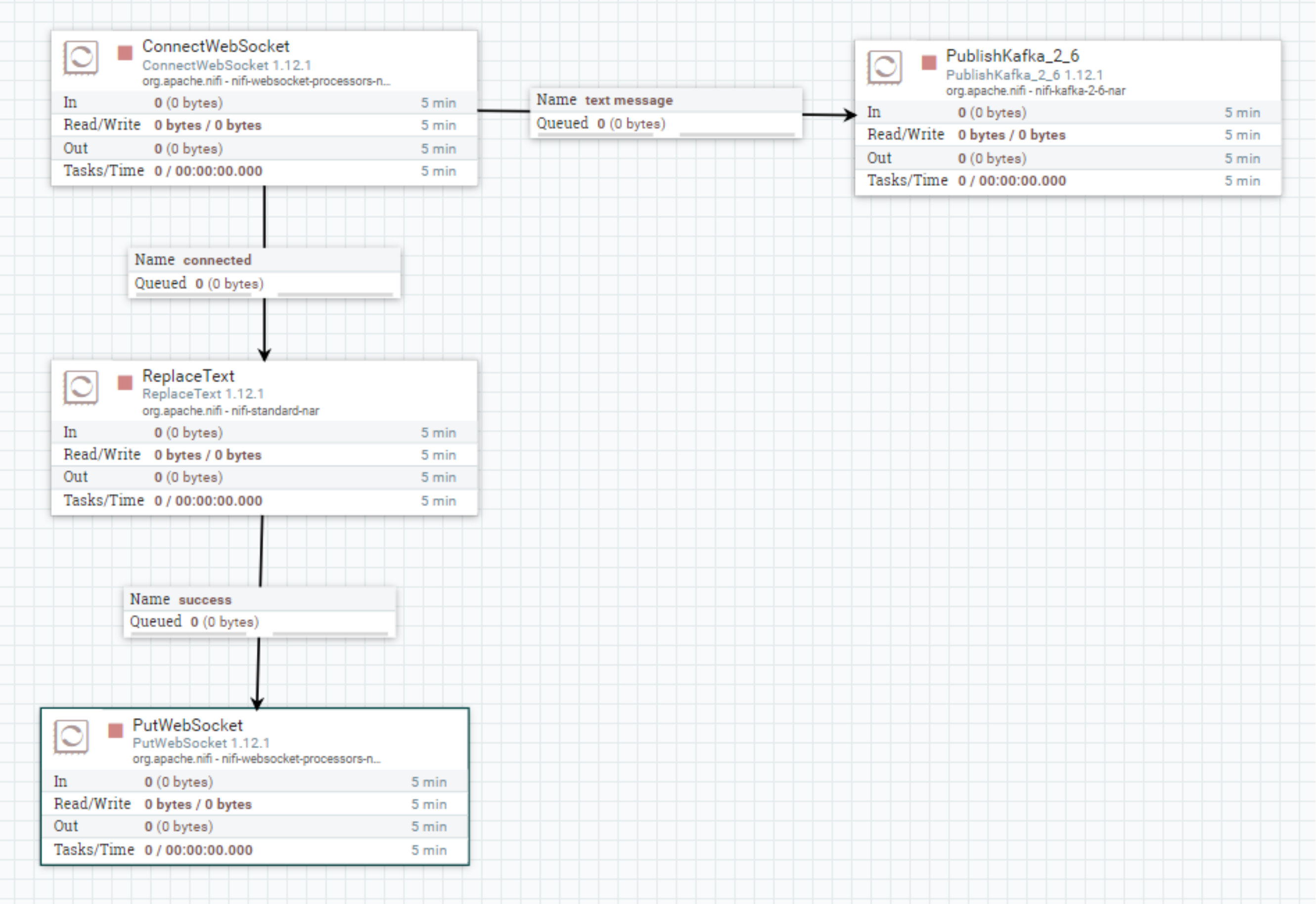
1. NIFI flow

1.1. Import template “sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2\nifi_kafka_hw2_template.xml”. Create flow with imported template.



ConnectWebSocket
ConnectWebSocket 1.12.1
org.apache.nifi - nifi-websocket-processors-n...

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name text message
Queued 0 (0 bytes)

PublishKafka_2_6
PublishKafka_2_6 1.12.1
org.apache.nifi - nifi-kafka-2-6-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name connected
Queued 0 (0 bytes)

ReplaceText
ReplaceText 1.12.1
org.apache.nifi - nifi-standard-nar

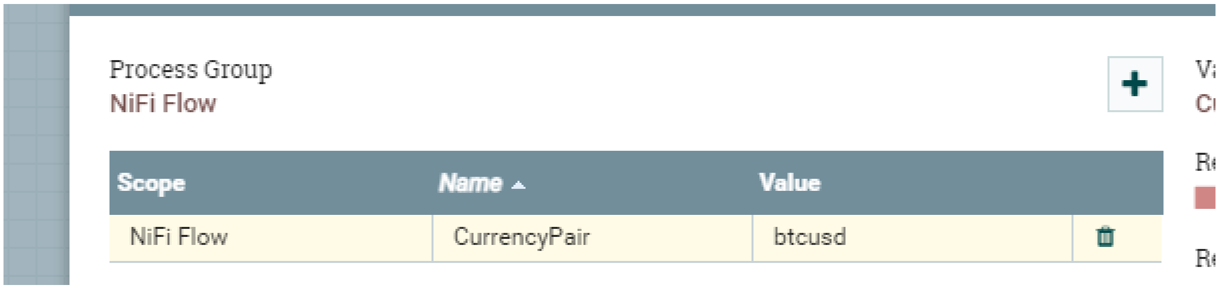
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

PutWebSocket
PutWebSocket 1.12.1
org.apache.nifi - nifi-websocket-processors-n...

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

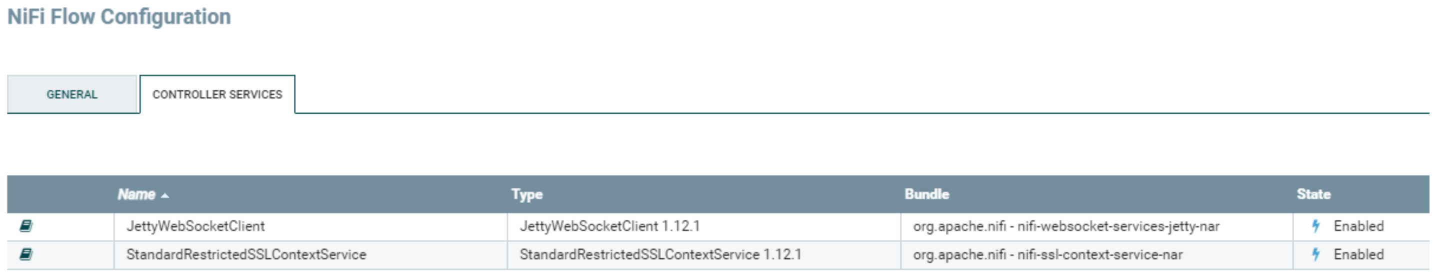
1.2. Create NIFI variable “CurrencyPair” with value “btcusd”



Process Group
NiFi Flow

Scope	Name ▲	Value
NiFi Flow	CurrencyPair	btcusd

1.3. Configure and enable controller services: “JetyWebSocketClient”, “StandardRegisteredSSLContextService”



NiFi Flow Configuration

Name ▲	Type	Bundle	State
JettyWebSocketClient	JettyWebSocketClient 1.12.1	org.apache.nifi - nifi-websocket-services-jetty-nar	Enabled
StandardRestrictedSSLContextService	StandardRestrictedSSLContextService 1.12.1	org.apache.nifi - nifi-ssl-context-service-nar	Enabled

2. Scripts preparation on GCP data proc node.

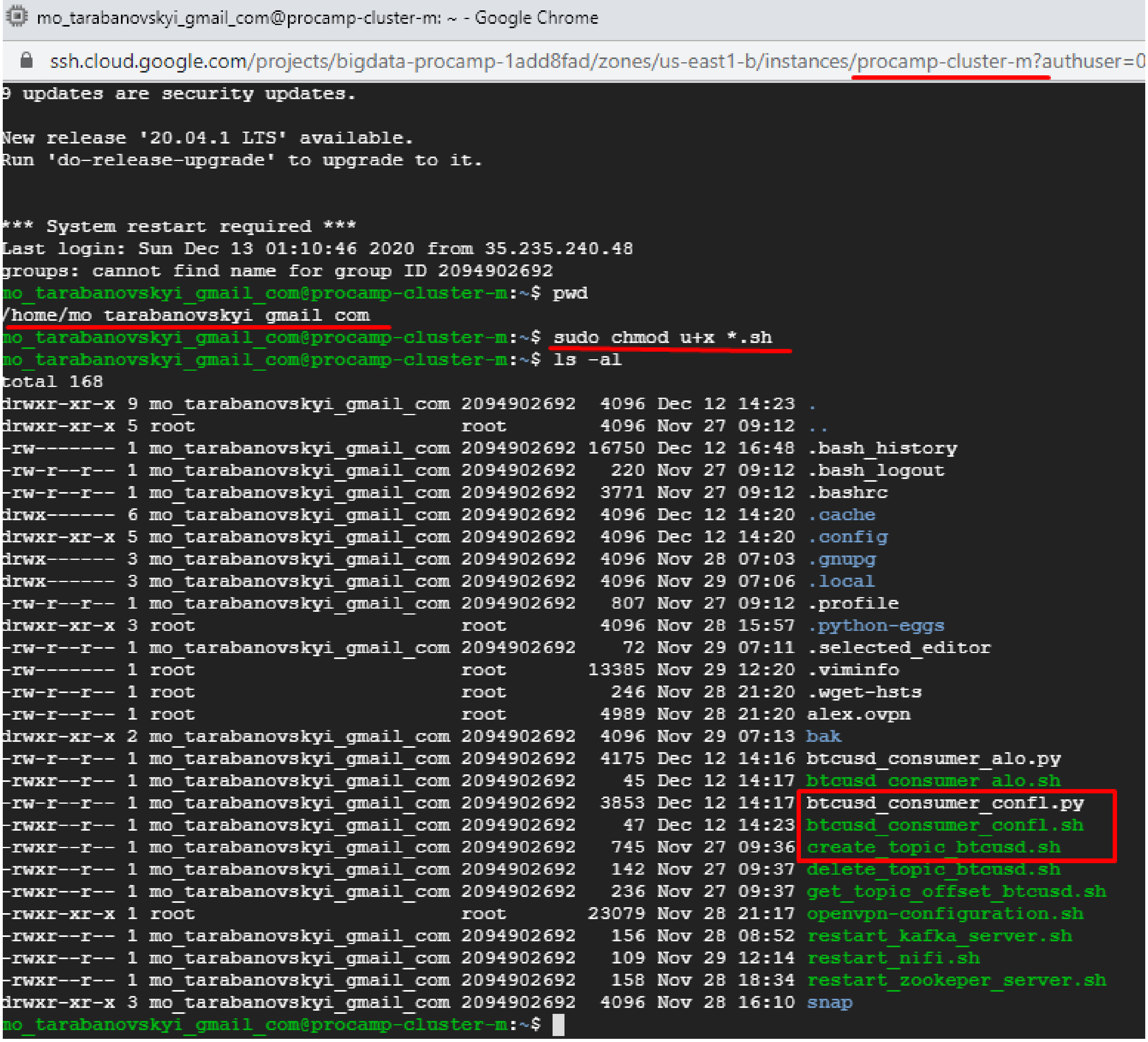
2.1. Copy shell and python scripts in home directory on GCP main node and grant permission:

Git Folder: “sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2”

Scripts:

- “btcusd_consumer_confl.py”
- “btcusd_consumer_confl.sh”
- “create_topic_btcusd.sh”

Execute command on main node (to grant needed permissions): “sudo chmod u+x *.sh”

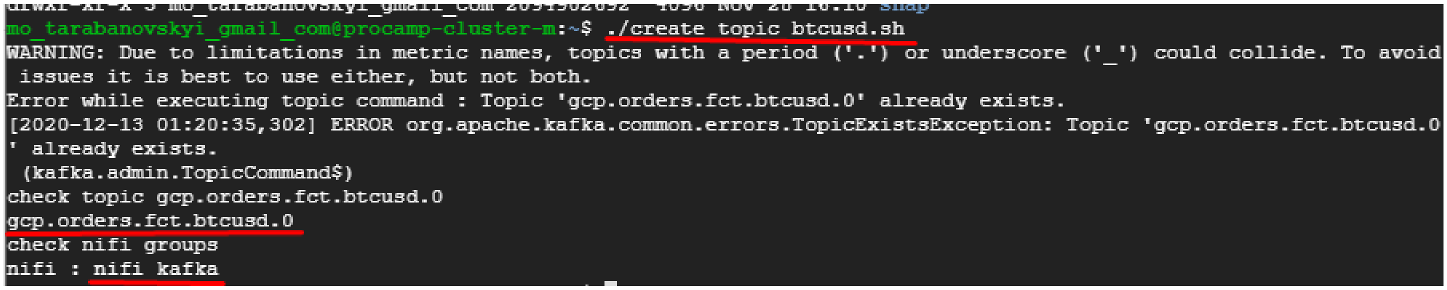


mo_tarabanovskyi_gmail_com@procamp-cluster-m: ~ - Google Chrome
ssh.cloud.google.com/projects/bigdata-procamp-1add8fad/zones/us-east1-b/instances/procamp-cluster-m?authuser=0

```
9 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** System restart required ***
Last login: Sun Dec 13 01:10:46 2020 from 35.235.240.48
groups: cannot find name for group ID 2094902692
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ pwd
/home/mo_tarabanovskyi_gmail_com
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ sudo chmod u+x *.sh
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ ls -al
total 168
drwxr-xr-x 9 mo_tarabanovskyi_gmail_com 2094902692  4096 Dec 12 14:23 .
drwxr-xr-x 5 root                       root        4096 Nov 27 09:12 ..
-rw------- 1 mo_tarabanovskyi_gmail_com 2094902692 16750 Dec 12 16:48 .bash_history
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   220 Nov 27 09:12 .bash_logout
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692  3771 Nov 27 09:12 .bashrc
drwx------ 6 mo_tarabanovskyi_gmail_com 2094902692  4096 Dec 12 14:20 .cache
drwxr-xr-x 5 mo_tarabanovskyi_gmail_com 2094902692  4096 Dec 12 14:20 .config
drwx------ 3 mo_tarabanovskyi_gmail_com 2094902692  4096 Nov 28 07:03 .gnupg
drwx------ 3 mo_tarabanovskyi_gmail_com 2094902692  4096 Nov 29 07:06 .local
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   807 Nov 27 09:12 .profile
drwxr-xr-x 3 root                       root        4096 Nov 28 15:57 .python-eggs
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692    72 Nov 29 07:11 .selected_editor
-rw------- 1 root                       root       13385 Nov 29 12:20 .viminfo
-rw-r--r-- 1 root                       root         246 Nov 28 21:20 .wget-hsts
-rw-r--r-- 1 root                       root        4989 Nov 28 21:20 alex.ovpn
drwxr-xr-x 2 mo_tarabanovskyi_gmail_com 2094902692  4096 Nov 29 07:13 bak
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692  4175 Dec 12 14:16 btcusd_consumer_alo.py
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692    45 Dec 12 14:17 btcusd_consumer_alo.sh
-rw-r--r-- 1 mo_tarabanovskyi_gmail_com 2094902692  3853 Dec 12 14:17 btcusd_consumer_confl.py
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692    47 Dec 12 14:23 btcusd_consumer_confl.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   745 Nov 27 09:36 create_topic_btcusd.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   142 Nov 27 09:37 delete_topic_btcusd.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   236 Nov 27 09:37 get_topic_offset_btcusd.sh
-rwxr-xr-x 1 root                       root       23079 Nov 28 21:17 openvpn-configuration.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   156 Nov 28 08:52 restart_kafka_server.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   109 Nov 29 12:14 restart_nifi.sh
-rwxr--r-- 1 mo_tarabanovskyi_gmail_com 2094902692   158 Nov 28 18:34 restart_zookeper_server.sh
drwxr-xr-x 3 mo_tarabanovskyi_gmail_com 2094902692  4096 Nov 28 16:10 snap
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$
```

3. Create Kafka topic.

Launch shell script “create_topic_btcusd.sh” to create topic “gcp.orders.fct.btcusd.0” and add current and “nifi” users to group “kafka”:



```
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ ./create_topic_btcusd.sh
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid
 issues it is best to use either, but not both.
Error while executing topic command : Topic 'gcp.orders.fct.btcusd.0' already exists.
[2020-12-13 01:20:35,302] ERROR org.apache.kafka.common.errors.TopicExistsException: Topic 'gcp.orders.fct.btcusd.0
' already exists.
 (kafka.admin.TopicCommand$)
check topic gcp.orders.fct.btcusd.0
gcp.orders.fct.btcusd.0
check nifi groups
nifi : nifi kafka
```

4. Install python modules “confluent-kafka”, “pandas”.

4.1. Install python module “kafka-confluence”:

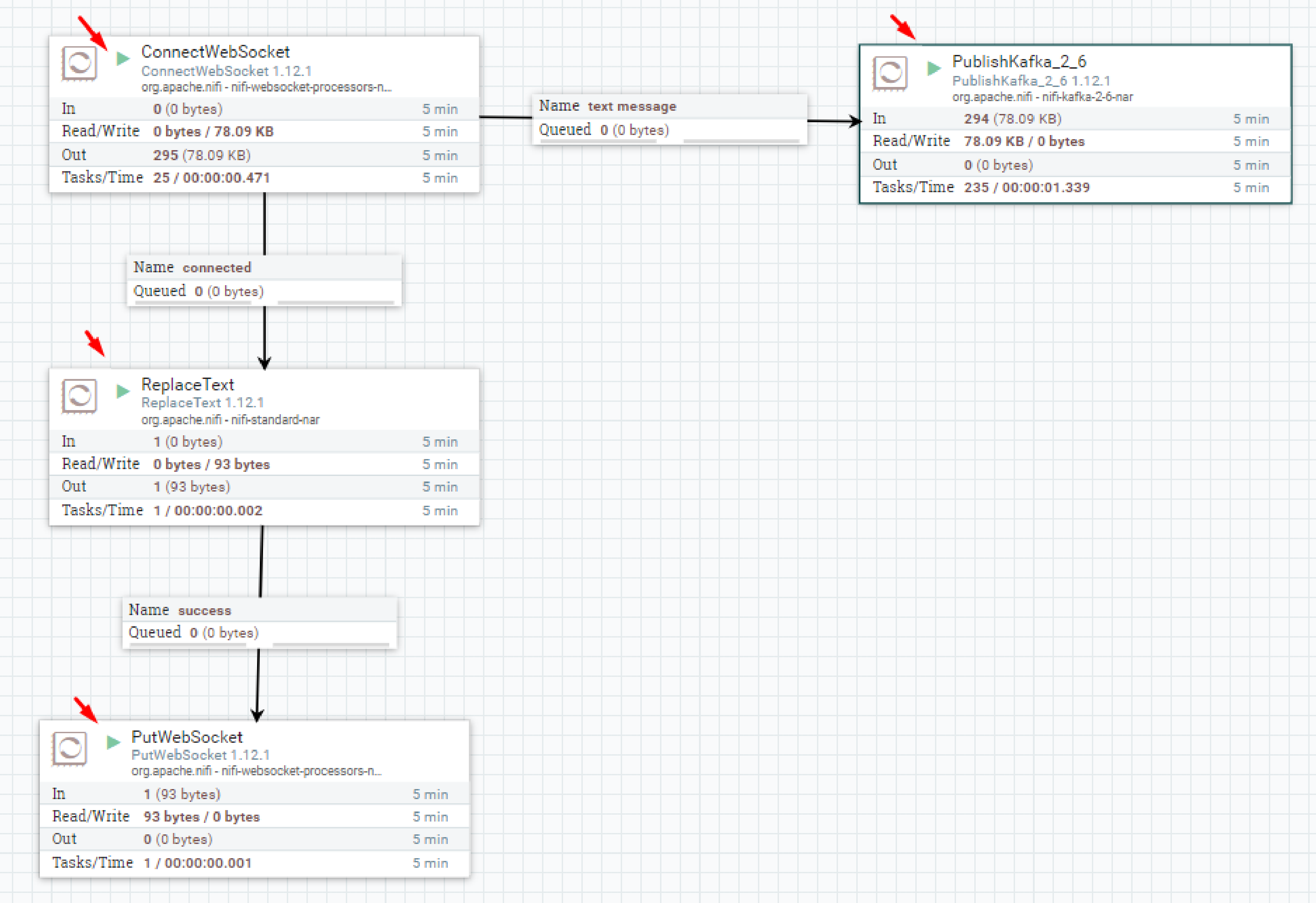
```
sudo /opt/conda/default/bin/python -m pip install confluent-kafka
```

4.2. Reinstall python module pandas (to install actual version)

```
sudo /opt/conda/default/bin/python -m pip uninstall pandas
sudo /opt/conda/default/bin/python -m pip install pandas
```

5. Start kafka producer

Start all Apache NIFI flow processors:



ConnectWebSocket
ConnectWebSocket 1.12.1
org.apache.nifi - nifi-websocket-processors-n...

```
In          0 (0 bytes)               5 min
Read/Write  0 bytes / 78.09 KB        5 min
Out         295 (78.09 KB)            5 min
Tasks/Time  25 / 00:00:00.471         5 min
```

Name text message
Queued 0 (0 bytes)

PublishKafka_2_6
PublishKafka_2_6 1.12.1
org.apache.nifi - nifi-kafka-2-6-nar

```
In          294 (78.09 KB)            5 min
Read/Write  78.09 KB / 0 bytes        5 min
Out         0 (0 bytes)               5 min
Tasks/Time  235 / 00:00:01.339        5 min
```

Name connected
Queued 0 (0 bytes)

ReplaceText
ReplaceText 1.12.1
org.apache.nifi - nifi-standard-nar

```
In          1 (0 bytes)               5 min
Read/Write  0 bytes / 93 bytes        5 min
Out         1 (93 bytes)              5 min
Tasks/Time  1 / 00:00:00.002          5 min
```

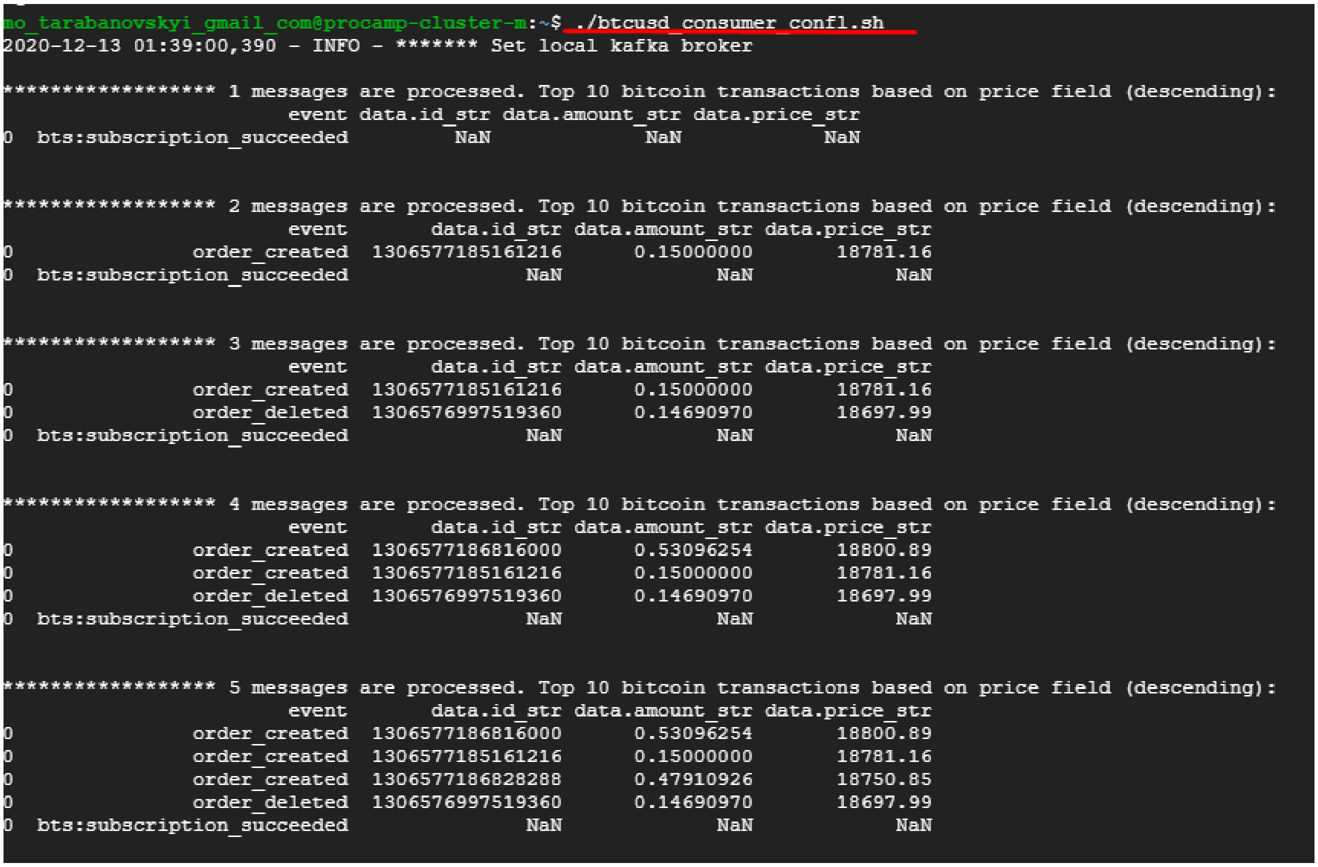
Name success
Queued 0 (0 bytes)

PutWebSocket
PutWebSocket 1.12.1
org.apache.nifi - nifi-websocket-processors-n...

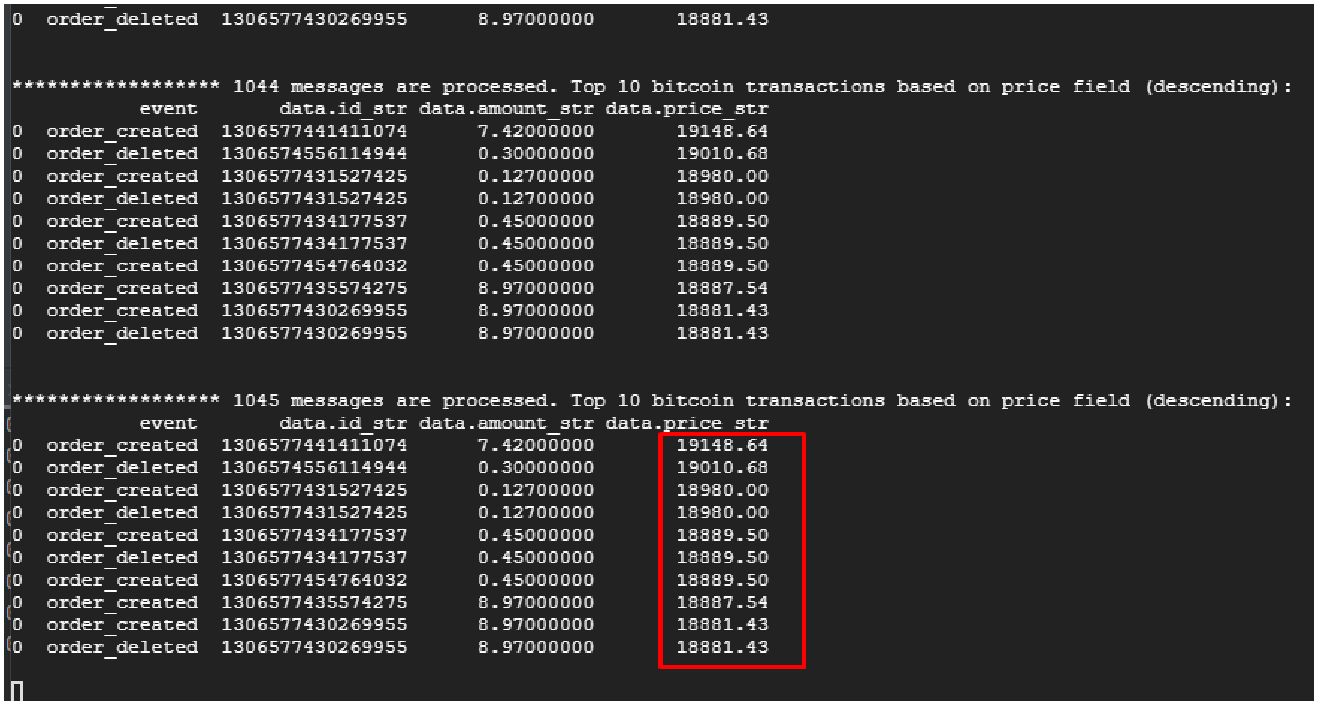
```
In          1 (93 bytes)              5 min
Read/Write  93 bytes / 0 bytes        5 min
Out         0 (0 bytes)               5 min
Tasks/Time  1 / 00:00:00.001          5 min
```

6. Launch kafka consumer on GCP node.

Execute shell script “`btcusd_consumer_confl.sh`” on GCP master node.

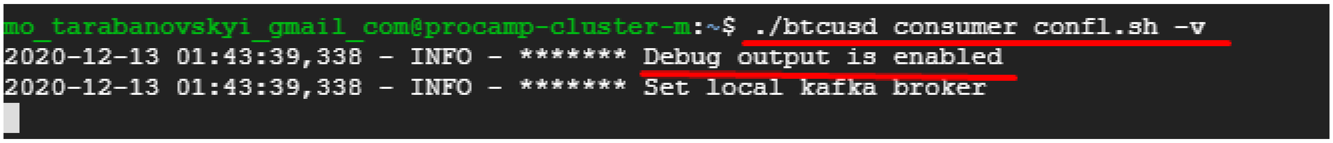


```
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ ./btcusd_consumer_confl.sh
2020-12-13 01:39:00,390 - INFO - ******* Set local kafka broker

****************** 1 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event data.id_str data.amount_str data.price_str
0  bts:subscription_succeeded         NaN             NaN            NaN


****************** 2 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_created  1306577185161216      0.15000000       18781.16
0  bts:subscription_succeeded               NaN             NaN            NaN


****************** 3 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_created  1306577185161216      0.15000000       18781.16
0               order_deleted  1306576997519360      0.14690970       18697.99
0  bts:subscription_succeeded               NaN             NaN            NaN


****************** 4 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_created  1306577186816000      0.53096254       18800.89
0               order_created  1306577185161216      0.15000000       18781.16
0               order_deleted  1306576997519360      0.14690970       18697.99
0  bts:subscription_succeeded               NaN             NaN            NaN


****************** 5 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_created  1306577186816000      0.53096254       18800.89
0               order_created  1306577185161216      0.15000000       18781.16
0               order_created  1306577186828288      0.47910926       18750.85
0               order_deleted  1306576997519360      0.14690970       18697.99
0  bts:subscription_succeeded               NaN             NaN            NaN
```



```
0  order_deleted  1306577430269955      8.97000000       18881.43


****************** 1044 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_created  1306577441411074      7.42000000       19148.64
0  order_deleted  1306574556114944      0.30000000       19010.68
0  order_created  1306577431527425      0.12700000       18980.00
0  order_deleted  1306577431527425      0.12700000       18980.00
0  order_created  1306577434177537      0.45000000       18889.50
0  order_deleted  1306577434177537      0.45000000       18889.50
0  order_created  1306577454764032      0.45000000       18889.50
0  order_created  1306577435574275      8.97000000       18887.54
0  order_created  1306577430269955      8.97000000       18881.43
0  order_deleted  1306577430269955      8.97000000       18881.43


****************** 1045 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_created  1306577441411074      7.42000000       19148.64
0  order_deleted  1306574556114944      0.30000000       19010.68
0  order_created  1306577431527425      0.12700000       18980.00
0  order_deleted  1306577431527425      0.12700000       18980.00
0  order_created  1306577434177537      0.45000000       18889.50
0  order_deleted  1306577434177537      0.45000000       18889.50
0  order_created  1306577454764032      0.45000000       18889.50
0  order_created  1306577435574275      8.97000000       18887.54
0  order_created  1306577430269955      8.97000000       18881.43
0  order_deleted  1306577430269955      8.97000000       18881.43
```

Note: if something went wrang, launch script with option “-v” for verbose output:



```
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ ./btcusd_consumer_confl.sh -v
2020-12-13 01:43:39,338 - INFO - ******* Debug output is enabled
2020-12-13 01:43:39,338 - INFO - ******* Set local kafka broker
```

7. Launch kafka consumer on local PC. Windows OS example.

(optional step, can be useful for debug).

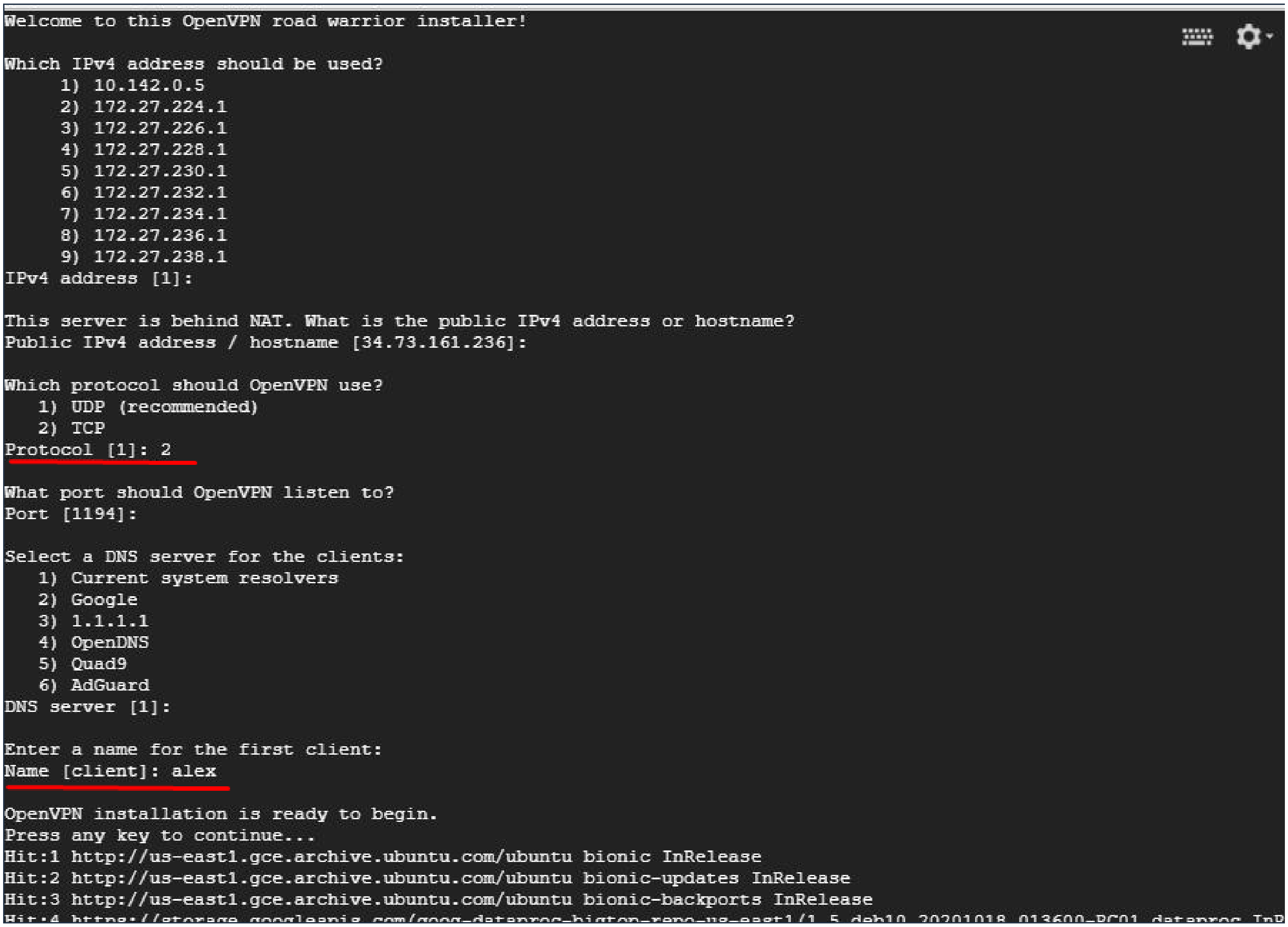
7.1.

7.2. Install python modules on local-PC:
“pandas”, “kafka-consumer”.
“kafka-consumer” can be installed on python 3.7 from “whl” file “sandbox_repo\training\GLBig_Data_ProCamp\infra\confluent\confluent_kafka-1.4.1-cp37-cp37m-win_amd64.whl” (modules for version < 3.7 can be downloaded from confluent site).

7.3. Up open VPN server on GCP master node

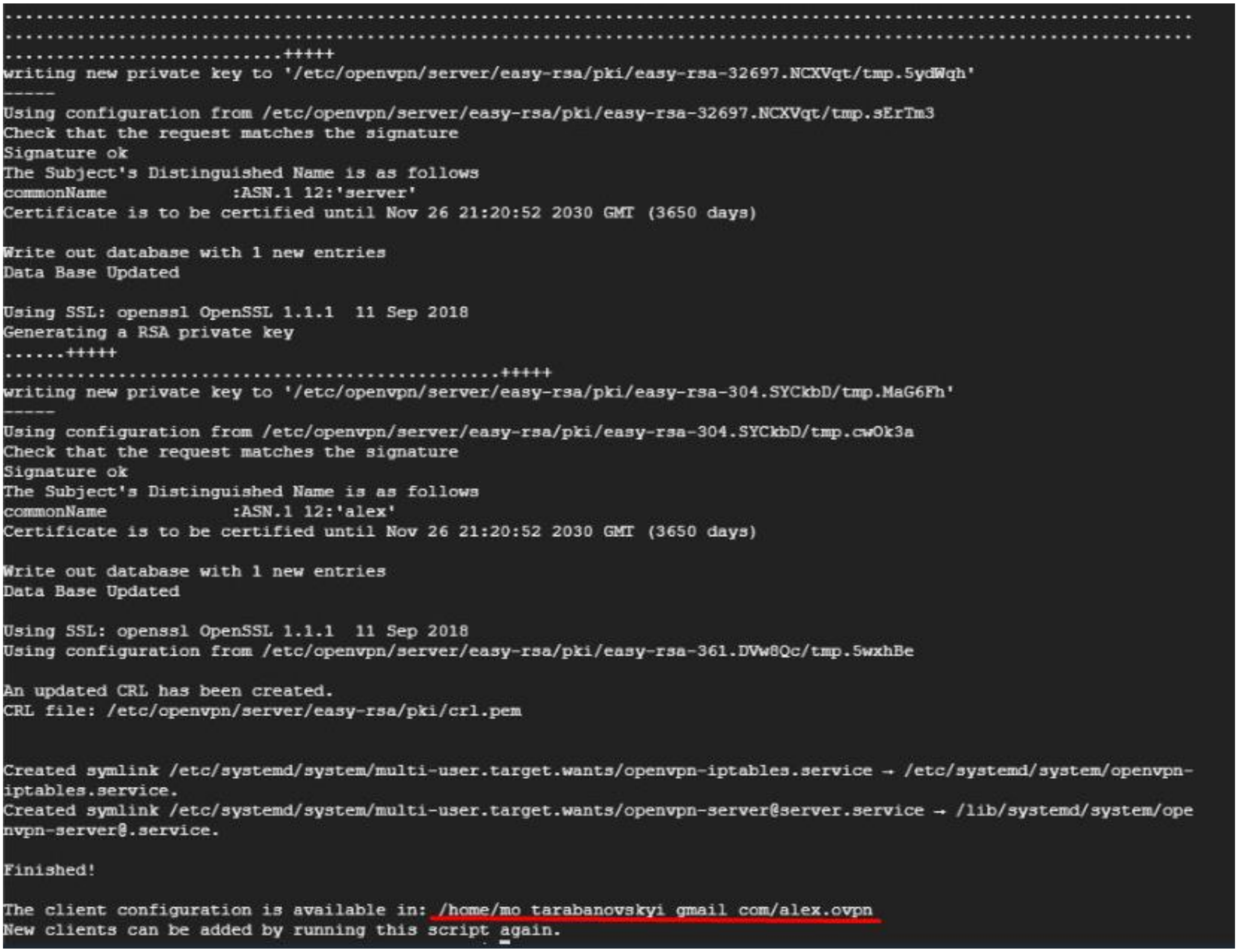
7.3.1. Copy script on master node in home directory “sandbox_repo\training\GLBig_Data_ProCamp\infra\vpn\openvpn-configuration.sh”

7.3.2. Launch script “sudo openvpn-configuration.sh” and fill parameters by default values (except: protocol - choose “TCP” and set client name)



```
Welcome to this OpenVPN road warrior installer!

Which IPv4 address should be used?
     1) 10.142.0.5
     2) 172.27.224.1
     3) 172.27.226.1
     4) 172.27.228.1
     5) 172.27.230.1
     6) 172.27.232.1
     7) 172.27.234.1
     8) 172.27.236.1
     9) 172.27.238.1
IPv4 address [1]:

This server is behind NAT. What is the public IPv4 address or hostname?
Public IPv4 address / hostname [34.73.161.236]:

Which protocol should OpenVPN use?
   1) UDP (recommended)
   2) TCP
Protocol [1]: 2

What port should OpenVPN listen to?
Port [1194]:

Select a DNS server for the clients:
   1) Current system resolvers
   2) Google
   3) 1.1.1.1
   4) OpenDNS
   5) Quad9
   6) AdGuard
DNS server [1]:

Enter a name for the first client:
Name [client]: alex

OpenVPN installation is ready to begin.
Press any key to continue...
Hit:1 http://us-east1.gce.archive.ubuntu.com/ubuntu bionic InRelease
Hit:2 http://us-east1.gce.archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:3 http://us-east1.gce.archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:4 https://storage.googleapis.com/goog-dataproc-bigtop-repo-us-east1/1_5_deb10_20201018_013600-RC01 dataproc InR
```

7.3.3. Copy client configuration on local PC

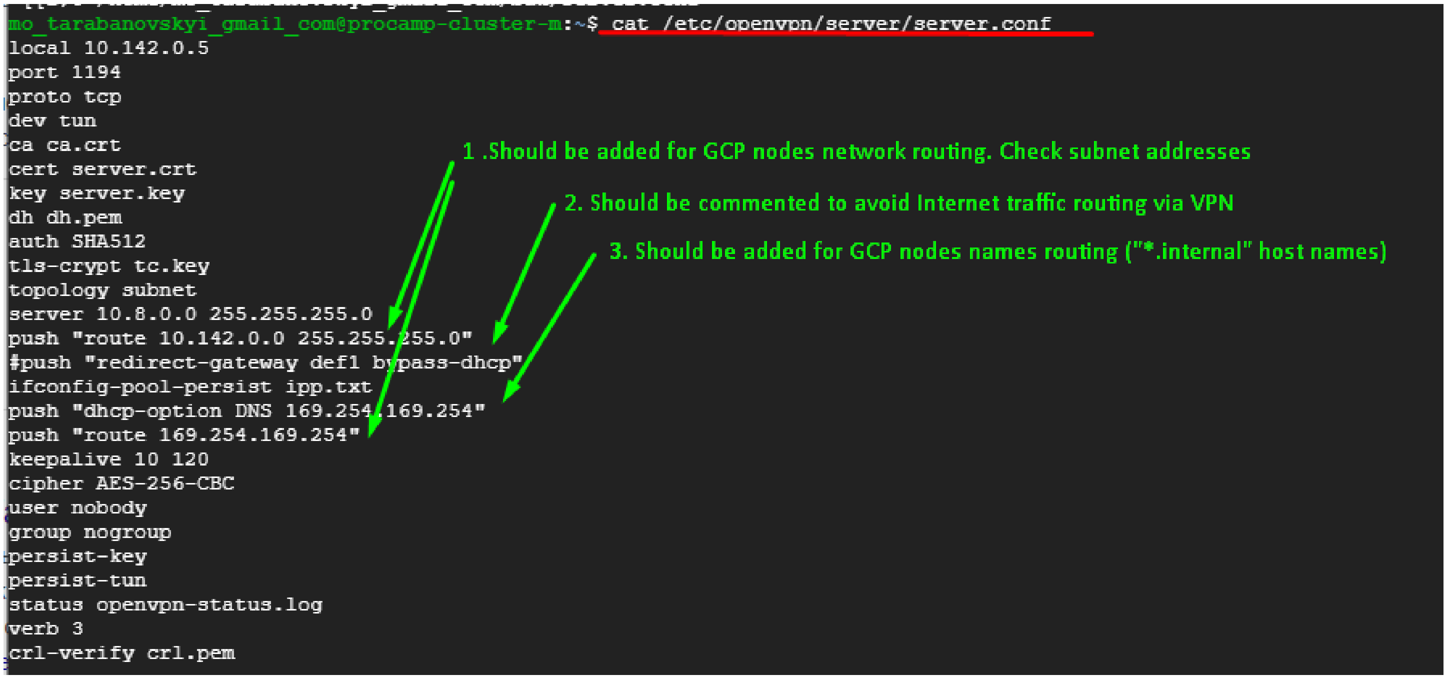


```
...................................................................................................................
...................................................................................................................
...........................+++++
writing new private key to '/etc/openvpn/server/easy-rsa/pki/easy-rsa-32697.NCXVqt/tmp.5ydWqh'
-----
Using configuration from /etc/openvpn/server/easy-rsa/pki/easy-rsa-32697.NCXVqt/tmp.sErTm3
Check that the request matches the signature
Signature ok
The Subject's Distinguished Name is as follows
commonName            :ASN.1 12:'server'
Certificate is to be certified until Nov 26 21:20:52 2030 GMT (3650 days)

Write out database with 1 new entries
Data Base Updated

Using SSL: openssl OpenSSL 1.1.1  11 Sep 2018
Generating a RSA private key
......+++++
................................................+++++
writing new private key to '/etc/openvpn/server/easy-rsa/pki/easy-rsa-304.SYCkbD/tmp.MaG6Fh'
-----
Using configuration from /etc/openvpn/server/easy-rsa/pki/easy-rsa-304.SYCkbD/tmp.cwOk3a
Check that the request matches the signature
Signature ok
The Subject's Distinguished Name is as follows
commonName            :ASN.1 12:'alex'
Certificate is to be certified until Nov 26 21:20:52 2030 GMT (3650 days)

Write out database with 1 new entries
Data Base Updated

Using SSL: openssl OpenSSL 1.1.1  11 Sep 2018
Using configuration from /etc/openvpn/server/easy-rsa/pki/easy-rsa-361.DVw8Qc/tmp.5wxhBe

An updated CRL has been created.
CRL file: /etc/openvpn/server/easy-rsa/pki/crl.pem


Created symlink /etc/systemd/system/multi-user.target.wants/openvpn-iptables.service → /etc/systemd/system/openvpn-
iptables.service.
Created symlink /etc/systemd/system/multi-user.target.wants/openvpn-server@server.service → /lib/systemd/system/ope
nvpn-server@.service.

Finished!

The client configuration is available in: /home/mo tarabanovskyi gmail com/alex.ovpn
New clients can be added by running this script again.
```

7.3.4. Change vpn server configuration and reboot master node

Server configuration example: “sandbox_repo\training\GLBig_Data_ProCamp\infra\vpn\server.conf”

The configuration location is “`/etc/openvpn/server/server.conf`” on GCP master node.



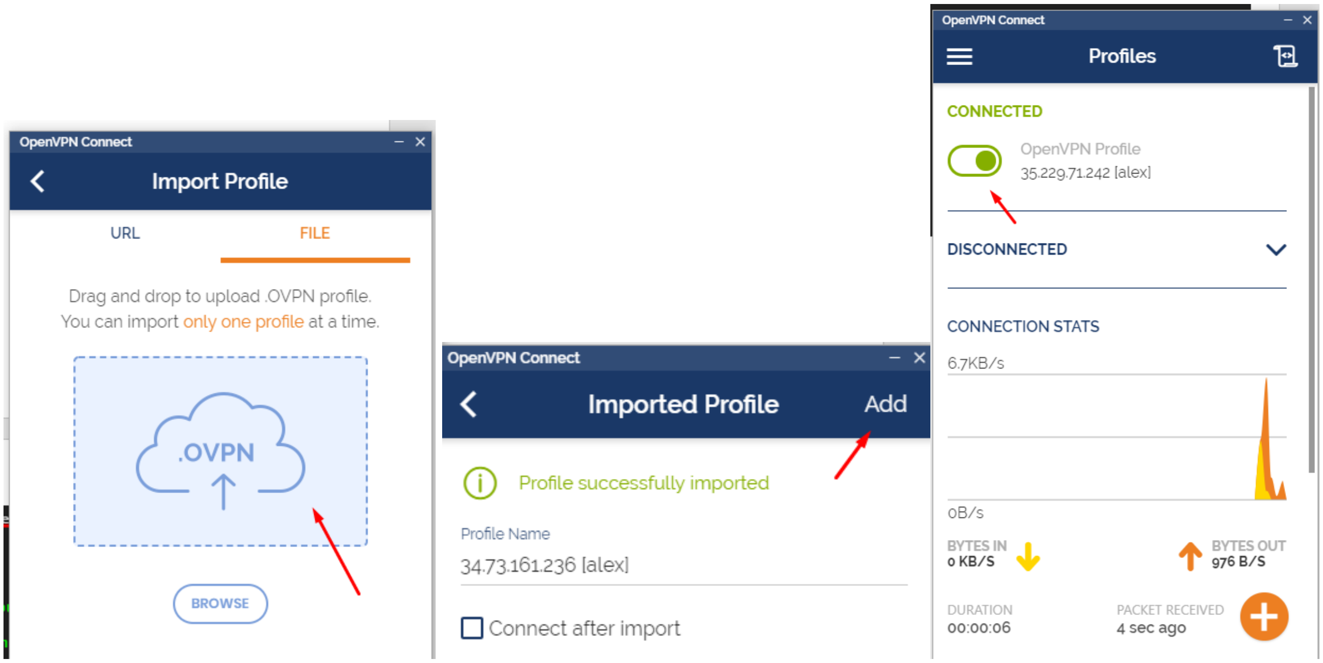
```
mo_tarabanovskyi_gmail_com@procamp-cluster-m:~$ cat /etc/openvpn/server/server.conf
local 10.142.0.5
port 1194
proto tcp
dev tun
ca ca.crt
cert server.crt
key server.key
dh dh.pem
auth SHA512
tls-crypt tc.key
topology subnet
server 10.8.0.0 255.255.255.0
push "route 10.142.0.0 255.255.255.0"
#push "redirect-gateway def1 bypass-dhcp"
ifconfig-pool-persist ipp.txt
push "dhcp-option DNS 169.254.169.254"
push "route 169.254.169.254"
keepalive 10 120
cipher AES-256-CBC
user nobody
group nogroup
persist-key
persist-tun
status openvpn-status.log
verb 3
crl-verify crl.pem
```

1 .Should be added for GCP nodes network routing. Check subnet addresses

2. Should be commented to avoid Internet traffic routing via VPN

3. Should be added for GCP nodes names routing ("*.internal" host names)

7.4. Install Open VPN client and import client configuration file (described in 7.2.3) and enable connection

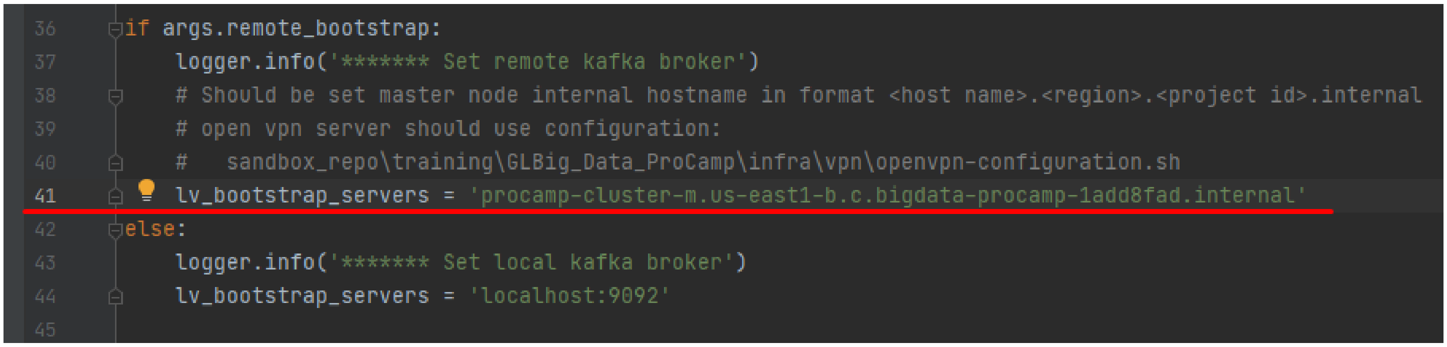


Profile successfully imported

Profile Name
34.73.161.236 [alex]

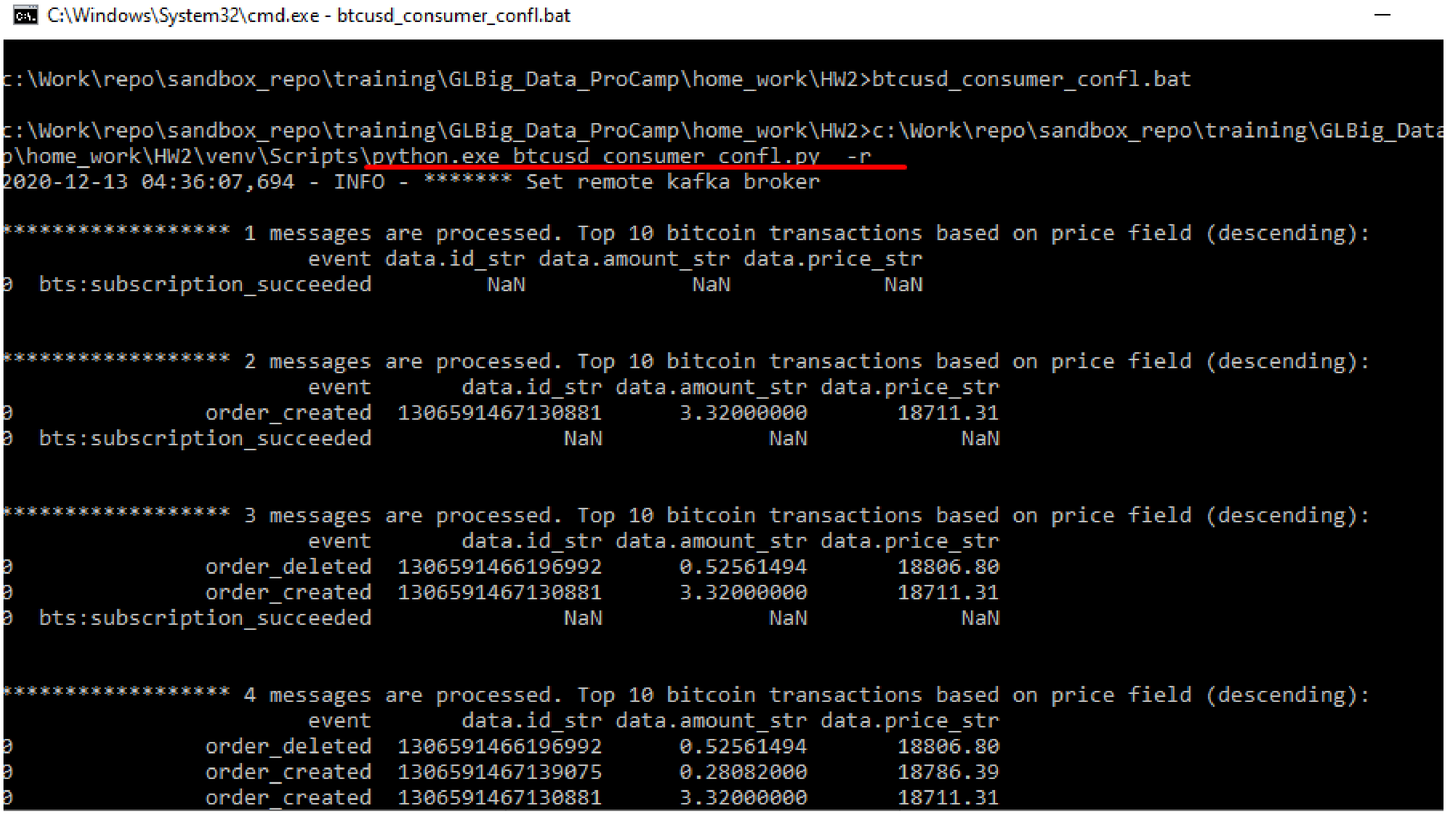
□ Connect after import

7.5. Change GCP master node name in python script "sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2\btcusd_consumer_confl.py":

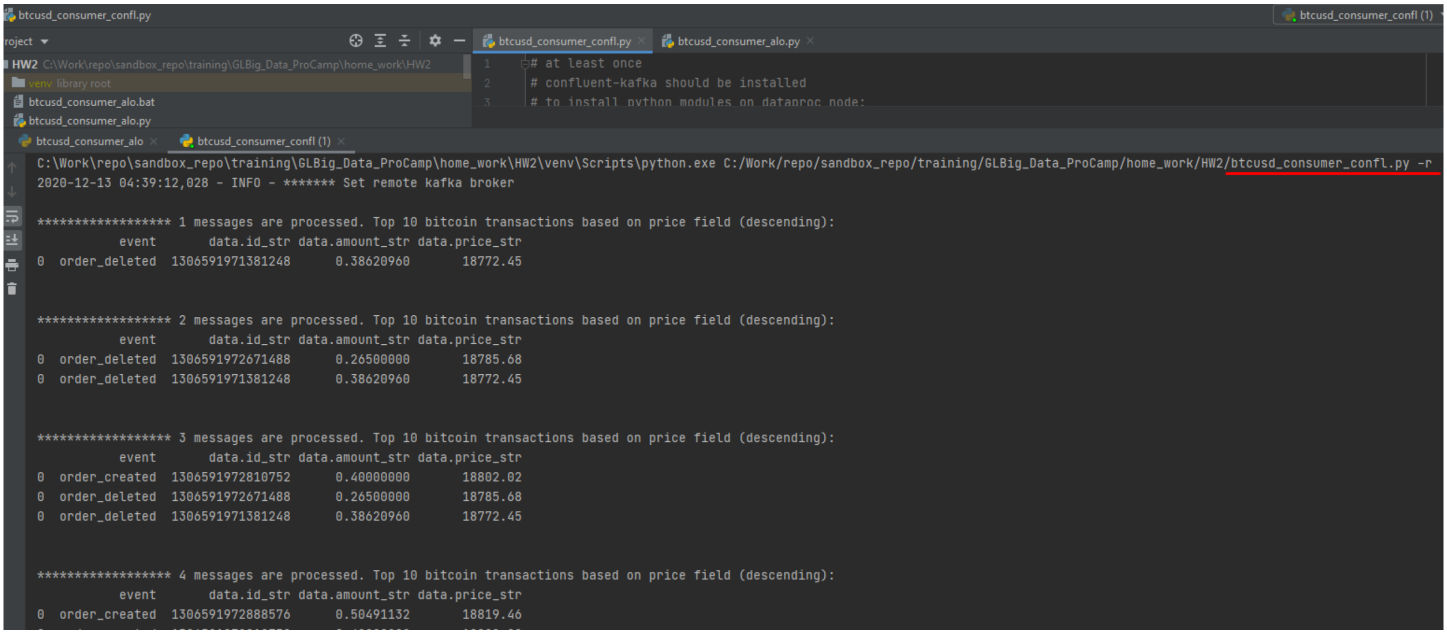


```
if args.remote_bootstrap:
    logger.info('******* Set remote kafka broker')
    # Should be set master node internal hostname in format <host name>.<region>.<project id>.internal
    # open vpn server should use configuration:
    #   sandbox_repo\training\GLBig_Data_ProCamp\infra\vpn\openvpn-configuration.sh
    lv_bootstrap_servers = 'procamp-cluster-m.us-east1-b.c.bigdata-procamp-1add8fad.internal'
else:
    logger.info('******* Set local kafka broker')
    lv_bootstrap_servers = 'localhost:9092'
```

7.6. Launch consumer with option “-r” (remote broker) – “sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2\btcusd_consumer_confl.py”



```
c:\Work\repo\sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2>btcusd_consumer_confl.bat

c:\Work\repo\sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2>c:\Work\repo\sandbox_repo\training\GLBig_Data
p\home_work\HW2\venv\Scripts\python.exe btcusd_consumer_confl.py  -r
2020-12-13 04:36:07,694 - INFO - ******* Set remote kafka broker

****************** 1 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event data.id_str data.amount_str data.price_str
0  bts:subscription_succeeded         NaN             NaN            NaN


****************** 2 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_created  1306591467130881      3.32000000       18711.31
0  bts:subscription_succeeded               NaN             NaN            NaN


****************** 3 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_deleted  1306591466196992      0.52561494       18806.80
0               order_created  1306591467130881      3.32000000       18711.31
0  bts:subscription_succeeded               NaN             NaN            NaN


****************** 4 messages are processed. Top 10 bitcoin transactions based on price field (descending):
                        event       data.id_str data.amount_str data.price_str
0               order_deleted  1306591466196992      0.52561494       18806.80
0               order_created  1306591467139075      0.28082000       18786.39
0               order_created  1306591467130881      3.32000000       18711.31
```



```
# at least once
# confluent-kafka should be installed
# to install python modules on dataproc node:
```

```
C:\Work\repo\sandbox_repo\training\GLBig_Data_ProCamp\home_work\HW2\venv\Scripts\python.exe C:/Work/repo/sandbox_repo/training/GLBig_Data_ProCamp/home_work/HW2/btcusd_consumer_confl.py -r
2020-12-13 04:39:12,028 - INFO - ******* Set remote kafka broker

****************** 1 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_deleted  1306591971381248      0.38620960       18772.45


****************** 2 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_deleted  1306591972671488      0.26500000       18785.68
0  order_deleted  1306591971381248      0.38620960       18772.45


****************** 3 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_created  1306591972810752      0.40000000       18802.02
0  order_deleted  1306591972671488      0.26500000       18785.68
0  order_deleted  1306591971381248      0.38620960       18772.45


****************** 4 messages are processed. Top 10 bitcoin transactions based on price field (descending):
           event       data.id_str data.amount_str data.price_str
0  order_created  1306591972888576      0.50491132       18819.46
```