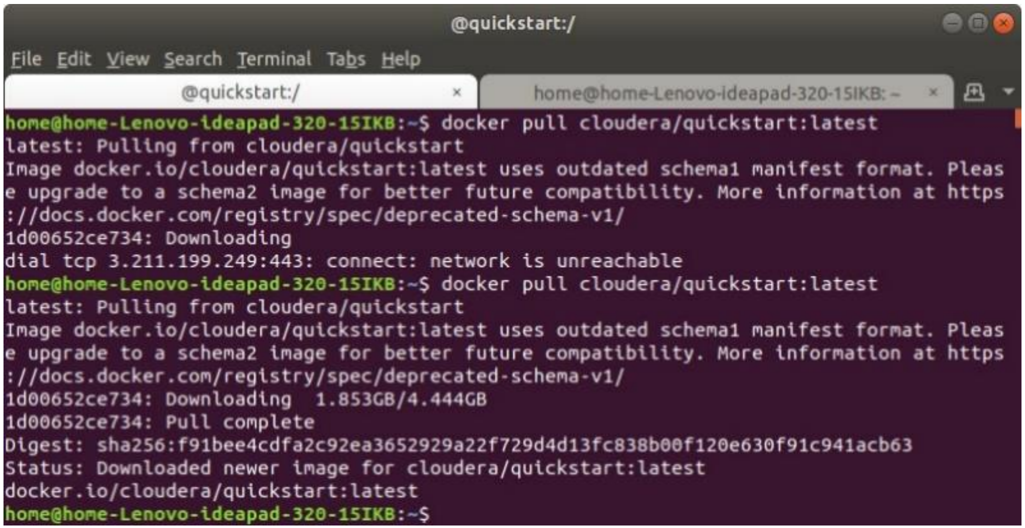
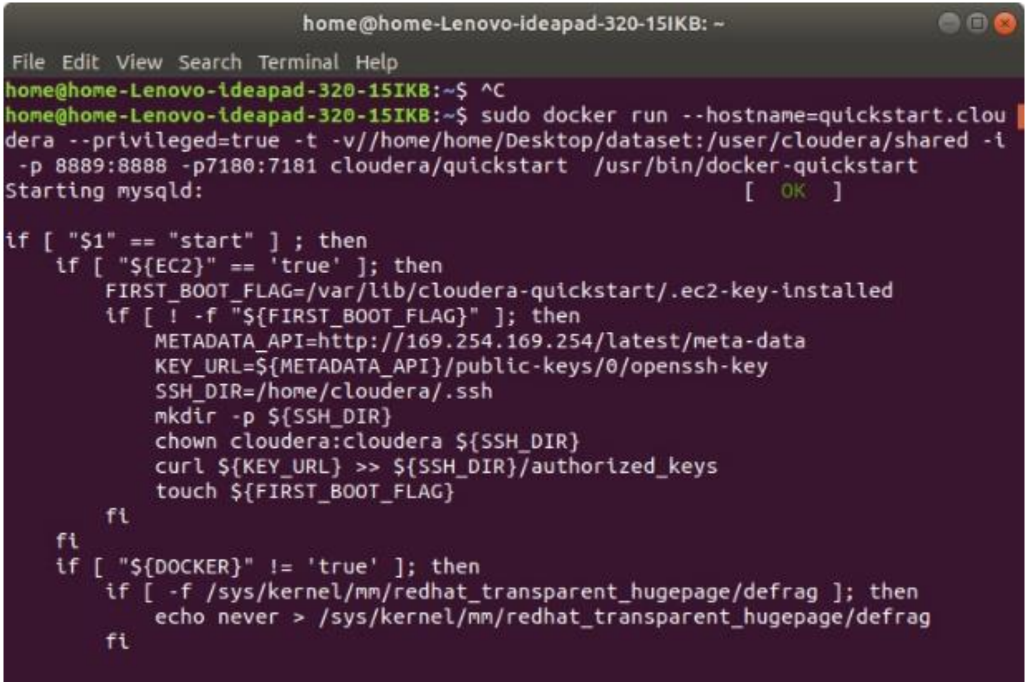
Map Reduce Lab
Marium Jamal – 14881

Before Tasks – downloading process:


@quickstart:/
File Edit View Search Terminal Tabs Help
@quickstart:/
home@home-Lenovo-ideapad-320-15IKB: ~

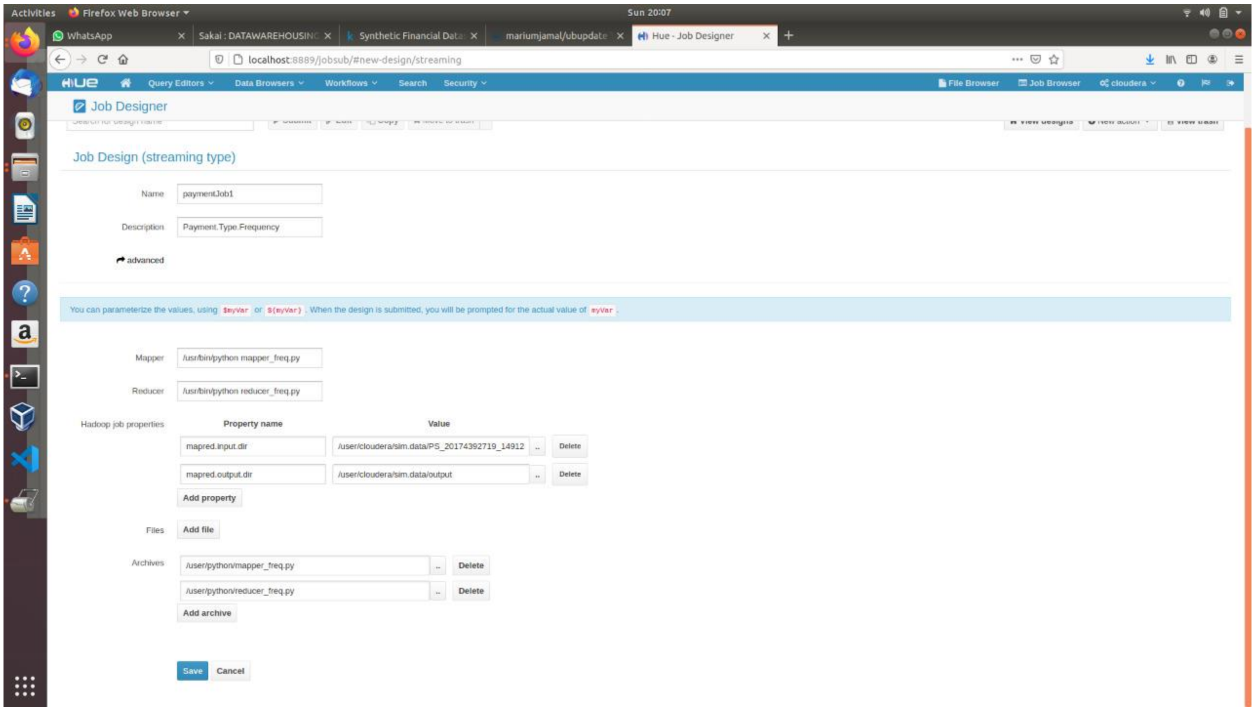
```
home@home-Lenovo-ideapad-320-15IKB:~$ docker pull cloudera/quickstart:latest
latest: Pulling from cloudera/quickstart
Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format. Pleas
e upgrade to a schema2 image for better future compatibility. More information at https
://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading
dial tcp 3.211.199.249:443: connect: network is unreachable
home@home-Lenovo-ideapad-320-15IKB:~$ docker pull cloudera/quickstart:latest
latest: Pulling from cloudera/quickstart
Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format. Pleas
e upgrade to a schema2 image for better future compatibility. More information at https
://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading  1.853GB/4.444GB
1d00652ce734: Pull complete
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Downloaded newer image for cloudera/quickstart:latest
docker.io/cloudera/quickstart:latest
home@home-Lenovo-ideapad-320-15IKB:~$
```


home@home-Lenovo-ideapad-320-15IKB: ~
File Edit View Search Terminal Help

```
home@home-Lenovo-ideapad-320-15IKB:~$ ^C
home@home-Lenovo-ideapad-320-15IKB:~$ sudo docker run --hostname=quickstart.clou
dera --privileged=true -t -v//home/home/Desktop/dataset:/user/cloudera/shared -i
 -p 8889:8888 -p7180:7181 cloudera/quickstart  /usr/bin/docker-quickstart
Starting mysqld:                                           [  OK  ]

if [ "$1" == "start" ] ; then
    if [ "${EC2}" == 'true' ]; then
        FIRST_BOOT_FLAG=/var/lib/cloudera-quickstart/.ec2-key-installed
        if [ ! -f "${FIRST_BOOT_FLAG}" ]; then
            METADATA_API=http://169.254.169.254/latest/meta-data
            KEY_URL=${METADATA_API}/public-keys/0/openssh-key
            SSH_DIR=/home/cloudera/.ssh
            mkdir -p ${SSH_DIR}
            chown cloudera:cloudera ${SSH_DIR}
            curl ${KEY_URL} >> ${SSH_DIR}/authorized_keys
            touch ${FIRST_BOOT_FLAG}
        fi
    fi
    if [ "${DOCKER}" != 'true' ]; then
        if [ -f /sys/kernel/mm/redhat_transparent_hugepage/defrag ]; then
            echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
        fi
```

Task 1:



Job Design (streaming type)

Name: paymentJob1

Description: Payment.Type.Frequency

advanced

You can parameterize the values, using `$myVar` or `${myVar}`. When the design is submitted, you will be prompted for the actual value of `myVar`.

Mapper: /usr/bin/python mapper_freq.py

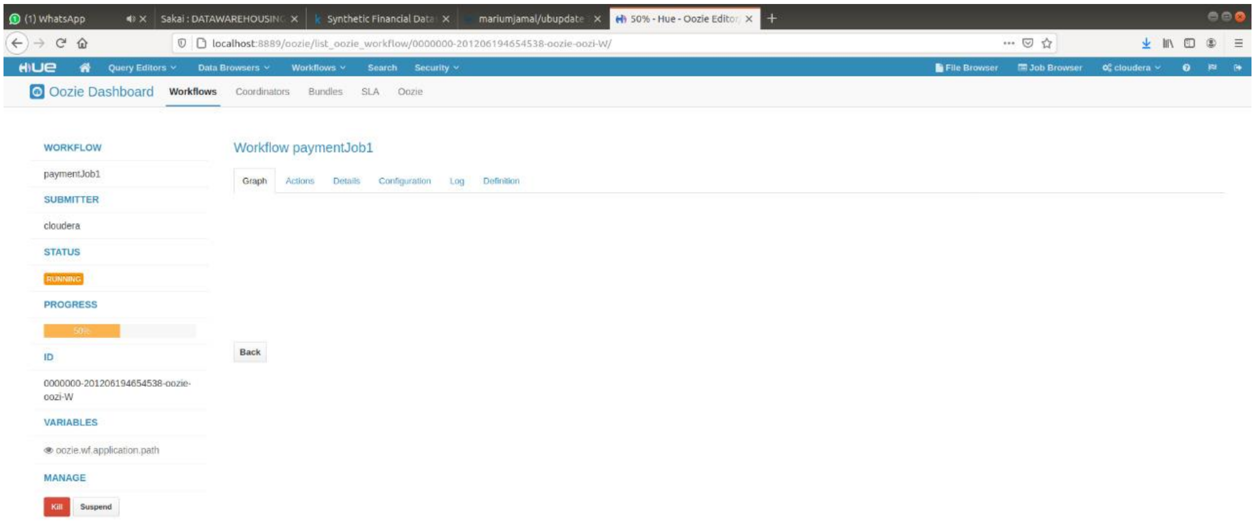
Reducer: /usr/bin/python reducer_freq.py

Hadoop job properties

Property name	Value
mapred.input.dir	/user/cloudera/sim.data/PS_20174392719_14912
mapred.output.dir	/user/cloudera/sim.data/output

Files

Archives:
/user/python/mapper_freq.py
/user/python/reducer_freq.py



Workflow paymentJob1

WORKFLOW
paymentJob1

SUBMITTER
cloudera

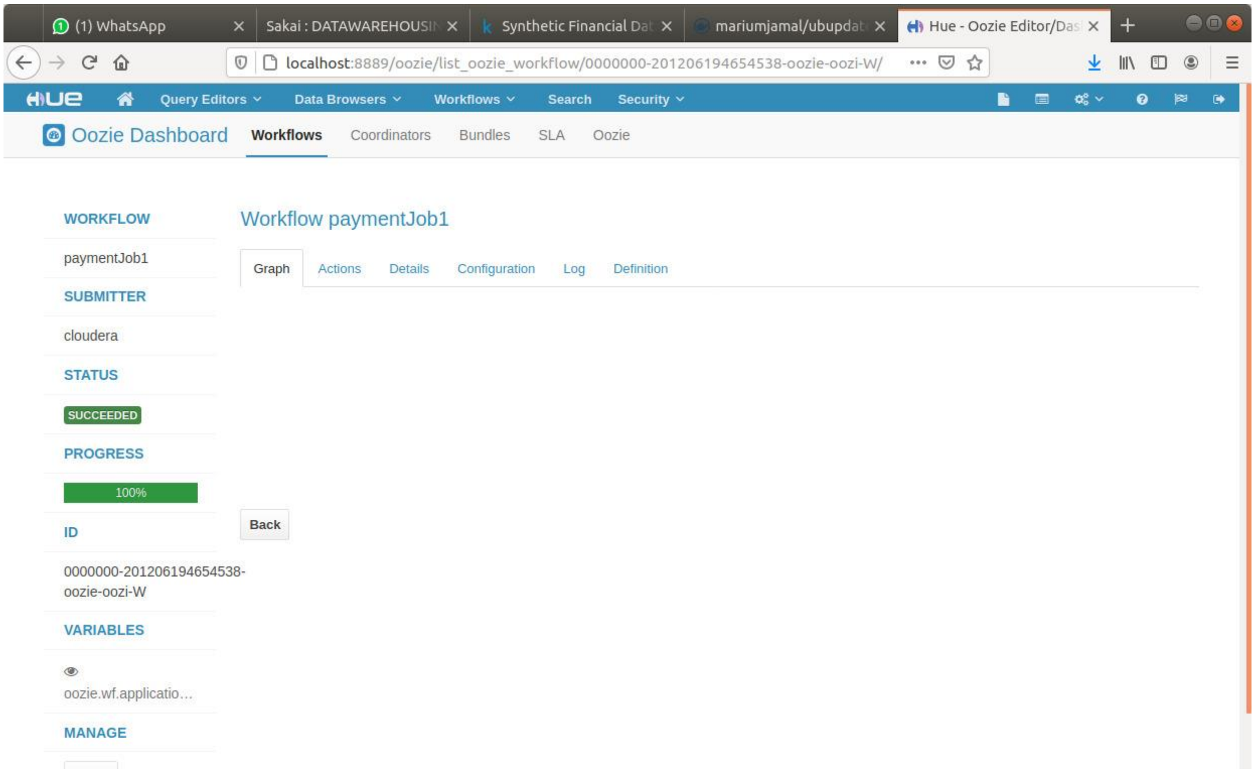
STATUS
RUNNING

PROGRESS
50%

ID
0000000-201206194654538-oozie-oozi-W

VARIABLES
oozie.wf.application.path

MANAGE



Workflow paymentJob1

WORKFLOW
paymentJob1

SUBMITTER
cloudera

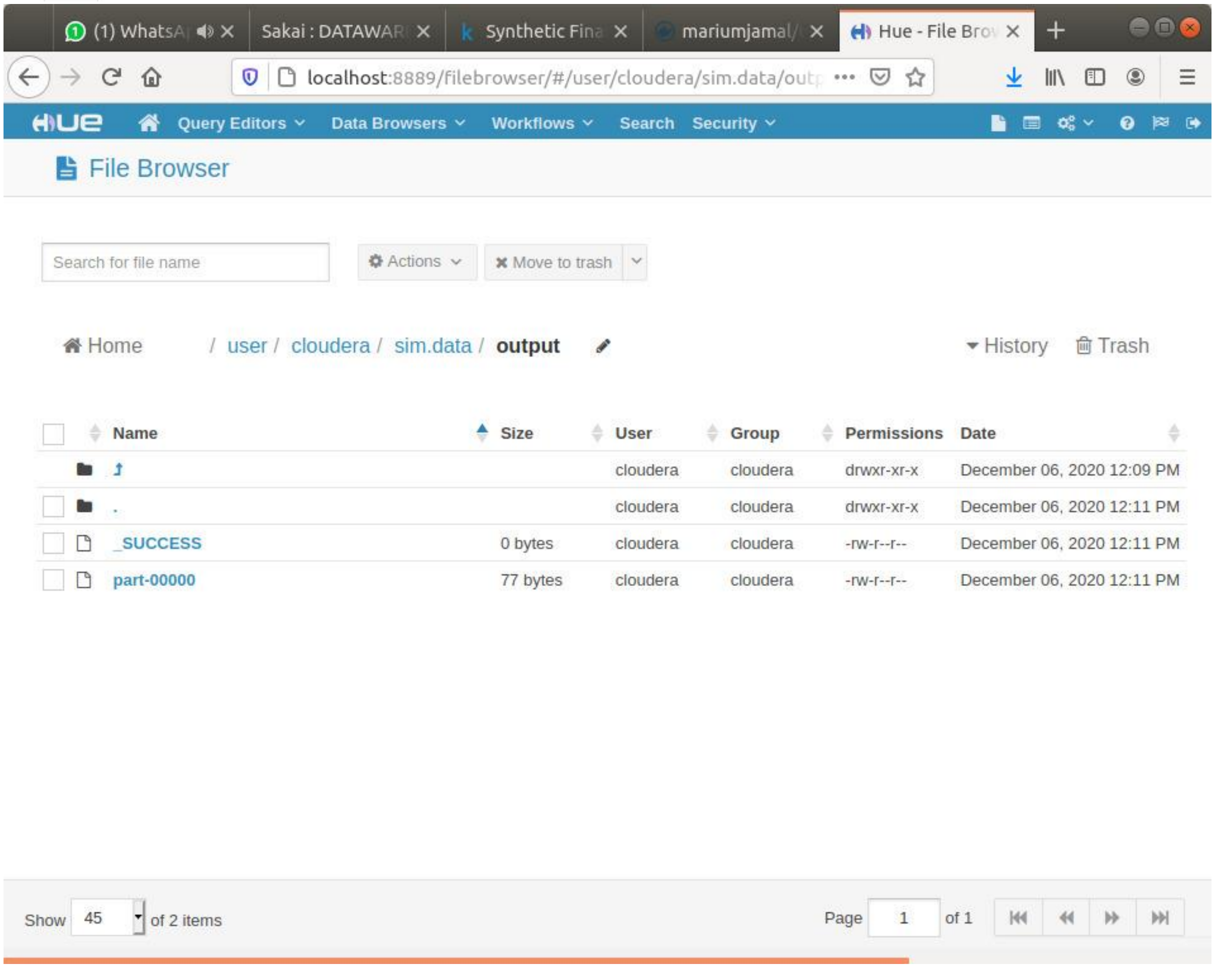
STATUS
SUCCEEDED

PROGRESS
100%

ID
0000000-201206194654538-oozie-oozi-W

VARIABLES
oozie.wf.applicatio…

MANAGE

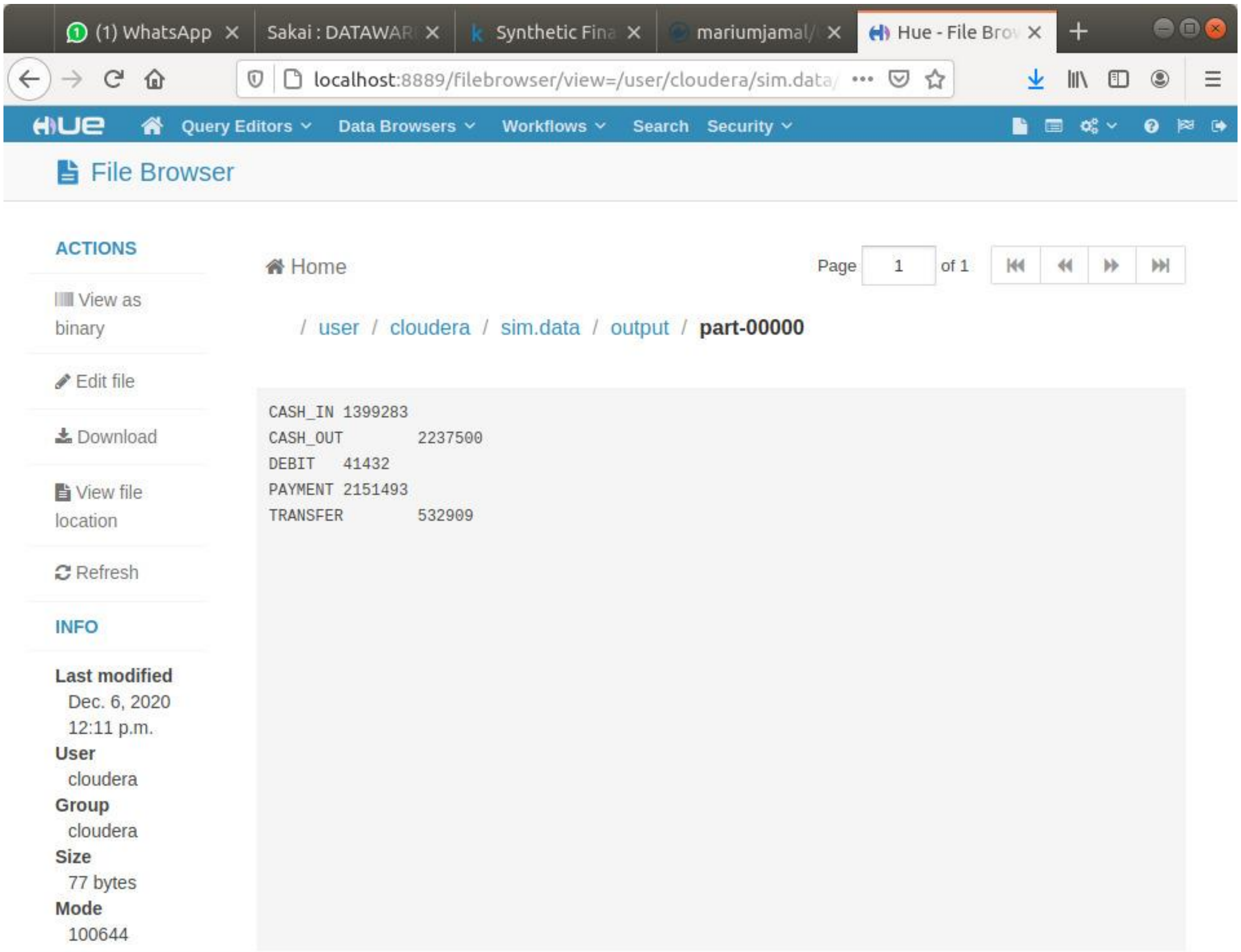


/ user / cloudera / sim.data / output

	Name	Size	User	Group	Permissions	Date
	↥		cloudera	cloudera	drwxr-xr-x	December 06, 2020 12:09 PM
□	.		cloudera	cloudera	drwxr-xr-x	December 06, 2020 12:11 PM
□	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 06, 2020 12:11 PM
□	part-00000	77 bytes	cloudera	cloudera	-rw-r--r--	December 06, 2020 12:11 PM

Show 45 of 2 items

Page 1 of 1

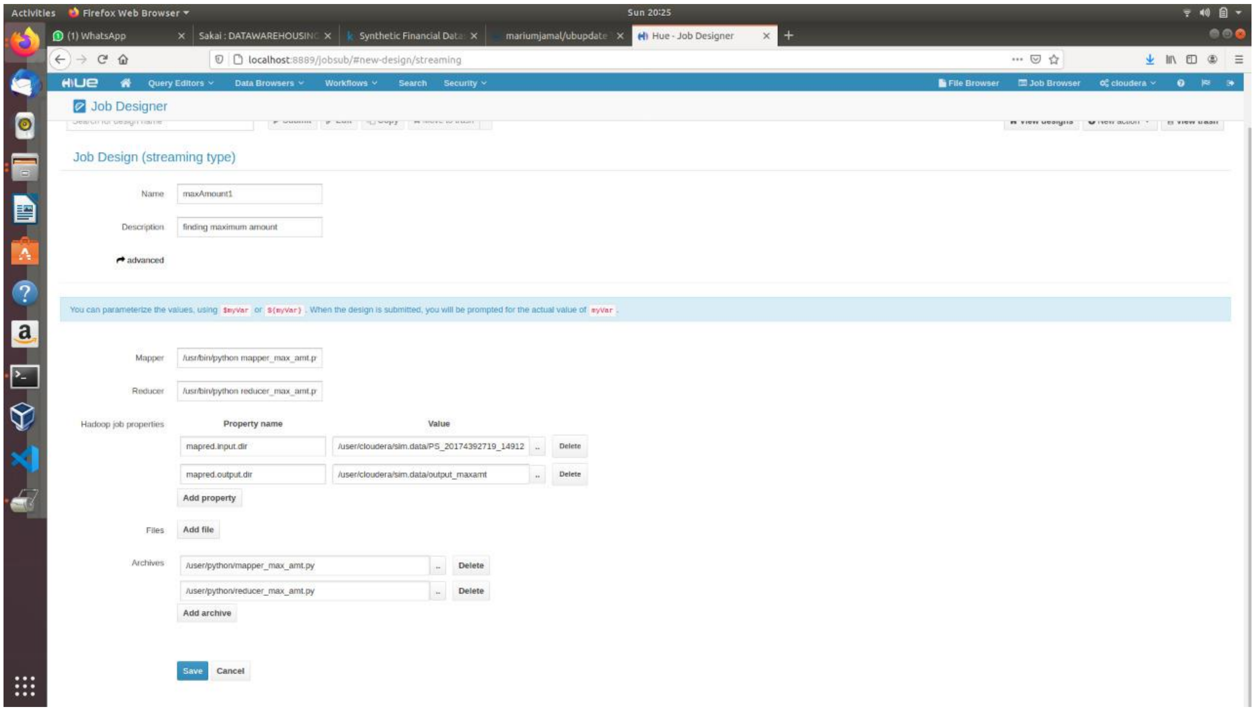


/ user / cloudera / sim.data / output / **part-00000**

```
CASH_IN 1399283
CASH_OUT        2237500
DEBIT   41432
PAYMENT 2151493
TRANSFER        532909
```

Last modified
Dec. 6, 2020 12:11 p.m.
User
cloudera
Group
cloudera
Size
77 bytes
Mode
100644

Task 2:



Job Designer

Job Design (streaming type)

Name: maxAmount1

Description: finding maximum amount

advanced

You can parameterize the values, using `$myVar` or `${myVar}`. When the design is submitted, you will be prompted for the actual value of `myVar`.

Mapper: /usr/bin/python mapper_max_amt.p

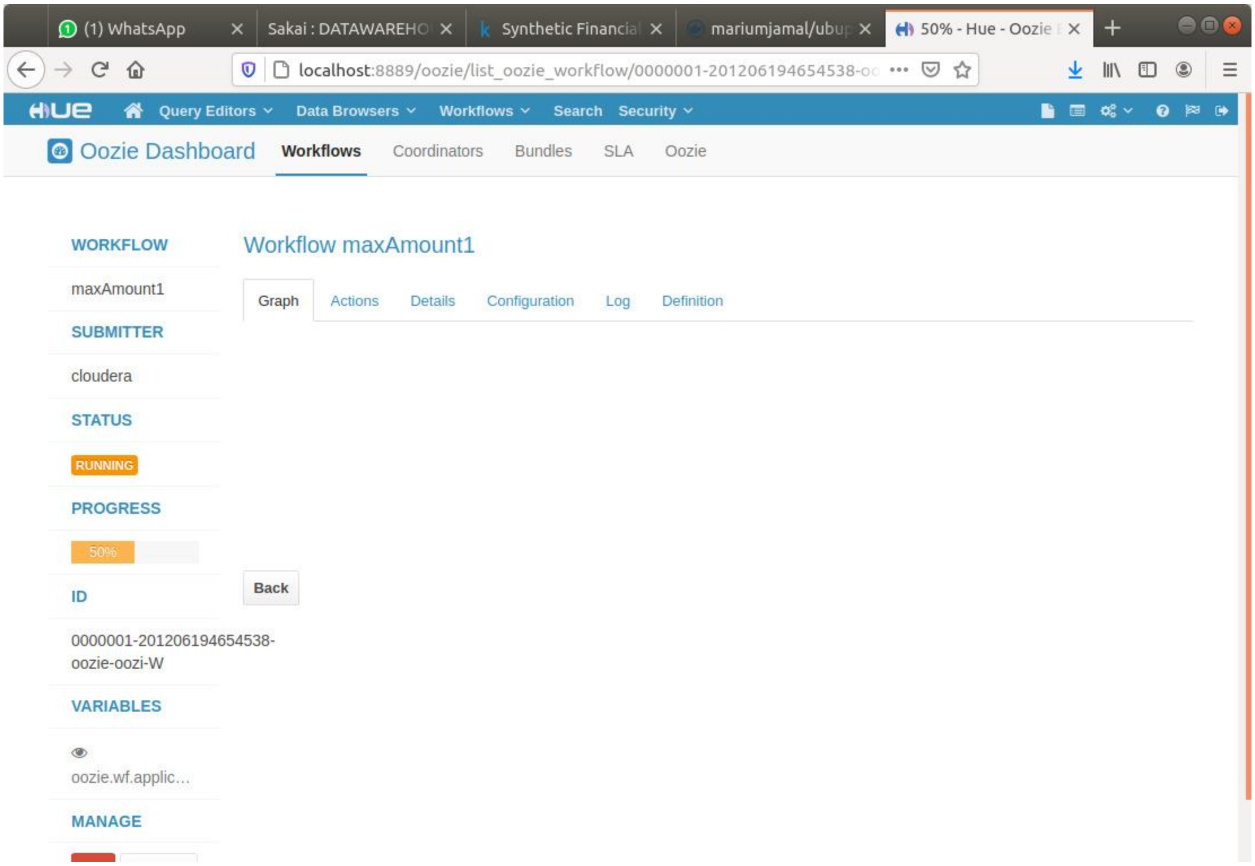
Reducer: /usr/bin/python reducer_max_amt.p

Hadoop job properties

Property name	Value
mapred.input.dir	/user/cloudera/sim.data/PS_20174392719_14912
mapred.output.dir	/user/cloudera/sim.data/output_maxamt

Files

Archives: /user/python/mapper_max_amt.py
/user/python/reducer_max_amt.py



Workflow maxAmount1

WORKFLOW
maxAmount1

SUBMITTER
cloudera

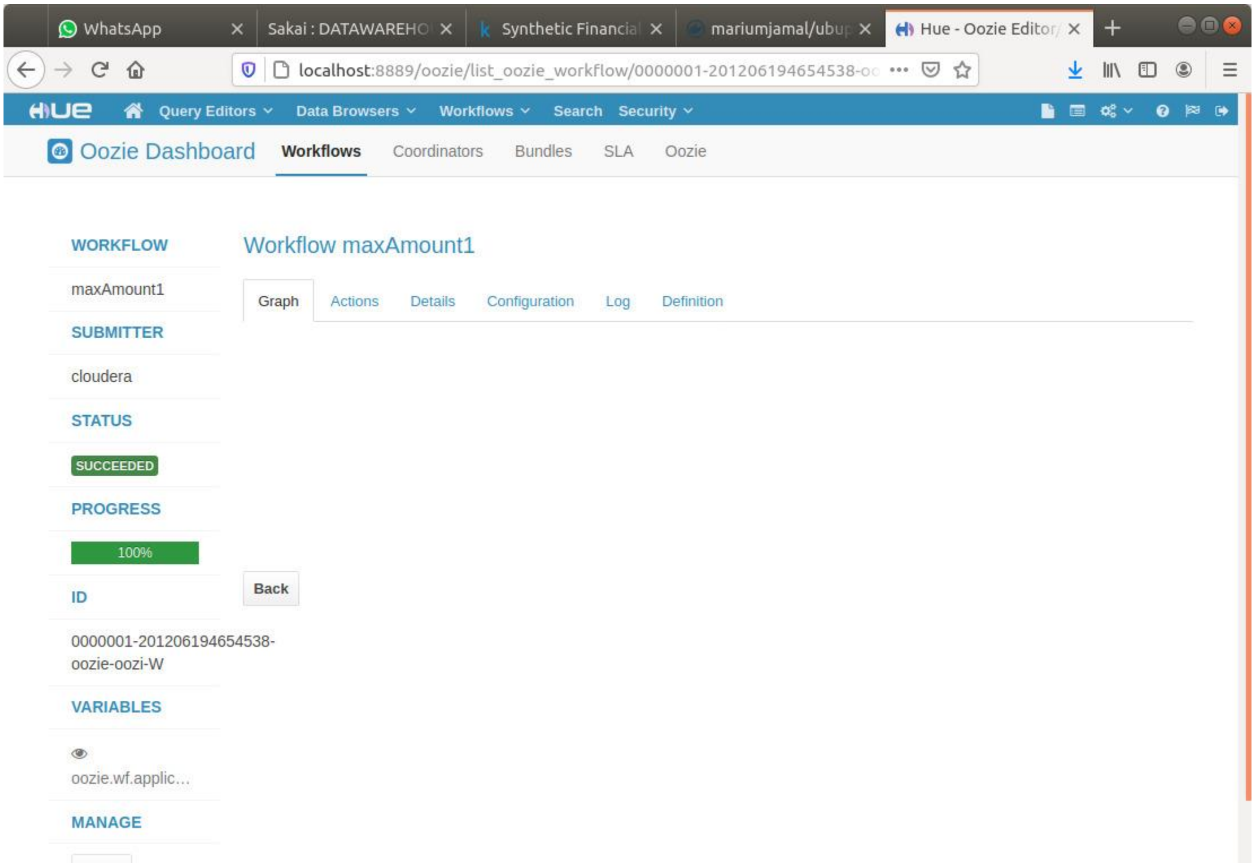
STATUS
RUNNING

PROGRESS
50%

ID
0000001-201206194654538-oozie-oozi-W

VARIABLES
oozie.wf.applic…

MANAGE



localhost:8889/oozie/list_oozie_workflow/0000001-201206194654538-oo

Workflow maxAmount1

WORKFLOW
maxAmount1

SUBMITTER
cloudera

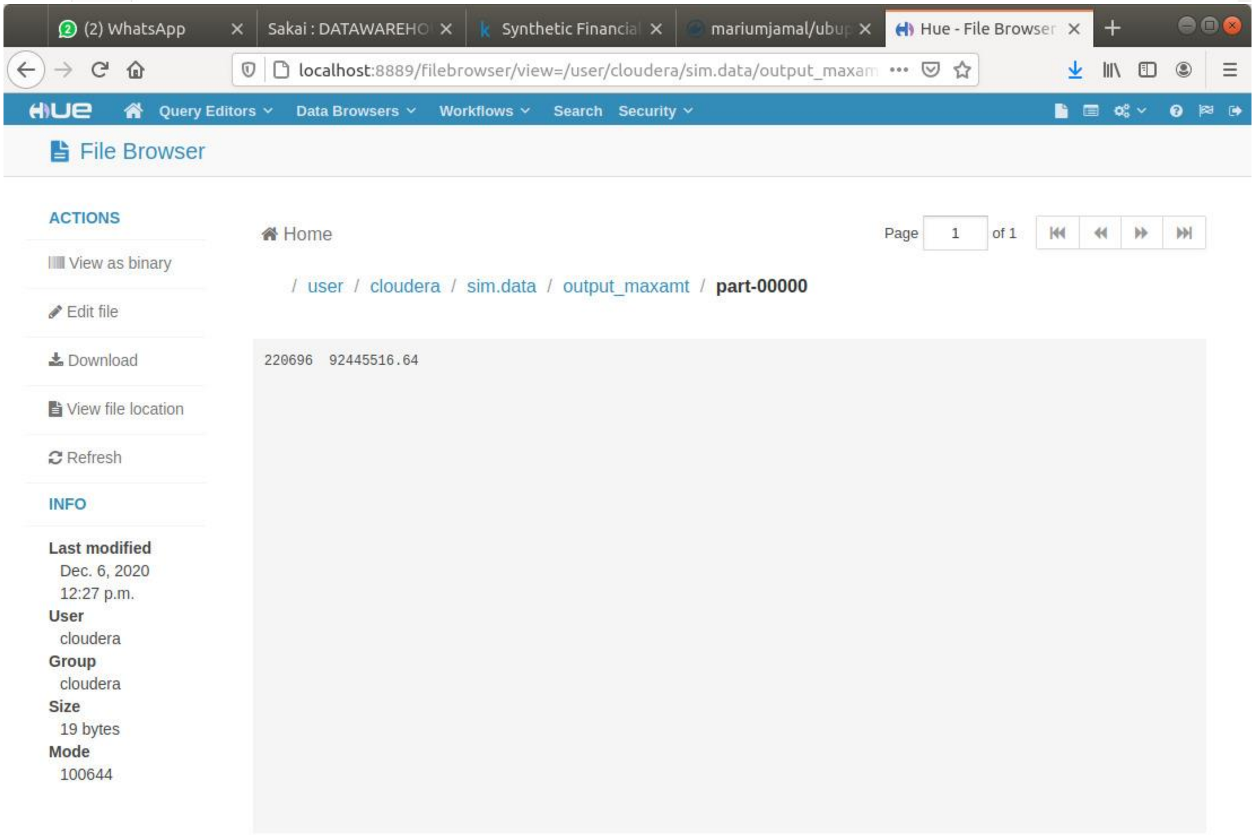
STATUS
SUCCEEDED

PROGRESS
100%

ID
0000001-201206194654538-oozie-oozi-W

VARIABLES
oozie.wf.applic...

MANAGE



localhost:8889/filebrowser/view=/user/cloudera/sim.data/output_maxam

/ user / cloudera / sim.data / output_maxamt / **part-00000**

```
220696  92445516.64
```

INFO

Last modified
Dec. 6, 2020
12:27 p.m.
User
cloudera
Group
cloudera
Size
19 bytes
Mode
100644

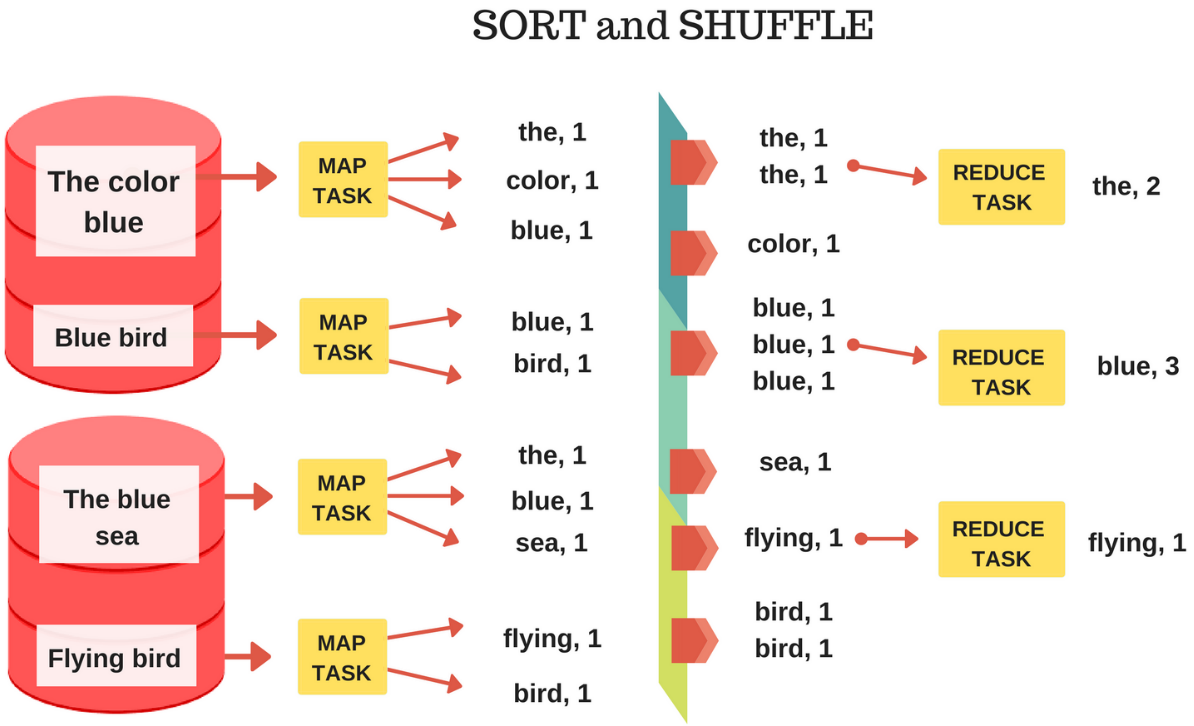
Task 3:

a) Why is shuffling needed in MR? Give an example to justify.

Shuffling is required in MR to transfer data from the mappers to the reducer because it is the process in which sorted intermediate output from mappers is transferred as the input to the reducer; otherwise, reducers would not have any input. An example of shuffling is given below.

b) Why is sorting needed in MR? Give an example to justify.

Sorting is needed in MR because it helps reducer to distiguish when a new reduce task should start i.e. it identifies when the next key in the reducer's input is different than the previous. An example for sorting is given below.



SORT and SHUFFLE
The color blue
Blue bird
The blue sea
Flying bird
MAP TASK
MAP TASK
MAP TASK
MAP TASK
the, 1
color, 1
blue, 1
blue, 1
bird, 1
the, 1
blue, 1
sea, 1
flying, 1
bird, 1
the, 1
the, 1
color, 1
blue, 1
blue, 1
blue, 1
sea, 1
flying, 1
bird, 1
bird, 1
REDUCE TASK
REDUCE TASK
REDUCE TASK
the, 2
blue, 3
flying, 1