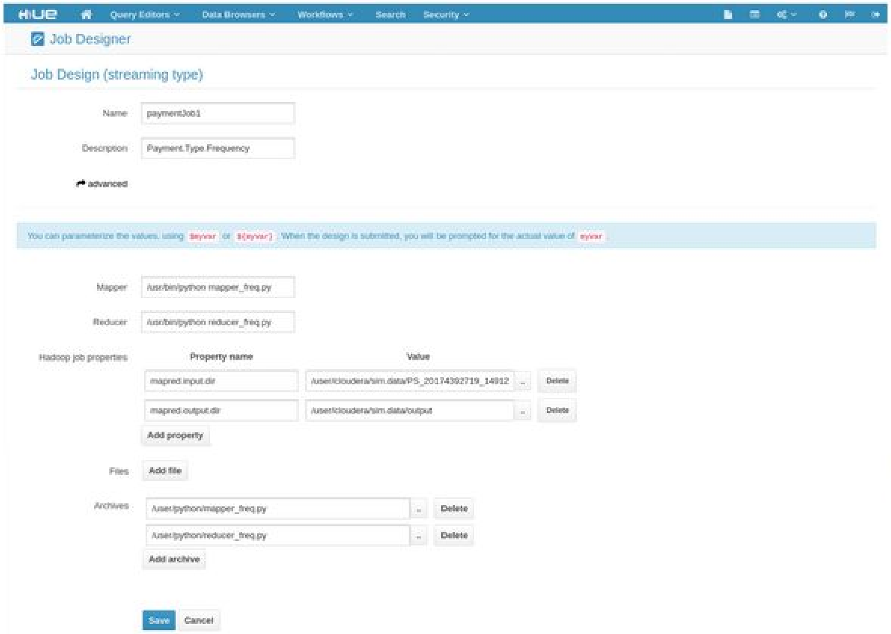
Map Reduce Lab

Haaniyah Muhammad Mundia 14804

MR Task 1: Payment Job

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

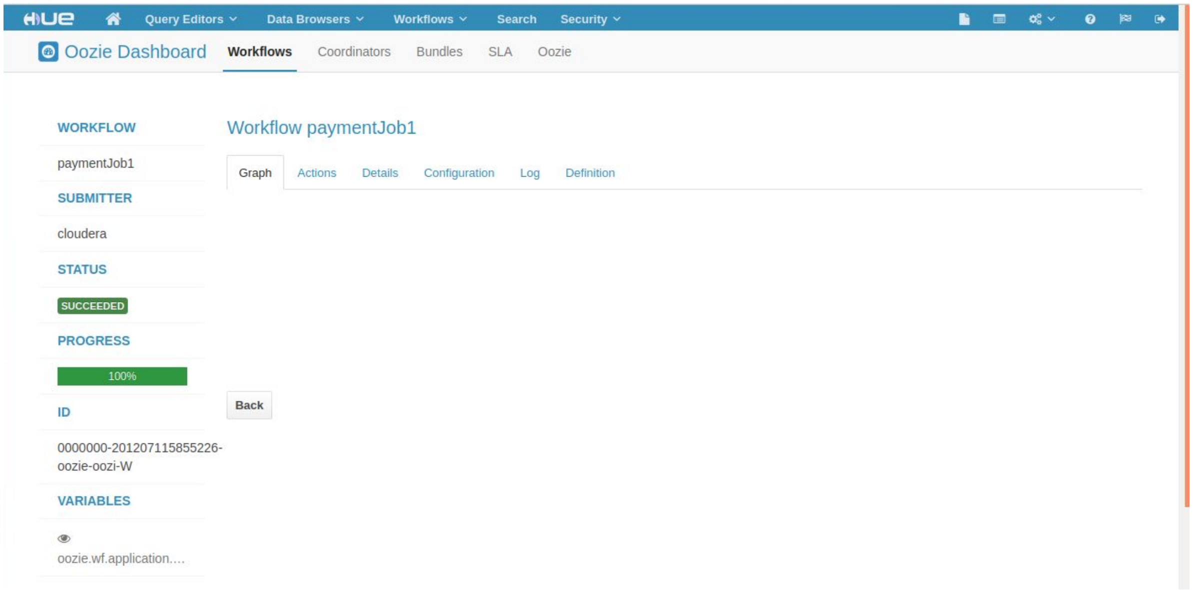
Hadoop job properties:

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

2-



Workflow paymentJob1

WORKFLOW: paymentJob1

SUBMITTER: cloudera

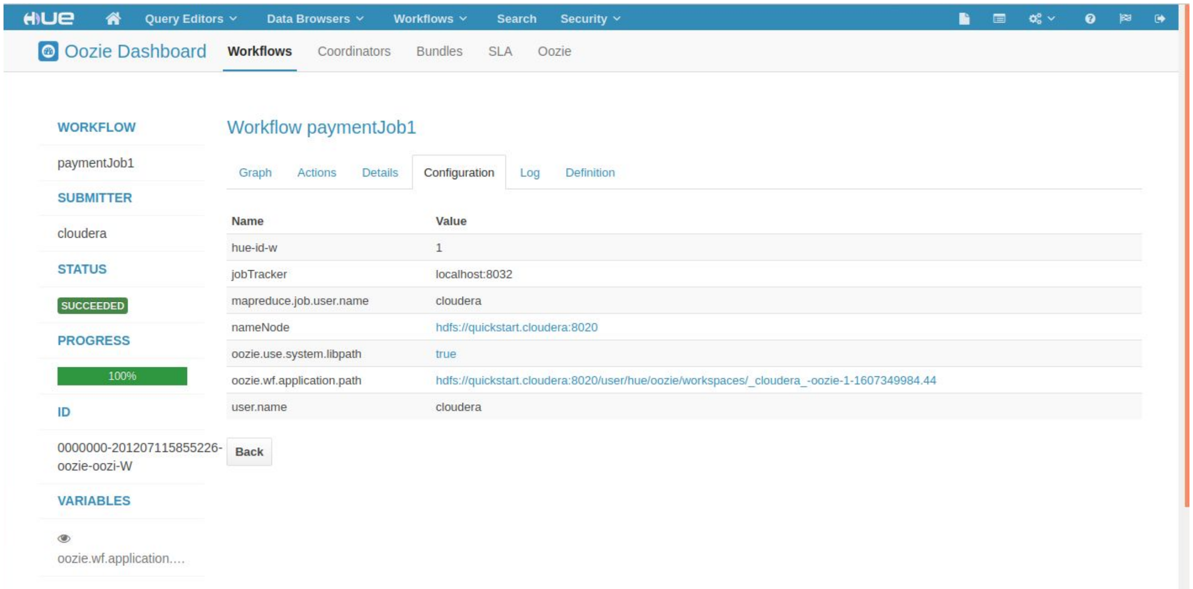
STATUS: SUCCEEDED

PROGRESS: 100%

ID: 0000000-201207115855226-oozie-oozi-W

VARIABLES: oozie.wf.application....

3-

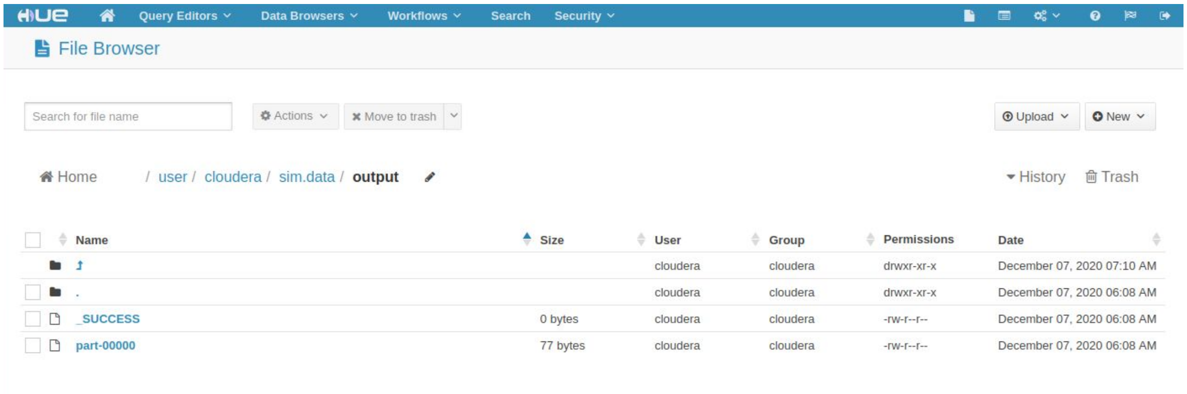


WORKFLOW
paymentJob1
SUBMITTER
cloudera
STATUS
SUCCEEDED
PROGRESS
100%
ID
0000000-201207115855226-oozie-oozi-W
VARIABLES
oozie.wf.application.…

Workflow paymentJob1

Name	Value
hue-id-w	1
jobTracker	localhost:8032
mapreduce.job.user.name	cloudera
nameNode	hdfs://quickstart.cloudera:8020
oozie.use.system.libpath	true
oozie.wf.application.path	hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/_cloudera_-oozie-1-1607349984.44
user.name	cloudera

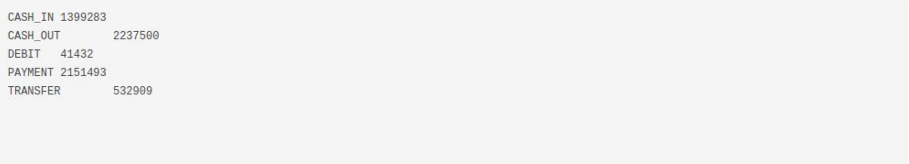
4-



Home / user / cloudera / sim.data / output

Name	Size	User	Group	Permissions	Date
↥		cloudera	cloudera	drwxr-xr-x	December 07, 2020 07:10 AM
.		cloudera	cloudera	drwxr-xr-x	December 07, 2020 06:08 AM
_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 06:08 AM
part-00000	77 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 06:08 AM

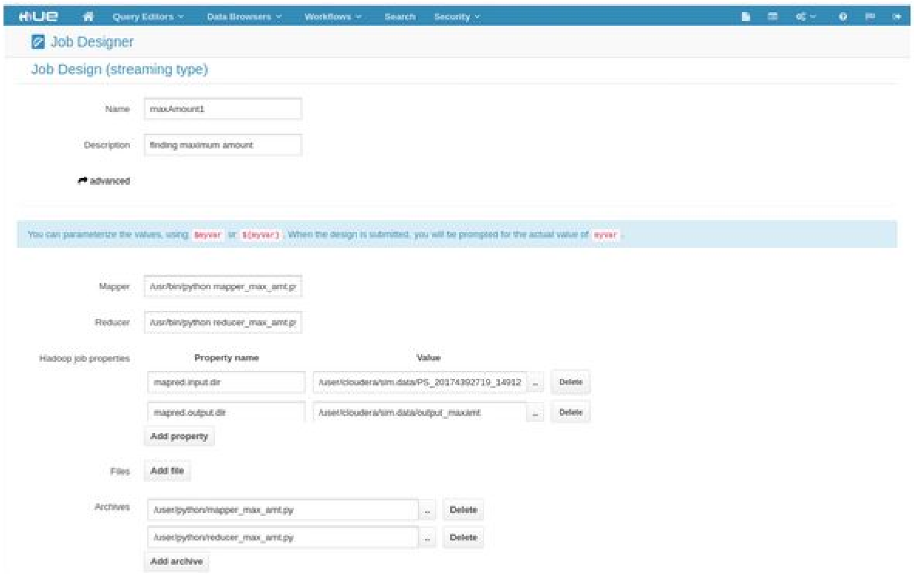
5- Output



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

MR Task 2: Maximum Amount

1-



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

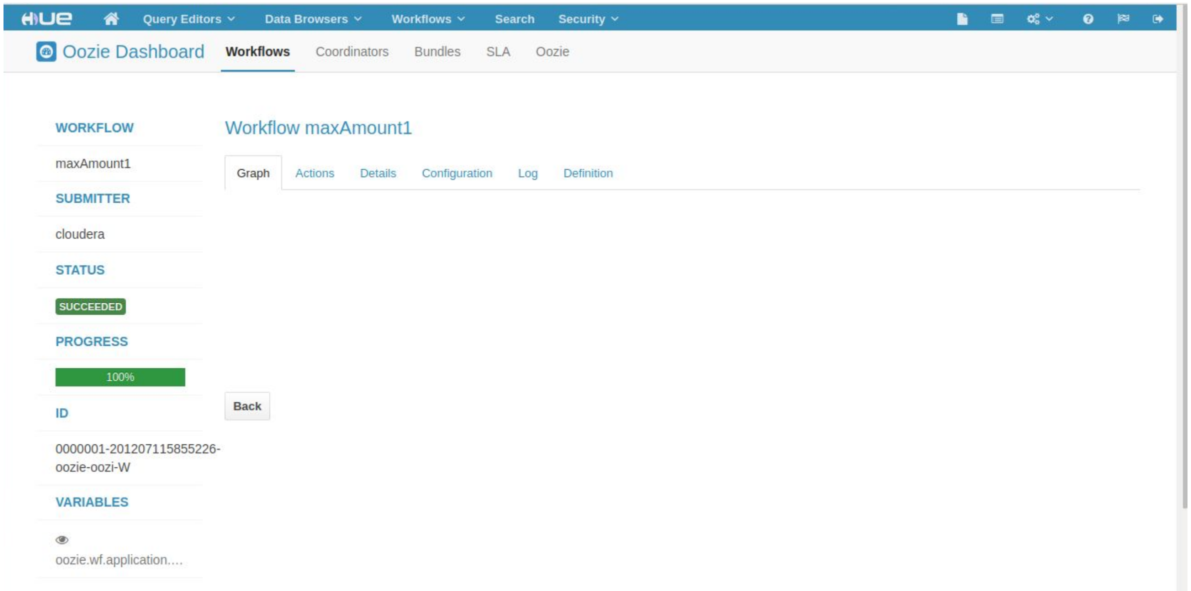
Hadoop job properties

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

2-



Workflow maxAmount1

WORKFLOW
maxAmount1

SUBMITTER
cloudera

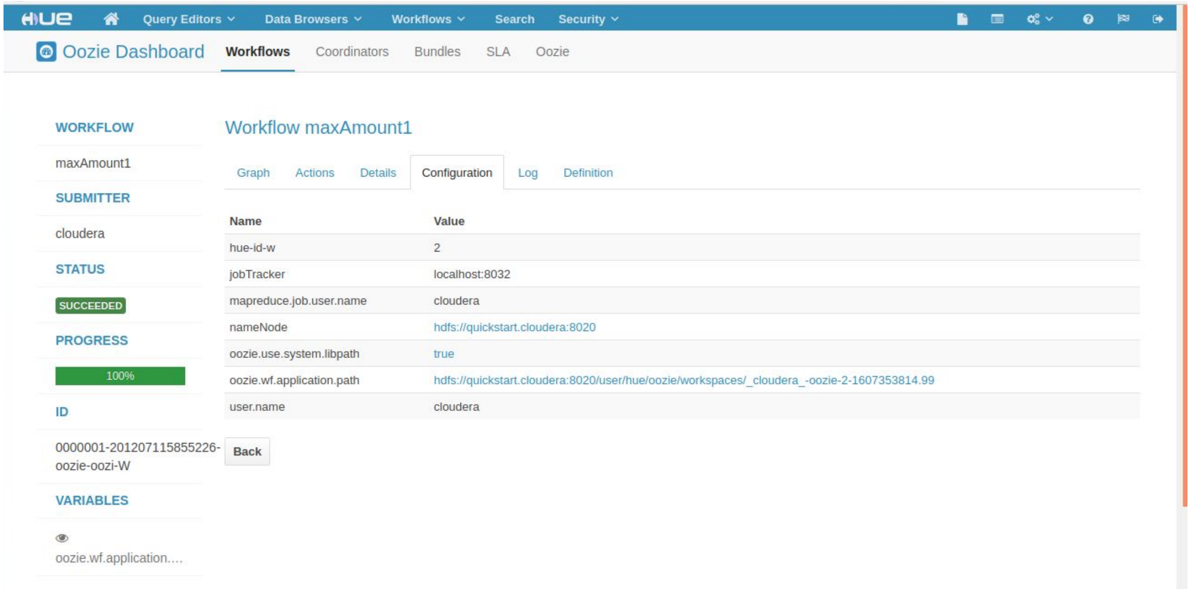
STATUS
SUCCEEDED

PROGRESS
100%

ID
0000001-201207115855226-oozie-oozi-W

VARIABLES
oozie.wf.application....

3-



WORKFLOW

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000001-201207115855226-oozie-oozi-W

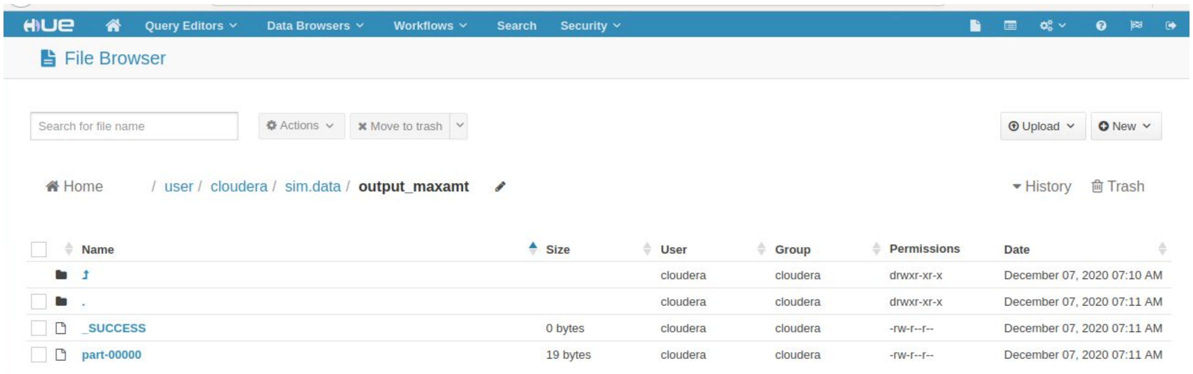
VARIABLES

oozie.wf.application....

Workflow maxAmount1

Name	Value
hue-id-w	2
jobTracker	localhost:8032
mapreduce.job.user.name	cloudera
nameNode	hdfs://quickstart.cloudera:8020
oozie.use.system.libpath	true
oozie.wf.application.path	hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/_cloudera_-oozie-2-1607353814.99
user.name	cloudera

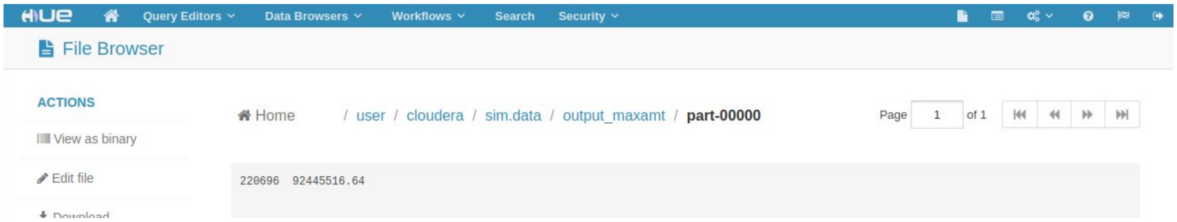
4-



Home / user / cloudera / sim.data / output_maxamt

	Name	Size	User	Group	Permissions	Date
	↥		cloudera	cloudera	drwxr-xr-x	December 07, 2020 07:10 AM
□	.		cloudera	cloudera	drwxr-xr-x	December 07, 2020 07:11 AM
□	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 07:11 AM
□	part-00000	19 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 07:11 AM

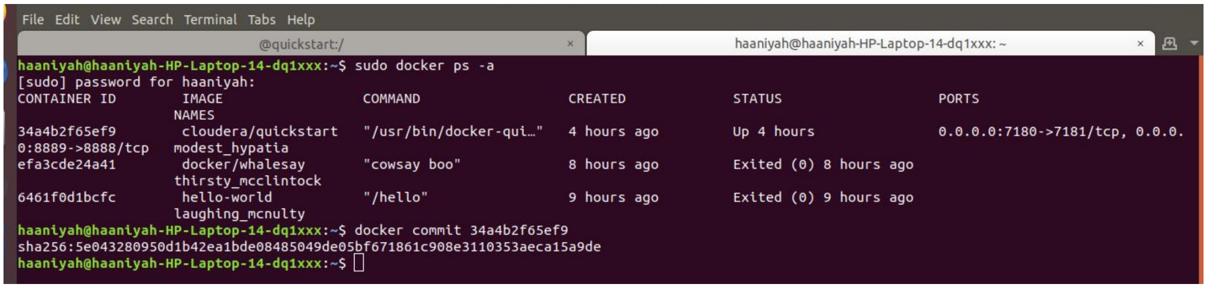
5- Output



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

```
220696  92445516.64
```

Docker Commit



File Edit View Search Terminal Tabs Help
@quickstart:/
haaniyah@haaniyah-HP-Laptop-14-dq1xxx: ~

```
haaniyah@haaniyah-HP-Laptop-14-dq1xxx:~$ sudo docker ps -a
[sudo] password for haaniyah:
CONTAINER ID        IMAGE                 COMMAND                  CREATED             STATUS                    PORTS
                   NAMES
34a4b2f65ef9        cloudera/quickstart   "/usr/bin/docker-qui…"   4 hours ago         Up 4 hours                0.0.0.0:7180->7181/tcp, 0.0.0.
0:8889->8888/tcp   modest_hypatia
efa3cde24a41        docker/whalesay       "cowsay boo"             8 hours ago         Exited (0) 8 hours ago
                   thirsty_mcclintock
6461f0d1bcfc        hello-world           "/hello"                 9 hours ago         Exited (0) 9 hours ago
                   laughing_mcnulty
haaniyah@haaniyah-HP-Laptop-14-dq1xxx:~$ docker commit 34a4b2f65ef9
sha256:5e043280950d1b42ea1bde08485049de05bf671861c908e3110353aeca15a9de
haaniyah@haaniyah-HP-Laptop-14-dq1xxx:~$ 
```

Shuffling:
The process of transferring data from the mapper to the reducer is known as shuffling. In simpler terms, the mapper processes and sorts the data as mapper output to feed it into the reducer as input. Example, to anonymize a dataset of employee information and then randomly shuffling the records within the dataset. Here the mapper code first removes the first and last name of the employee to attain data anonymity and then shuffles by assigning the key to a random number.

Sorting:
Values passed to the sorter are not sorted. Sorting map reduce jobs helps reducer to distinguish when a new job starts. Example, to sort the data of employees with respect to their salary. The mapper fetches employees according to their salary and the reducer in the sort phase, sorts the keys and outputs the results.