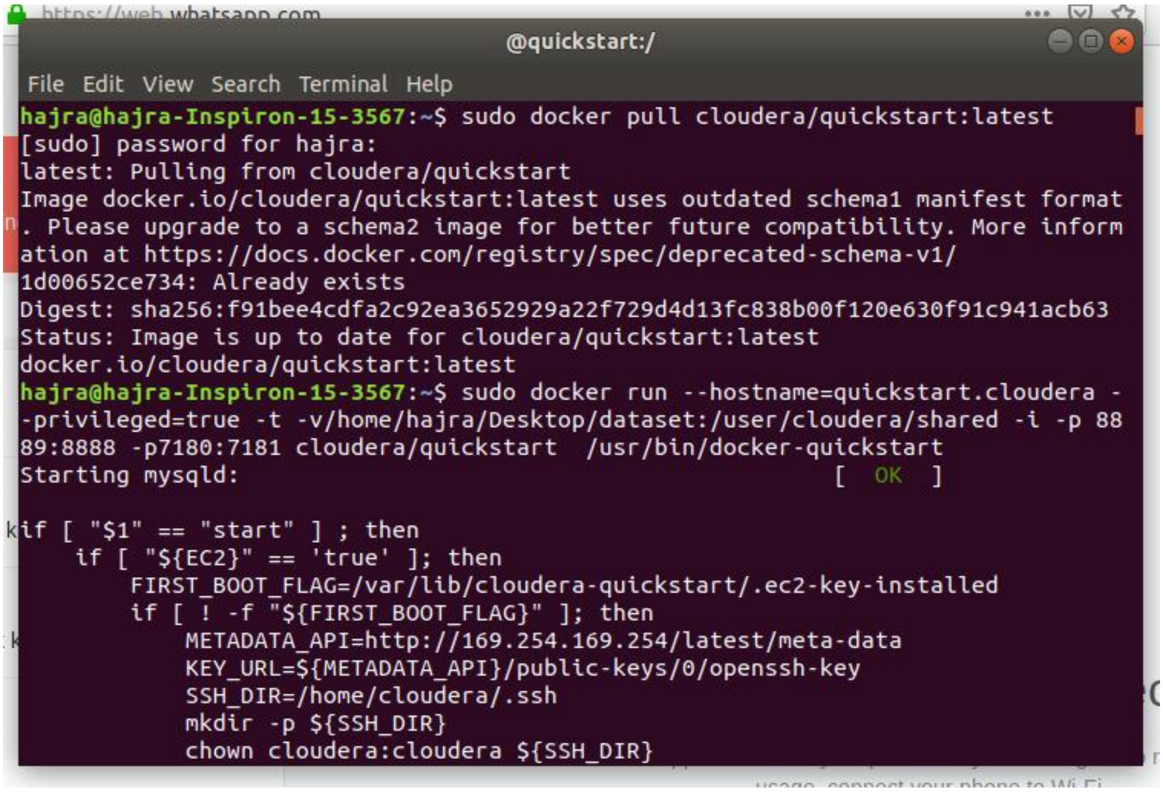
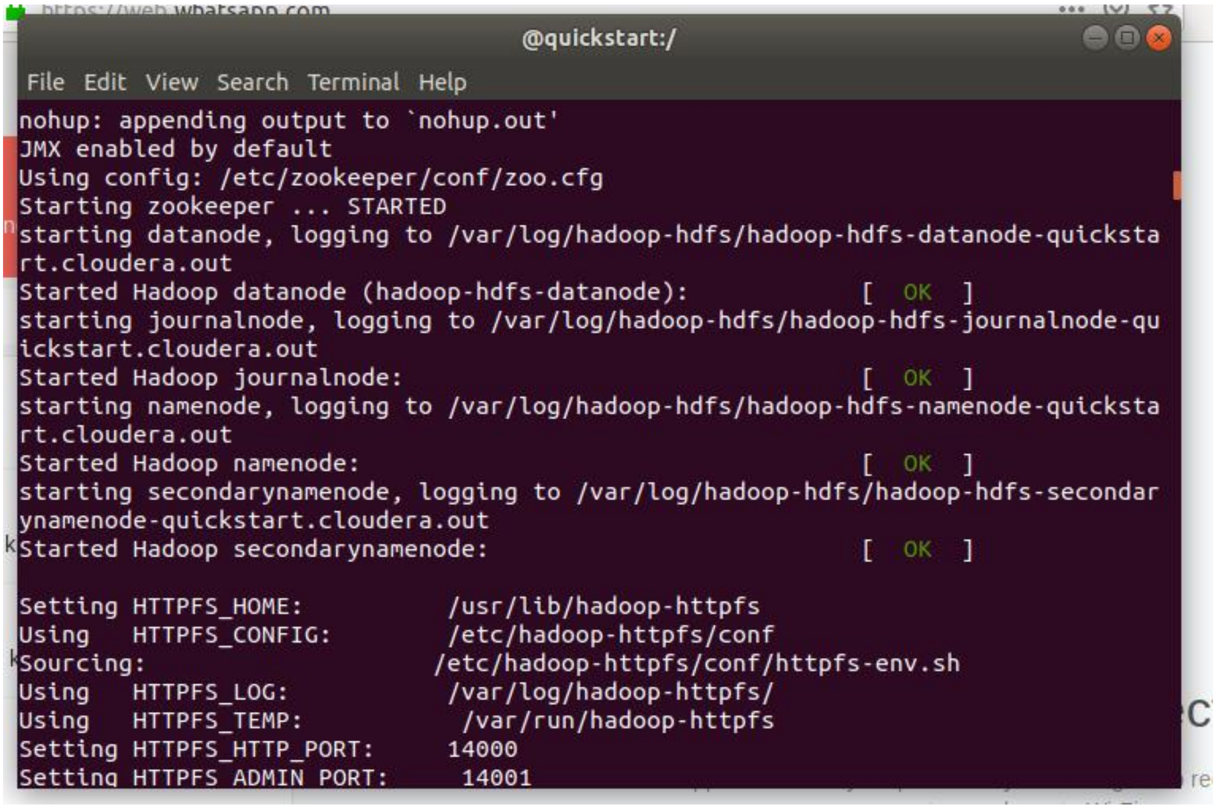
Assignment – Map Reduce Lab

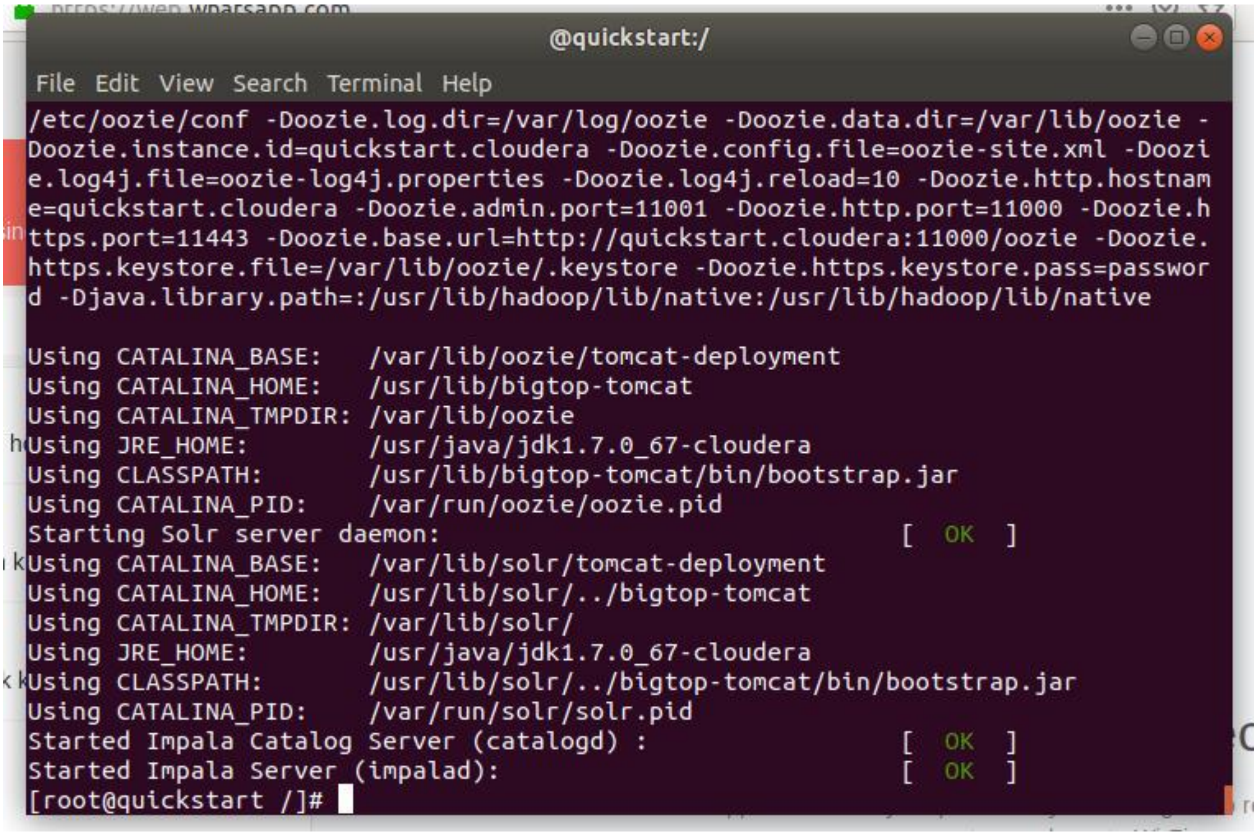
Hajra Abdul Hai 14893



```
hajra@hajra-Inspiron-15-3567:~$ sudo docker pull cloudera/quickstart:latest
[sudo] password for hajra:
latest: Pulling from cloudera/quickstart
Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format
. Please upgrade to a schema2 image for better future compatibility. More inform
ation at https://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Already exists
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Image is up to date for cloudera/quickstart:latest
docker.io/cloudera/quickstart:latest
hajra@hajra-Inspiron-15-3567:~$ sudo docker run --hostname=quickstart.cloudera -
-privileged=true -t -v/home/hajra/Desktop/dataset:/user/cloudera/shared -i -p 88
89:8888 -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}
```

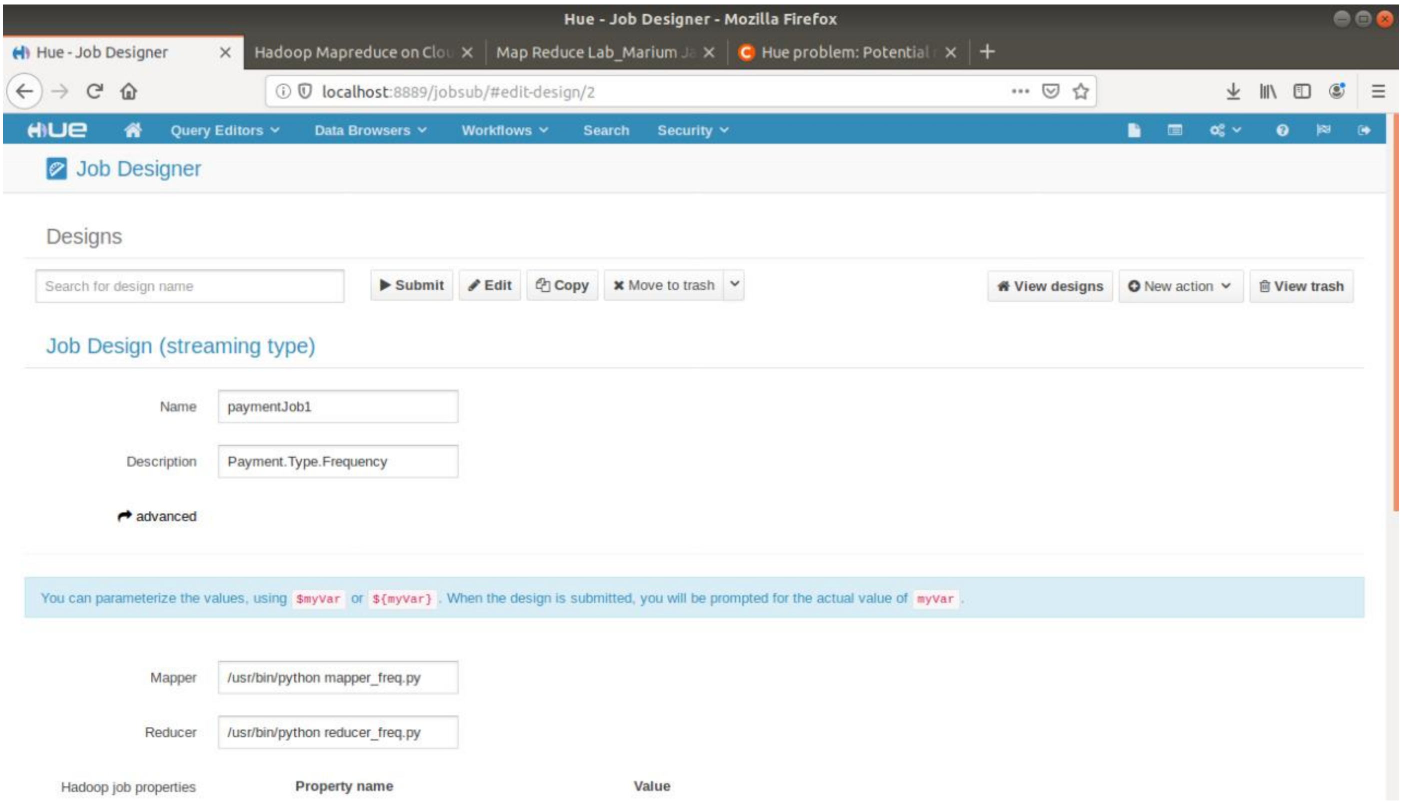


```
nohup: appending output to `nohup.out'
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... STARTED
starting datanode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-datanode-quicksta
rt.cloudera.out
Started Hadoop datanode (hadoop-hdfs-datanode):            [  OK  ]
starting journalnode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-journalnode-qu
ickstart.cloudera.out
Started Hadoop journalnode:                                [  OK  ]
starting namenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-namenode-quicksta
rt.cloudera.out
Started Hadoop namenode:                                   [  OK  ]
starting secondarynamenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-secondar
ynamenode-quickstart.cloudera.out
Started Hadoop secondarynamenode:                          [  OK  ]

Setting HTTPFS_HOME:          /usr/lib/hadoop-httpfs
Using   HTTPFS_CONFIG:        /etc/hadoop-httpfs/conf
Sourcing:                    /etc/hadoop-httpfs/conf/httpfs-env.sh
Using   HTTPFS_LOG:           /var/log/hadoop-httpfs/
Using   HTTPFS_TEMP:           /var/run/hadoop-httpfs
Setting HTTPFS_HTTP_PORT:     14000
Setting HTTPFS_ADMIN_PORT:     14001
```



```
/etc/oozie/conf -Doozie.log.dir=/var/log/oozie -Doozie.data.dir=/var/lib/oozie -
Doozie.instance.id=quickstart.cloudera -Doozie.config.file=oozie-site.xml -Doozi
e.log4j.file=oozie-log4j.properties -Doozie.log4j.reload=10 -Doozie.http.hostnam
e=quickstart.cloudera -Doozie.admin.port=11001 -Doozie.http.port=11000 -Doozie.h
ttps.port=11443 -Doozie.base.url=http://quickstart.cloudera:11000/oozie -Doozie.
https.keystore.file=/var/lib/oozie/.keystore -Doozie.https.keystore.pass=passwor
d -Djava.library.path=:/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native

Using CATALINA_BASE:   /var/lib/oozie/tomcat-deployment
Using CATALINA_HOME:   /usr/lib/bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/oozie
Using JRE_HOME:        /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH:       /usr/lib/bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID:    /var/run/oozie/oozie.pid
Starting Solr server daemon:                               [  OK  ]
Using CATALINA_BASE:   /var/lib/solr/tomcat-deployment
Using CATALINA_HOME:   /usr/lib/solr/../bigtop-tomcat
Using CATALINA_TMPDIR: /var/lib/solr/
Using JRE_HOME:        /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH:       /usr/lib/solr/../bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID:    /var/run/solr/solr.pid
Started Impala Catalog Server (catalogd) :                 [  OK  ]
Started Impala Server (impalad):                           [  OK  ]
[root@quickstart /]#
```

Task 1



Job Designer

Designs

Job Design (streaming type)

Name: paymentJob1

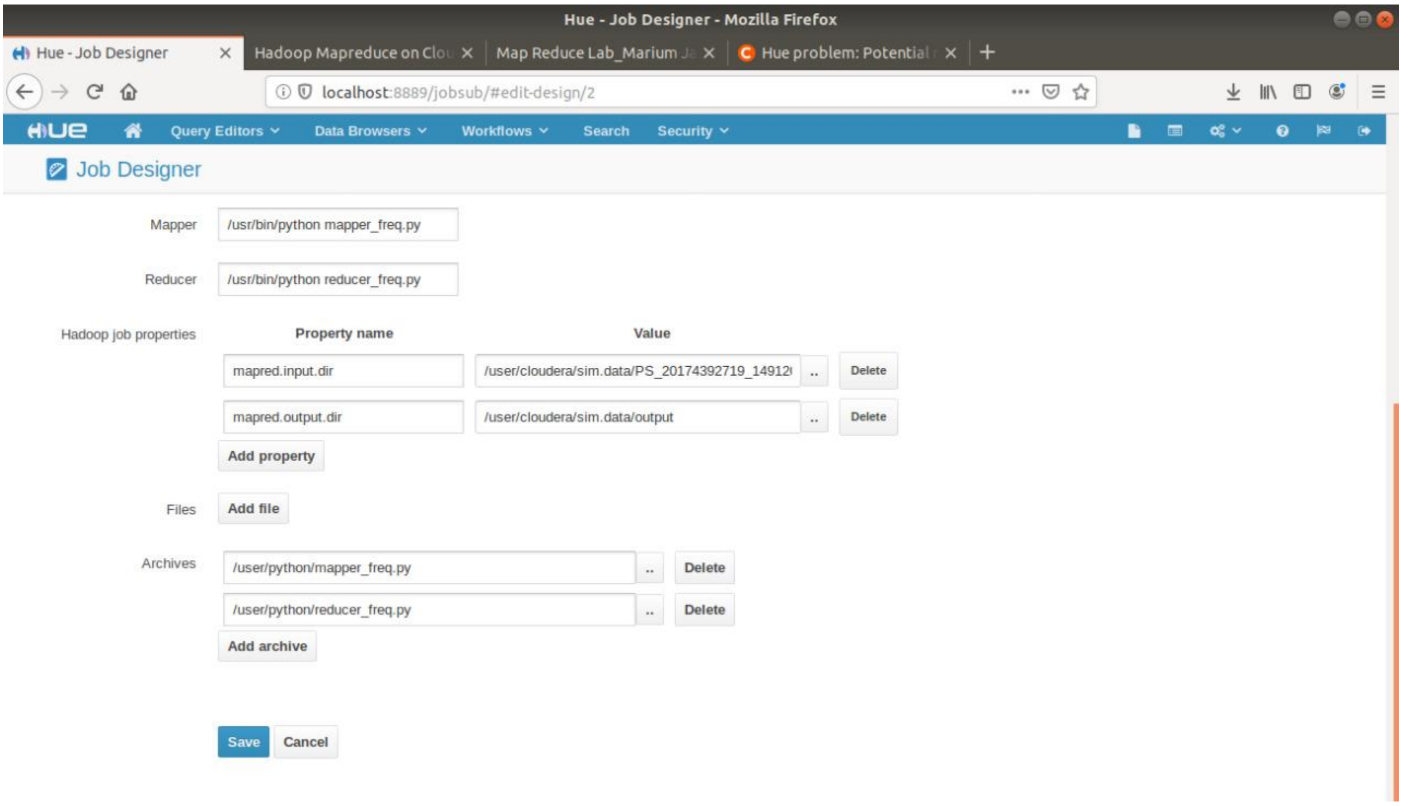
Description: Payment.Type.Frequency

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



Mapper: /usr/bin/python mapper_freq.py

Reducer: /usr/bin/python reducer_freq.py

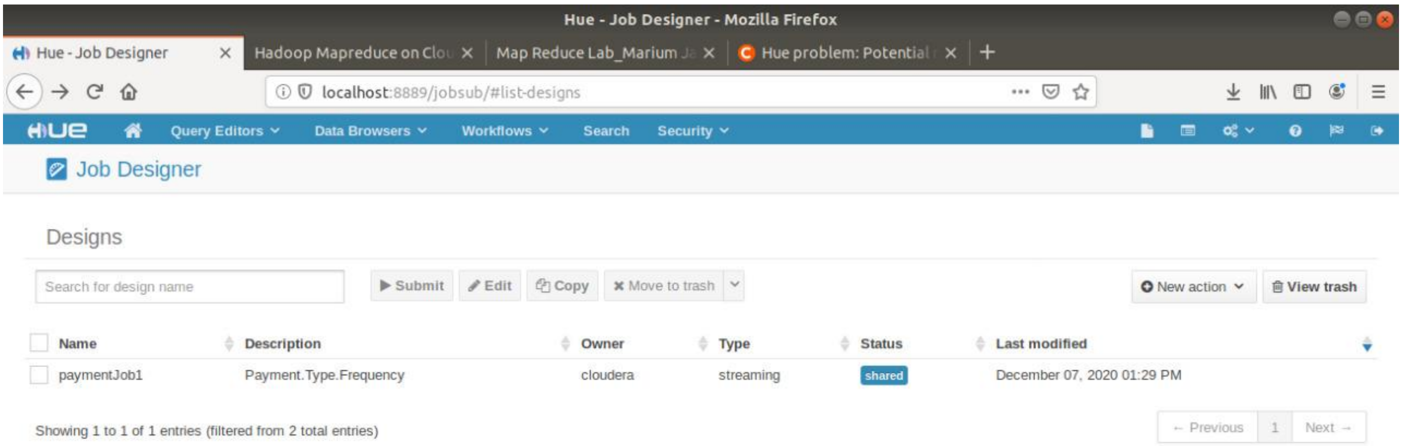
Hadoop job properties

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

Files

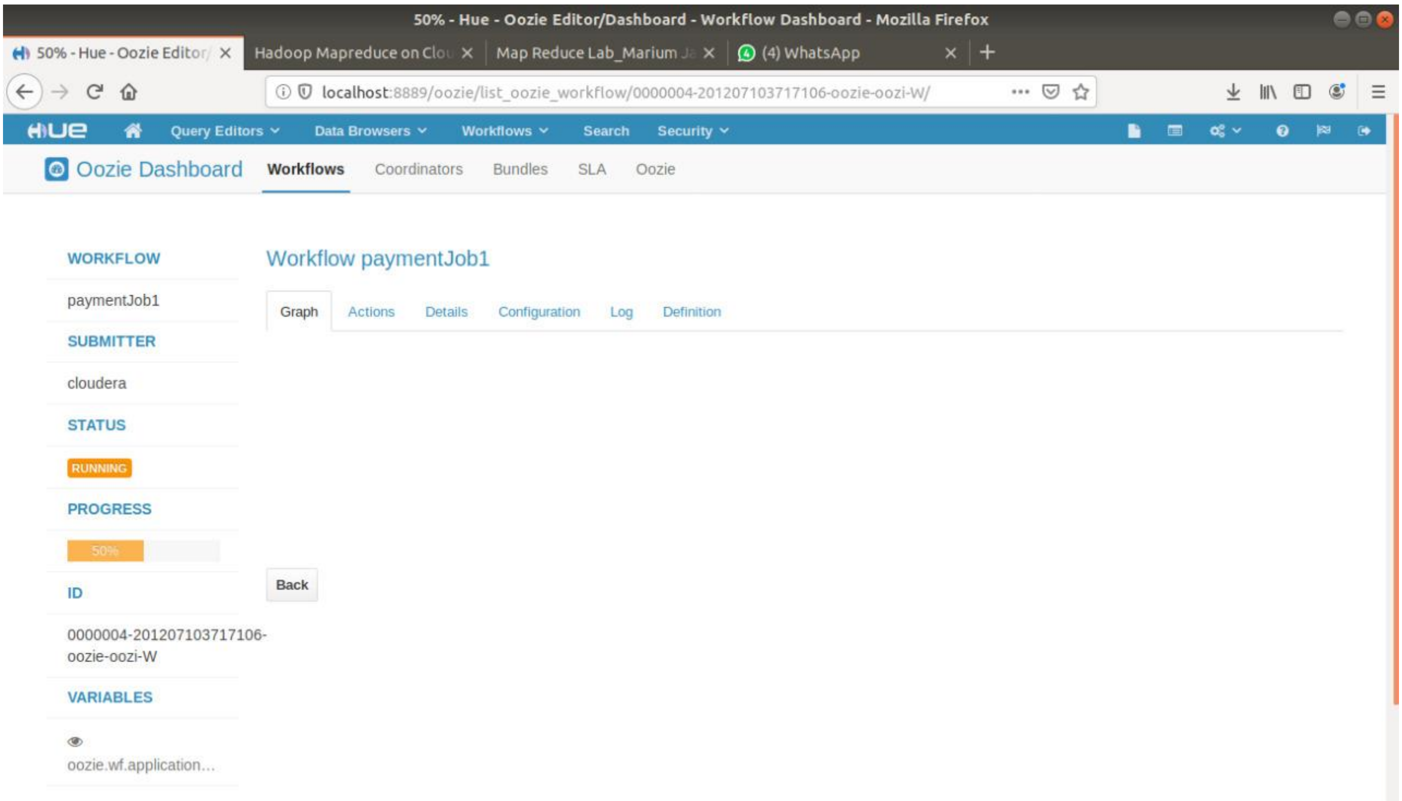
Archives

- /user/python/mapper_freq.py
- /user/python/reducer_freq.py



Name	Description	Owner	Type	Status	Last modified
paymentJob1	Payment.Type.Frequency	cloudera	streaming	shared	December 07, 2020 01:29 PM

Showing 1 to 1 of 1 entries (filtered from 2 total entries)



Workflow paymentJob1

WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

RUNNING

PROGRESS

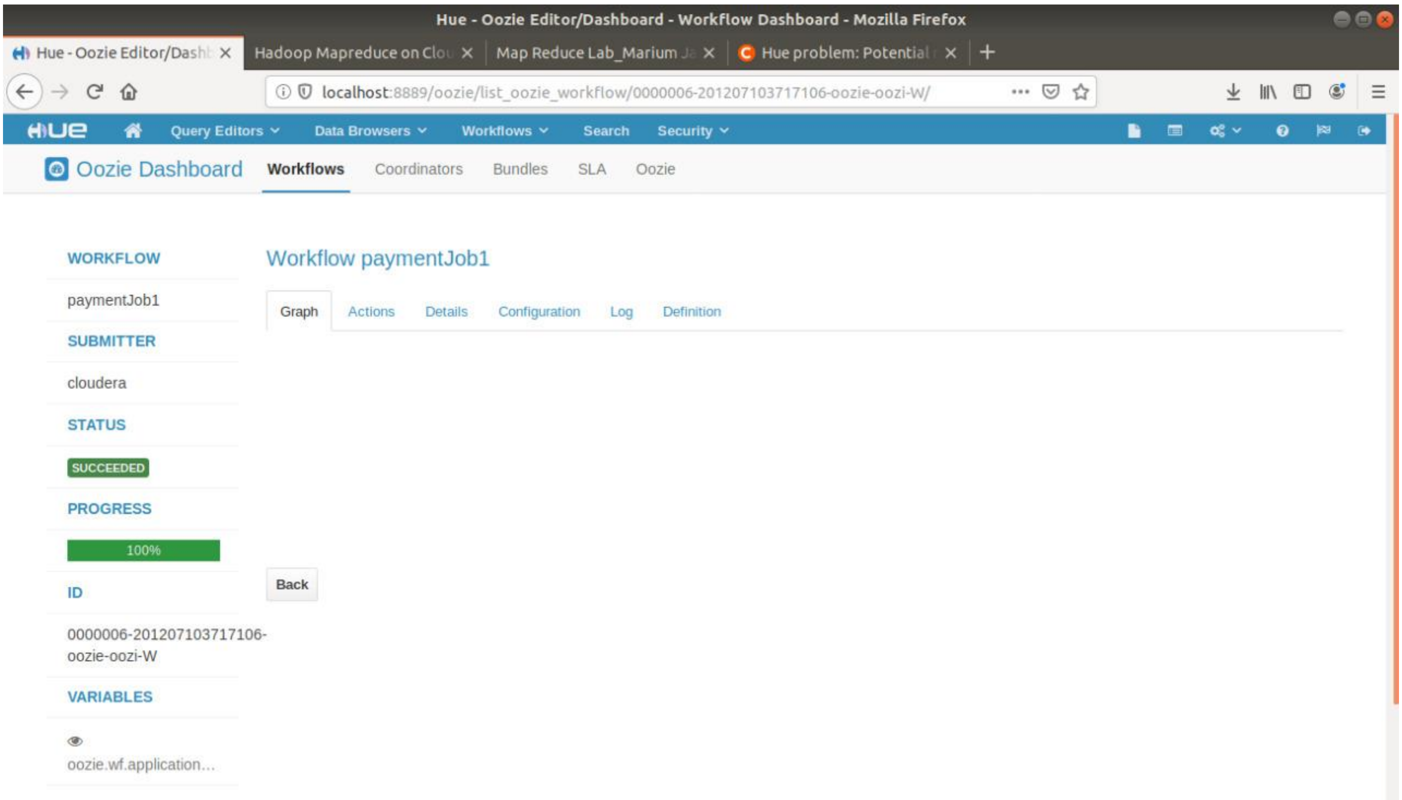
50%

ID

0000004-201207103717106-oozie-oozi-W

VARIABLES

oozie.wf.application...



Workflow paymentJob1

WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

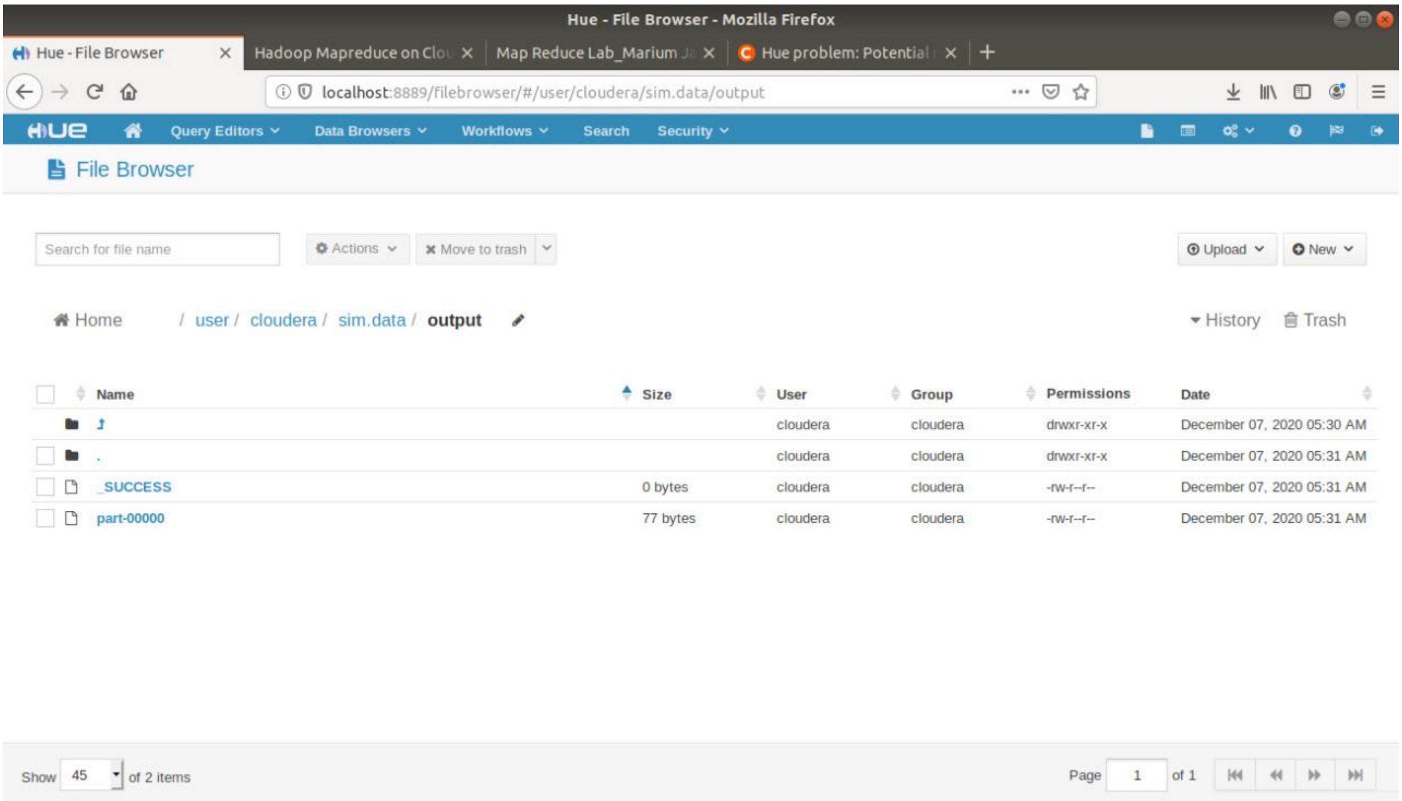
100%

ID

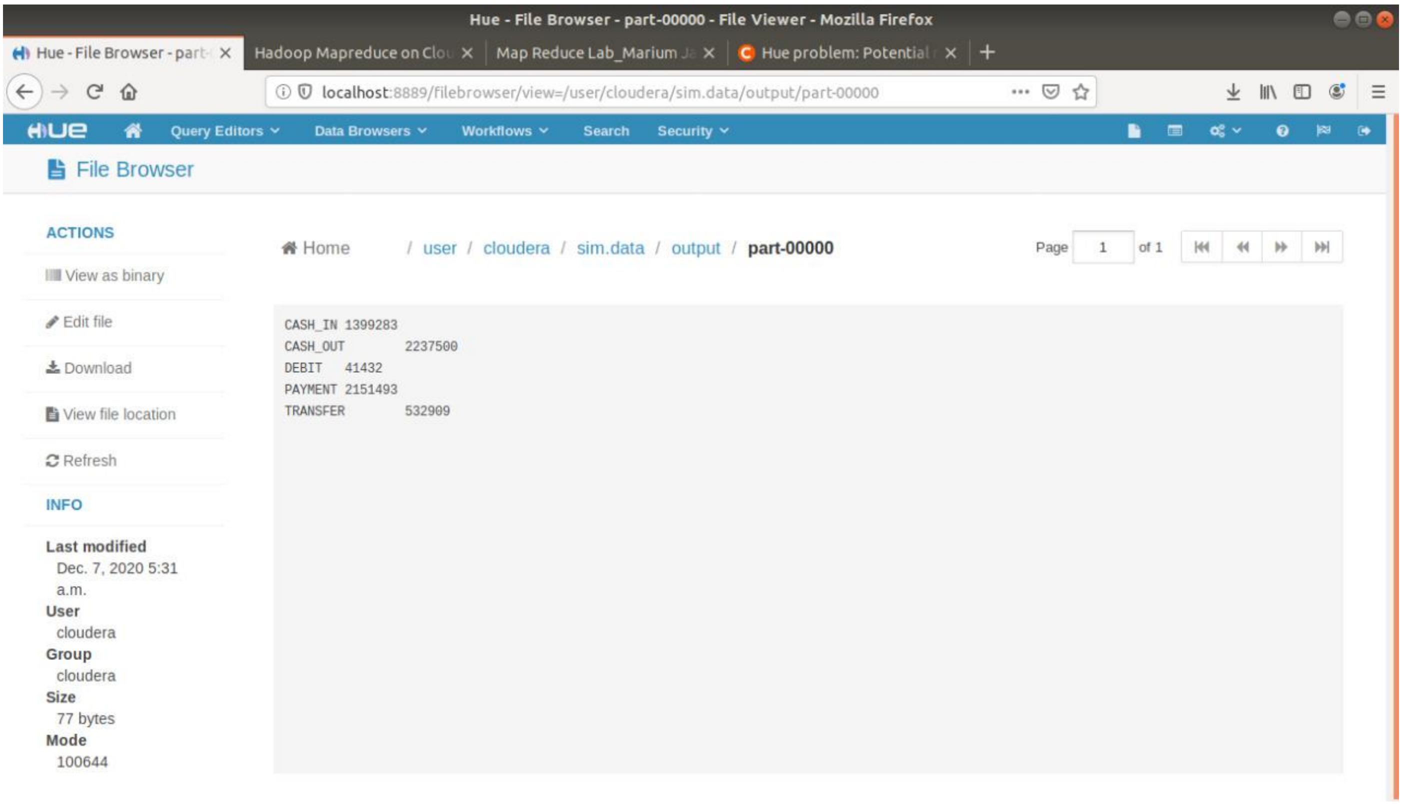
0000006-201207103717106-oozie-oozi-W

VARIABLES

oozie.wf.application...

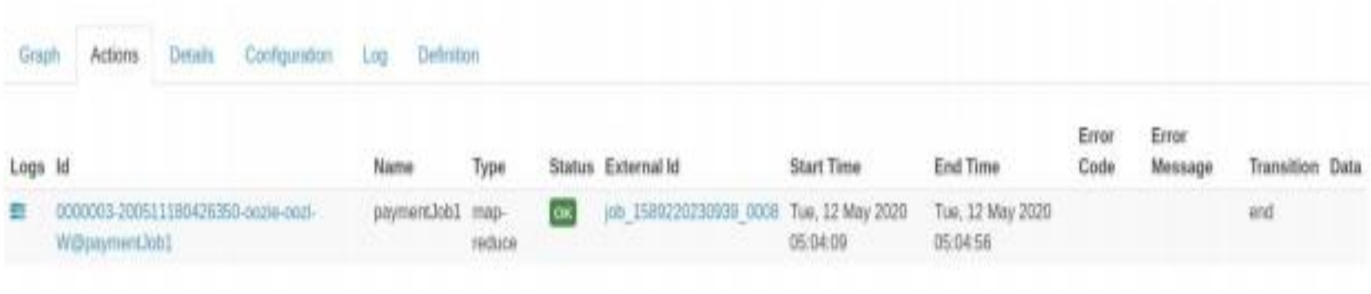


□	Name	Size	User	Group	Permissions	Date
	↥		cloudera	cloudera	drwxr-xr-x	December 07, 2020 05:30 AM
□	.		cloudera	cloudera	drwxr-xr-x	December 07, 2020 05:31 AM
□	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 05:31 AM
□	part-00000	77 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 05:31 AM

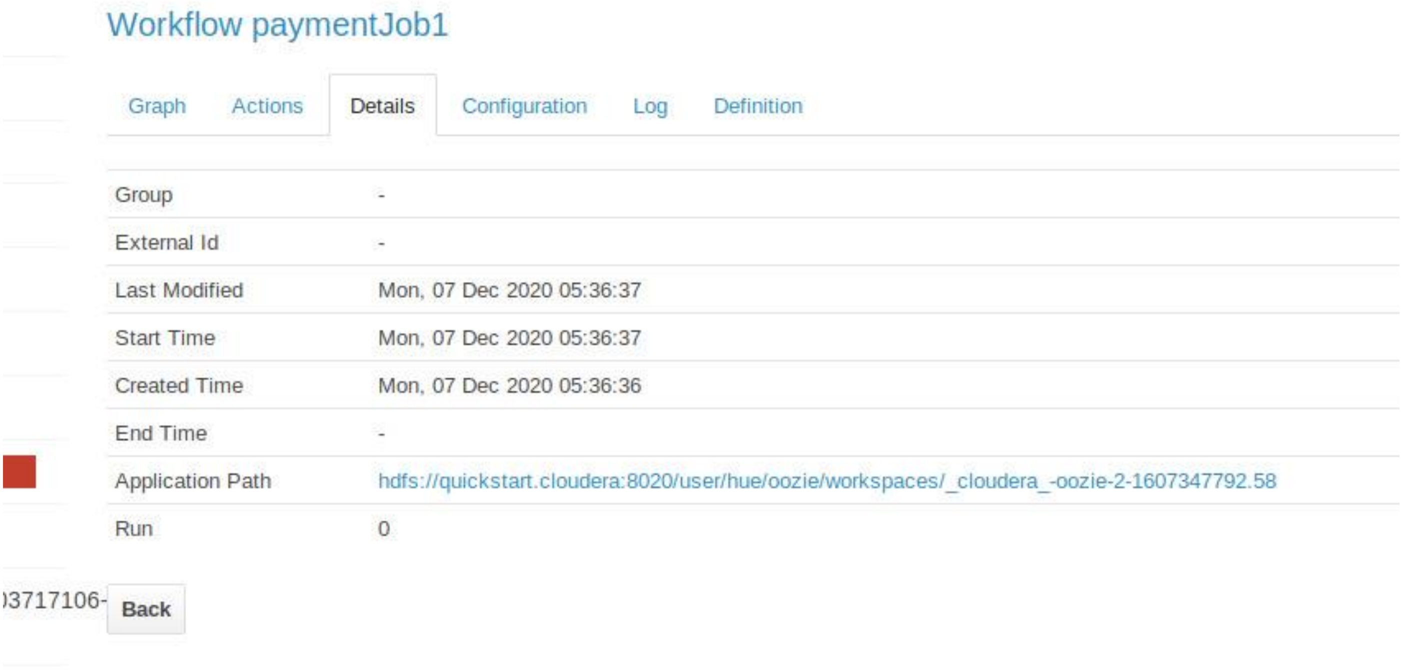


```
CASH_IN 1399283
CASH_OUT        2237500
DEBIT   41432
PAYMENT 2151493
TRANSFER        532909
```

Last modified
Dec. 7, 2020 5:31 a.m.
User
cloudera
Group
cloudera
Size
77 bytes
Mode
100644

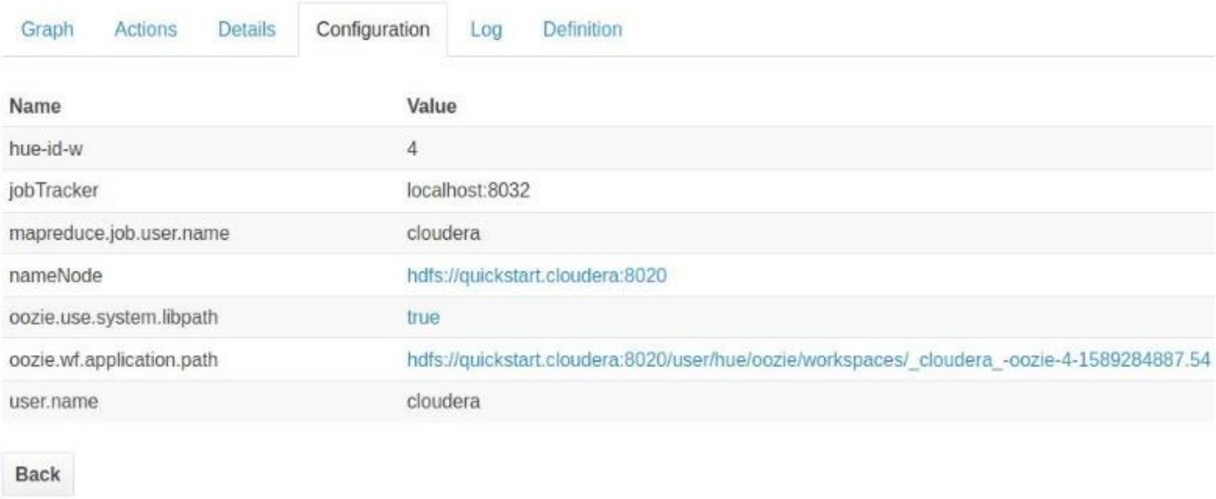


Logs	Id	Name	Type	Status	External Id	Start Time	End Time	Error Code	Error Message	Transition	Data
≣	0000003-200511180426350-oozie-oozi-W@paymentJob1	paymentJob1	map-reduce	OK	job_1589220230939_0008	Tue, 12 May 2020 05:04:09	Tue, 12 May 2020 05:04:56			end	

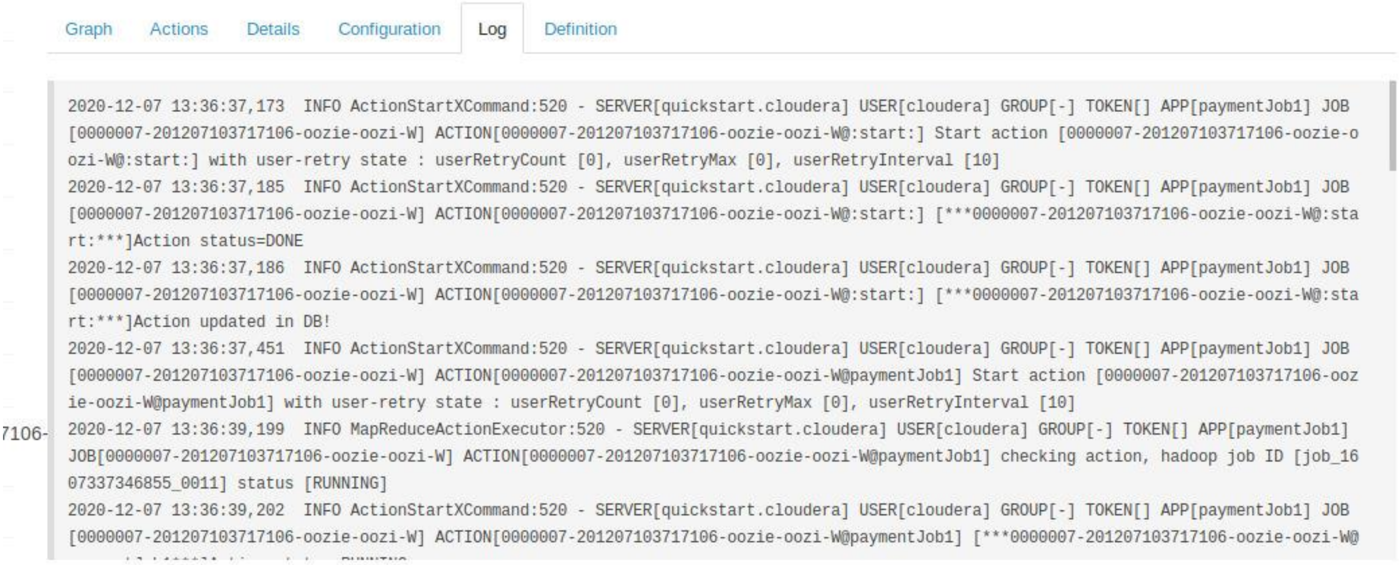


Workflow paymentJob1

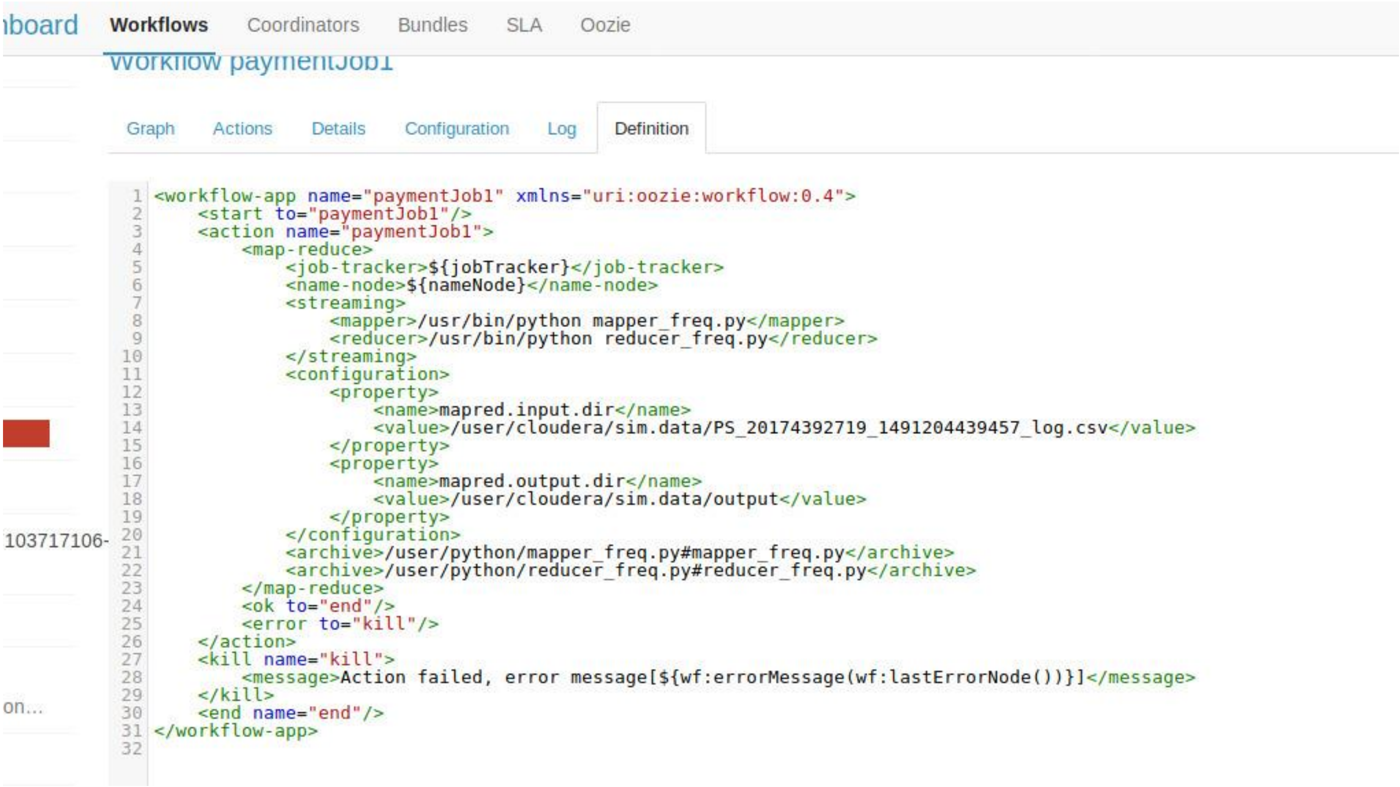
Group	-
External Id	-
Last Modified	Mon, 07 Dec 2020 05:36:37
Start Time	Mon, 07 Dec 2020 05:36:37
Created Time	Mon, 07 Dec 2020 05:36:36
End Time	-
Application Path	hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/_cloudera_-oozie-2-1607347792.58
Run	0



Name	Value
hue-id-w	4
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-4-1589284887.54
user.name	cloudera

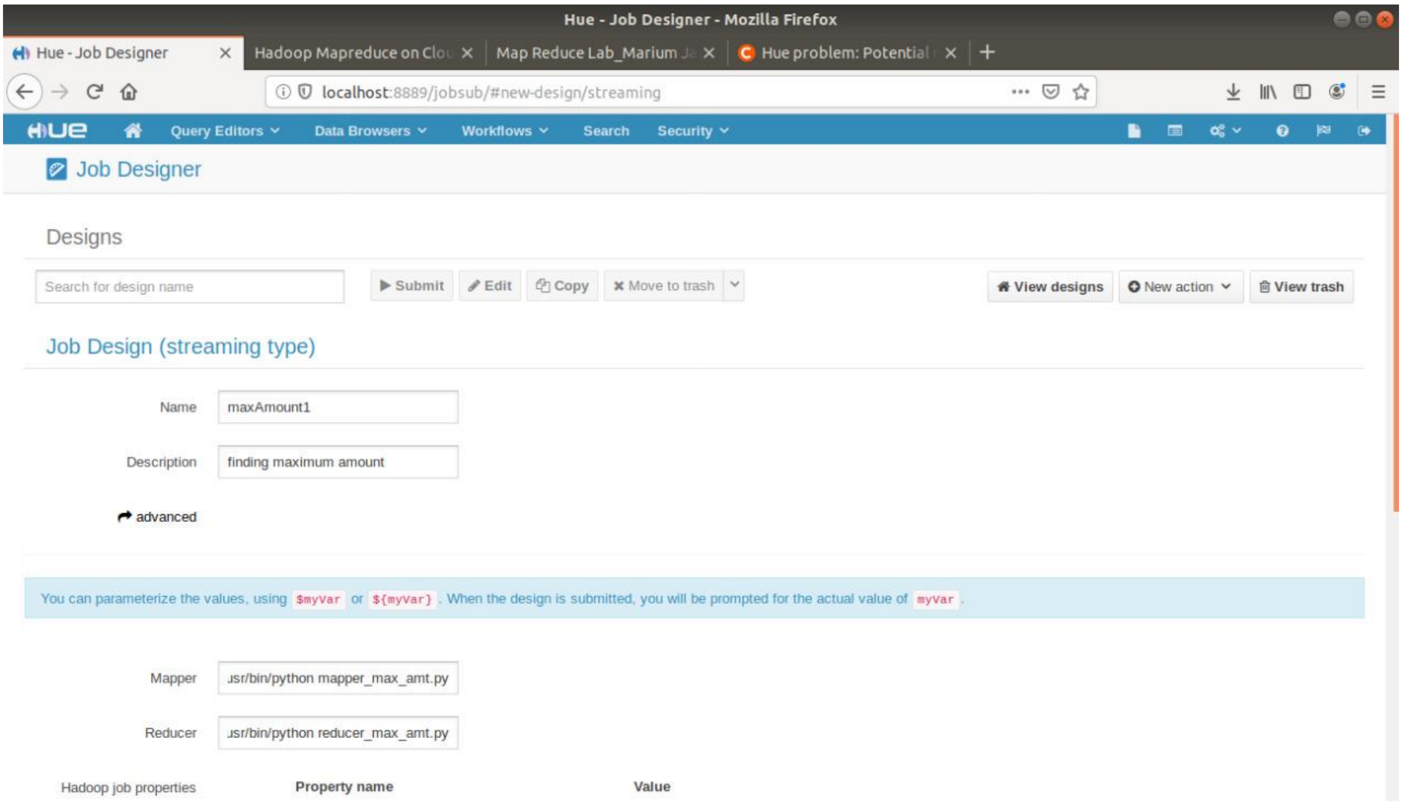


```
2020-12-07 13:36:37,173  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB
[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@:start:] Start action [0000007-201207103717106-oozie-o
ozi-W@:start:] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-07 13:36:37,185  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB
[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@:start:] [***0000007-201207103717106-oozie-oozi-W@:sta
rt:***]Action status=DONE
2020-12-07 13:36:37,186  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB
[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@:start:] [***0000007-201207103717106-oozie-oozi-W@:sta
rt:***]Action updated in DB!
2020-12-07 13:36:37,451  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB
[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@paymentJob1] Start action [0000007-201207103717106-ooz
ie-oozi-W@paymentJob1] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-07 13:36:39,199  INFO MapReduceActionExecutor:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1]
JOB[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@paymentJob1] checking action, hadoop job ID [job_16
07337346855_0011] status [RUNNING]
2020-12-07 13:36:39,202  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB
[0000007-201207103717106-oozie-oozi-W] ACTION[0000007-201207103717106-oozie-oozi-W@paymentJob1] [***0000007-201207103717106-oozie-oozi-W@
```



```xml
<workflow-app name="paymentJob1" xmlns="uri:oozie:workflow:0.4">
    <start to="paymentJob1"/>
    <action name="paymentJob1">
        <map-reduce>
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <streaming>
                <mapper>/usr/bin/python mapper_freq.py</mapper>
                <reducer>/usr/bin/python reducer_freq.py</reducer>
            </streaming>
            <configuration>
                <property>
                    <name>mapred.input.dir</name>
                    <value>/user/cloudera/sim.data/PS_20174392719_1491204439457_log.csv</value>
                </property>
                <property>
                    <name>mapred.output.dir</name>
                    <value>/user/cloudera/sim.data/output</value>
                </property>
            </configuration>
            <archive>/user/python/mapper_freq.py#mapper_freq.py</archive>
            <archive>/user/python/reducer_freq.py#reducer_freq.py</archive>
        </map-reduce>
        <ok to="end"/>
        <error to="kill"/>
    </action>
    <kill name="kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
    </kill>
    <end name="end"/>
</workflow-app>
```

Task 2



Job Design (streaming type)

Name: maxAmount1

Description: finding maximum amount

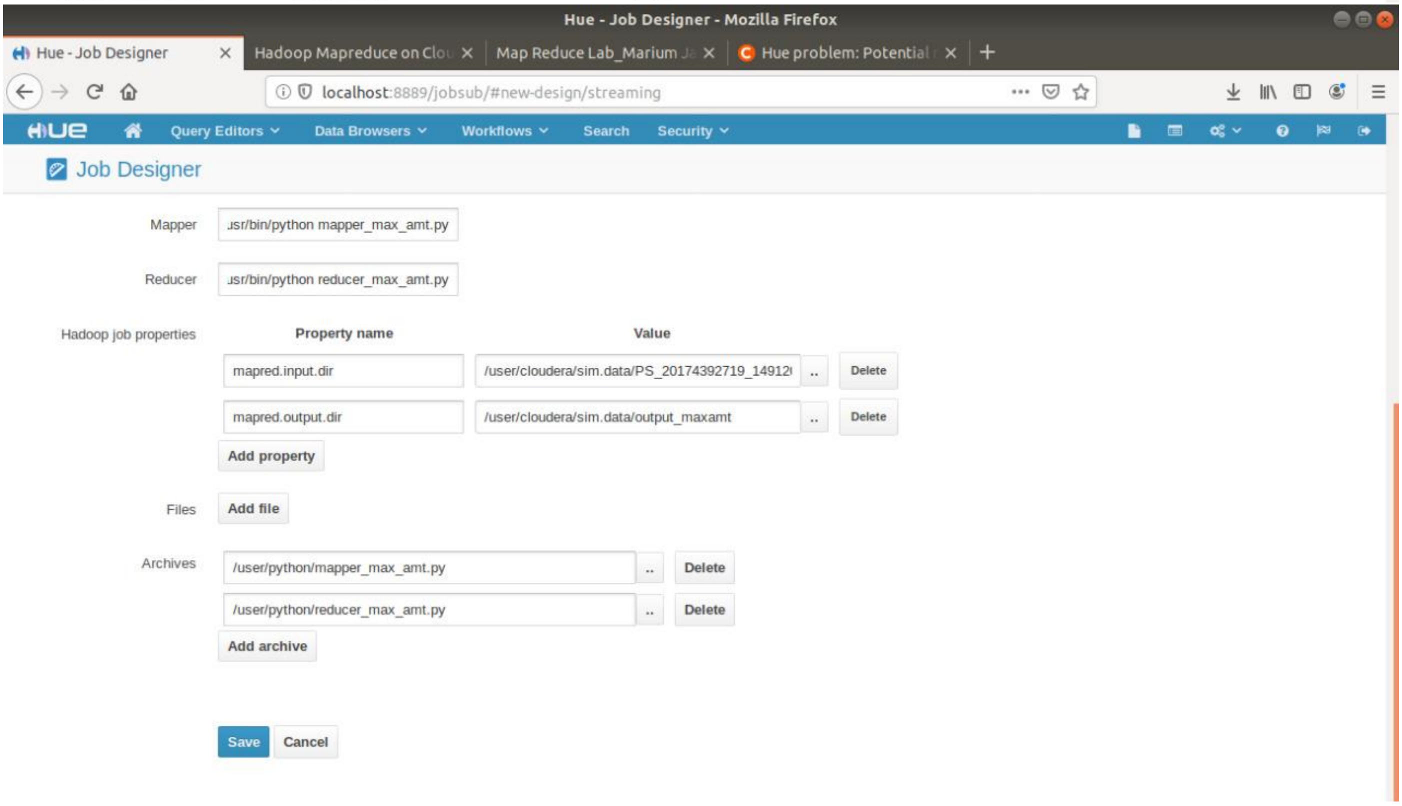
advanced

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.py

Reducer: usr/bin/python reducer_max_amt.py

Hadoop job properties — Property name — Value



Mapper: usr/bin/python mapper_max_amt.py

Reducer: usr/bin/python reducer_max_amt.py

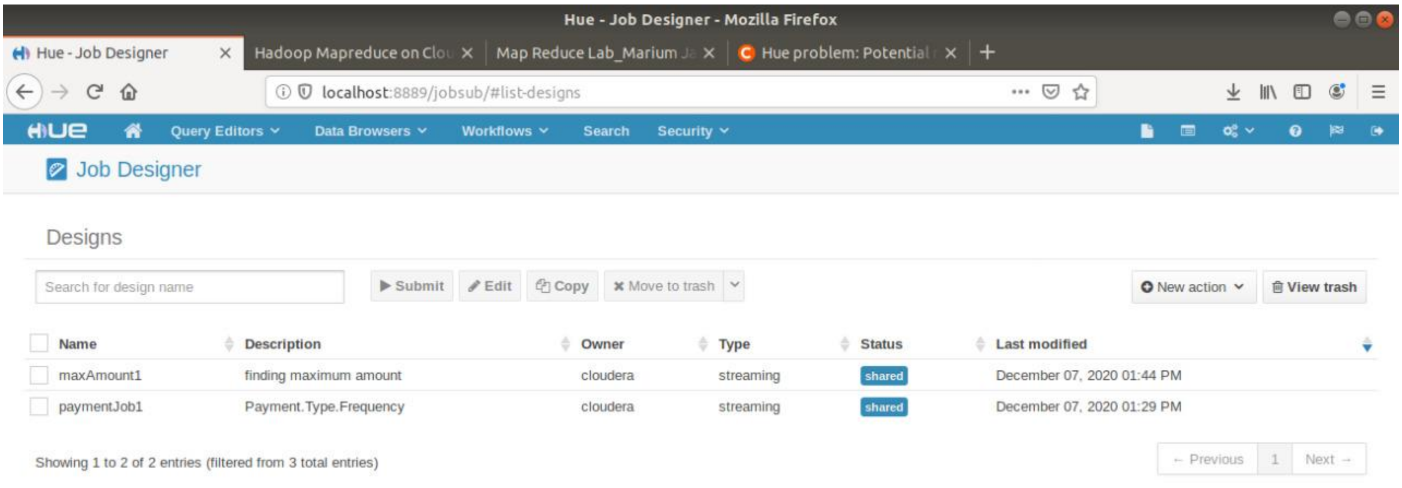
Hadoop job properties

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

Files

Archives:

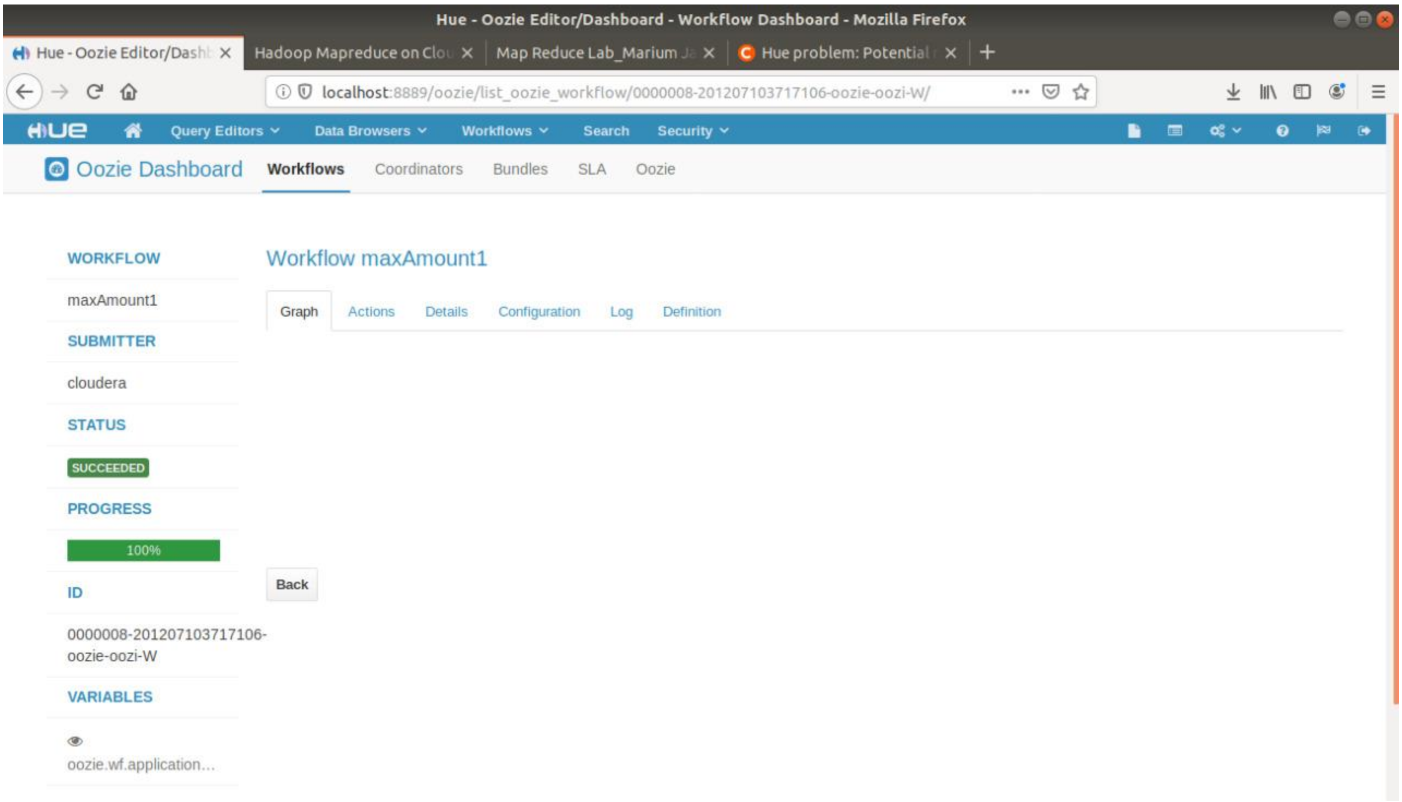
- /user/python/mapper_max_amt.py
- /user/python/reducer_max_amt.py



Designs

Name	Description	Owner	Type	Status	Last modified
maxAmount1	finding maximum amount	cloudera	streaming	shared	December 07, 2020 01:44 PM
paymentJob1	Payment.Type.Frequency	cloudera	streaming	shared	December 07, 2020 01:29 PM

Showing 1 to 2 of 2 entries (filtered from 3 total entries)



Workflow maxAmount1

WORKFLOW
maxAmount1

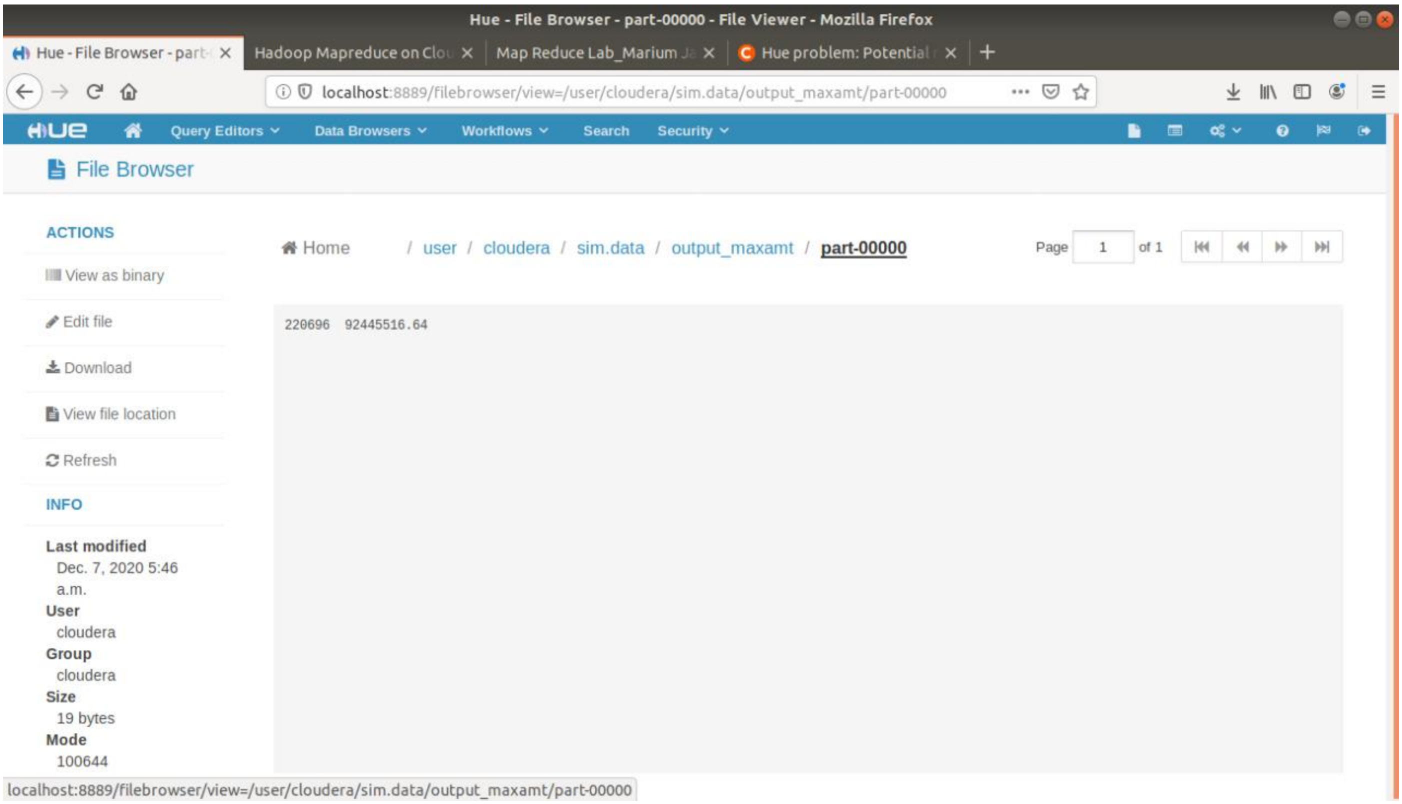
SUBMITTER
cloudera

STATUS
SUCCEEDED

PROGRESS
100%

ID
0000008-201207103717106-oozie-oozi-W

VARIABLES
oozie.wf.application...


Hue - File Browser - part-00000 - File Viewer - Mozilla Firefox
Hue - File Browser - part
Hadoop Mapreduce on Clou
Map Reduce Lab_Marium Ja
Hue problem: Potential
localhost:8889/filebrowser/view=/user/cloudera/sim.data/output_maxamt/part-00000
Query Editors
Data Browsers
Workflows
Search
Security
File Browser
ACTIONS
View as binary
Edit file
Download
View file location
Refresh
INFO
Home
user / cloudera / sim.data / output_maxamt / part-00000
Page 1 of 1

```
220696  92445516.64
```

Last modified
Dec. 7, 2020 5:46 a.m.
User
cloudera
Group
cloudera
Size
19 bytes
Mode
100644

localhost:8889/filebrowser/view=/user/cloudera/sim.data/output_maxamt/part-00000

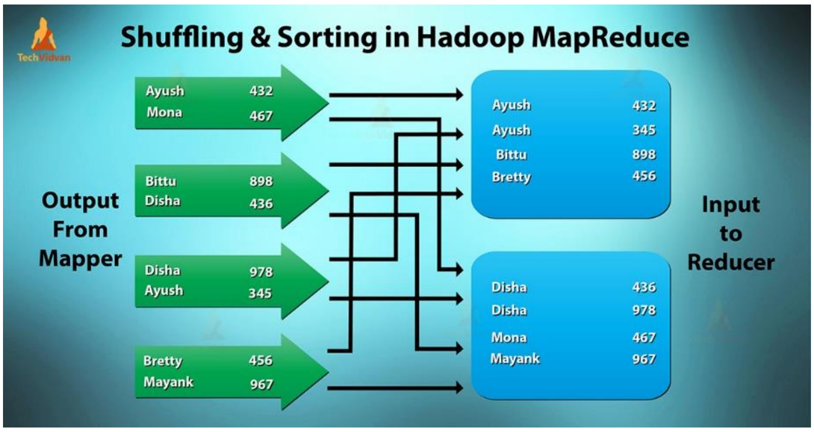
Task 3

Q1: Why is shuffling needed in MR? Give an example to justify.

The process of transferring data from the mappers to reducers is known as shuffling i.e. the process by which the system performs the sort, and transfers the map output to the reducer as input. So, MapReduce shuffle is necessary for the reducers, otherwise, they would not have any input (or input from every mapper).

Q2: Why is sorting needed in MR? Give an example to justify.

The keys generated by the mapper are automatically sorted by MapReduce Framework, i.e. before starting of reducer, all intermediate key-value pairs in MapReduce that are generated by mapper get sorted by key and not by value. Values passed to each reducer are not sorted; they can be in any order.
Sorting in Hadoop helps reducer to easily distinguish when a new reduce task should start. This saves time for the reducer. Reducer starts a new reduce task when the next key in the sorted input data is different than the previous.


Shuffling & Sorting in Hadoop MapReduce
TechVidvan
Output From Mapper
Ayush 432
Mona 467
Bittu 898
Disha 436
Disha 978
Ayush 345
Bretty 456
Mayank 967
Ayush 432
Ayush 345
Bittu 898
Bretty 456
Disha 436
Disha 978
Mona 467
Mayank 967
Input to Reducer