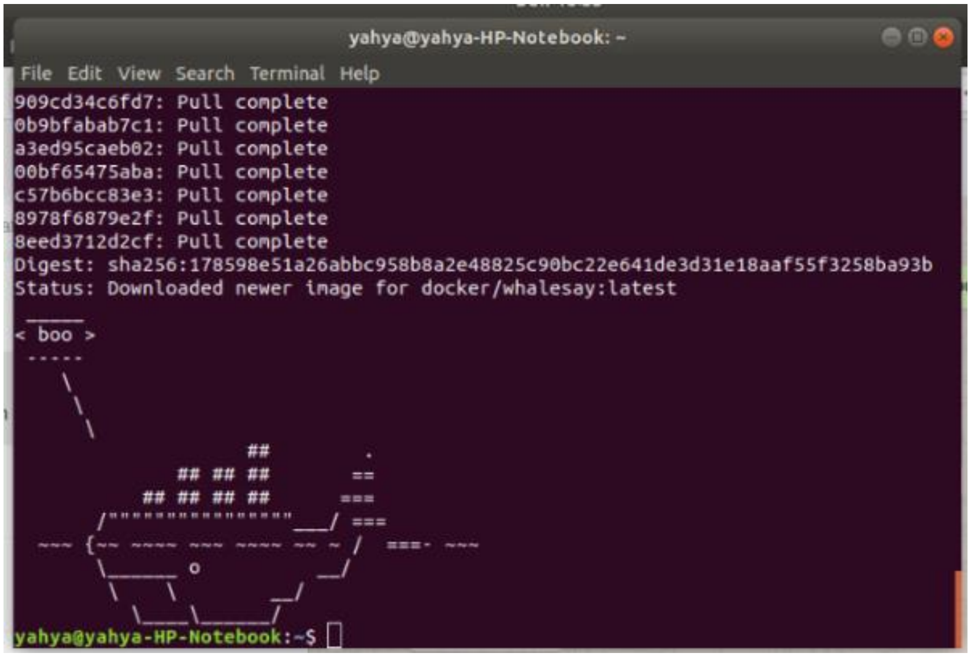
Map Reducer Lab

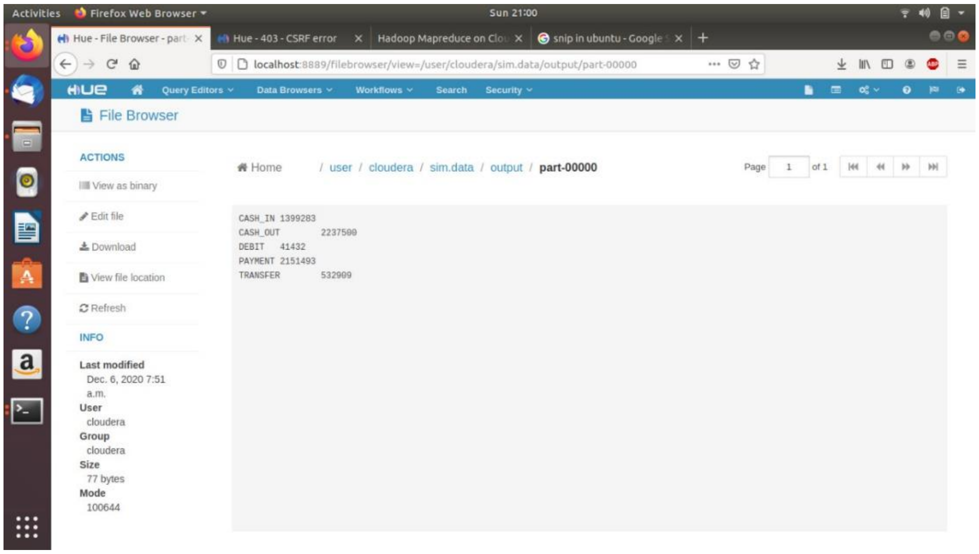
Docker Installation



yahya@yahya-HP-Notebook: ~
File Edit View Search Terminal Help

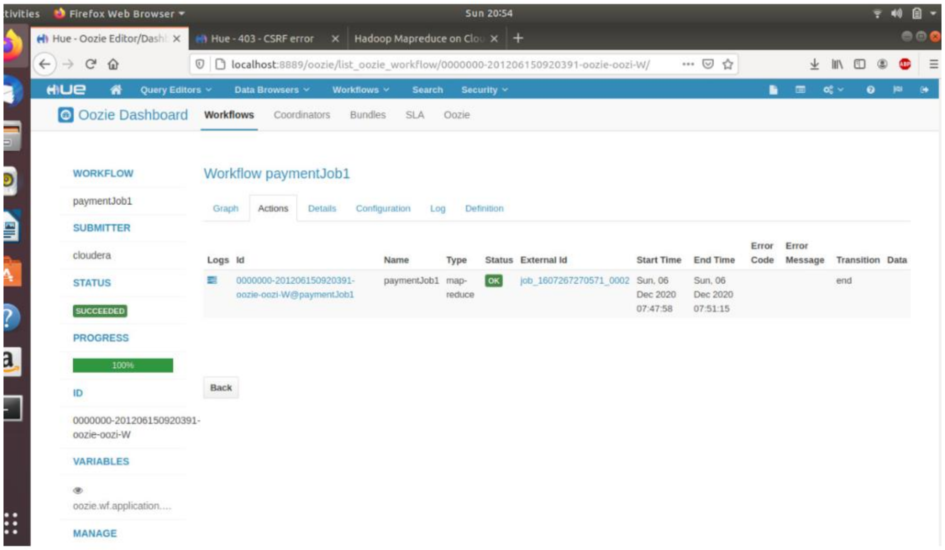
```
909cd34c6fd7: Pull complete
0b9bfabab7c1: Pull complete
a3ed95caeb02: Pull complete
00bf65475aba: Pull complete
c57b6bcc83e3: Pull complete
8978f6879e2f: Pull complete
8eed3712d2cf: Pull complete
Digest: sha256:178598e51a26abbc958b8a2e48825c90bc22e641de3d31e18aaf55f3258ba93b
Status: Downloaded newer image for docker/whalesay:latest
 _____
< boo >
 -----
    \
     \
      \
                    ##        .
              ## ## ##       ==
           ## ## ## ##      ===
       /""""""""""""""""___/ ===
  ~~~ {~~ ~~~~ ~~~ ~~~~ ~~ ~ /  ===- ~~~
       \______ o          __/
        \    \        __/
          \____\______/
yahya@yahya-HP-Notebook:~$
```

1st Map Reducer Output



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

Last modified
Dec. 6, 2020 7:51 a.m.
User
cloudera
Group
cloudera
Size
77 bytes
Mode
100644



WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000000-201206150920391-oozie-oozi-W

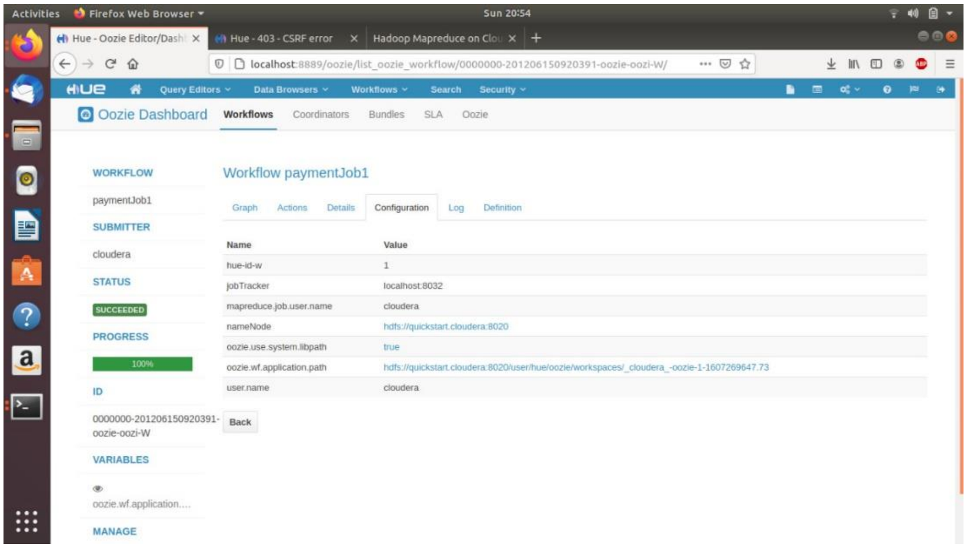
VARIABLES

oozie.wf.application....

MANAGE

Workflow paymentJob1

Logs	Id	Name	Type	Status	External Id	Start Time	End Time	Error Code	Error Message	Transition	Data
☰	0000000-201206150920391-oozie-oozi-W@paymentJob1	paymentJob1	map-reduce	OK	job_1607267270571_0002	Sun, 06 Dec 2020 07:47:58	Sun, 06 Dec 2020 07:51:15			end	



WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000000-201206150920391-oozie-oozi-W

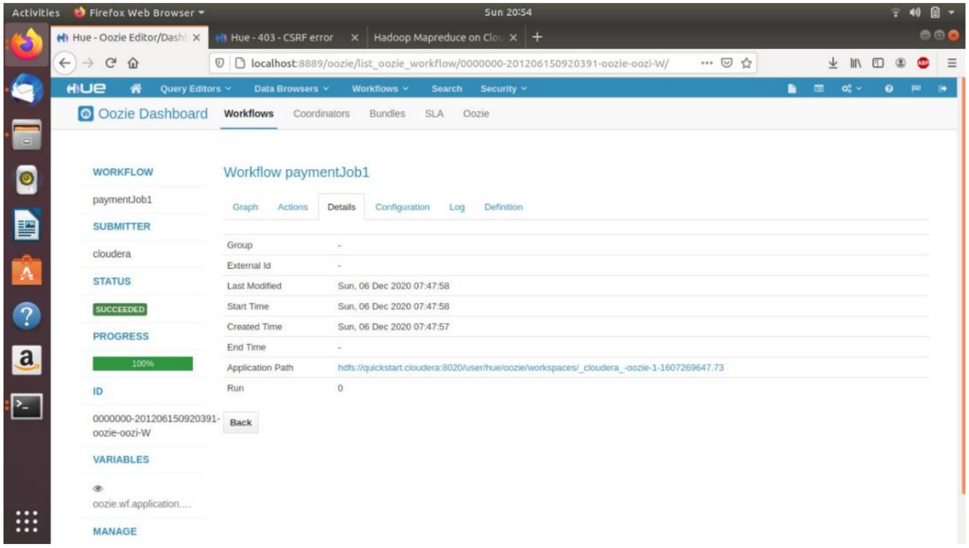
VARIABLES

oozie.wf.application....

MANAGE

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-1-1607269647.73
user.name	cloudera



WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000000-201206150920391-oozie-oozi-W

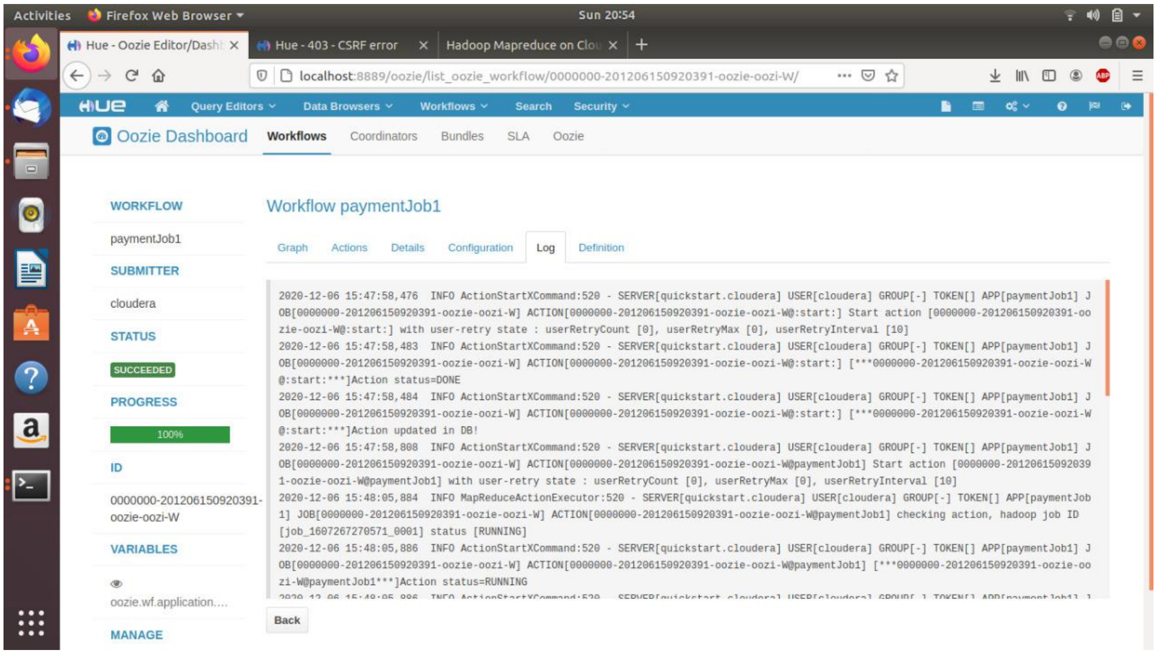
VARIABLES

oozie.wf.application....

MANAGE

Workflow paymentJob1

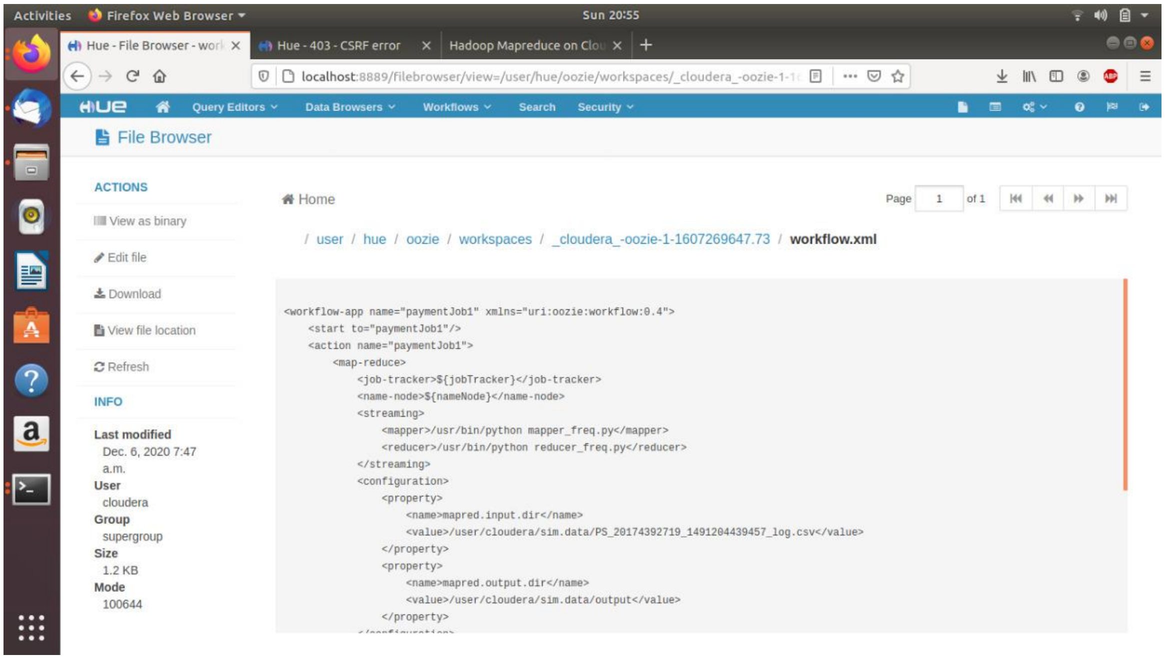
Group	-
External Id	-
Last Modified	Sun, 06 Dec 2020 07:47:58
Start Time	Sun, 06 Dec 2020 07:47:58
Created Time	Sun, 06 Dec 2020 07:47:57
End Time	-
Application Path	hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/_cloudera_-oozie-1-1607269647.73
Run	0



WORKFLOW
paymentJob1
SUBMITTER
cloudera
STATUS
SUCCEEDED
PROGRESS
100%
ID
0000000-201206150920391-oozie-oozi-W
VARIABLES
oozie.wf.application....
MANAGE

Workflow paymentJob1

```
2020-12-06 15:47:58,476  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] J
OB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@:start:] Start action [0000000-201206150920391-oo
zie-oozi-W@:start:] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-06 15:47:58,483  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] J
OB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@:start:] [***0000000-201206150920391-oozie-oozi-W
@:start:***]Action status=DONE
2020-12-06 15:47:58,484  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] J
OB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@:start:] [***0000000-201206150920391-oozie-oozi-W
@:start:***]Action updated in DB!
2020-12-06 15:47:58,808  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] J
OB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] Start action [0000000-2012061509203
91-oozie-oozi-W@paymentJob1] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-06 15:48:05,884  INFO MapReduceActionExecutor:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob
1] JOB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] checking action, hadoop job ID
[job_1607267270571_0001] status [RUNNING]
2020-12-06 15:48:05,886  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] J
OB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] [***0000000-201206150920391-oozie-oo
zi-W@paymentJob1***]Action status=RUNNING
```

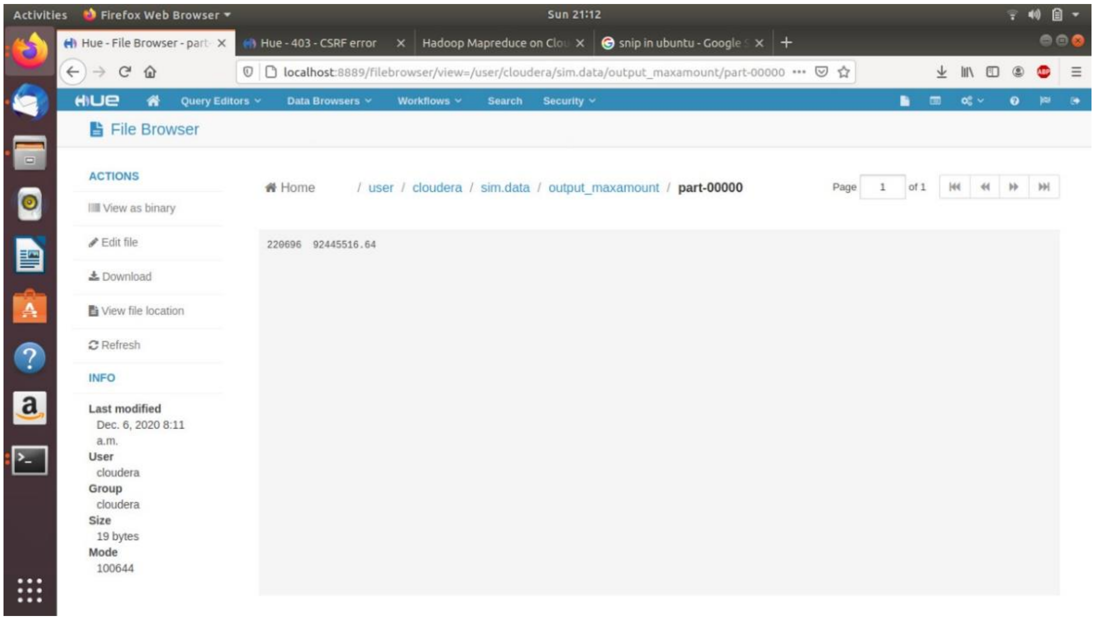


INFO
Last modified
Dec. 6, 2020 7:47 a.m.
User
cloudera
Group
supergroup
Size
1.2 KB
Mode
100644

/ user / hue / oozie / workspaces / _cloudera_-oozie-1-1607269647.73 / **workflow.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>
```

2^{nd} Map Reducer Output

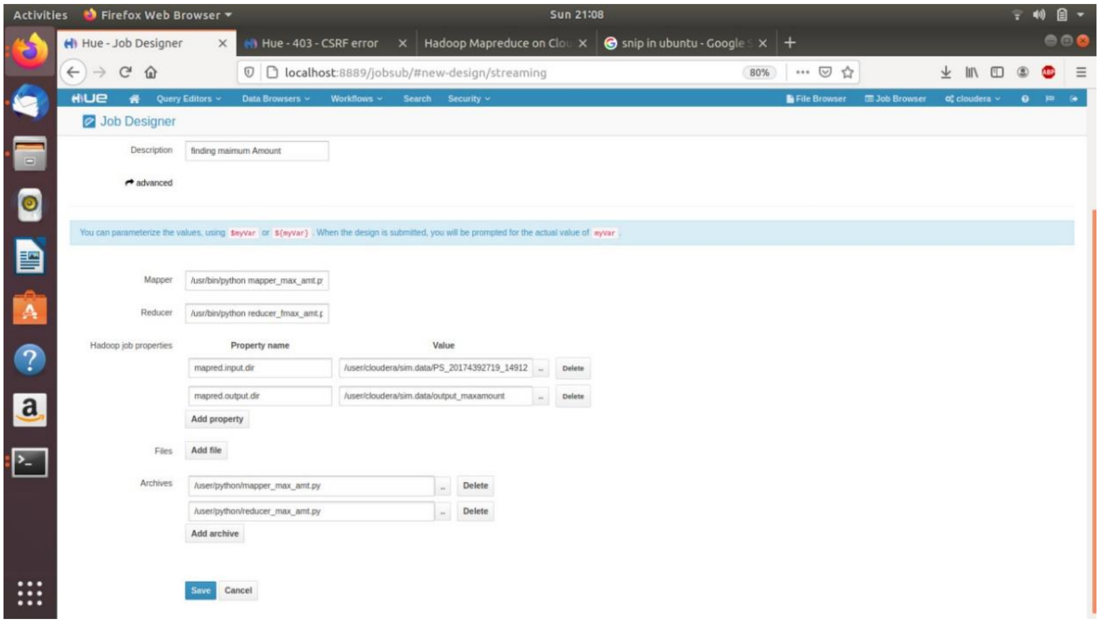


localhost:8889/filebrowser/view=/user/cloudera/sim.data/output_maxamount/part-00000

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

```
220696  92445516.64
```

Last modified
Dec. 6, 2020 8:11 a.m.
User
cloudera
Group
cloudera
Size
19 bytes
Mode
100644



localhost:8889/jobsub/#new-design/streaming

Description: finding maimum Amount

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

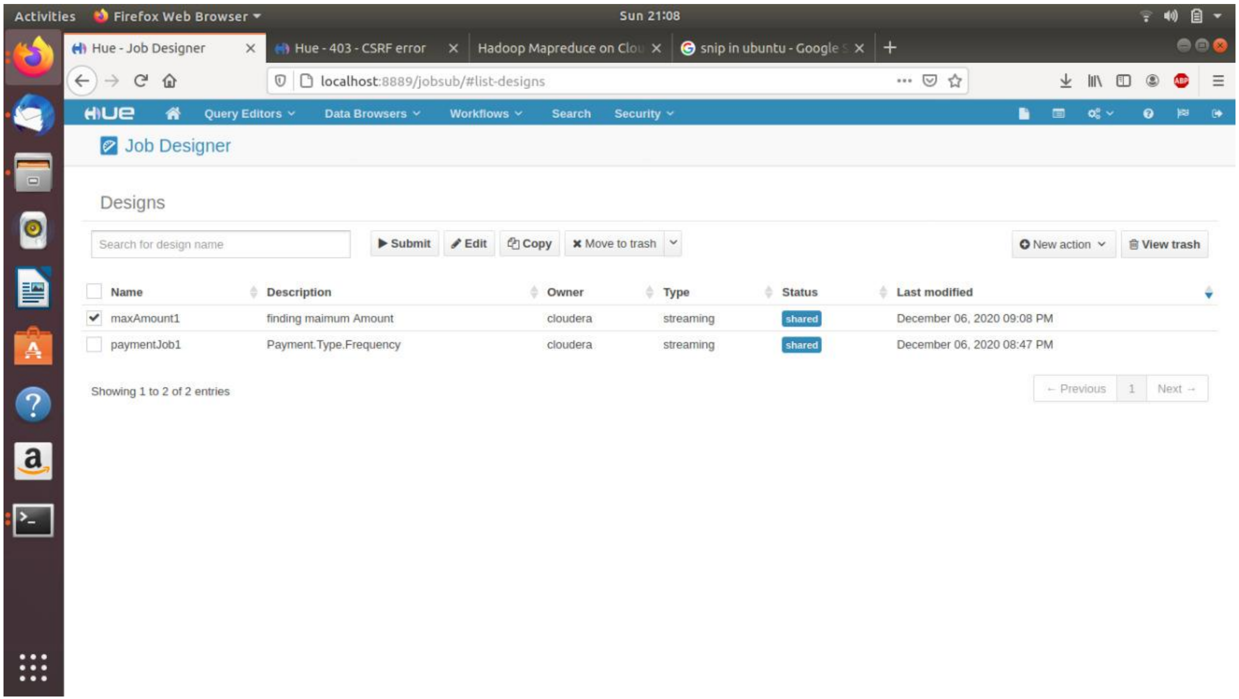
Reducer: /usr/bin/python reducer_fmax_amt.p

Hadoop job properties

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

Files

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



localhost:8889/jobsub/#list-designs

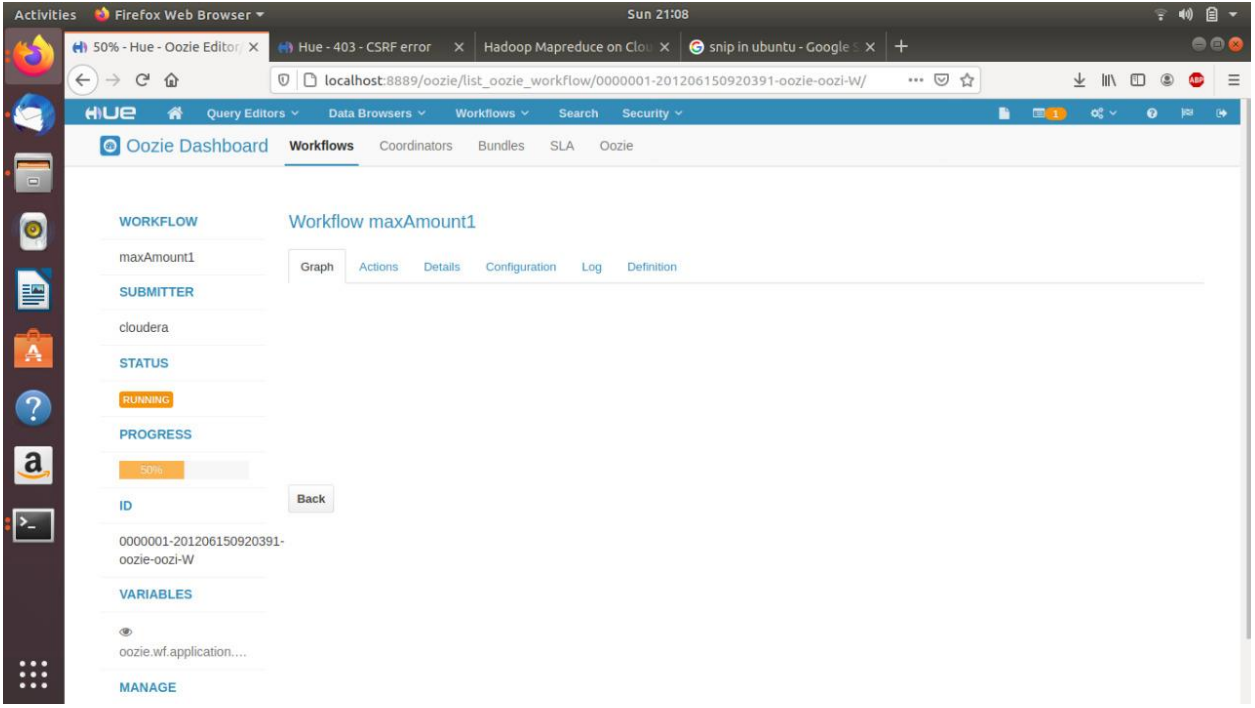
Job Designer

Designs

Search for design name

	Name	Description	Owner	Type	Status	Last modified
☑	maxAmount1	finding maimum Amount	cloudera	streaming	shared	December 06, 2020 09:08 PM
□	paymentJob1	Payment.Type.Frequency	cloudera	streaming	shared	December 06, 2020 08:47 PM

Showing 1 to 2 of 2 entries



localhost:8889/oozie/list_oozie_workflow/0000001-201206150920391-oozie-oozi-W/

Workflow maxAmount1

WORKFLOW
maxAmount1

SUBMITTER
cloudera

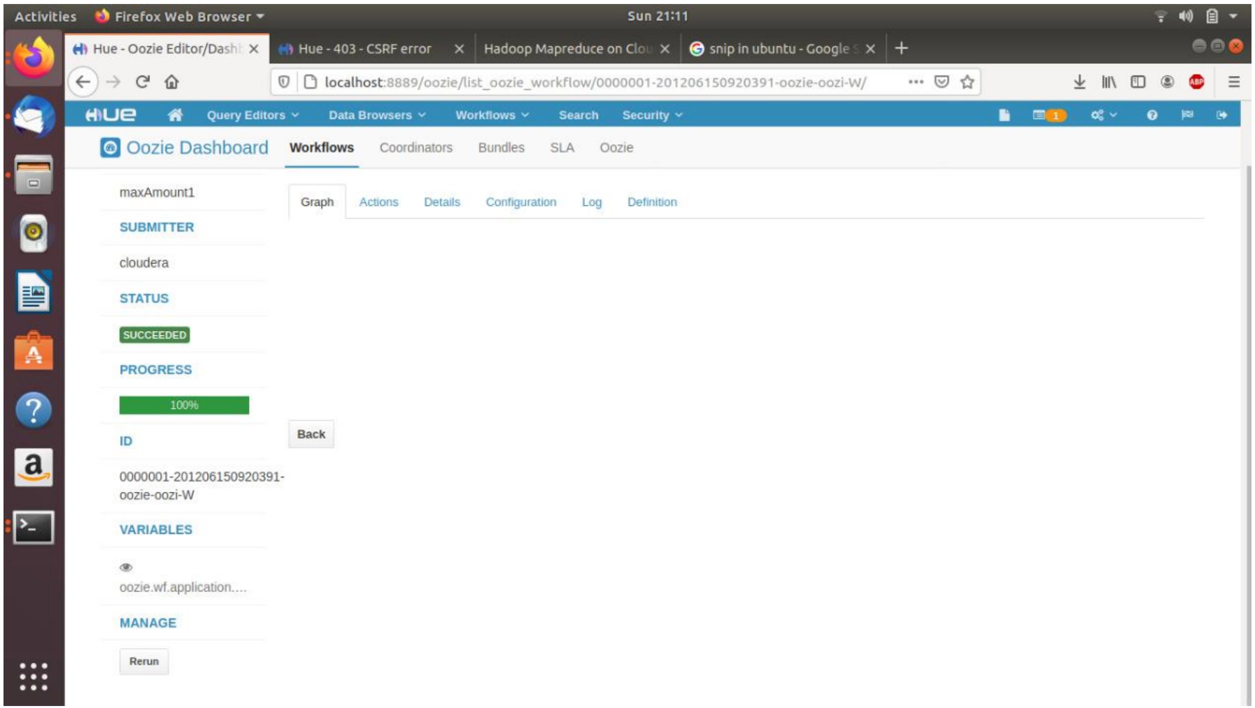
STATUS
RUNNING

PROGRESS
50%

ID
0000001-201206150920391-oozie-oozi-W

VARIABLES
oozie.wf.application....

MANAGE



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

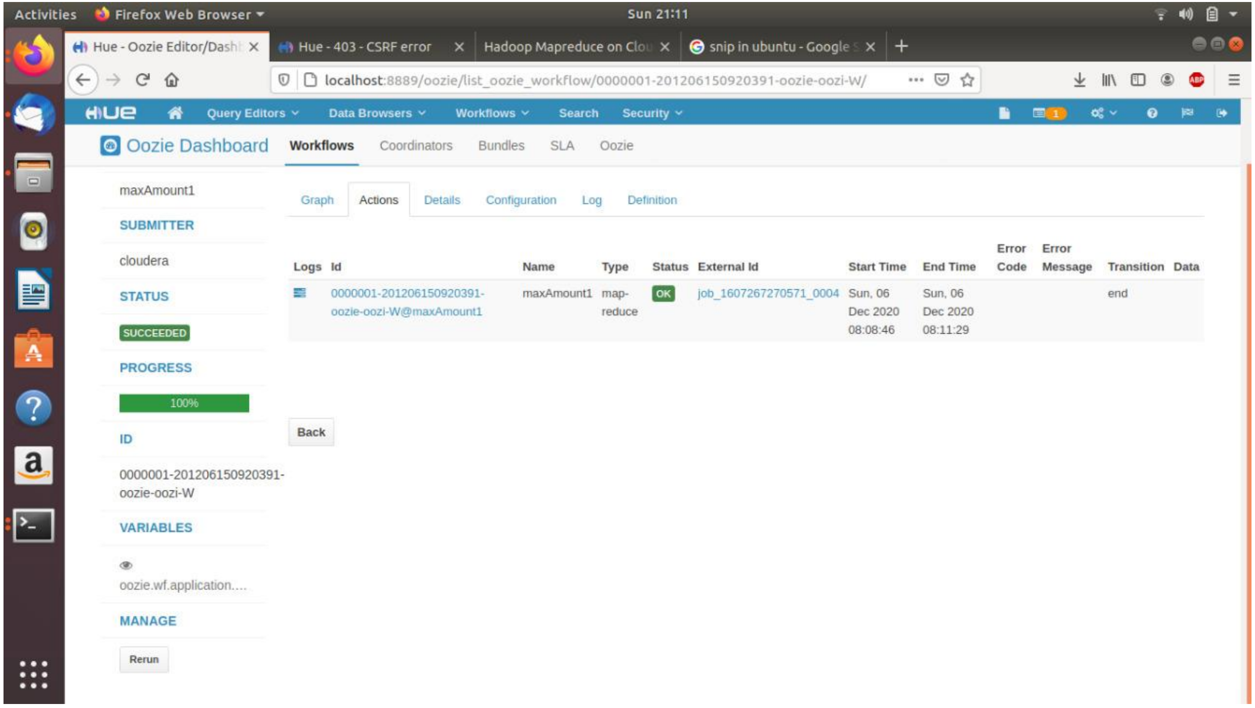
ID

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

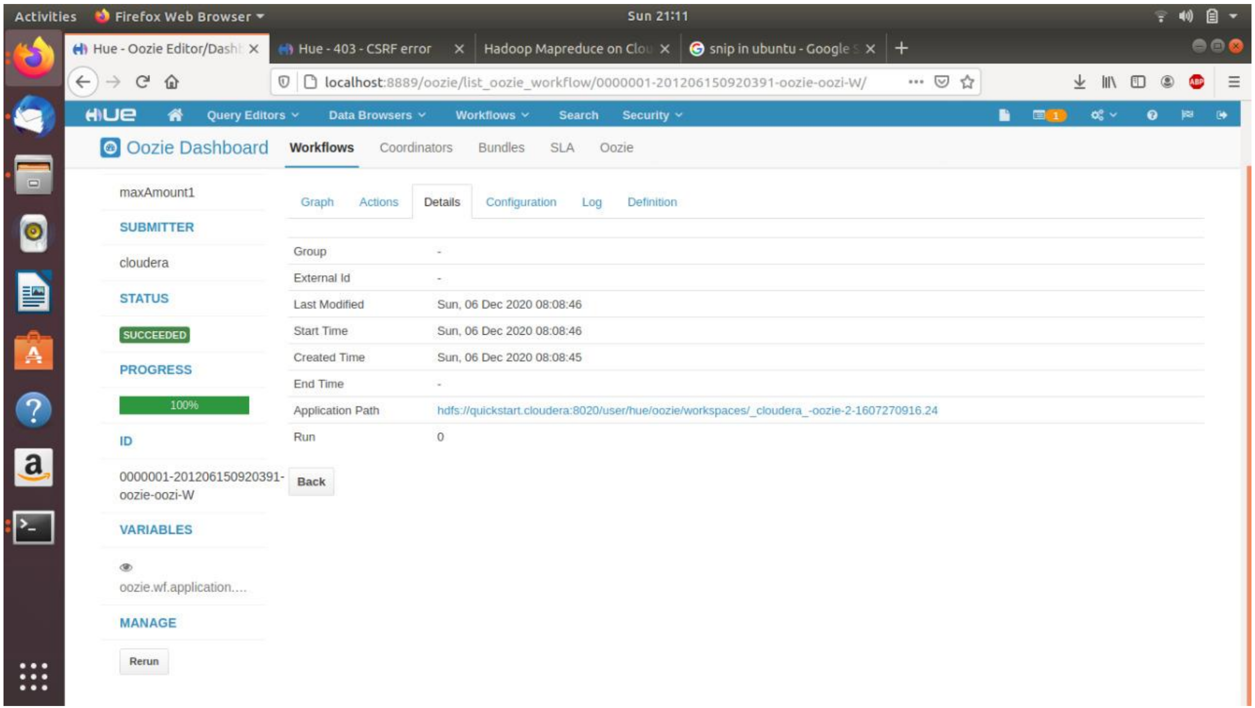
0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Logs	Id	Name	Type	Status	External Id	Start Time	End Time	Error Code	Error Message	Transition	Data
☰	0000001-201206150920391-oozie-oozi-W@maxAmount1	maxAmount1	map-reduce	OK	job_1607267270571_0004	Sun, 06 Dec 2020 08:08:46	Sun, 06 Dec 2020 08:11:29			end	



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

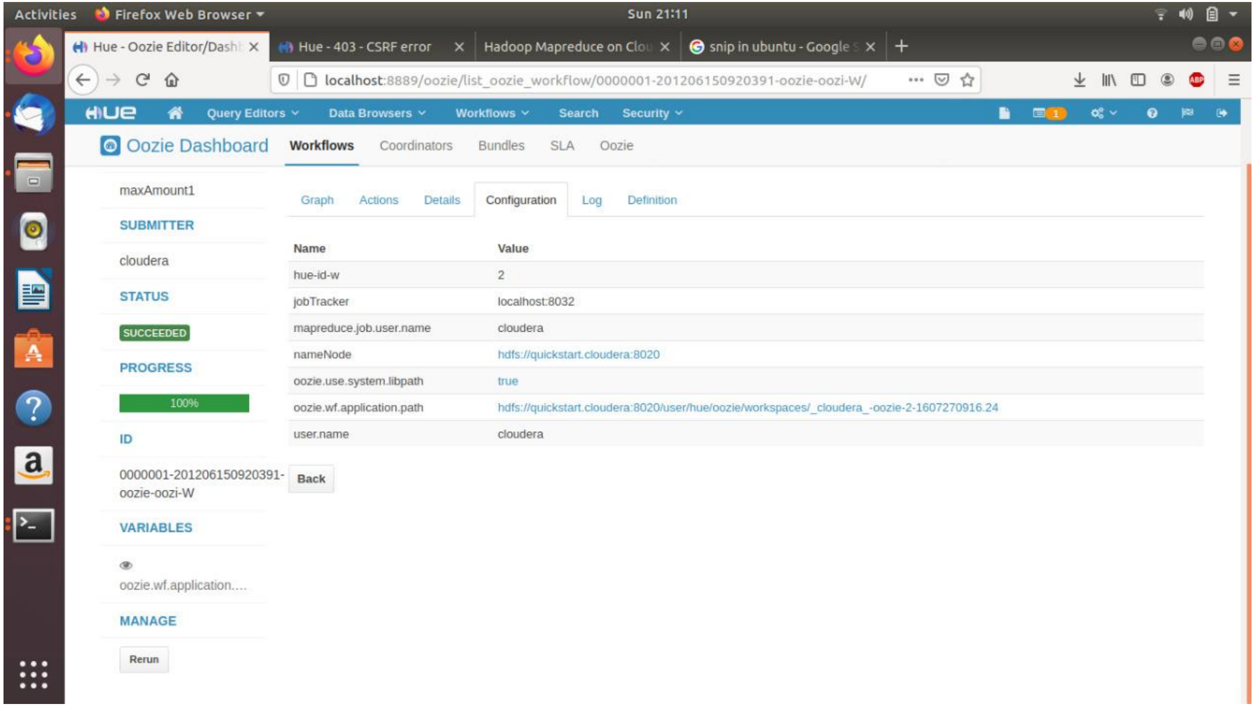
0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Group	-
External Id	-
Last Modified	Sun, 06 Dec 2020 08:08:46
Start Time	Sun, 06 Dec 2020 08:08:46
Created Time	Sun, 06 Dec 2020 08:08:45
End Time	-
Application Path	hdfs://quickstart.cloudera:8020/user/hue/oozie/workspaces/_cloudera_-oozie-2-1607270916.24
Run	0



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

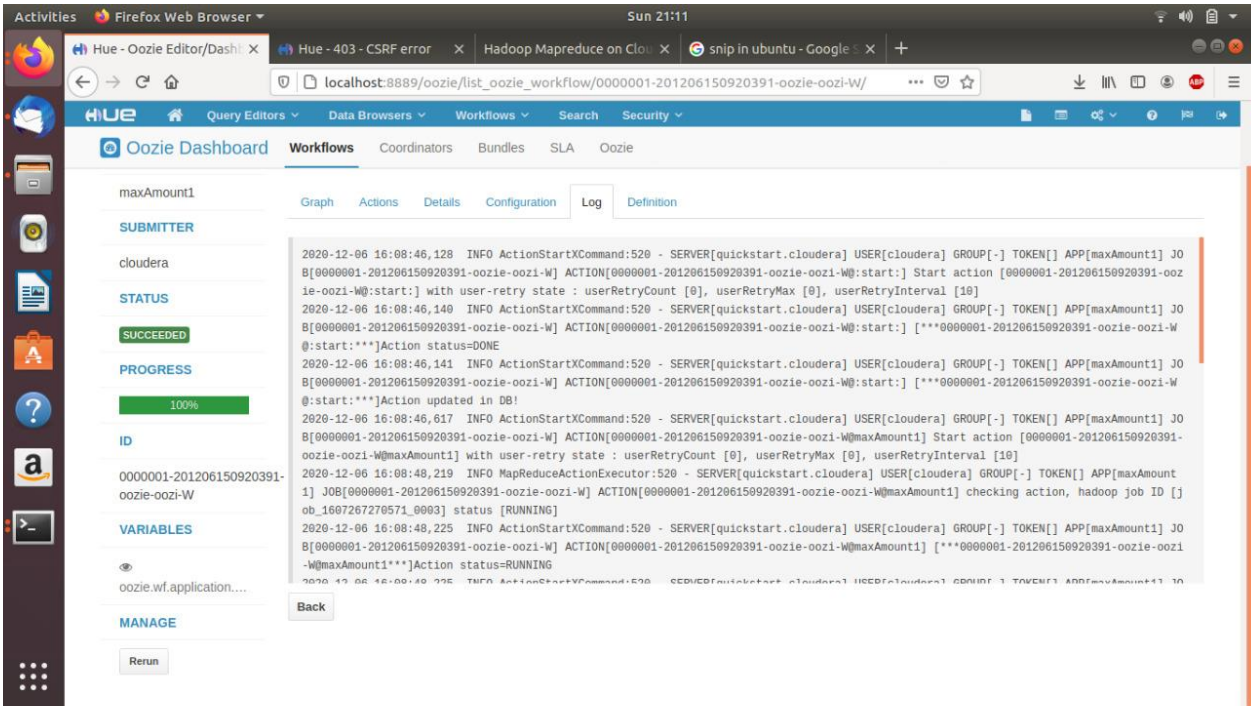
0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

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-2-1607270916.24
user.name	cloudera



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

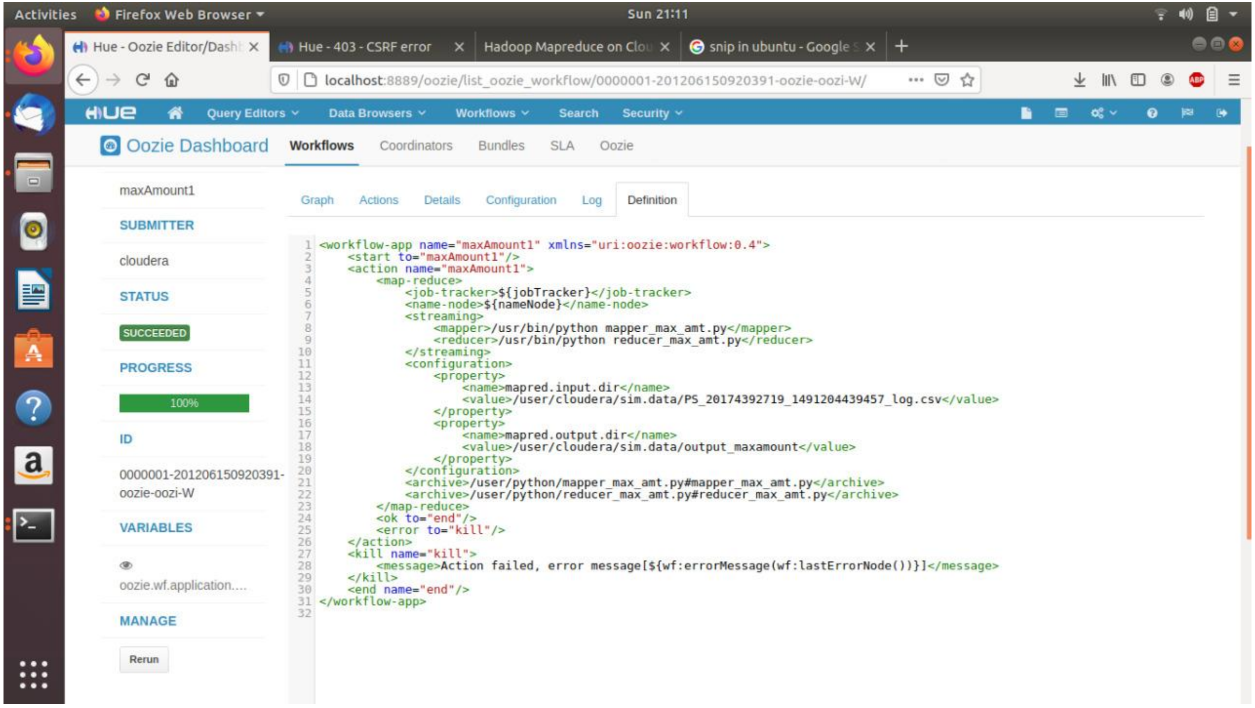
0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

```
2020-12-06 16:08:46,128  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
B[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@:start:] Start action [0000001-201206150920391-ooz
ie-oozi-W@:start:] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-06 16:08:46,140  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
B[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@:start:] [***0000001-201206150920391-oozie-oozi-W
@:start:***]Action status=DONE
2020-12-06 16:08:46,141  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
B[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@:start:] [***0000001-201206150920391-oozie-oozi-W
@:start:***]Action updated in DB!
2020-12-06 16:08:46,617  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
B[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@maxAmount1] Start action [0000001-201206150920391-
oozie-oozi-W@maxAmount1] with user-retry state : userRetryCount [0], userRetryMax [0], userRetryInterval [10]
2020-12-06 16:08:48,219  INFO MapReduceActionExecutor:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount
1] JOB[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@maxAmount1] checking action, hadoop job ID [j
ob_1607267270571_0003] status [RUNNING]
2020-12-06 16:08:48,225  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
B[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W@maxAmount1] [***0000001-201206150920391-oozie-oozi
-W@maxAmount1***]Action status=RUNNING
2020-12-06 16:08:48,225  INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JO
```



maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

```
<workflow-app name="maxAmount1" xmlns="uri:oozie:workflow:0.4">
    <start to="maxAmount1"/>
    <action name="maxAmount1">
        <map-reduce>
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <streaming>
                <mapper>/usr/bin/python mapper_max_amt.py</mapper>
                <reducer>/usr/bin/python reducer_max_amt.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_maxamount</value>
                </property>
            </configuration>
            <archive>/user/python/mapper_max_amt.py#mapper_max_amt.py</archive>
            <archive>/user/python/reducer_max_amt.py#reducer_max_amt.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>
```

Q3: The process of transferring data from mappers to reducers is known as shuffling, the process the process by which the system performs the sort and transfers the map output to the reducer as input. This step is necessary for reducers otherwise they wont have any input to work on.

Q4: Sorting in Hadoop helps reducer to easily distinguish when a new reduce task should start. It does it by checking when the next key is different from the previous one in input sorted data.