

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

The screenshot shows the Hue File Browser interface. The top navigation bar includes tabs for 'Hue - File Browser - part...', 'Hue - 403 - CSRF error', 'Hadoop Mapreduce on Clou...', and 'snip in ubuntu - Google'. The address bar shows the URL 'localhost:8889/filebrowser/view=/user/cloudera/sim.data/output/part-00000'. The main header displays 'File Browser' and the breadcrumb path 'Home / user / cloudera / sim.data / output / part-00000'. The left sidebar contains a list of actions: 'View as binary', 'Edit file', 'Download', 'View file location', 'Refresh', and 'INFO'. The 'Last modified' section shows the file was modified on 'Dec. 6, 2020 7:51 a.m.' by user 'cloudera' and group 'cloudera'. The file size is '77 bytes' and the mode is '100644'. The main content area displays a table of transaction data:

CASH_IN	1399283
CASH_OUT	2237590
DEBIT	41432
PAYMENT	2151493
TRANSFER	532909

Firefox Web Browser - Sun 20:54

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error Hadoop Mapreduce on Clou...

localhost:8889/oozie/list_oozie_workflow/0000000-201206150920391-oozie-oozi-W/

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

Workflow paymentJob1

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000000-201206150920391-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
	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	

Firefox Web Browser - Sun 20:54

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error Hadoop Mapreduce on Clou...

localhost:8889/oozie/list_oozie_workflow/0000000-201206150920391-oozie-oozi-W/

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

Workflow paymentJob1

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000000-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application...

MANAGE

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

Firefox Web Browser - Sun 20:54

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error Hadoop Mapreduce on Clou...

localhost:8889/oozie/list_oozie_workflow/0000000-201206150920391-oozie-oozi-W/

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

Workflow paymentJob1

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000000-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application...

MANAGE

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

Activities Firefox Web Browser Sun 20:54

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error Hadoop Mapreduce on Clou... +

localhost:8889/oozie/list_oozie_workflow/0000000-201206150920391-oozie-oozi-W/

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000000-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Workflow paymentJob1

Graph Actions Details Configuration Log Definition

2020-12-06 15:47:58,476 INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W:start:] Start action [0000000-201206150920391-oozie-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] JOB[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] JOB[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] JOB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] Start action [0000000-201206150920391-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[paymentJob1] 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] JOB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] [***0000000-201206150920391-oozie-oozi-W@paymentJob1***]Action status=RUNNING

2020-12-06 15:48:06,004 INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[paymentJob1] JOB[0000000-201206150920391-oozie-oozi-W] ACTION[0000000-201206150920391-oozie-oozi-W@paymentJob1] [***0000000-201206150920391-oozie-oozi-W@paymentJob1***]Action status=RUNNING

Back

Activities Firefox Web Browser Sun 20:55

Hue - File Browser - work... Hue - 403 - CSRF error Hadoop Mapreduce on Clou... +

localhost:8889/filebrowser/view=/user/hue/oozie/workspaces/_cloudera_oozie-1-1607269647.73

HUE Query Editors Data Browsers Workflows Search Security

File Browser

ACTIONS

View as binary

Edit file

Download

View file location

Refresh

INFO

Last modified

Dec. 6, 2020 7:47 a.m.

User

cloudera

Group

supergroup

Size

1.2 KB

Mode

100644

Home

Page 1 of 1

/ 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.pyc</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>
    </map-reduce>
  </action>
</workflow-app>
```

2nd Map Reducer Output

The screenshot shows the Hue File Browser interface. The breadcrumb path is `/ user / cloudera / sim.data / output_maxamount / part-00000`. The file `part-00000` is selected, and its content is displayed in a large text area. The content is a single line of text: `220696 9245516.64`. The left sidebar shows the file's metadata: Last modified Dec. 6, 2020 8:11 a.m., User cloudera, Group cloudera, Size 19 bytes, and Mode 100644.

File Browser

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

Page 1 of 1

ACTIONS

- View as binary
- Edit file
- Download
- View file location
- Refresh

INFO

Last modified: Dec. 6, 2020 8:11 a.m.

User: cloudera

Group: cloudera

Size: 19 bytes

Mode: 100644

220696 9245516.64

The screenshot shows the Hue Job Designer interface. The job is titled "Finding maximum Amount". The Mapper is `usr/bin/python mapper_max_amt.py` and the Reducer is `usr/bin/python reducer_max_amt.py`. The Hadoop job properties are configured as follows:

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

The job is configured with the following files and archives:

- Files: `usr/bin/python mapper_max_amt.py`, `usr/bin/python reducer_max_amt.py`
- Archives: `usr/bin/python mapper_max_amt.py`, `usr/bin/python reducer_max_amt.py`

The job is ready to be saved.

Job Designer

Description: Finding maximum Amount

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_maxamount

Files

- `usr/bin/python mapper_max_amt.py`
- `usr/bin/python reducer_max_amt.py`

Archives

- `usr/bin/python mapper_max_amt.py`
- `usr/bin/python reducer_max_amt.py`

Save Cancel

Activities Firefox Web Browser Sun 21:08

Hue - Job Designer x Hue - 403 - CSRF error x Hadoop Mapreduce on Clou x snip in ubuntu - Google x +

localhost:8889/jobsub/#list-designs

Job Designer

Designs

Search for design name Submit Edit Copy Move to trash New action View trash

<input type="checkbox"/>	Name	Description	Owner	Type	Status	Last modified
<input checked="" type="checkbox"/>	maxAmount1	finding maimum Amount	cloudera	streaming	shared	December 06, 2020 09:08 PM
<input type="checkbox"/>	paymentJob1	Payment.Type.Frequency	cloudera	streaming	shared	December 06, 2020 08:47 PM

Showing 1 to 2 of 2 entries Previous 1 Next

Activities Firefox Web Browser Sun 21:08

50% - Hue - Oozie Editor / X Hue - 403 - CSRF error x Hadoop Mapreduce on Clou x snip in ubuntu - Google x +

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

Oozie Dashboard

Workflows Coordinators Bundles SLA Oozie

Workflow maxAmount1

Graph Actions Details Configuration Log Definition

maxAmount1

cloudera

RUNNING

PROGRESS 50%

ID 0000001-201206150920391-oozie-oozi-W Back

VARIABLES

oozie.wf.application...

MANAGE

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dashl X Hue - 403 - CSRF error X Hadoop Mapreduce on Clou X snip in ubuntu - Google X +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Rerun

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dashl X Hue - 403 - CSRF error X Hadoop Mapreduce on Clou X snip in ubuntu - Google X +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Rerun

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	

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error x Hadoop Mapreduce on Clou... snip in ubuntu - Google x +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Rerun

Graph Actions Details Configuration Log Definition

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

Back

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error x Hadoop Mapreduce on Clou... snip in ubuntu - Google x +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000001-201206150920391-oozie-oozi-W

VARIABLES

oozie.wf.application....

MANAGE

Rerun

Graph Actions Details Configuration Log Definition

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

Back

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error X Hadoop Mapreduce on Clou... snip in ubuntu - Google X +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER cloudera

STATUS SUCCEEDED

PROGRESS 100%

ID 0000001-201206150920391-oozie-oozi-W

VARIABLES oozie.wf.application....

MANAGE Rerun

Graph Actions Details Configuration Log Definition

```
2020-12-06 16:08:46,128 INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JOB[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W:start:] Start action [0000001-201206150920391-oozie-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] JOB[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] JOB[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] JOB[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[maxAmount1] JOB[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W|maxAmount1] checking action, hadoop job ID [job_1607267270571_0003] status [RUNNING]
2020-12-06 16:08:48,225 INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JOB[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W|maxAmount1] [***0000001-201206150920391-oozie-oozi-W|maxAmount1***]Action status=RUNNING
2020-12-06 16:08:48,928 INFO ActionStartXCommand:520 - SERVER[quickstart.cloudera] USER[cloudera] GROUP[-] TOKEN[] APP[maxAmount1] JOB[0000001-201206150920391-oozie-oozi-W] ACTION[0000001-201206150920391-oozie-oozi-W|maxAmount1] [***0000001-201206150920391-oozie-oozi-W|maxAmount1***]Action status=RUNNING
```

Back

Activities Firefox Web Browser Sun 21:11

Hue - Oozie Editor/Dash... Hue - 403 - CSRF error X Hadoop Mapreduce on Clou... snip in ubuntu - Google X +

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

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

maxAmount1

SUBMITTER cloudera

STATUS SUCCEEDED

PROGRESS 100%

ID 0000001-201206150920391-oozie-oozi-W

VARIABLES oozie.wf.application....

MANAGE Rerun

Graph Actions Details Configuration Log Definition

```
1 <workflow-app name="maxAmount1" xmlns="uri:oozie:workflow:0.4">
2   <start to="maxAmount1"/>
3   <action name="maxAmount1">
4     <map-reduce>
5       <job-tracker>${jobTracker}</job-tracker>
6       <name-node>${nameNode}</name-node>
7       <streaming>
8         <mapper>/usr/bin/python mapper_max_amt.py</mapper>
9         <reducer>/usr/bin/python reducer_max_amt.py</reducer>
10      </streaming>
11      <configuration>
12        <property>
13          <name>mapred.input.dir</name>
14          <value>/user/cloudera/sim.data/PS_20174392719_1491204439457_log.csv</value>
15        </property>
16        <property>
17          <name>mapred.output.dir</name>
18          <value>/user/cloudera/sim.data/output_maxamount</value>
19        </property>
20      </configuration>
21      <archive>/user/python/mapper_max_amt.py#mapper_max_amt.py</archive>
22      <archive>/user/python/reducer_max_amt.py#reducer_max_amt.py</archive>
23    </map-reduce>
24    <ok to="end"/>
25    <error to="kill"/>
26  </action>
27  <kill name="kill">
28    <message>Action failed, error message[${wf.errorMessage(wf.lastErrorNode())}]</message>
29  </kill>
30  <end name="end"/>
31 </workflow-app>
32
```


Q3: The process of transferring data from mappers to reducers is known as shuffling, 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 won't have any input to work on.

Q4: Sorting in Hadoop helps the 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.