

Map Reduce Lab

Haaniyah Muhammad Mundia 14804

MR Task 1: Payment Job

1-

The Hive Job Designer interface shows a job named 'paymentJob1' with the description 'Payment.Type.Frequency'. It is configured as a streaming type job. The Mapper is set to 'Asr/bin/python mapper_freq.py' and the Reducer is 'Asr/bin/python reducer_freq.py'. Under 'Hadoop job properties', there are two properties: 'mapred.input.dir' with value 'Asr/cloudera/sim.data/PS_20174392719_14812' and 'mapred.output.dir' with value 'Asr/cloudera/sim.data/output'. There are also sections for 'Files' and 'Archives' with 'Add' and 'Delete' buttons. At the bottom, there are 'Save' and 'Cancel' buttons.

2-

The Oozie Dashboard shows the 'Workflows' tab. A workflow named 'paymentJob1' is displayed. The status is 'SUCCEEDED'. The progress bar is at 100%. The ID is '0000000-201207115855226-oozie-oozi-W'. The variables section shows 'oozie.wf.application....'. There are tabs for 'Graph', 'Actions', 'Details', 'Configuration', 'Log', and 'Definition'. A 'Back' button is visible.

3-

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

WORKFLOW

paymentJob1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

0000000-201207115855226-oozie-oozi-W [Back](#)

VARIABLES

oozie.wf.application,...

Workflow paymentJob1

Graph Actions Details **Configuration** Log Definition

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-

HUE Query Editors Data Browsers Workflows Search Security

File Browser

Search for file name Actions Move to trash Upload New

Home / user / cloudera / sim.data / **output** History Trash

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 2237590
DEBIT 41432
PAYMENT 2151493
TRANSFER 532989
```

MR Task 2: Maximum Amount

1-

The screenshot shows the HUE Job Designer interface for a streaming job named 'maxAmount1'. The job description is 'finding maximum amount'. The Mapper is set to 'usr/bin/python mapper_max_amt.py' and the Reducer is 'usr/bin/python reducer_max_amt.py'. Under 'Hadoop job properties', there are two properties: 'mapred.input.dir' with value 'hdfs://cloudera-him.data/PS_20174392719_14912' and 'mapred.output.dir' with value 'hdfs://cloudera-him.data/output_maxamt'. There are also sections for 'Files' and 'Archives' with 'Add' and 'Delete' buttons.

Job Designer

Job Design (streaming type)

Name: maxAmount1

Description: finding maximum amount

advanced

You can parameterize the values, using `$key1` or `$key2`. When the design is submitted, you will be prompted for the actual value of `key1`.

Mapper: usr/bin/python mapper_max_amt.py

Reducer: usr/bin/python reducer_max_amt.py

Hadoop job properties

Property name	Value		Delete
mapred.input.dir	hdfs://cloudera-him.data/PS_20174392719_14912	--	Delete
mapred.output.dir	hdfs://cloudera-him.data/output_maxamt	--	Delete

Add property

Files: Add file

Archives

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

Add archive

2-

The screenshot shows the HUE Oozie Dashboard for a workflow named 'maxAmount1'. The workflow is in a 'SUCCEEDED' state with 100% progress. The dashboard includes sections for 'WORKFLOW', 'SUBMITTER' (cloudera), 'STATUS' (SUCCEEDED), 'PROGRESS' (100%), 'ID' (0000001-201207115855226-oozie-oozi-W), and 'VARIABLES' (oozie.wf.application,...).

Oozie Dashboard

Workflows

Workflow maxAmount1

Graph Actions Details Configuration Log Definition

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000001-201207115855226-oozie-oozi-W

VARIABLES

oozie.wf.application,...

3-

The screenshot shows the Hue Oozie Dashboard. On the left sidebar, under 'WORKFLOW', 'maxAmount1' is selected. Below it, 'SUBMITTER' is 'cloudera', 'STATUS' is 'SUCCEEDED', and 'PROGRESS' is '100%'. The 'ID' is '0000001-201207115855226-oozie-oozi-W'. Under 'VARIABLES', 'oozie.wf.application....' is visible. The main panel shows the 'Configuration' tab for 'Workflow maxAmount1'. It contains a table with the following data:

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-

The screenshot shows the Hue File Browser. The breadcrumb path is 'Home / user / cloudera / sim.data / output_maxamt'. The file list is as follows:

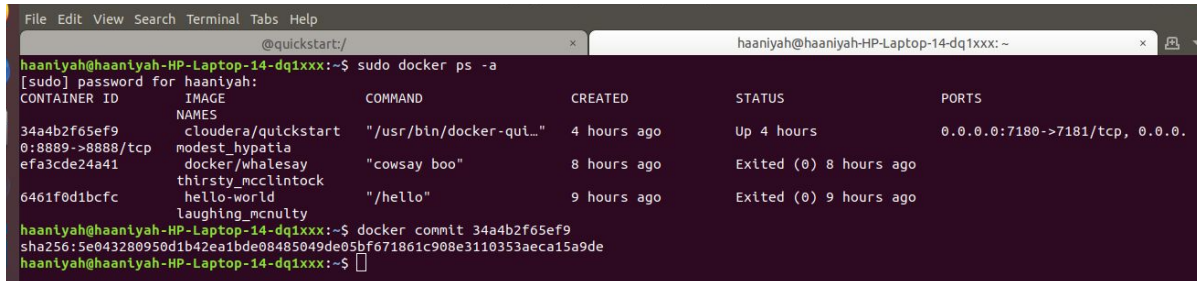
Name	Size	User	Group	Permissions	Date
.		cloudera	cloudera	drwxr-xr-x	December 07, 2020 07:10 AM
..		cloudera	cloudera	drwxr-xr-x	December 07, 2020 07:11 AM
._SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 07:11 AM
part-00000	19 bytes	cloudera	cloudera	-rw-r--r--	December 07, 2020 07:11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

5- Output

The screenshot shows the Hue File Browser with the file 'part-00000' selected. The breadcrumb path is 'Home / user / cloudera / sim.data / output_maxamt / part-00000'. The file content is displayed as a single line of text:

```
220696 92445516.64
```

Docker Commit



```
File Edit View Search Terminal Tabs Help
@quickstart/
haaniyah@haaniyah-HP-Laptop-14-dq1xxx:~$ sudo docker ps -a
[sudo] password for haaniyah:
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS              PORTS
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
efa3cde24a41        docker/whalesay     "cowsay boo"         8 hours ago         Exited (0) 8 hours ago
6461f0d1bcfc        thirsty_mcclintock   "hello-world"         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.

