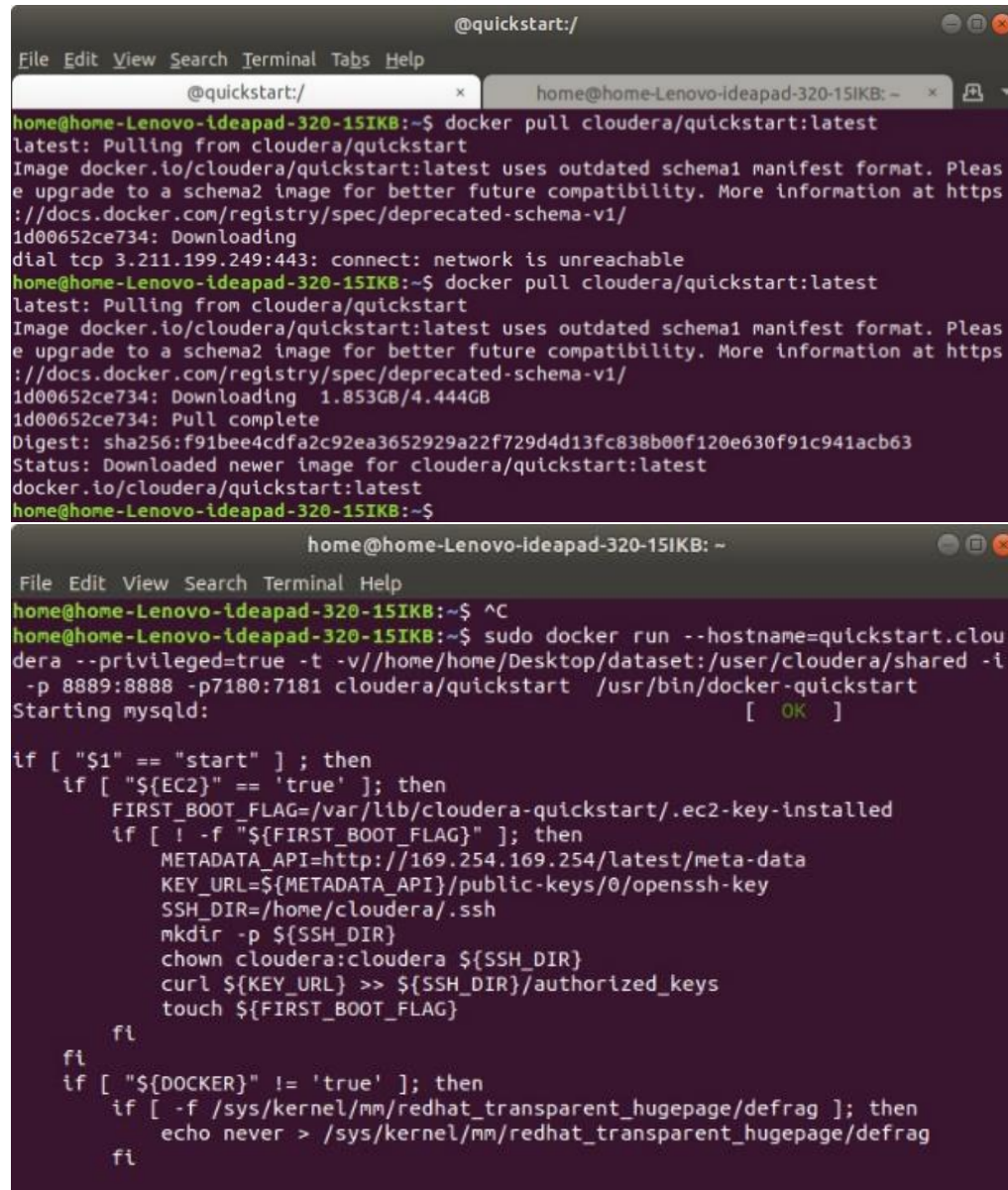


## Map Reduce Lab

### Marium Jamal – 14881

#### Before Tasks – downloading process:



```
@quickstart:/
File Edit View Search Terminal Tabs Help
@quickstart:/ x home@home-Lenovo-ideapad-320-15IKB: ~ x
home@home-Lenovo-ideapad-320-15IKB:~$ docker pull cloudera/quickstart:latest
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Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format.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 information at https://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading
dial tcp 3.211.199.249:443: connect: network is unreachable
home@home-Lenovo-ideapad-320-15IKB:~$ docker pull cloudera/quickstart:latest
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Image docker.io/cloudera/quickstart:latest uses outdated schema1 manifest format.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 information at https://docs.docker.com/registry/spec/deprecated-schema-v1/
1d00652ce734: Downloading 1.853GB/4.444GB
1d00652ce734: Pull complete
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Downloaded newer image for cloudera/quickstart:latest
docker.io/cloudera/quickstart:latest
home@home-Lenovo-ideapad-320-15IKB:~$

home@home-Lenovo-ideapad-320-15IKB: ~
File Edit View Search Terminal Help
home@home-Lenovo-ideapad-320-15IKB:~$ ^C
home@home-Lenovo-ideapad-320-15IKB:~$ sudo docker run --hostname=quickstart.cloudera --privileged=true -t -v//home/home/Desktop/dataset:/user/cloudera/shared -i -p 8889:8888 -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}
            curl ${KEY_URL} >> ${SSH_DIR}/authorized_keys
            touch ${FIRST_BOOT_FLAG}
        fi
    fi
    if [ "${DOCKER}" != 'true' ]; then
        if [ -f /sys/kernel/mm/redhat_transparent_hugepage/defrag ]; then
            echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
        fi
    fi
fi
```

## Task 1:

The image shows two screenshots from a web browser interface, likely a Linux desktop environment with Firefox.

**Top Screenshot: Hue Job Designer**

The browser window shows the Hue Job Designer interface. The URL is `localhost:8889/jobs/sub/new-design/streaming`. The page title is "Job Designer". The job is named "paymentJob1" with the description "Payment.Type.Frequency". The job type is "Job Design (streaming type)".

Below the job details, there is a section for "Hadoop job properties" with a table of properties:

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

There are also sections for "Files" and "Archives" with "Add file" and "Add archive" buttons respectively.

**Bottom Screenshot: Oozie Dashboard**

The browser window shows the Oozie Dashboard interface. The URL is `localhost:8889/oozie/list_oozie_workflow/0000000-201206194654538-oozie-oozi-W/`. The page title is "Oozie Dashboard". The workflow is named "Workflow paymentJob1".

The dashboard shows the workflow status as "SUCCEEDED" and "50%". The "PROGRESS" bar is at 50%. The "ID" is "0000000-201206194654538-oozie-oozi-W". The "VARIABLES" section shows "oozie.wf.application.path". The "MANAGE" section has "Kill" and "Suspend" buttons.

Browser tabs: (1) WhatsApp, Sakai : DATAWAREHOUSE, Synthetic Financial Data, mariumjamal/ubupdat, Hue - Oozie Editor/Dashboard

URL: localhost:8889/oozie/list\_oozie\_workflow/0000000-201206194654538-oozie-oozi-W/

HUE Query Editors Data Browsers Workflows Search Security

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

### Workflow paymentJob1

paymentJob1

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000000-201206194654538-oozie-oozi-W

VARIABLES

oozie.wf.applicatio...

MANAGE

Browser tabs: (1) WhatsApp, Sakai : DATAWAREHOUSE, Synthetic Financial Data, mariumjamal/, Hue - File Browser

URL: localhost:8889/filebrowser/#/user/cloudera/sim.data/output

HUE Query Editors Data Browsers Workflows Search Security

### File Browser

Search for file name Actions Move to trash

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

	Name	Size	User	Group	Permissions	Date
	↑		cloudera	cloudera	drwxr-xr-x	December 06, 2020 12:09 PM
	.		cloudera	cloudera	drwxr-xr-x	December 06, 2020 12:11 PM
	_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	December 06, 2020 12:11 PM
	part-00000	77 bytes	cloudera	cloudera	-rw-r--r--	December 06, 2020 12:11 PM

Show 45 of 2 items Page 1 of 1

## ACTIONS

View as  
binary

Edit file

Download

View file  
location

Refresh

## INFO

**Last modified**  
Dec. 6, 2020  
12:11 p.m.

**User**  
cloudera

**Group**  
cloudera

**Size**  
77 bytes

**Mode**  
100644

Home

Page 1 of 1

/ user / cloudera / sim.data / output / part-00000

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

## Task 2:

The image shows two screenshots of the Hue interface. The top screenshot is the 'Job Designer' for a 'Job Design (streaming type)'. The job is named 'maxAmount1' with the description 'finding maximum amount'. It is configured with a 'Mapper' and 'Reducer' both set to 'userbinpython mapper\_max\_amt.py' and 'userbinpython reducer\_max\_amt.py' respectively. Hadoop job properties are listed, including 'mapred.input.dir' and 'mapred.output.dir'. The bottom screenshot is the 'Oozie Dashboard' showing the 'Workflow maxAmount1'. The workflow is in a 'RUNNING' state, with a progress bar at 50%. The dashboard includes sections for 'WORKFLOW', 'SUBMITTER' (cloudera), 'STATUS' (RUNNING), 'PROGRESS' (50%), 'ID' (0000001-201206194654538-oozie-oozi-W), 'VARIABLES' (oozie.wf.applic...), and 'MANAGE'.

**Job Designer (streaming type)**

Name: maxAmount1  
Description: finding maximum amount  
advanced

You can parameterize the values, using `sysVar` for `$(sysVar)`. When the design is submitted, you will be prompted for the actual value of `sysVar`.

Mapper: userbinpython mapper\_max\_amt.py  
Reducer: userbinpython 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: Add file  
Archives: userpythonmapper\_max\_amt.py, userpythonreducer\_max\_amt.py

Save Cancel

**Oozie Dashboard**

Workflow maxAmount1

maxAmount1

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

RUNNING

PROGRESS

50%

ID

Back

0000001-201206194654538-oozie-oozi-W

VARIABLES

oozie.wf.applic...

MANAGE

WhatsApp x Sakai : DATAWAREHO x Synthetic Financial x mariumjamal/ubuntu x Hue - Oozie Editor x

localhost:8889/oozie/list\_oozie\_workflow/0000001-201206194654538-oo ...

HUE Query Editors v Data Browsers v Workflows v Search Security v

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

Workflow maxAmount1

maxAmount1

SUBMITTER

cloudera

STATUS

SUCCEEDED

PROGRESS

100%

ID

Back

0000001-201206194654538-oozie-oozi-W

VARIABLES

oozie.wf.applic...

MANAGE

(2) WhatsApp x Sakai : DATAWAREHO x Synthetic Financial x mariumjamal/ubuntu x Hue - File Browser x

localhost:8889/filebrowser/view=/user/cloudera/sim.data/output\_maxam...

HUE Query Editors v Data Browsers v Workflows v Search Security v

File Browser

ACTIONS

View as binary

Edit file

Download

View file location

Refresh

INFO

Last modified

Dec. 6, 2020 12:27 p.m.

User

cloudera

Group

cloudera

Size

19 bytes

Mode

100644

Home

Page 1 of 1

/ user / cloudera / sim.data / output\_maxam / part-00000

220696 92445516.64

### Task 3:

#### **a) Why is shuffling needed in MR? Give an example to justify.**

Shuffling is required in MR to transfer data from the mappers to the reducer because it is the process in which sorted intermediate output from mappers is transferred as the input to the reducer; otherwise, reducers would not have any input. An example of shuffling is given below.

#### **b) Why is sorting needed in MR? Give an example to justify.**

Sorting is needed in MR because it helps reducer to distinguish when a new reduce task should start i.e. it identifies when the next key in the reducer's input is different than the previous. An example for sorting is given below.

### **SORT and SHUFFLE**

