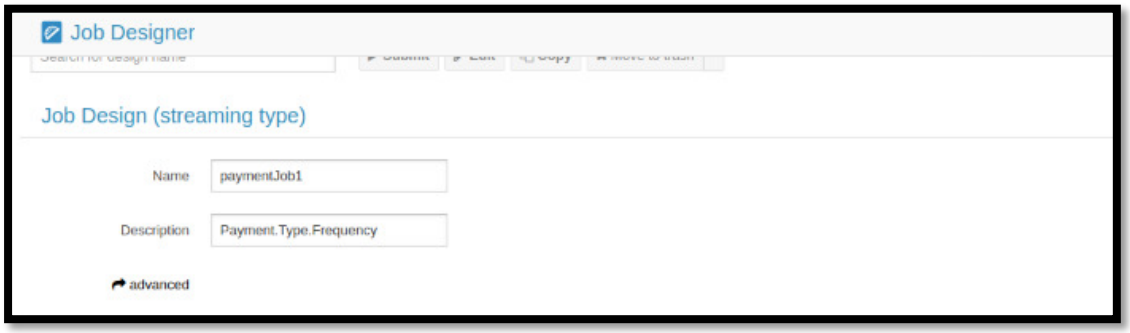
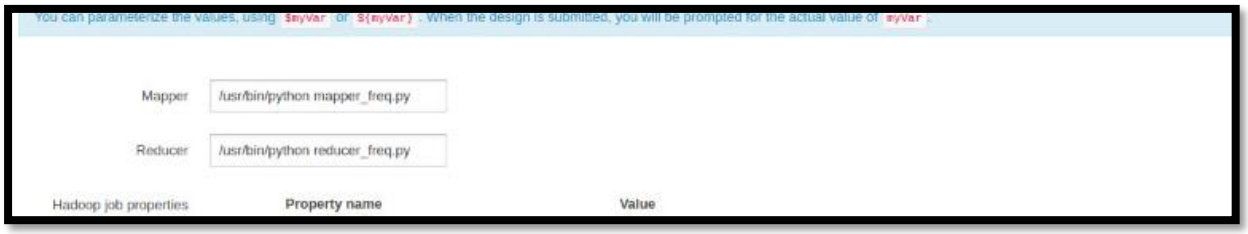
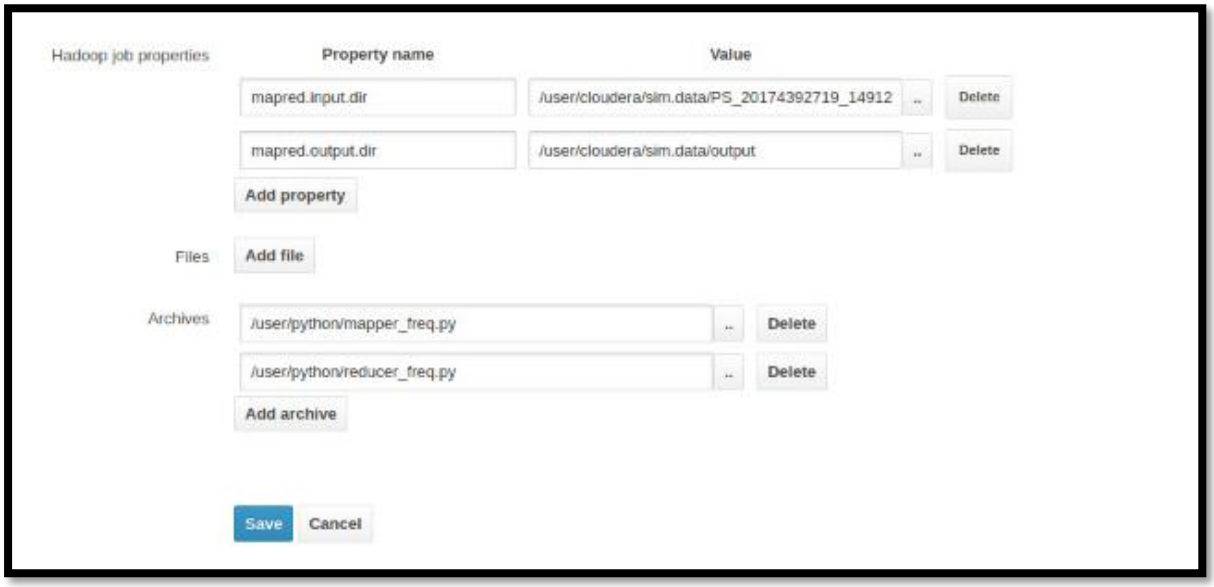
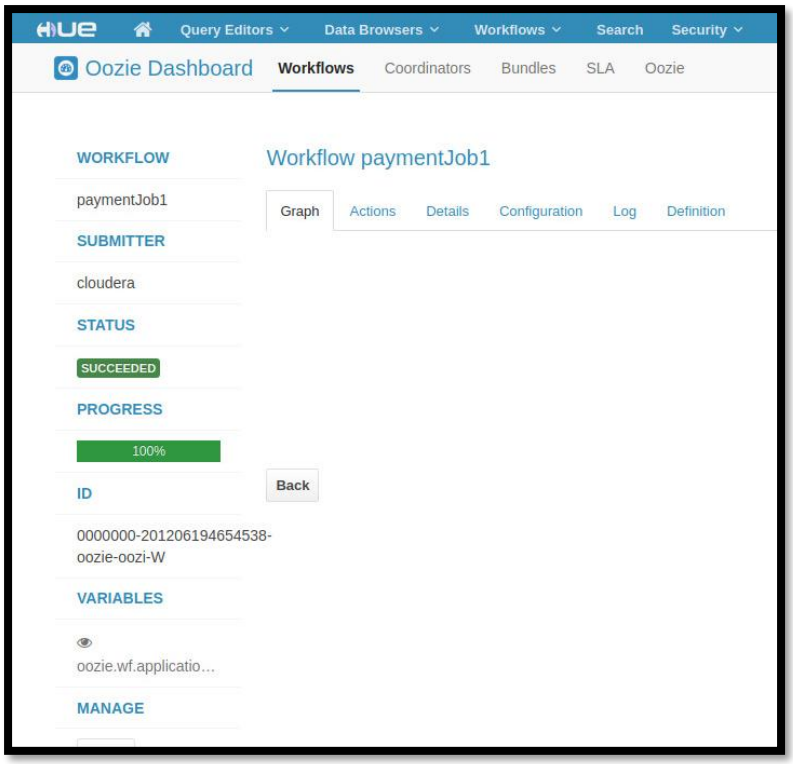
Map Reduce: Hurma Mahmood-14855

1. Task 1:
   Previous steps:

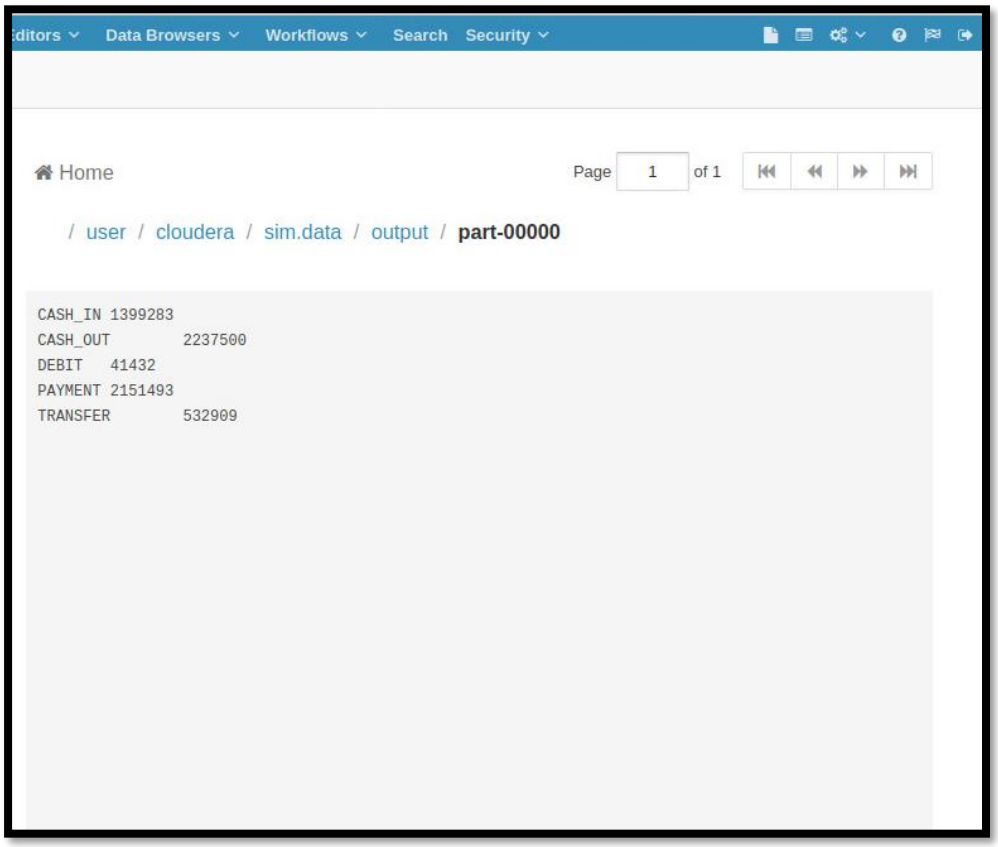






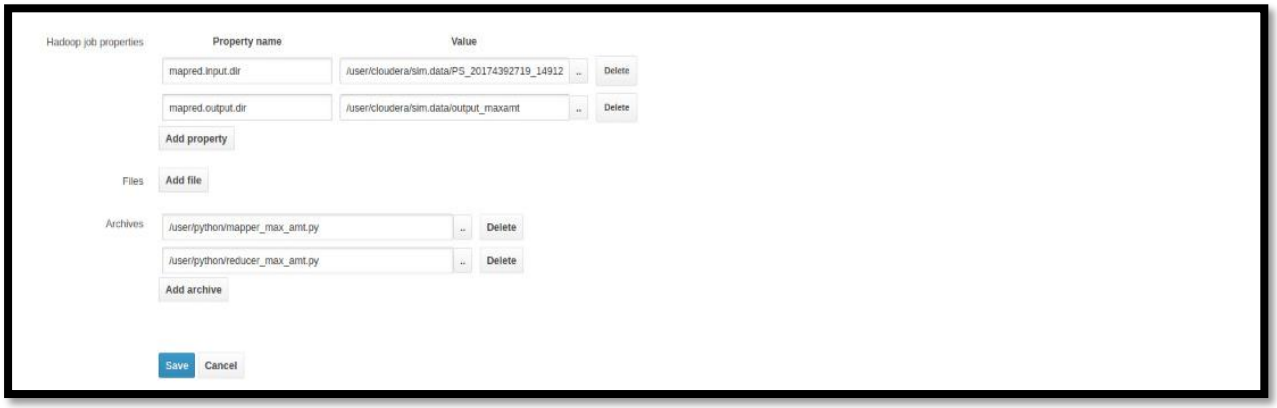

Query Editors
Data Browsers
Workflows
Search
Security
Oozie Dashboard
Workflows
Coordinators
Bundles
SLA
Oozie
WORKFLOW
paymentJob1
SUBMITTER
cloudera
STATUS
SUCCEEDED
PROGRESS
100%
ID
0000000-201206194654538-oozie-oozi-W
VARIABLES
oozie.wf.applicatio…
MANAGE
Workflow paymentJob1
Graph
Actions
Details
Configuration
Log
Definition
Back

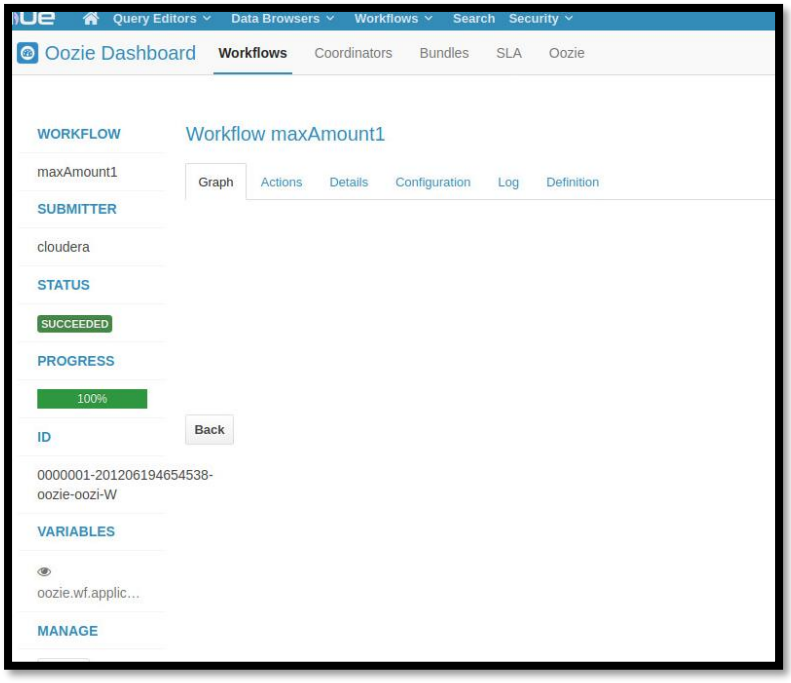
Output:


Data Browsers
Workflows
Search
Security
Home
Page 1 of 1
/ user / cloudera / sim.data / output / part-00000
CASH_IN 1399283
CASH_OUT 2237500
DEBIT 41432
PAYMENT 2151493
TRANSFER 532909

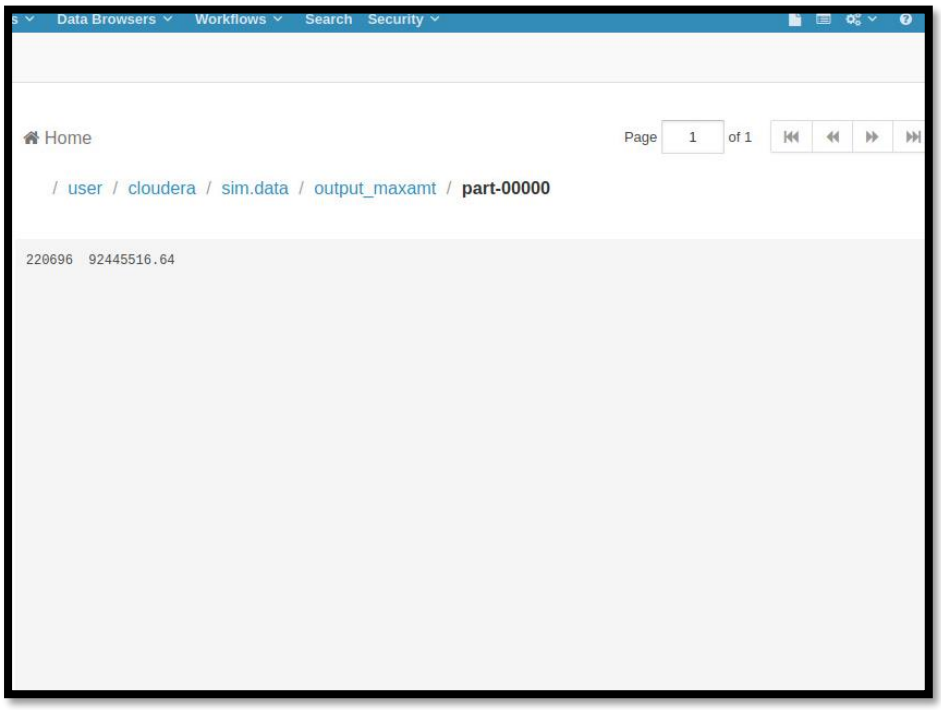
2. TASK 2:

Previous Steps:





Output:



3a. Why is shuffling needed in MR? Give an example to justify.

Shuffling is necessary in MapReduce because it is the process of transferring data from mappers to reducers through which the system performs the sort and transfers the map output to the reducer as input so it is necessary for reducers otherwise they would not have any input.

3b. Why is sorting needed in MR? Give an example to justify.

Sorting is necessary in MapReduce because sorting in Hadoop helps reducer to easily distinguish when a new reduce task should start. It basically saves time by stating a new reduce task when the next key in the sorted input data is different than previous.

Here in the given example both shuffling and sorting can be seen as explained above.