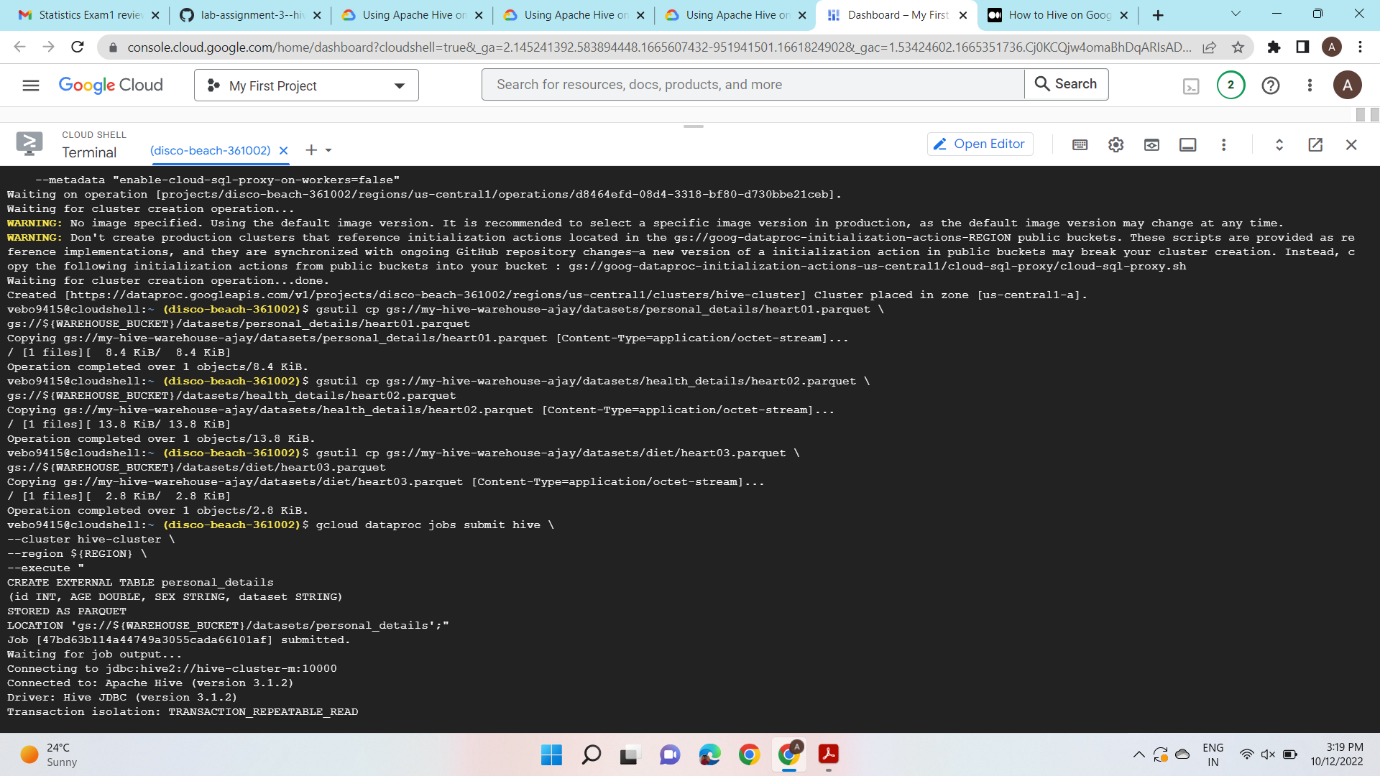
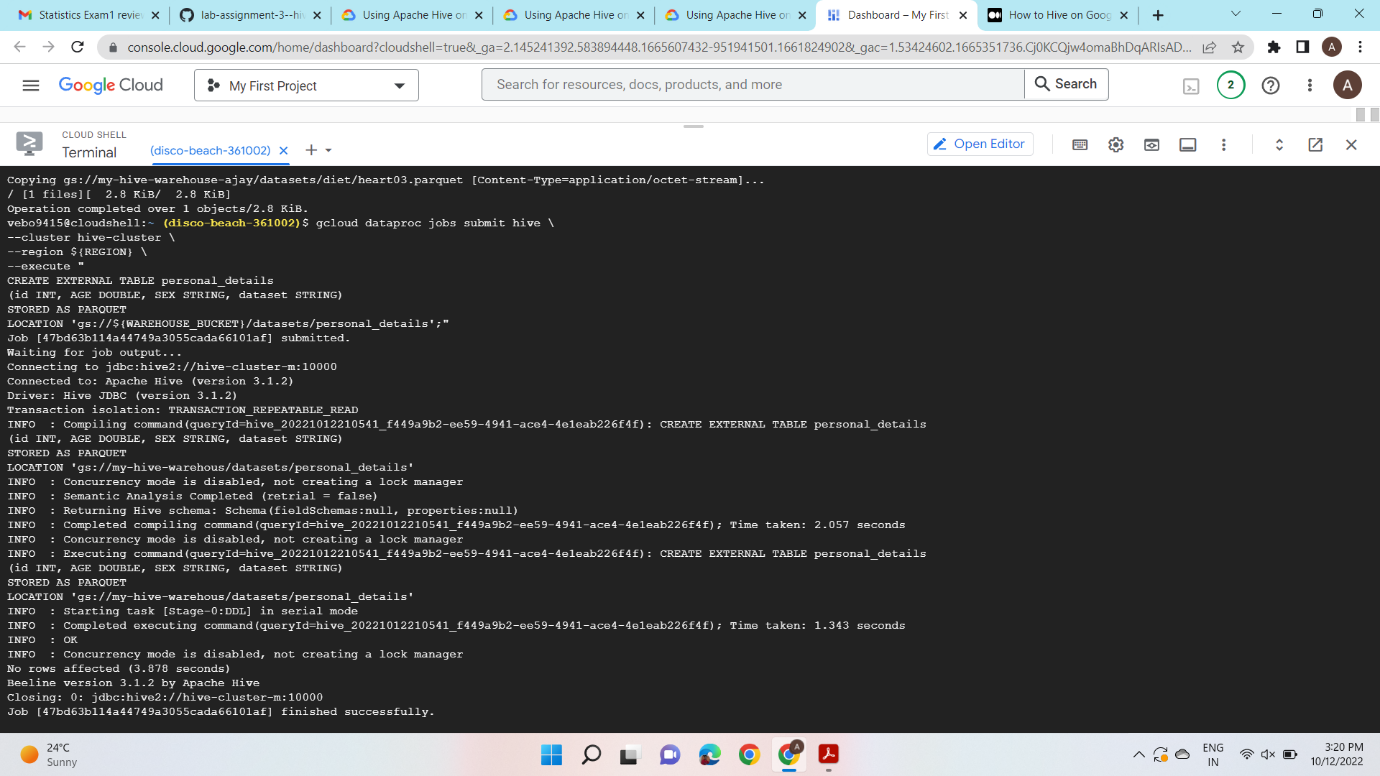
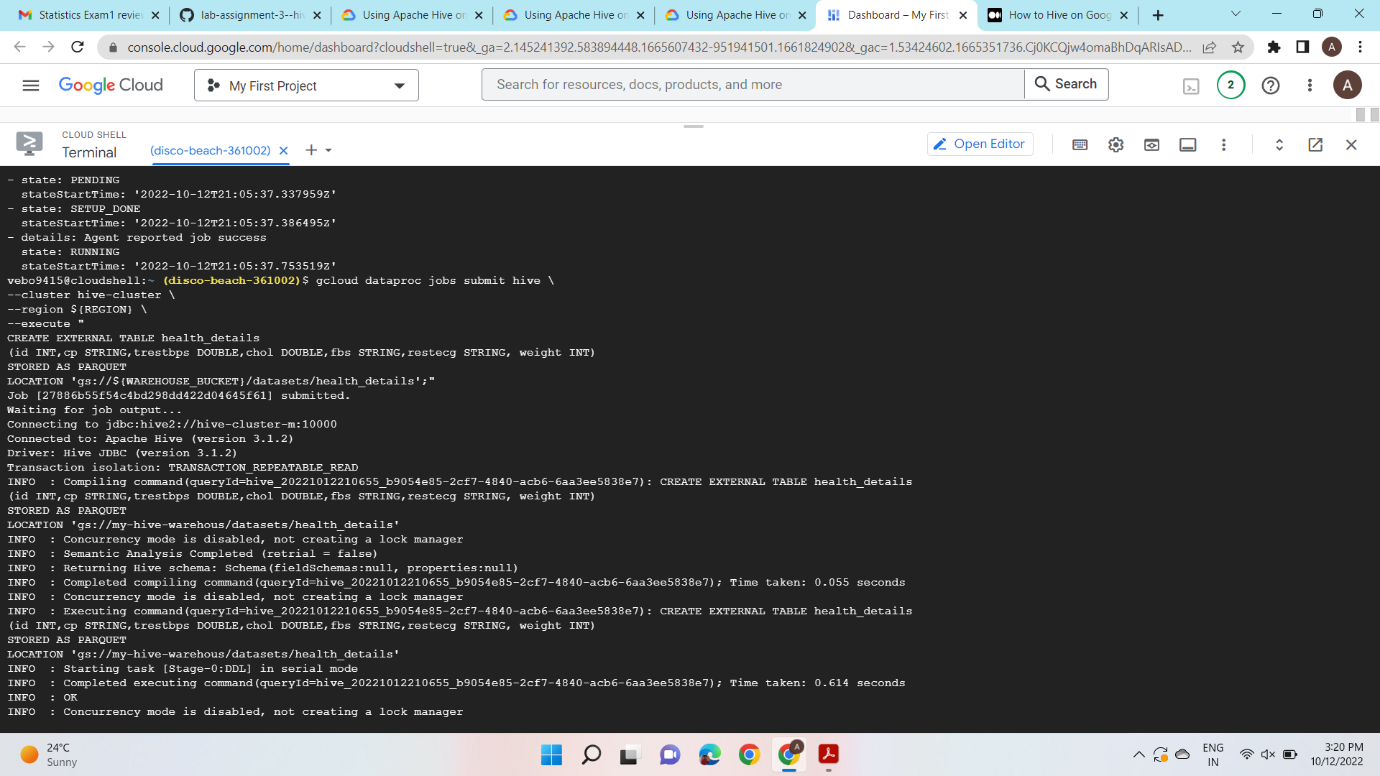
**Datacenter Scale Computing**

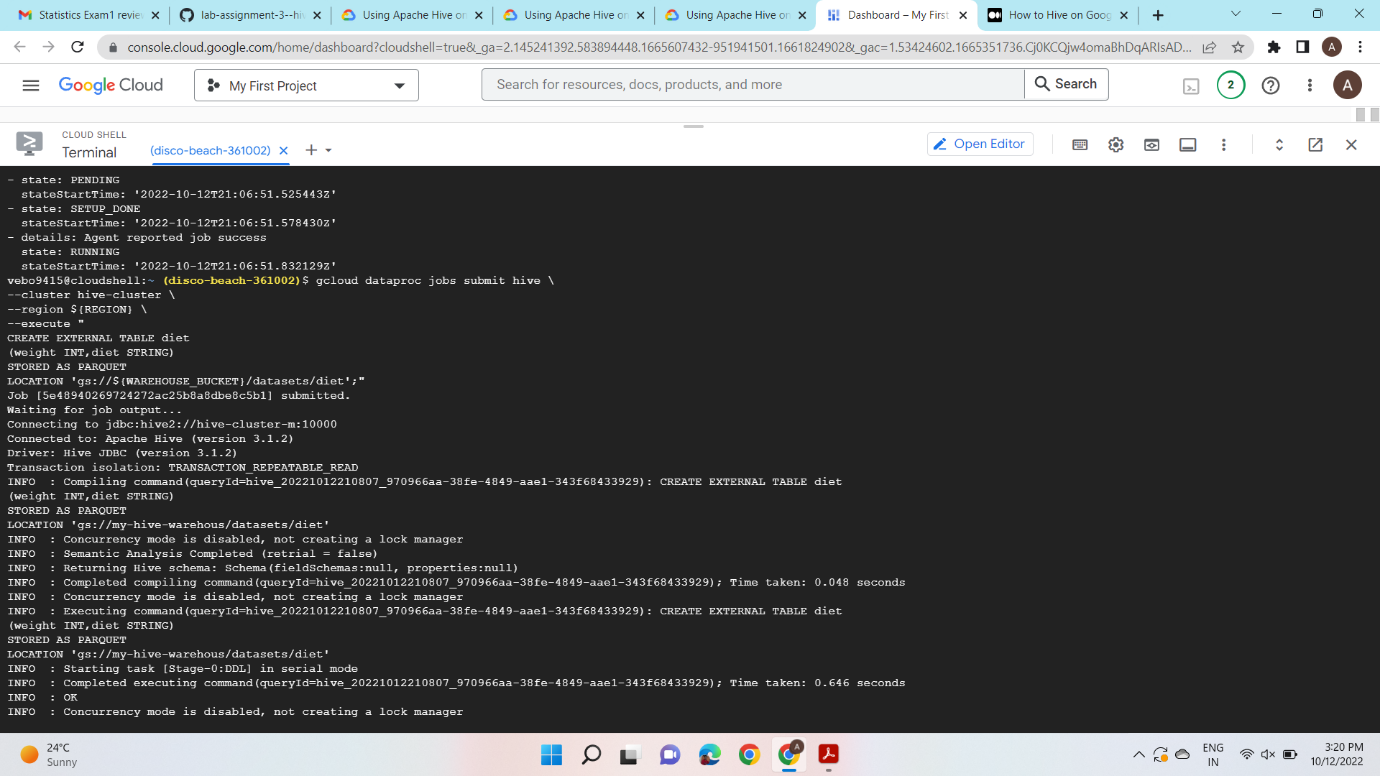
**Copying the datasets**



**Creating Hive tables**







1. **Write a Hive query to retrieve id, age and dataset where the dataset value is “Hungary”.**

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

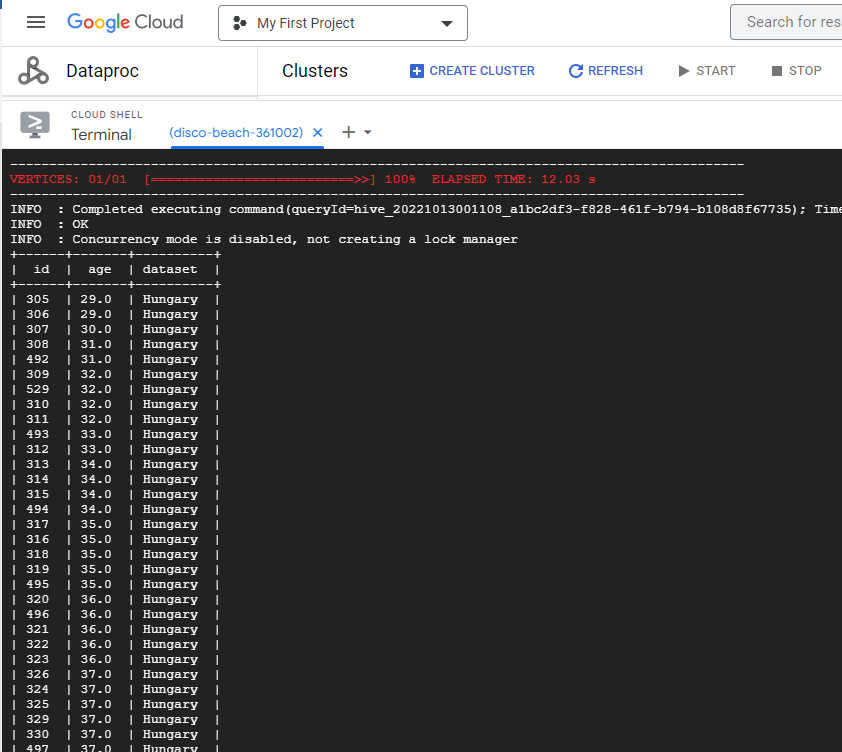
--region ${REGION} \

--execute "

SELECT id,age,dataset

FROM personal\_details

WHERE dataset = 'Hungary';"



2. Write a Hive query to retrieve id, age, dataset, chol and fbs and sort the values in

ascending order of *id.*

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

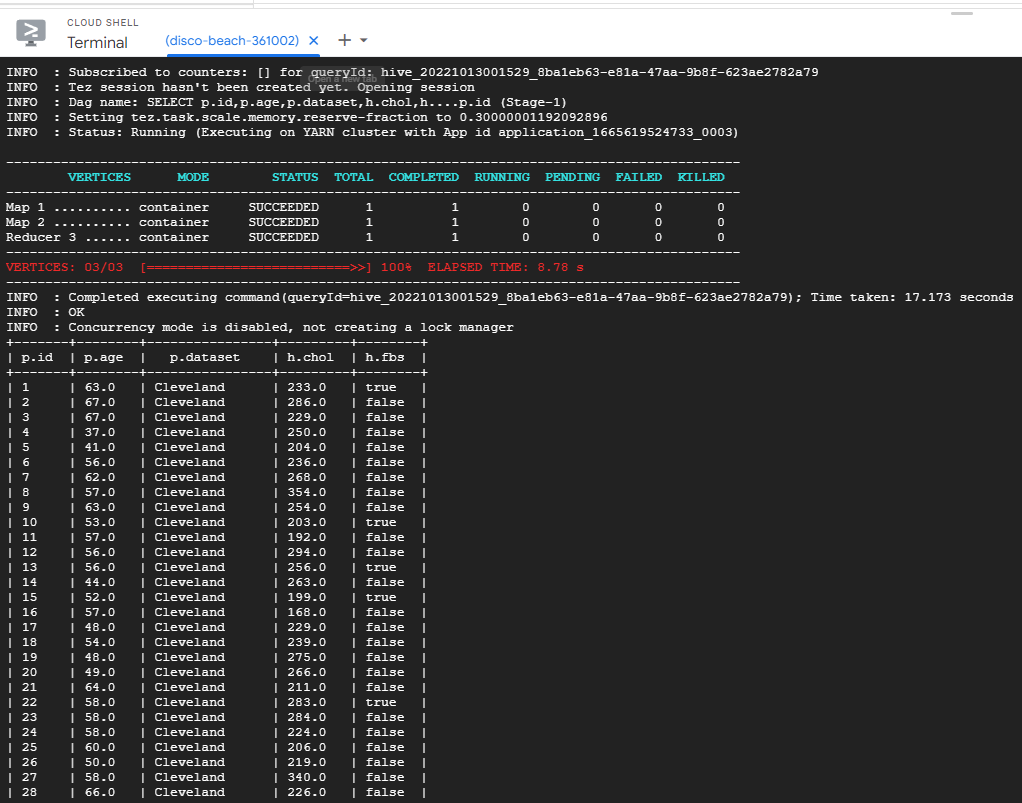
--region ${REGION} \

--execute "

SELECT p.id,p.age,p.dataset,h.chol,h.fbs

FROM personal\_details p JOIN health\_details h

ON (p.id = h.id) ORDER BY p.id;"



1. Modify the query in Q2 by using “DISTRIBUTE BY” and explain the difference.

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

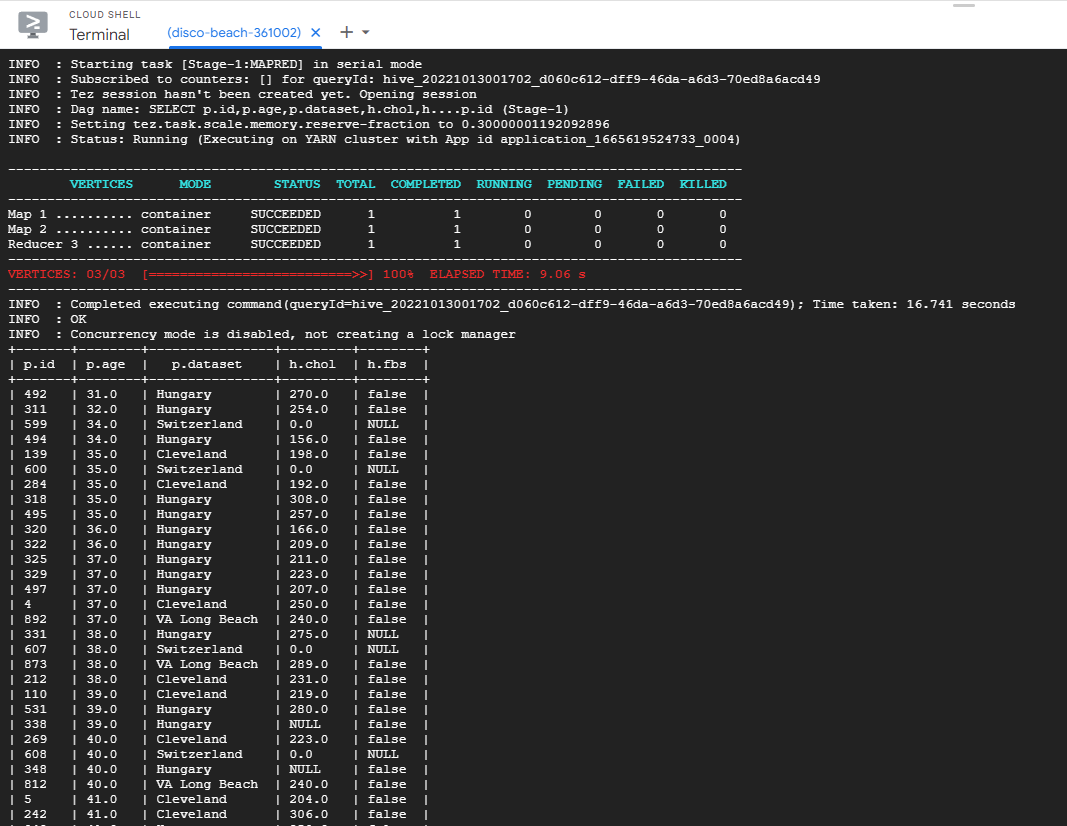
--execute "

SELECT p.id,p.age,p.dataset,h.chol,h.fbs

FROM personal\_details p JOIN health\_details h

ON (p.id = h.id) DISTRIBUTE BY p.id;"

Distribute by in hive is used to distribute the rows among reducers. So here all distribute by columns goes to same reducer so here the distribute by is p.id so here all the p.ids goes to the same reducer.



4. Modify the query in Q2 by using “CLUSTER BY” and explain the difference between Q2,

Q3 and Q4.

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

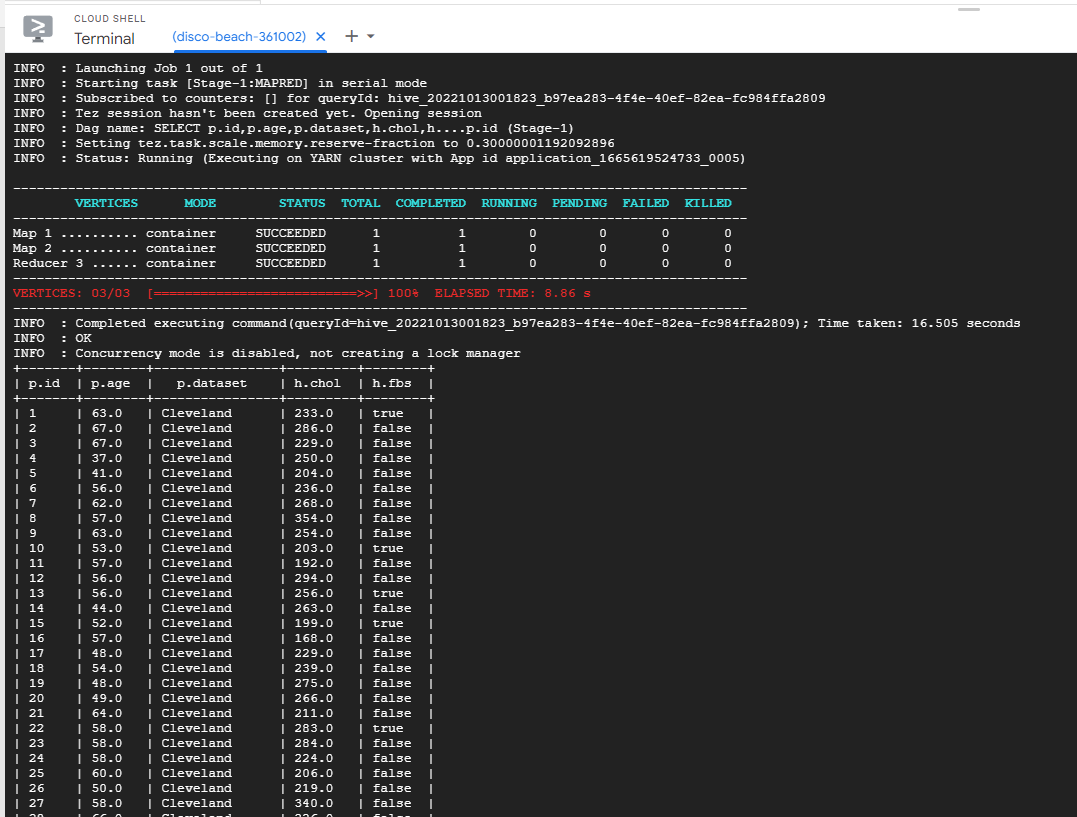
--execute "

SELECT p.id,p.age,p.dataset,h.chol,h.fbs

FROM personal\_details p JOIN health\_details h

ON (p.id = h.id) CLUSTER BY p.id;"

In Q2, sort by is used to sort by ascending or descending where as distributed by and cluster by deal with reducer so if we pass cluster by to a column then it divides it by multiple reducers where as distribute by sends it to same reducer.



5. Write a query to join tables *personal\_details, health\_details* and *diet*. Observe the results

and point out the error/issue if any.

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

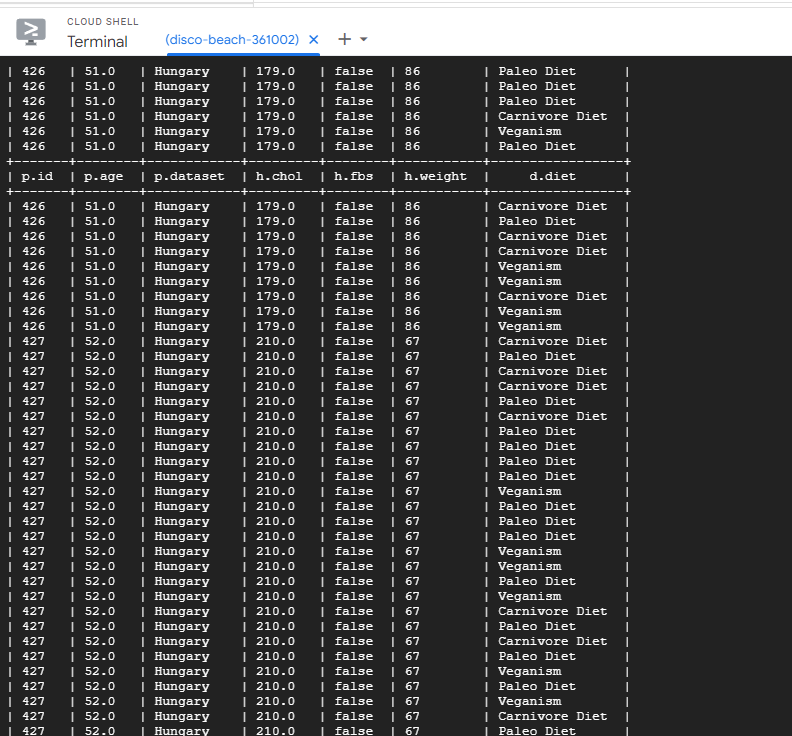
--execute "

SELECT p.id,p.age,p.dataset,h.chol,h.fbs,h.weight,d.diet

FROM personal\_details p JOIN health\_details h JOIN diet d

ON ((p.id = h.id) AND (h.weight = d.weight)) ORDER BY p.id;"

Here when we joined all 3 tables we had a lot of data redundancy. Because there is no primary key so when we joined then there are duplicates generated.



**Theory Questions - (10 points each)**

1. In your own words, describe the working of Hive. (Hint - how hive is on top of hadoop

and internally what techniques are used for querying)

Apache Hive was designed to eliminate the job of writing the map-reduce Java programs. The programs are written in Hive Query Language but its similar to SQL. Hive translates hive queries into MapReduce programs.

1. List out the advantages and disadvantages of HIVE

Advantages

1. Easy to use
2. Quick experience which enables bulk processing means analysing data on bits and parts.
3. Fault tolerance
4. Takes very little time to write hive queries in comparison to mapreduce code.
5. Multiple users can query the data with help of HiveQL

Disadvantages

1. Hive is not designed for the OLTP (Online transaction processing). We can use it for OLAP.

2. It does not offer real-time queries.

3. It provides limited subquery support.

4. Latency of Hive is generally very high.