**LAB 3 - HIVE**

**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 \

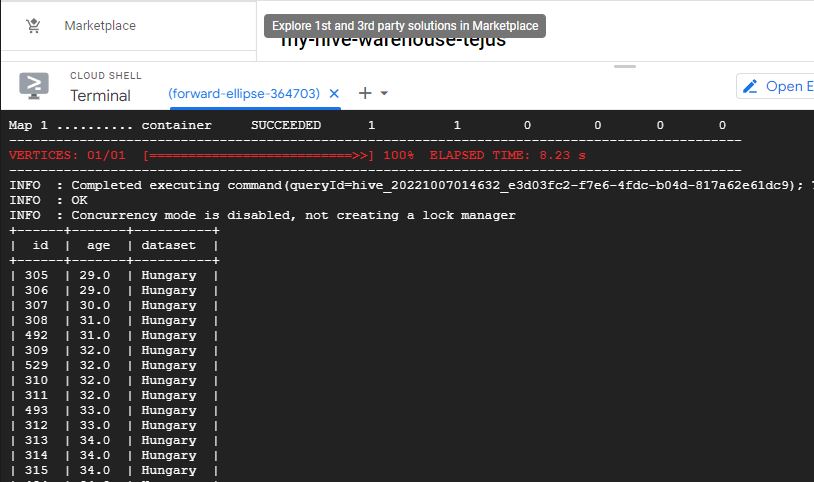
--region ${REGION} \

--execute "

SELECT id,age,dataset

FROM personal\_details

WHERE dataset = 'Hungary';"



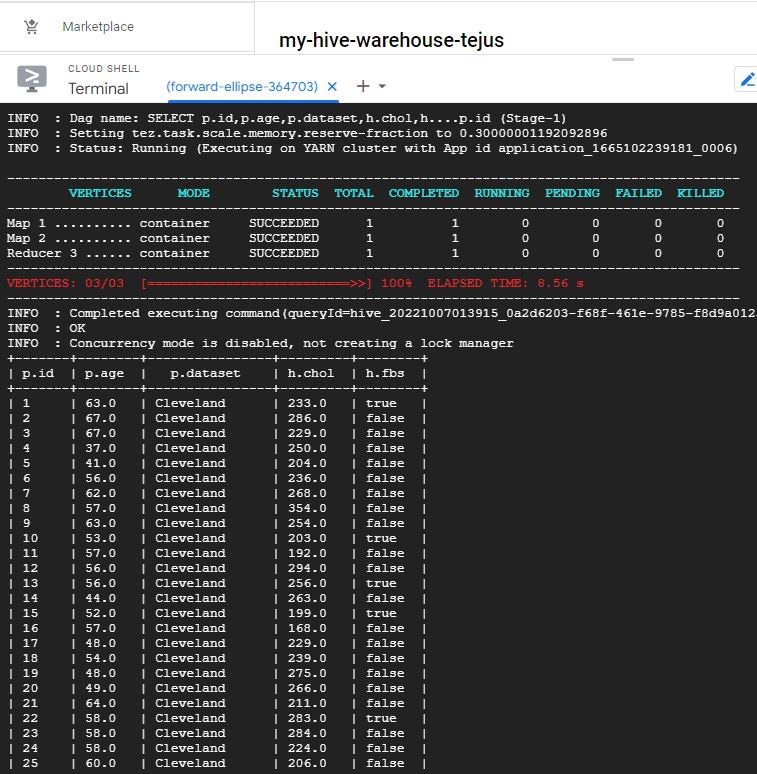
**2. Write a Hive query to retrieve id, age, dataset, chol and fbs and sort the values in ascending order of id.**

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\_deatils h

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



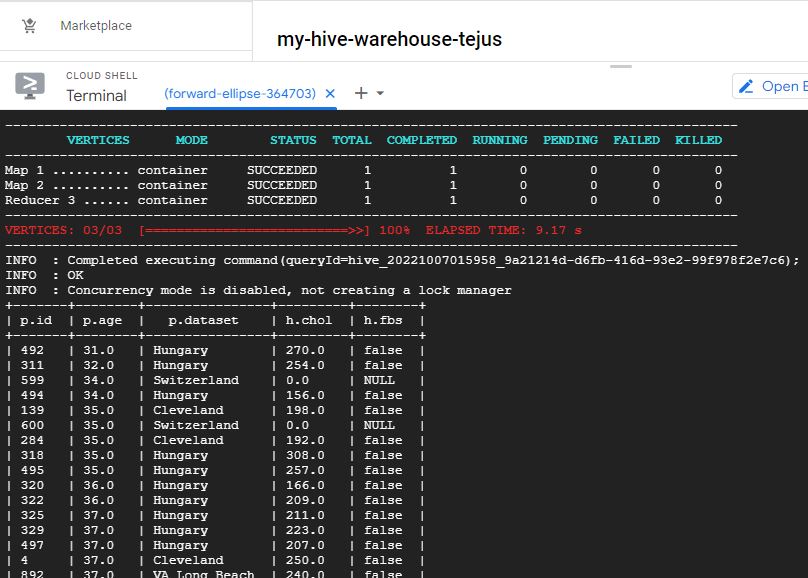
**3. Modify the query in Q2 by using “DISTRIBUTE BY” and explain the difference.**

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\_deatils h

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



**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\_deatils h

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

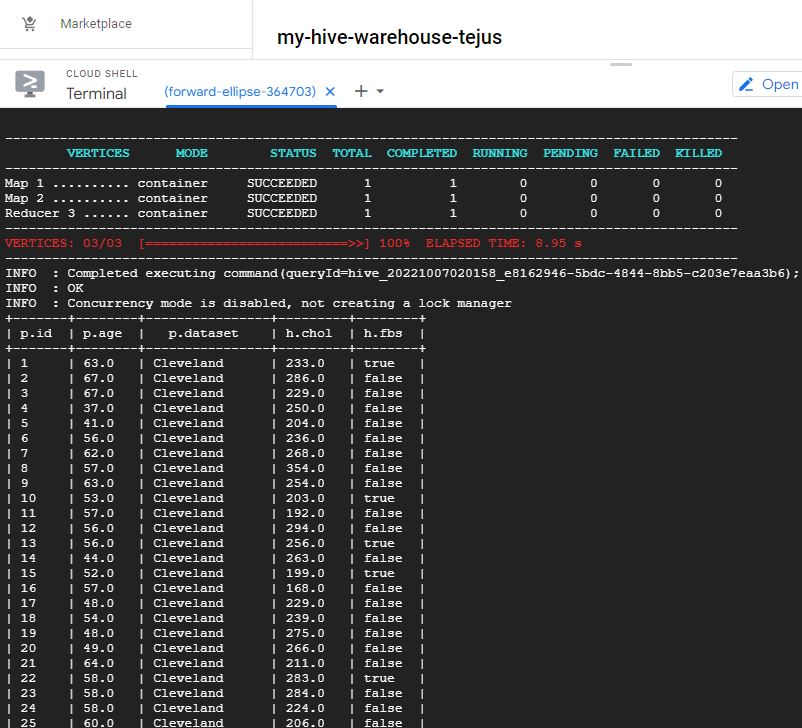
Sort by and order by are used to sort the output data in ascending or descending order. Whereas cluster by and distribute by are used to distribute the data to reducers.

When using sort by the data is sorted in the reducers only, they are not sorted globally. So when we get the output files from the reducers, each file will be sorted but both the files together will not be sorted. And there might be overlapping data ranges as well. For example when we sort by name, same names might be in both the reducers.

Order by passes all the data to one reducer, so the data is sorted globally. This might impact the performance negatively the file size is large.

Distribute by distributes the data to all reducers such that there is no overlapping data range, i.e when we distribute by name, all the same names will be sent to the same reducer. For example, if the name Regina is sent to reducer 1, then reducer 2 ….. n will not have that name. Distribute by neither sorts the data in each reducer nor globally.

Cluster by sorts the data in each reducers and makes sure that there is no overlapping data ranges but it does not sort the output data globally.



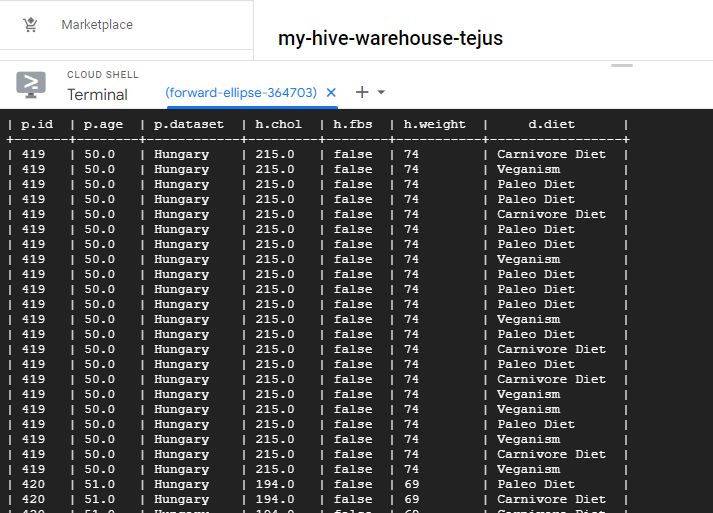
**5. Write a query to join tables personal\_details, health\_details and diet. Observe the results and point out the error/issue if any.**

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\_deatils h JOIN diet d

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



When we join all the 3 tables we get a lot of data redundancy. This is because the diet table and health\_details table have only one column in common which is the weight column, and it is not a primary key in either table. So when we join these two tables based on weight column we get redundant data. The personal\_details and health\_details have the id column which is a primary key in both the tables, so joining these two doesn’t give us any duplicate data.

**Theory Questions**

**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)**

Hive is a data warehousing system which uses HiveQL an SQL like language to manage and analyze massive amount of data. This enable people who are familiar with SQL to manage petabytes of data using HiveQL. Hive uses batch processing to operate on a large and distributed database. Hive transforms the HQL queries into map reduce or tez jobs which run on the HDFS framework, YARN.

**2. List out the advantages and disadvantages of HIVE**

Advantages:

* Hive is easy to code using HiveQL, which is similar to SQL
* Hive allows users to store and access large amounts of data from different platforms like HDFS, HBase, Amazon S3,etc.
* Can easily embed MapReduce code to process unstructured data.
* Since Hive is stored on top of HDFS, Hadoop provides fault tolerance.

Disadvantages:

* Hive doesn’t support online transaction processing (OLTP)
* Latency is very high
* Hive can’t be used for real-time data querying.