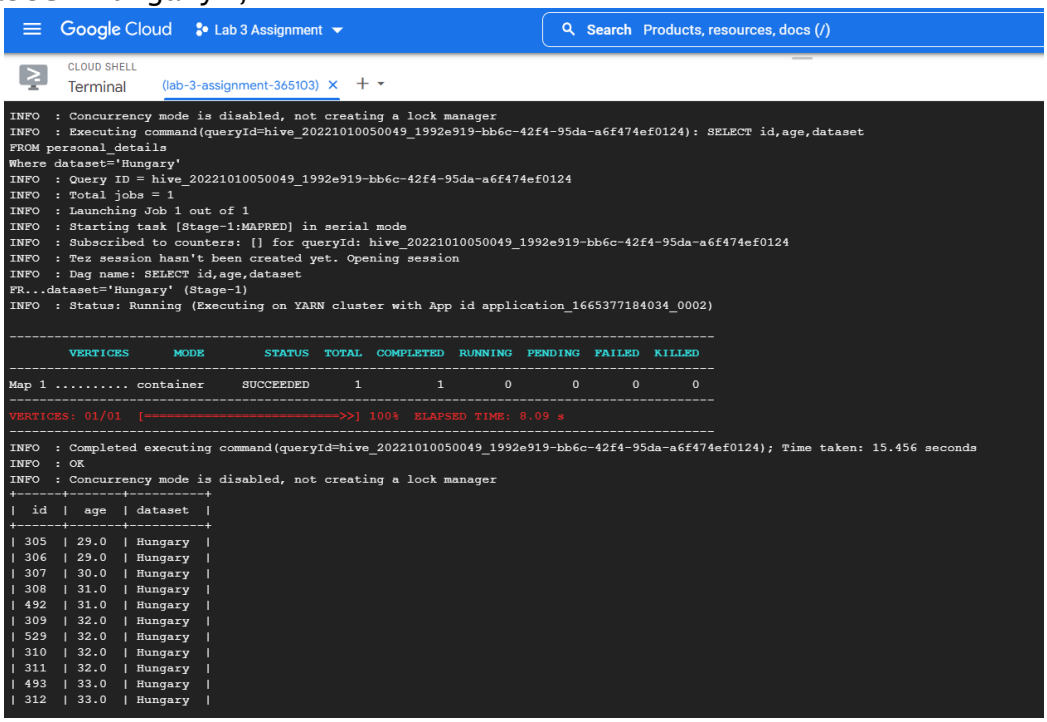
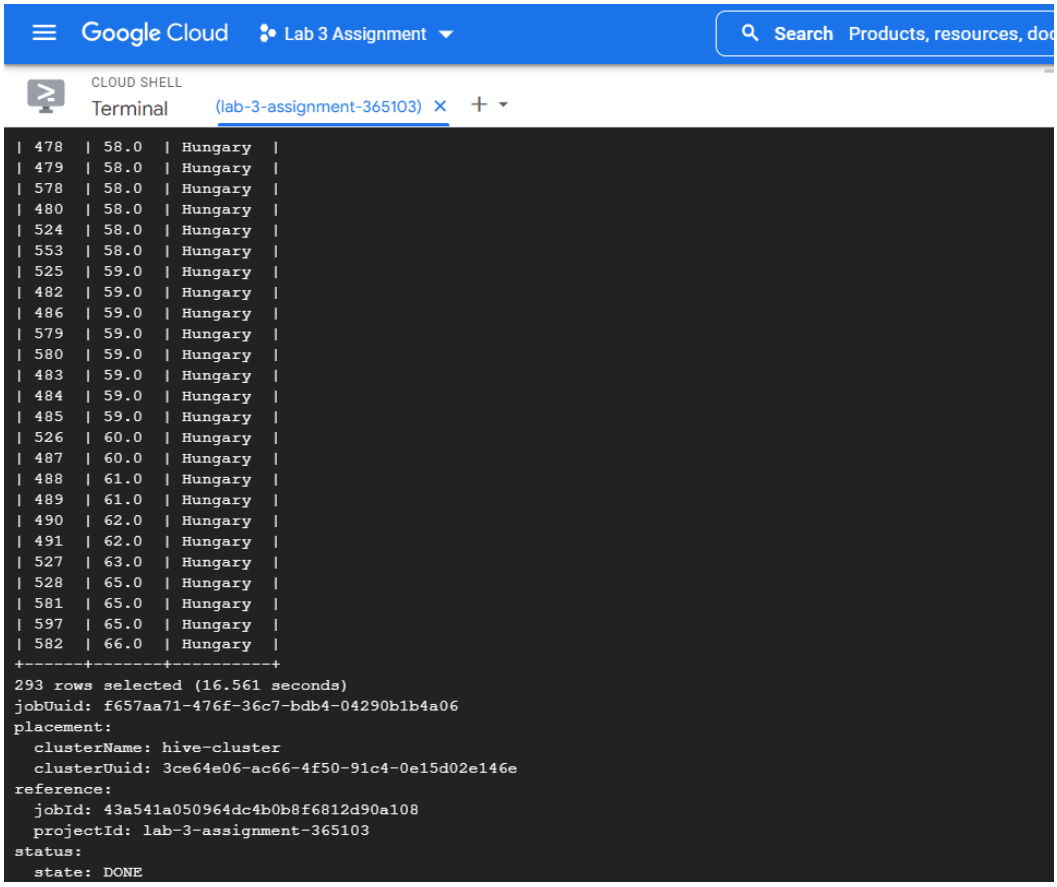
DATA CENTER SCALE COMPUTING – LAB 3

Query Questions:

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

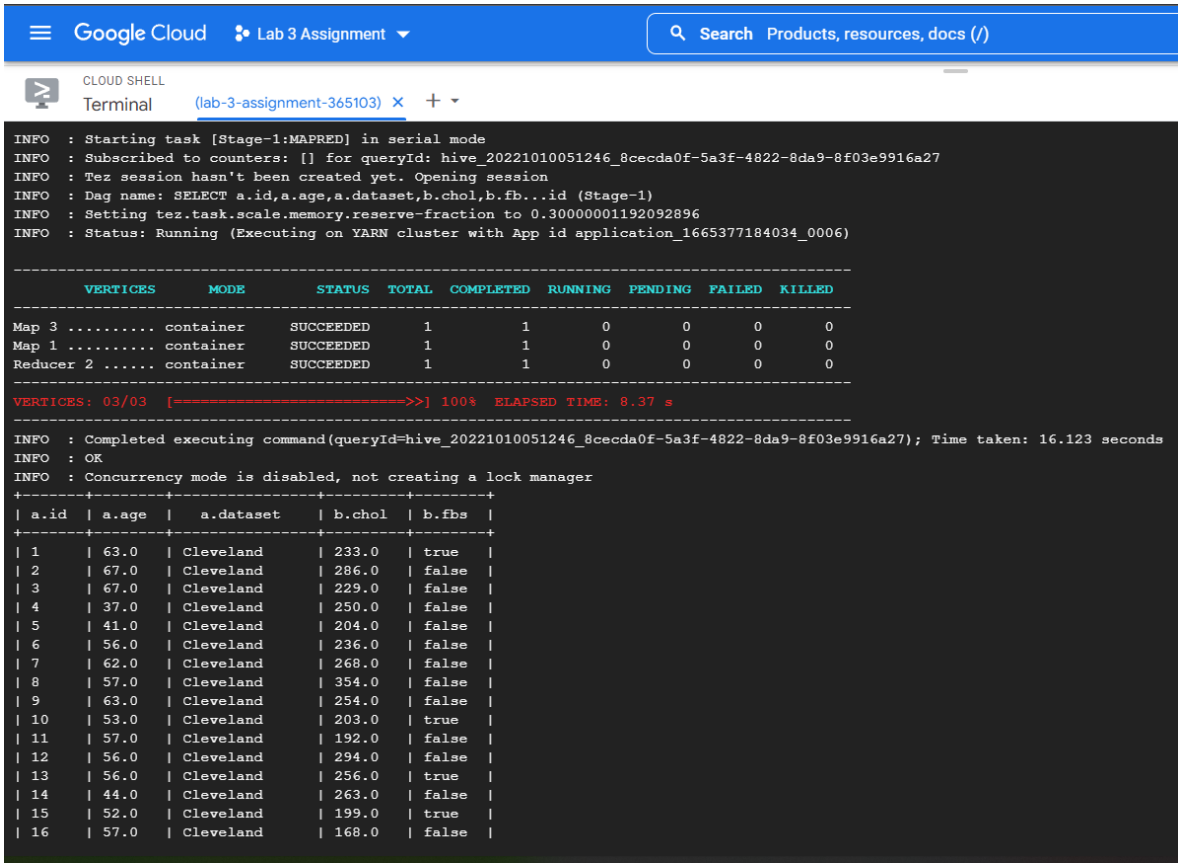
Query Used:
```
SELECT id, age, dataset
FROM personal_details
WHERE dataset='Hungary';
```

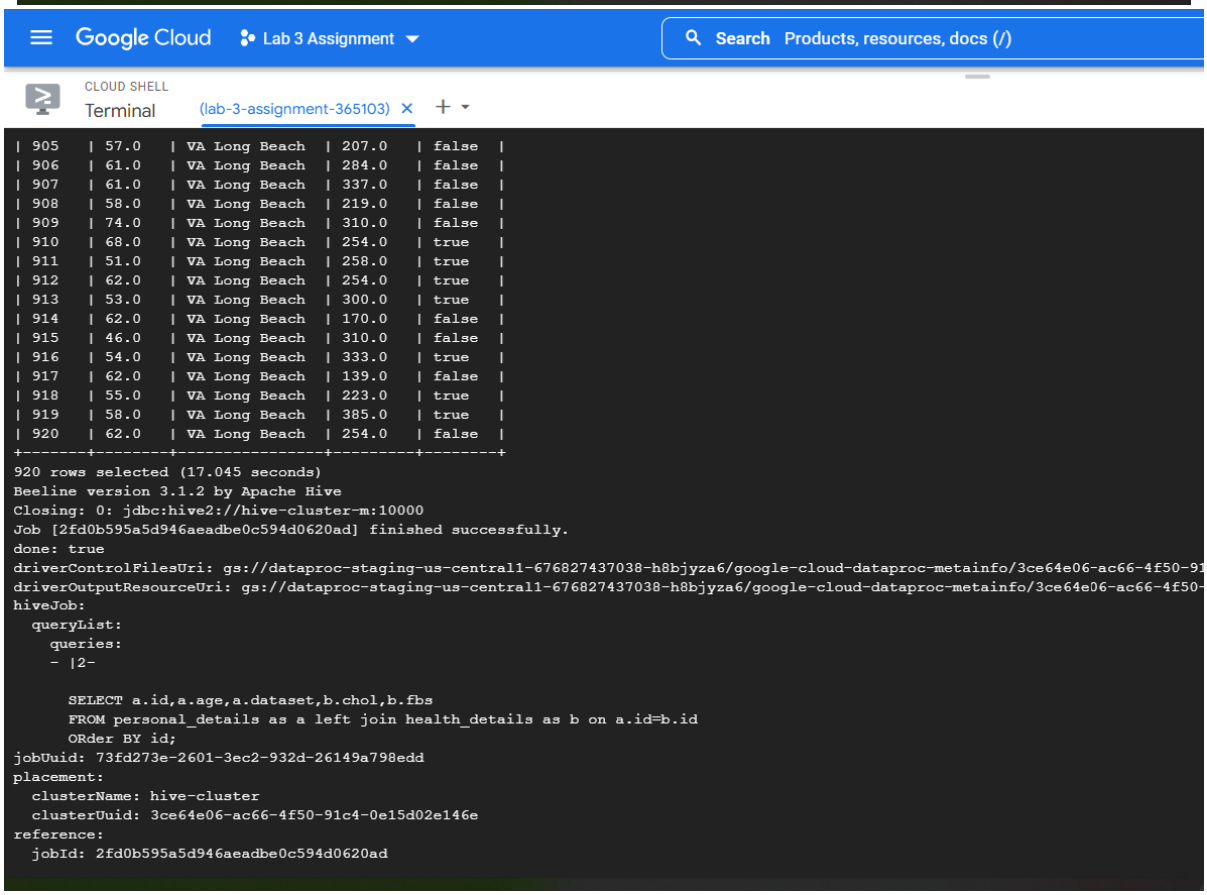




2) Write a Hive query to retrieve id, age, dataset, chol and fbs and sort the values in ascending order of id.

Query Used:
```
SELECT a.id, a.age, a.dataset, b.chol, b.fbs
FROM personal_details as a left join health _details as b on a.id=b.id
ORDER BY id;
```

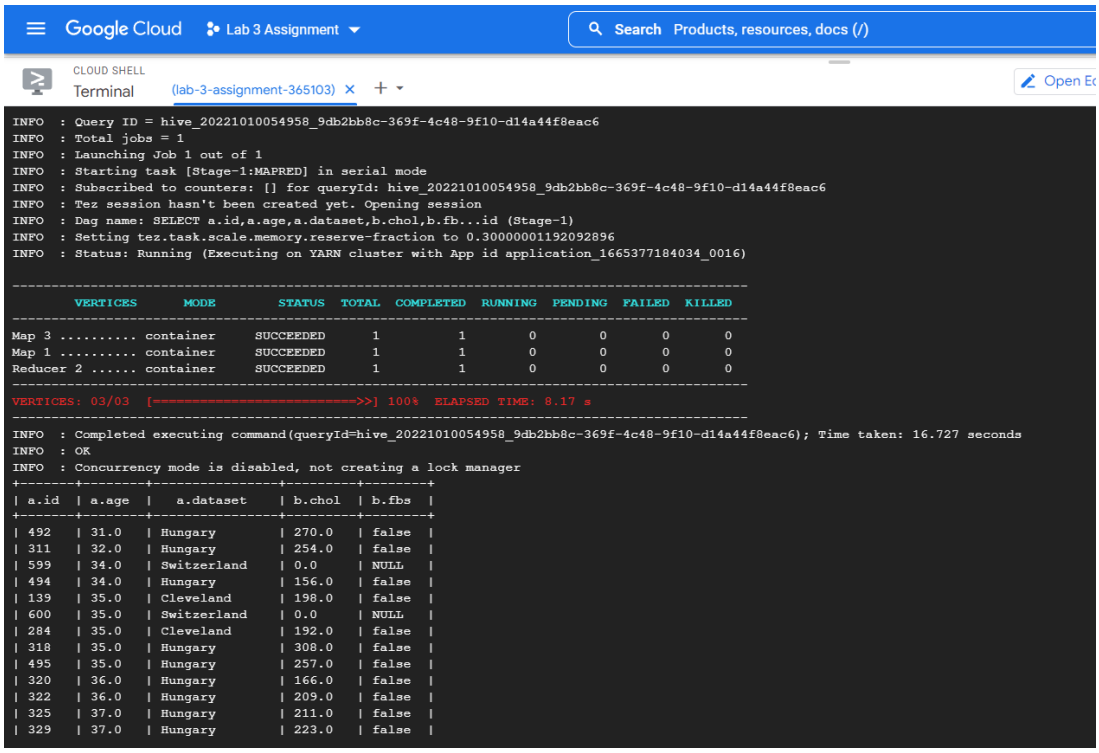


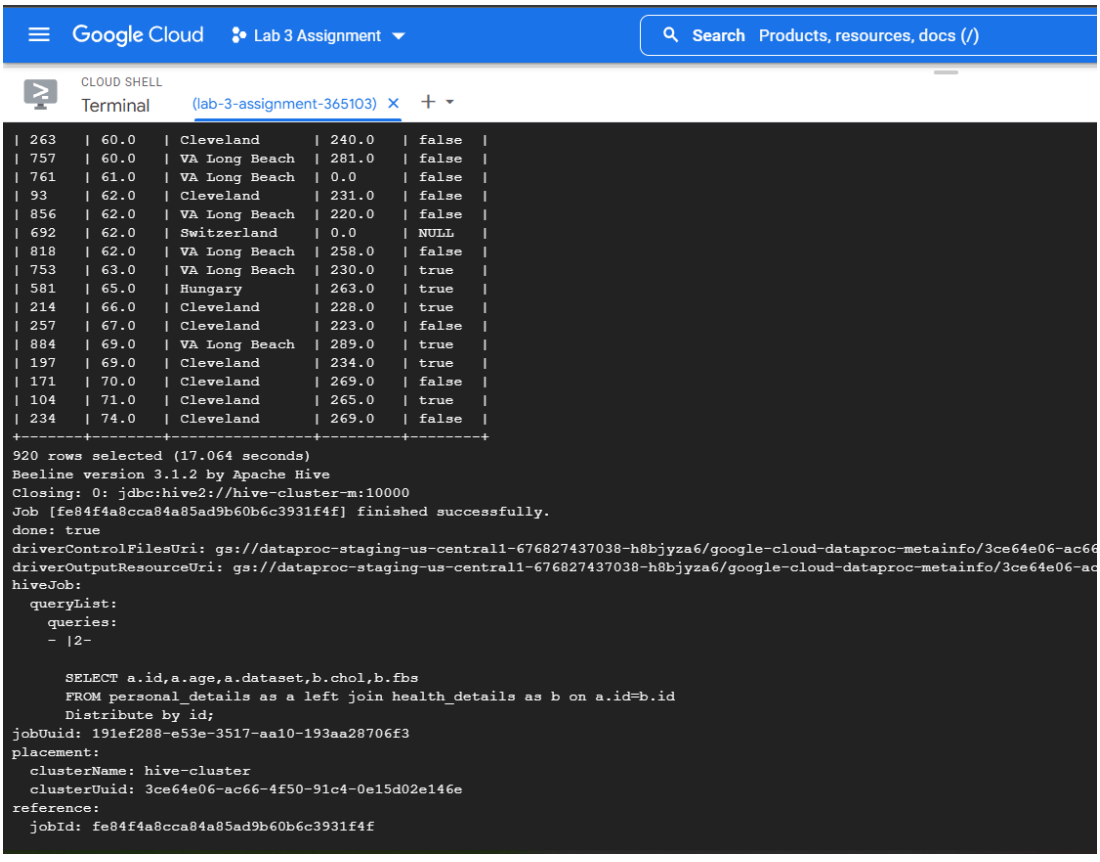


3) Modify the query in Q2 by using “DISTRIBUTE BY” and explain the difference.

Query Used:
```
SELECT a.id, a.age, a.dataset, b.chol, b.fbs
FROM personal_details as a left join health _details as b on a.id=b.id
DISTRIBUTE BY id;
```

The difference between Order by and Distribute By is that the data is sorted globally by the order by clause. And due to this all of the data must come from a single reducer. Whereas in Distribute by distributes the input rows among reducers and does not sort the data either at the reducer level or globally. We see below that the data is not sorted.

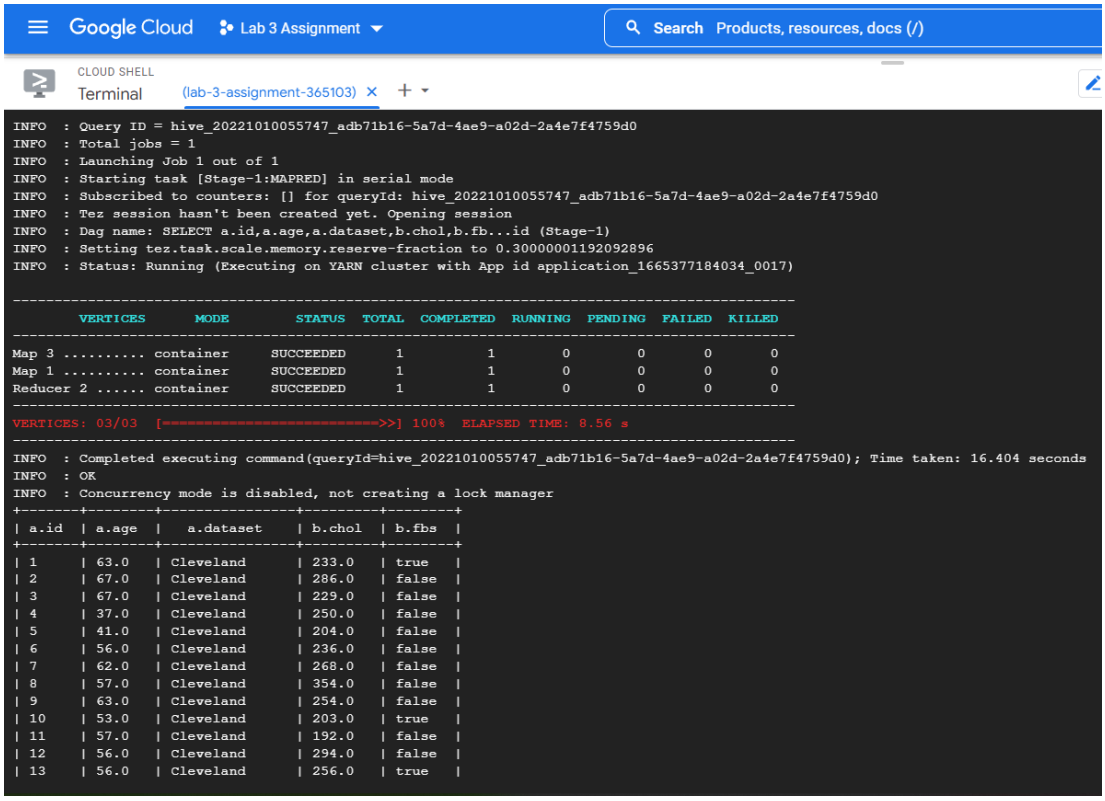


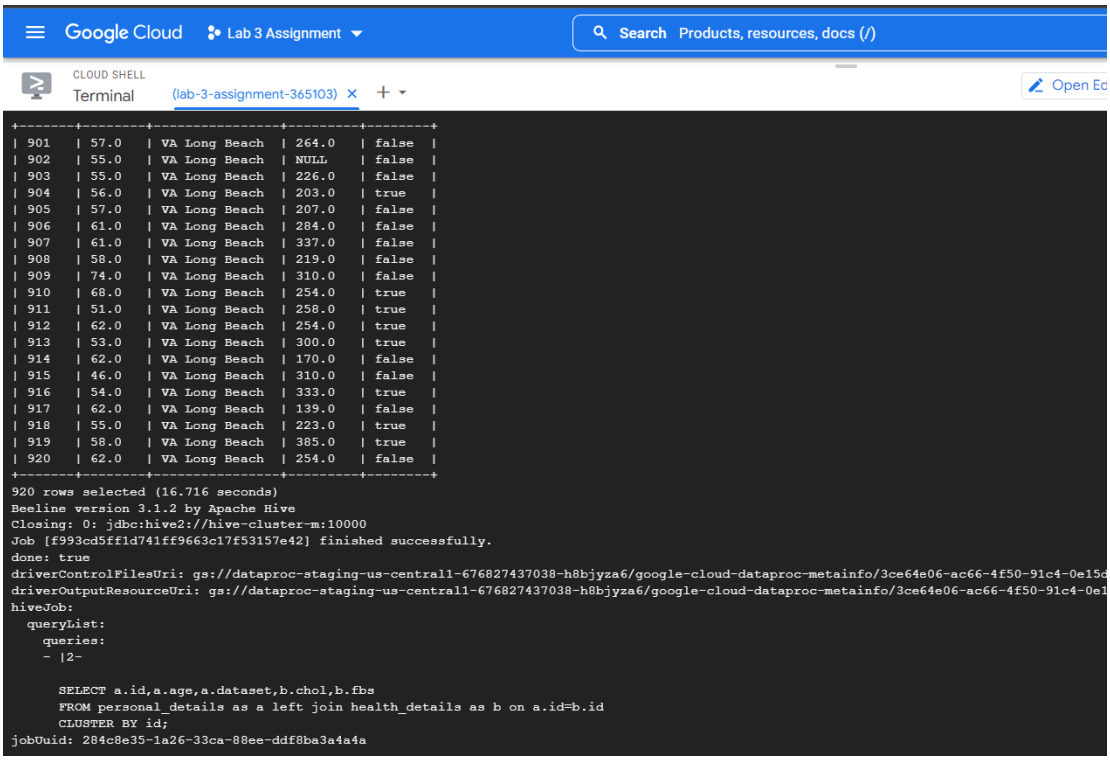


4) Modify the query in Q2 by using “CLUSTER BY” and explain the difference between Q2, Q3 and Q4..

Query Used:
```
SELECT a.id, a.age, a.dataset, b.chol, b.fbs
FROM personal_details as a left join health _details as b on a.id=b.id
CLUSTER BY id;
```

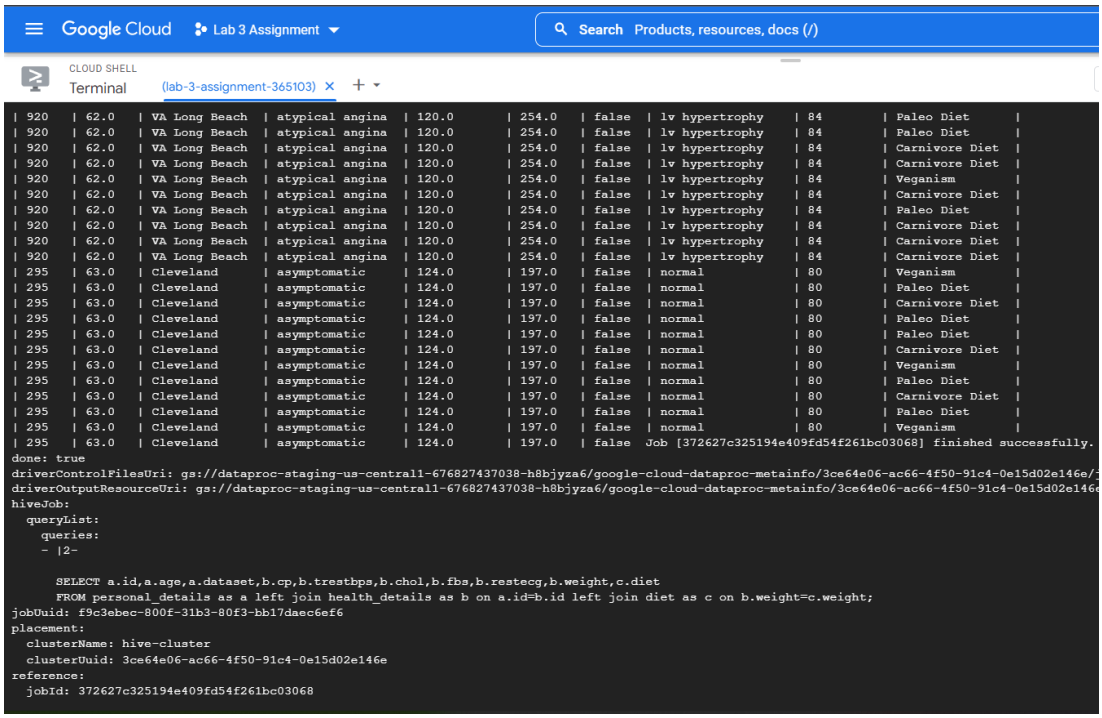
The difference between Q2, Q3 and Q4 is that the data is sorted globally in Q2(Order by). In Q3(Distribute by), we use multiple reducers, but data is not sorted at any level. In Q4(Cluster by), we use multiple reducers, and the data is sorted on the reducer level. If only one reducer is available, then the output is same as Q2.

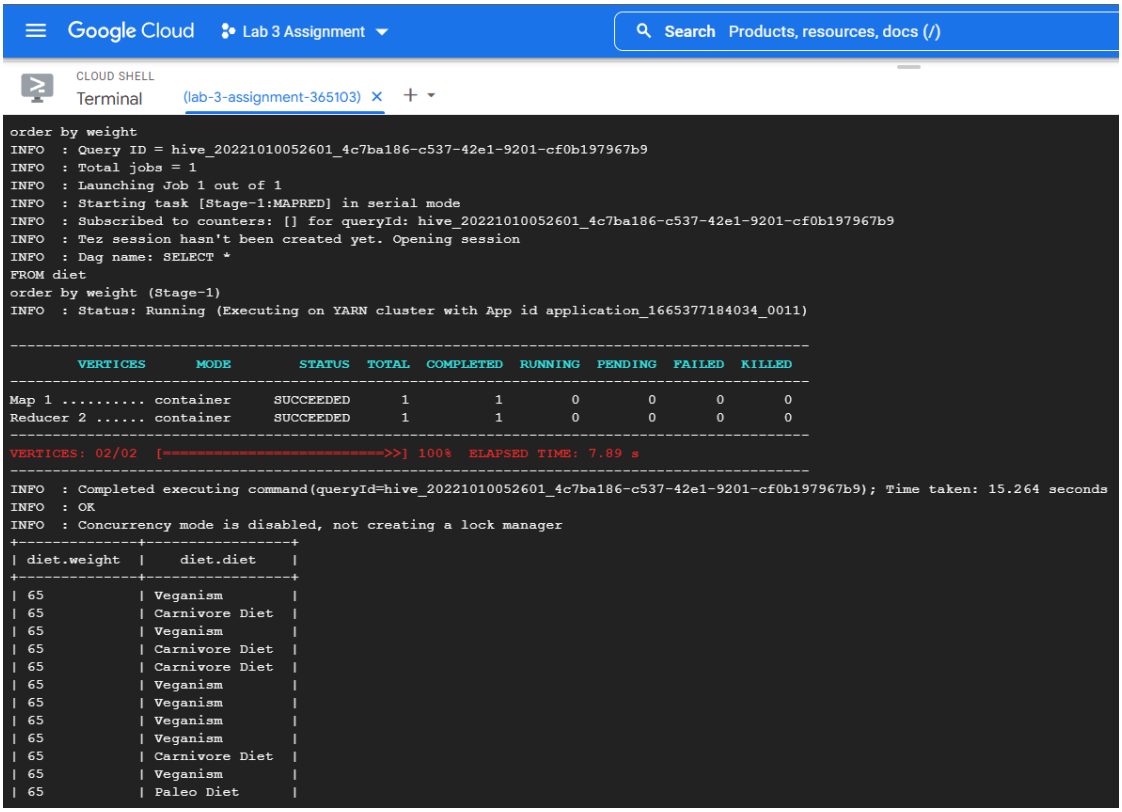




5) Write a query to join tables personal_details, health_details and diet. Observe the results and point out the error/issue if any.

Query Used:
```
SELECT a.id, a.age, a.dataset, b.cp, b.trestbps, b.chol, b.fbs,
b.restecg,  b.weight,c.diet
FROM personal details as a left join health details as b on a.id-b.id
left join diet as c on b.weight=c.weight;
```





Upon checking the output of the query, I see that the for one value of id we have multiple weight. Thus, upon looking at the diet table, I see that for one value of weight we have multiple values of diet, and this is causing an error in the output. Thus, the error is the diet table is not normalized.

Theory Questions:

1) In your own words, describe the working of Hive.

Answer: Large datasets stored in Hadoop files can be accessed and analyzed using the open-source data warehousing technology known as Apache Hive. Hive use language called HiveQL (HQL), which is similar to SQL. HiveQL automatically translates SQL-like queries into MapReduce jobs. Your SQL query is transformed into a series of tasks by the Hive, which typically runs on your computer and sends them to a Hadoop cluster for execution. Data are arranged into tables using Apache Hive. This offers a method for connecting the structure to HDFS data. Hive internally uses a MapReduce framework as a defacto engine for executing the queries.

The Processing framework, Resource Management, Distributed Storage, and Hive clients and services are Apache Hive's main constituents. Through the user interface, the user communicates with the Hive by sending Hive queries. The compiler receives the Hive query from the driver. The execution plan is produced by the compiler. The plan is carried out by the execution engine. Techniques used for querying are Hive Indexing, Vectorization, Bucketing and Partitioning.

2) List out the advantages and disadvantages of HIVE.

Answer:

Advantages of HIVE:

1. Keeps queries operating quickly.
2. Compared to writing MapReduce code, writing a Hive query takes very little time.
3. Like SQL, HiveQL is a declarative language.
4. Gives framework for a wide range of data formats.
5. Supports automation partition.
6. Stores both normalized and denormalized data.
7. Maintains a data warehouse.

Disadvantages of HIVE:

1. Useful when the data is structured.
2. By using MR programming, you may do any analytical procedure.
3. Code debugging is quite challenging.
4. You cannot do difficult operations.
5. Hive is not designed for Online transaction processing.
6. In Hive, subqueries are not supported.