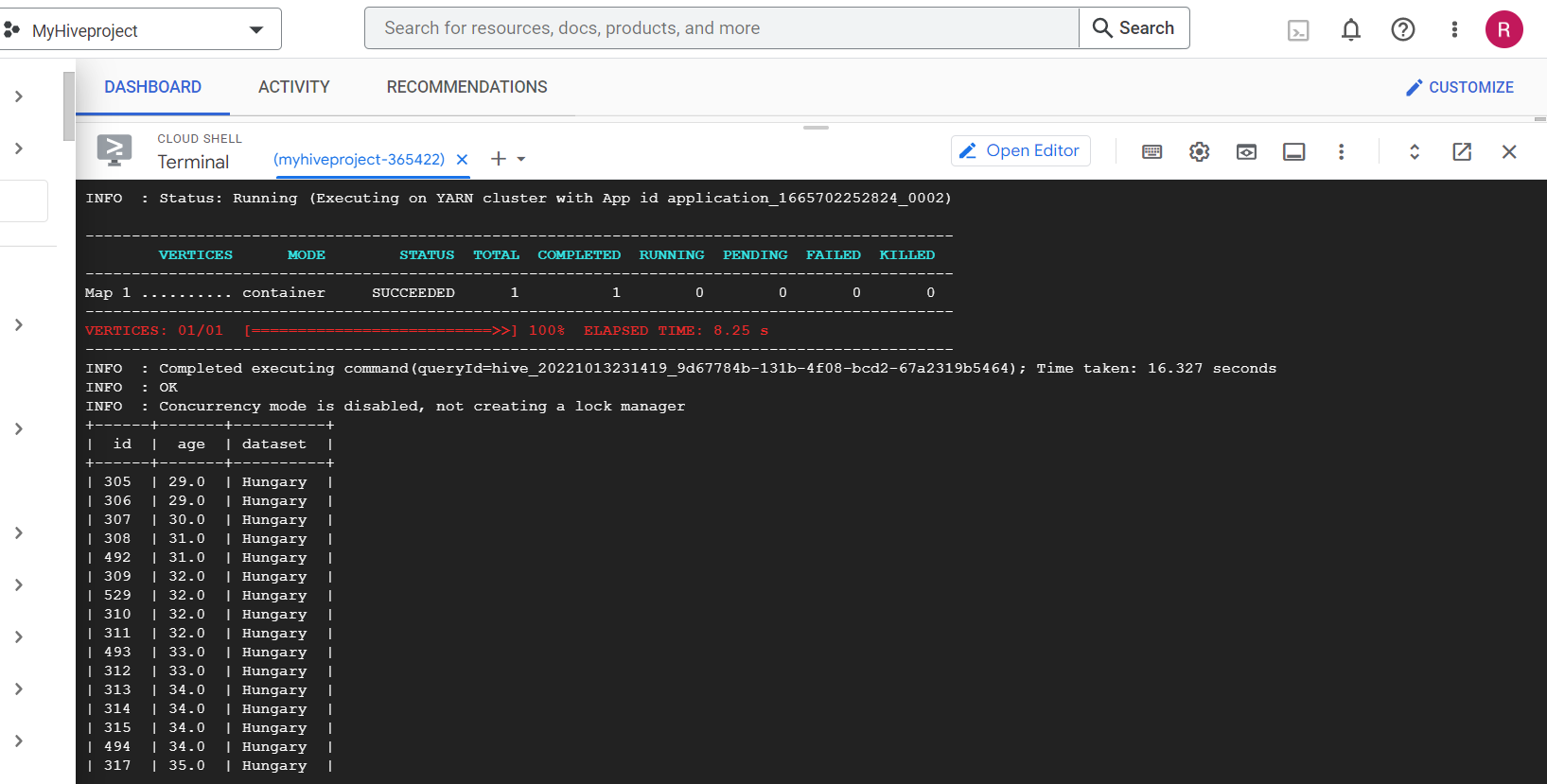
Rucha Dandavate-Lab\_3

Q1.) Write a Hive query to retrieve id, age and dataset where the dataset value is “Hungary”.

SELECT id,age,dataset

FROM personal\_details

WHERE dataset= 'Hungary';"

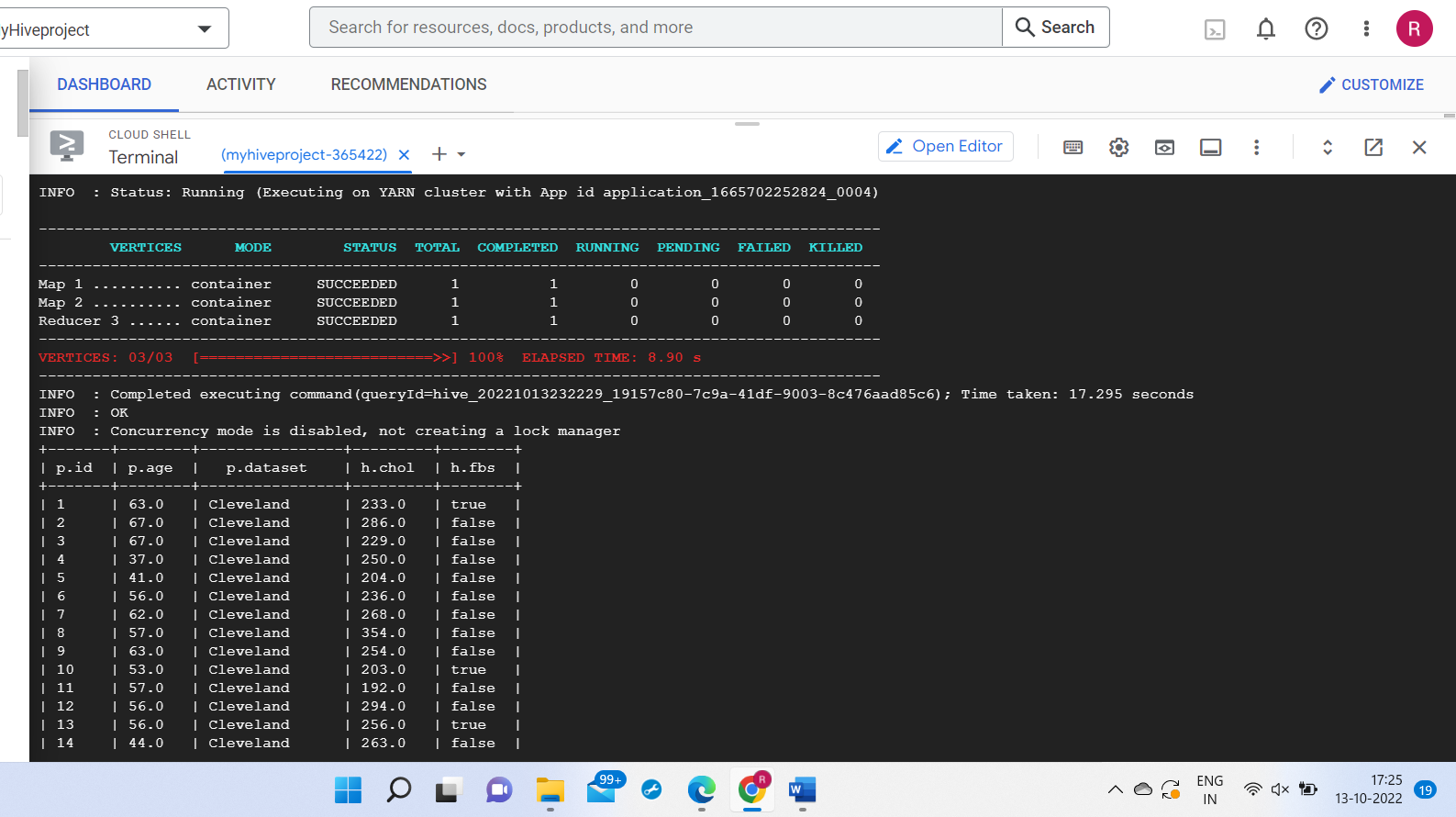


Q2.) Write a Hive query to retrieve id, age, dataset, chol and fbs and sort the values in ascending order of id.

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;"

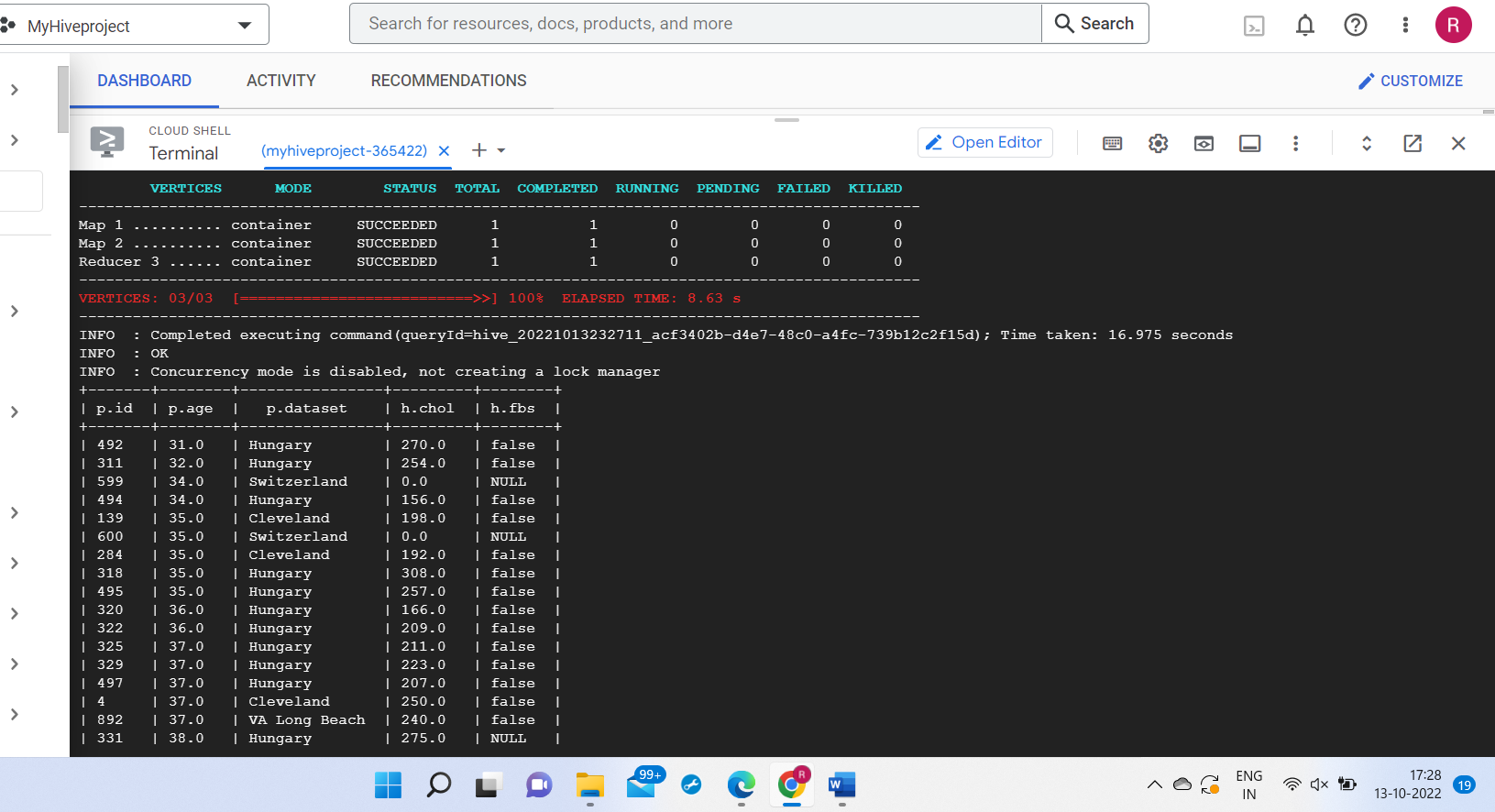


Q3.) Modify the query in Q2 by using “DISTRIBUTE BY” and explain the difference.

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;"



ORDER BY is used to set the data to either ascending order or descending order. Gets a full sorted output across the dataset. In order by the data is stored globally.

DISTRIBUTE BY is used for the same values in a distribute by column go to the same reducer, but doesn’t sort the output of each reducer.

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

ORDER BY is used to set the data to either ascending order or descending order. Gets a full sorted output across the dataset. In order by the data is stored globally.

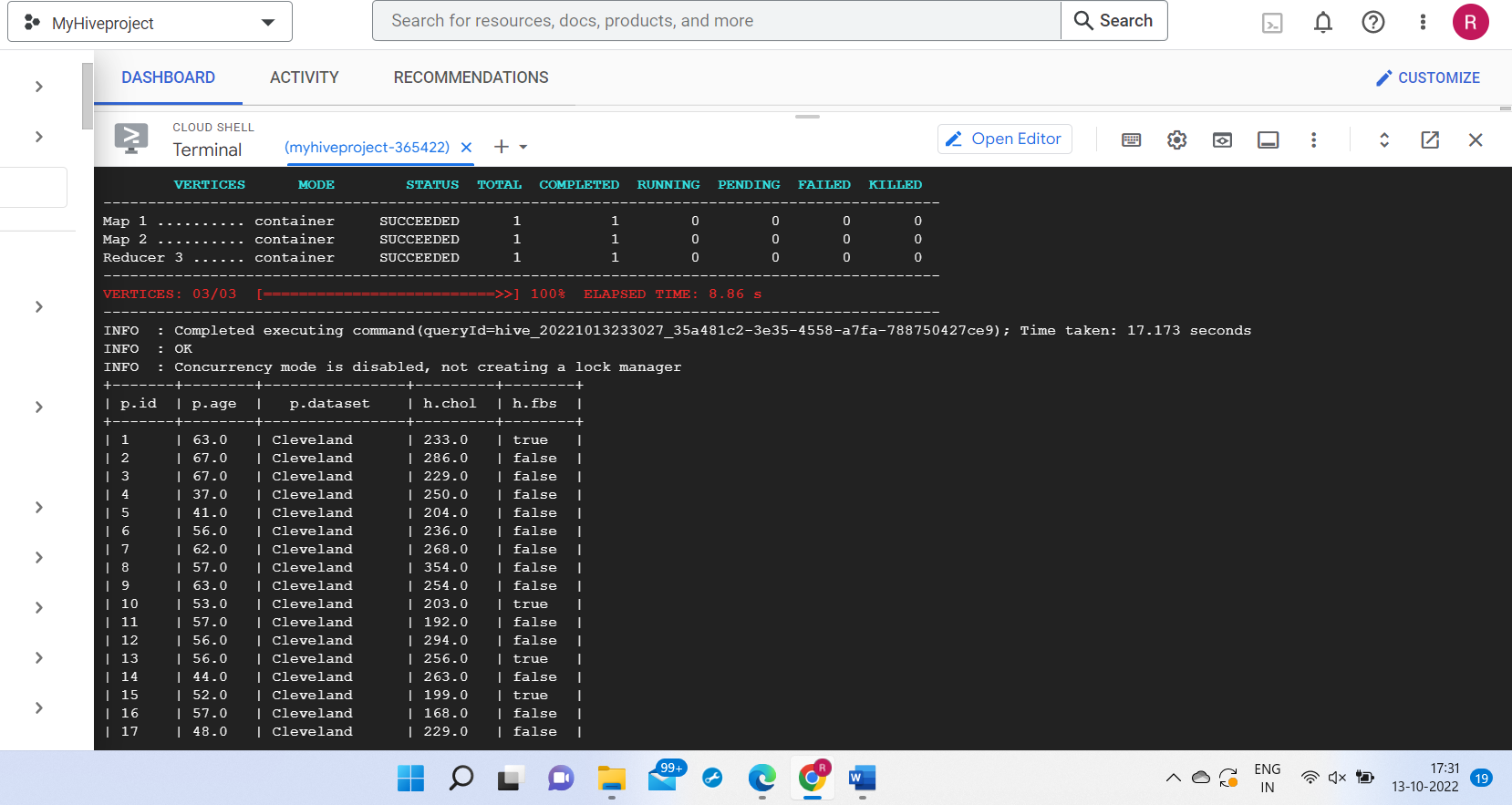
DISTRIBUTE BY is used for the same values in a distribute by column go to the same reducer, but doesn’t sort the output of each reducer. Distribute by clause is used for to repartition the data based on the input expressions. Distribute by unlike the cluster by clause, does not sort the data within each position.

Cluster by repartitions the data based on the input expressions and then sorts the data within each partition. Semantically equivalent to performing a Distribute by followed by sort by.

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;"



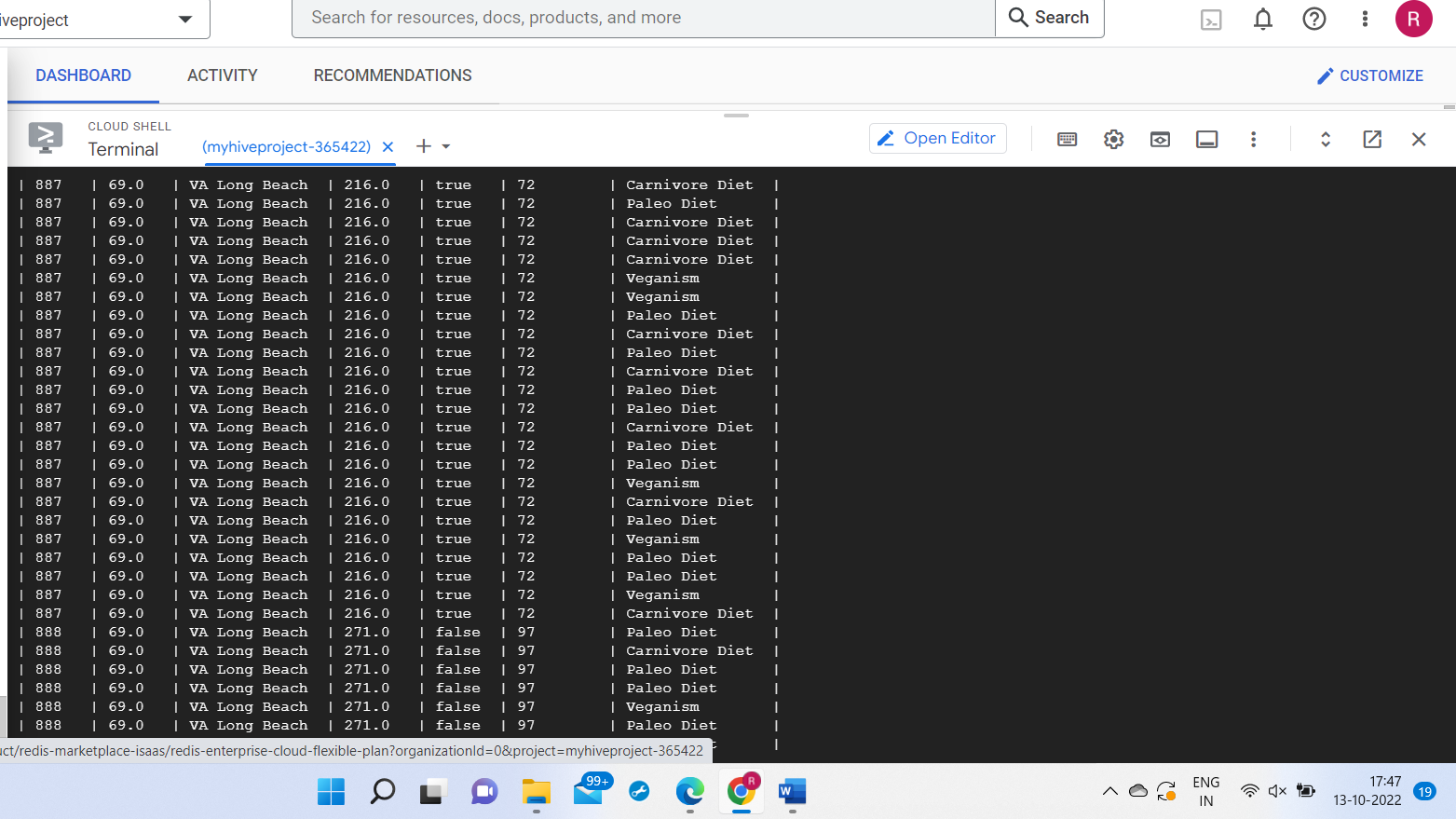
Q5.) Write a query to join tables personal\_details, health\_details and diet. Observe the results and point out the error/issue if any.

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;"

There is only 1 common column between health\_details and diet which is not a primary key. I’m not sure but I think there might be Partial dependency. Proper subset of a key determines a non-prime attribute. The table should have prime attributes(unique values) and eliminate m=non-prime attributes.



**Theory Questions**

Q1.) **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)

The major components of Hive and its interaction with Hadoop are User interface (UI) – It provides an interface between user and hive. Hiver server- Provides requests to hive driver. Driver- Driver receives the queries of user after the interface. Compiler- Execution plan with the help of table in database are generated by compiler. Metastore-Different level of information (structured data) is stored in metastore. Execution engine- It execute plan made by complier.

Working- 1.) execute query (UI) 2.) Get plan (transfer of query to complier) 3.) get Metadata (complier receives the metadata from metastore. 4.) Send metadata. 5.) Send plan (Complier to execute query communicating with driver) 6.) Execute Plan 7.) Send results (the results are sent to driver and to user interface)

Hive queries used such as Partitioning Hive Tables, Bucketing in Hive.

Q2.) **Advantages of Hive: -**

Apache Hive is cost effective option and it provides cheaper options for big data analysis.

Hive can perform data analysis in much faster manner.

Anyone who knows SQL can easily work with Hive.

**Disadvantages of Hive: -**

Hive is challenging for beginners and takes time to adapt with the applications.

Some projects can recur requiring the users to create dependent tasks.

Hive can’t support unstructured data even if it’s using SQL queires.