1

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

--execute "

SELECT id, age, dataset

FROM personal\_details

WHERE dataset='Hungary'

LIMIT 10;"

Text

Description automatically generated

2

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

--execute "

SELECT personal\_details.id, personal\_details.age, personal\_details.dataset, health\_details.chol, health\_details.fbs

FROM personal\_details, health\_details

WHERE personal\_details.id = health\_details.id

Order by id ASC

LIMIT 10;"

Graphical user interface, text

Description automatically generated

3

Distribute by does not sort. It ensures that all rows for the same key columns (id) are going to the same reducer. Aka portioning the data based on id

Difference: Q2 does not distribute but does sort

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

--execute "

SELECT personal\_details.id, personal\_details.age, personal\_details.dataset, health\_details.chol, health\_details.fbs

FROM personal\_details, health\_details

WHERE personal\_details.id = health\_details.id

DISTRIBUTE BY id

LIMIT 10;"

Text

Description automatically generated

4

Cluster by sorts and distributes. Based on the key column (id) it will distribute the data and then sort the output data by putting the same key column values adjacent to each other

Difference: Q2 does not distribute the data and Q3 does not sort

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

--execute "

SELECT personal\_details.id, personal\_details.age, personal\_details.dataset, health\_details.chol, health\_details.fbs

FROM personal\_details, health\_details

WHERE personal\_details.id = health\_details.id

CLUSTER BY id

LIMIT 10;"

A picture containing graphical user interface

Description automatically generated

5

HUGE ERROR! Can not join diet. Diet only has weight and diet. No other tables have these columns so can not match values to join correctly.

Will join health and personal based on id first then health and diet based on weight

gcloud dataproc jobs submit hive \

--cluster hive-cluster \

--region ${REGION} \

--execute "

SELECT \*

FROM health\_details

JOIN personal\_details ON health\_details.id=personal\_details.id

JOIN diet ON health\_details.weight=diet.weight

LIMIT 10;"

A picture containing graphical user interface

Description automatically generated

Theory Questions

1

Hive run on Hadoops framework.

It transforms HiveQL queries into Map Reduce

Queries data stored in a distributed storage solution (the warehouse bucket). This is similar to HDFS

You can quickly write SQL like queries to extract data from Hadoop

2

Advantages of HIVE

Fast-quickly handles petabytes of data using batch processing

Familiar – SQL like interface

Scalable – easy to distribute and scale based on needs

Disadvantages of HIVE

Does not support online transaction processing

Subqueries are not supported

Latency is very high

Not used for real-time data querying