Question

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

```
keyu8117@cloudshell:~ (hive-project-364721)$ gcloud dataproc jobs submit hive \
--cluster hive-cluster \
--region ${REGION} \
--execute "
SELECT id, age,dataset
FROM personal_details
WHERE dataset = 'Hungary';"
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 8.86 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20221006221900_a864a175-b210-4459-8a88-558b2823f07f); Time taken: 16.068 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+-------+----------+
|  id  |  age  | dataset  |
+------+-------+----------+
| 305  | 29.0  | Hungary  |
| 306  | 29.0  | Hungary  |
| 307  | 30.0  | Hungary  |
| 308  | 31.0  | Hungary  |
| 492  | 31.0  | Hungary  |
| 309  | 32.0  | Hungary  |
| 529  | 32.0  | Hungary  |
| 310  | 32.0  | Hungary  |
| 311  | 32.0  | Hungary  |
| 493  | 33.0  | Hungary  |
| 312  | 33.0  | Hungary  |
| 313  | 34.0  | Hungary  |
| 314  | 34.0  | Hungary  |
| 315  | 34.0  | Hungary  |
| 494  | 34.0  | Hungary  |
| 317  | 35.0  | Hungary  |
| 316  | 35.0  | Hungary  |
| 318  | 35.0  | Hungary  |
| 319  | 35.0  | Hungary  |
```

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

```
keyu8117@cloudshell:~ (hive-project-364721)$ gcloud dataproc jobs submit hive \
--cluster hive-cluster \
--region ${REGION} \
--execute "
SELECT personal_details.id, personal_details.age, personal_details.dataset, health_deatils.chol, health_deatils.fbs
FROM personal_details
JOIN health_deatils
> ON personal_details.id = health_deatils.id
> ORDER BY personal_details.id;"
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.20 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20221006223706_77a07b5f-b59d-4b17-bdb7-71e03cb52283); Time taken: 17.406 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------------+----------------------+--------------------------+---------------------+---------------------+
| personal_details.id | personal_details.age | personal_details.dataset | health_deatils.chol | health_deatils.fbs  |
+---------------------+----------------------+--------------------------+---------------------+---------------------+
| 1                   | 63.0                 | Cleveland                | 233.0               | true                |
| 2                   | 67.0                 | Cleveland                | 286.0               | false               |
| 3                   | 67.0                 | Cleveland                | 229.0               | false               |
| 4                   | 37.0                 | Cleveland                | 250.0               | false               |
| 5                   | 41.0                 | Cleveland                | 204.0               | false               |
| 6                   | 56.0                 | Cleveland                | 236.0               | false               |
| 7                   | 62.0                 | Cleveland                | 268.0               | false               |
| 8                   | 57.0                 | Cleveland                | 354.0               | false               |
| 9                   | 63.0                 | Cleveland                | 254.0               | false               |
| 10                  | 53.0                 | Cleveland                | 203.0               | true                |
| 11                  | 57.0                 | Cleveland                | 192.0               | false               |
| 12                  | 56.0                 | Cleveland                | 294.0               | false               |
| 13                  | 56.0                 | Cleveland                | 256.0               | true                |
| 14                  | 44.0                 | Cleveland                | 263.0               | false               |
| 15                  | 52.0                 | Cleveland                | 199.0               | true                |
| 16                  | 57.0                 | Cleveland                | 168.0               | false               |
| 17                  | 48.0                 | Cleveland                | 229.0               | false               |
```

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

```
keyu8117@cloudshell:~ (hive-project-364721)$ gcloud dataproc jobs submit hive \
--cluster hive-cluster \
--region ${REGION} \
--execute "
SELECT personal_details.id, personal_details.age, personal_details.dataset, health_deatils.chol, health_deatils.fbs
FROM personal_details
JOIN health_deatils
ON personal_details.id = health_deatils.id
DISTRIBUTE BY personal_details.id;"
```

```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.23 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20221006224455_8b70b8fe-8472-4570-bf5b-c9ccbed91ec3); Time taken: 16.839 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------+-----------------------+--------------------------+---------------------+---------------------+
| personal_details.id  | personal_details.age  | personal_details.dataset | health_deatils.chol | health_deatils.fbs  |
+----------------------+-----------------------+--------------------------+---------------------+---------------------+
| 492                  | 31.0                  | Hungary                  | 270.0               | false               |
| 311                  | 32.0                  | Hungary                  | 254.0               | false               |
| 599                  | 34.0                  | Switzerland              | 0.0                 | NULL                |
| 494                  | 34.0                  | Hungary                  | 156.0               | false               |
| 139                  | 35.0                  | Cleveland                | 198.0               | false               |
| 600                  | 35.0                  | Switzerland              | 0.0                 | NULL                |
| 284                  | 35.0                  | Cleveland                | 192.0               | false               |
| 318                  | 35.0                  | Hungary                  | 308.0               | false               |
| 495                  | 35.0                  | Hungary                  | 257.0               | false               |
| 320                  | 36.0                  | Hungary                  | 166.0               | false               |
| 322                  | 36.0                  | Hungary                  | 209.0               | false               |
| 325                  | 37.0                  | Hungary                  | 211.0               | false               |
| 329                  | 37.0                  | Hungary                  | 223.0               | false               |
| 497                  | 37.0                  | Hungary                  | 207.0               | false               |
| 4                    | 37.0                  | Cleveland                | 250.0               | false               |
| 892                  | 37.0                  | VA Long Beach            | 240.0               | false               |
| 331                  | 38.0                  | Hungary                  | 275.0               | NULL                |
| 607                  | 38.0                  | Switzerland              | 0.0                 | NULL                |
| 873                  | 38.0                  | VA Long Beach            | 289.0               | false               |
```

When we use "ORDER BY", the result of this sequence is because of the value of id.

When we use "DISTRUBUTE BY", the result of this sequence is becuase Hive uses the columns in "DISTRUBUTE BY" to distribute the rows among reducers, all "DISTRUBUTE BY" columns will go to the same reducer.

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

```
keyu8117@cloudshell:~ (hive-project-364721)$ gcloud dataproc jobs submit hive \
--cluster hive-cluster \
--region ${REGION} \
--execute "
SELECT personal_details.id, personal_details.age, personal_details.dataset, health_deatils.chol, health_deatils.fbs
FROM personal_details
JOIN health_deatils
ON personal_details.id = health_deatils.id
CLUSTER BY personal_details.id;"
```

```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.47 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20221006224705_b2e0cc5e-1ac2-4435-924c-815c767f7c60); Time taken: 16.932 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------+-----------------------+--------------------------+---------------------+---------------------+
| personal_details.id  | personal_details.age  | personal_details.dataset | health_deatils.chol | health_deatils.fbs  |
+----------------------+-----------------------+--------------------------+---------------------+---------------------+
| 1                    | 63.0                  | Cleveland                | 233.0               | true                |
| 2                    | 67.0                  | Cleveland                | 286.0               | false               |
| 3                    | 67.0                  | Cleveland                | 229.0               | false               |
| 4                    | 37.0                  | Cleveland                | 250.0               | false               |
| 5                    | 41.0                  | Cleveland                | 204.0               | false               |
| 6                    | 56.0                  | Cleveland                | 236.0               | false               |
| 7                    | 62.0                  | Cleveland                | 268.0               | false               |
| 8                    | 57.0                  | Cleveland                | 354.0               | false               |
| 9                    | 63.0                  | Cleveland                | 254.0               | false               |
| 10                   | 53.0                  | Cleveland                | 203.0               | true                |
| 11                   | 57.0                  | Cleveland                | 192.0               | false               |
| 12                   | 56.0                  | Cleveland                | 294.0               | false               |
| 13                   | 56.0                  | Cleveland                | 256.0               | true                |
| 14                   | 44.0                  | Cleveland                | 263.0               | false               |
| 15                   | 52.0                  | Cleveland                | 199.0               | true                |
| 16                   | 57.0                  | Cleveland                | 168.0               | false               |
| 17                   | 48.0                  | Cleveland                | 229.0               | false               |
| 18                   | 54.0                  | Cleveland                | 239.0               | false               |
| 19                   | 48.0                  | Cleveland                | 275.0               | false               |
```

When we use "CLUSTER BY", the result of this sequence is because Hive uses the columns in "CLUSTER BY" to distribute the rows among reducers, "CLUSTER BY" columns will go to the multiple reducers.

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

We cannot do this, there are only two columns in diet, they are weight and diet, neither of them is unique. There are many same value of weight, so we cannot match the weight of diet and health_deatils, which means we cannot do this join operation.

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)

Firstly, Hive client submits a query to a Hive server that runs in an ephemeral cluster. Then the server processes the query and requests metadata from the metastore service. The server loads data from the Hive warehouse located in HDFS.

Hive runs its query using HiveQL. Hive's query first get converted into Map Reduce than processed by Hadoop to query the data.

2. List out the advantages and disadvantages of HIVE

- advantages:
  - HiveQL is a language that is similar to SQL
  - It is a comparatively cheaper option.
  - Hive is a productive software.
  - Fault Tolerance Software.
- disadvantages:
  - Latency of Hive is generally very high.
  - Subqueries are not supported.