

# Naveenkumar

## Data Scale Computing Lab3: -

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_detail  
WHERE dataset = 'Hungary';"
```

```
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0  
VERTICES: 01/01 [=====] 100% ELAPSED TIME: 8.87 s  
INFO : Completed executing command(queryId=hive_20221013184550_5bfd7e84-9486-4b3c-b1cc-d71697cd87a8); Time taken: 17.354 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 |  
| 495 | 35.0 | Hungary |  
| 320 | 36.0 | Hungary |  
| 496 | 36.0 | Hungary |  
| 321 | 36.0 | Hungary |  
| 322 | 36.0 | Hungary |  
| 323 | 36.0 | Hungary |  
| 326 | 37.0 | Hungary |  
| 324 | 37.0 | 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_detail p JOIN health_details h  
ON (p.id = h.id) ORDER BY p.id;"
```

```

INFO : Task session hasn't been created yet. Opening session
INFO : Map name: SELECT p.id,p.age,p.dataset,h.chol,h.fbs FROM personal_detail p JOIN health_details h ON (p.id = h.id) (Stage-1)
INFO : Setting task.task.scale.memory.reserve-fraction to 0.10000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1665637537338_0009)

-----
VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 3 ..... container SUCCEEDED 1 1 0 0 0 0
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICES 03/03 [=====] 100% ELAPSED TIME: 10.09 s
INFO : Completed executing command(queryId=hive_20221013184741_dc58ba42-9419-4e51-b2ee-810daa9f0ea0); Time taken: 18.332 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

p.id      p.age      p.dataset      h.chol      h.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 36.0 Cleveland 236.0 false
7 62.0 Cleveland 268.0 false
8 57.0 Cleveland 354.0 false
9 62.0 Cleveland 254.0 false
10 53.0 Cleveland 203.0 true
11 57.0 Cleveland 192.0 false
12 56.0 Cleveland 284.0 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 true
18 54.0 Cleveland 139.0 false
19 48.0 Cleveland 275.0 false
20 49.0 Cleveland 166.0 false
21 64.0 Cleveland 211.0 false
22 58.0 Cleveland 224.0 true
23 58.0 Cleveland 284.0 false
24 58.0 Cleveland 224.0 false
25 60.0 Cleveland 206.0 false
26 50.0 Cleveland 219.0 false
27 58.0 Cleveland 226.0 false
28 66.0 Cleveland 226.0 false

```

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\_detail p JOIN health\_details h
ON (p.id = h.id) DISTRIBUTE BY p.id;"

```

INFO : Status: Running (Executing on YARN cluster with App id application_1665637537338_0010)

-----
VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 3 ..... container SUCCEEDED 1 1 0 0 0 0
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICES 03/03 [=====] 100% ELAPSED TIME: 8.88 s
INFO : Completed executing command(queryId=hive_20221013184848_0fcd273e-c8d0-4558-aa43-fec0ebf5aa0f); Time taken: 16.908 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

p.id      p.age      p.dataset      h.chol      h.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 36.0 Cleveland 236.0 false
7 62.0 Cleveland 268.0 false
8 57.0 Cleveland 354.0 false
9 62.0 Cleveland 254.0 false
10 53.0 Cleveland 203.0 true
11 57.0 Cleveland 192.0 false
12 56.0 Cleveland 284.0 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 true
18 54.0 Cleveland 139.0 false
19 48.0 Cleveland 275.0 false
20 49.0 Cleveland 166.0 false
21 64.0 Cleveland 211.0 false
22 58.0 Cleveland 224.0 true
23 58.0 Cleveland 284.0 false
24 58.0 Cleveland 224.0 false
25 60.0 Cleveland 206.0 false
26 50.0 Cleveland 219.0 false
27 58.0 Cleveland 226.0 false
28 66.0 Cleveland 226.0 false
29 63.0 Cleveland 233.0 true
30 67.0 Cleveland 286.0 false
31 67.0 Cleveland 229.0 false
32 37.0 Cleveland 250.0 false
33 41.0 Cleveland 204.0 false
34 36.0 Cleveland 236.0 false
35 62.0 Cleveland 268.0 false
36 57.0 Cleveland 354.0 false
37 62.0 Cleveland 254.0 false
38 53.0 Cleveland 203.0 true
39 57.0 Cleveland 192.0 false
40 56.0 Cleveland 284.0 false
41 56.0 Cleveland 256.0 true
42 44.0 Cleveland 263.0 false
43 52.0 Cleveland 199.0 true
44 57.0 Cleveland 168.0 false
45 48.0 Cleveland 229.0 true
46 54.0 Cleveland 139.0 false
47 48.0 Cleveland 275.0 false
48 49.0 Cleveland 166.0 false
49 64.0 Cleveland 211.0 false
50 58.0 Cleveland 224.0 true
51 58.0 Cleveland 284.0 false
52 58.0 Cleveland 224.0 false
53 60.0 Cleveland 206.0 false
54 50.0 Cleveland 219.0 false
55 58.0 Cleveland 226.0 false
56 66.0 Cleveland 226.0 false
57 63.0 Cleveland 233.0 true
58 67.0 Cleveland 286.0 false
59 67.0 Cleveland 229.0 false
60 37.0 Cleveland 250.0 false
61 41.0 Cleveland 204.0 false
62 36.0 Cleveland 236.0 false
63 62.0 Cleveland 268.0 false
64 57.0 Cleveland 354.0 false
65 62.0 Cleveland 254.0 false
66 53.0 Cleveland 203.0 true
67 57.0 Cleveland 192.0 false
68 56.0 Cleveland 284.0 false
69 56.0 Cleveland 256.0 true
70 44.0 Cleveland 263.0 false
71 52.0 Cleveland 199.0 true
72 57.0 Cleveland 168.0 false
73 48.0 Cleveland 229.0 true
74 54.0 Cleveland 139.0 false
75 48.0 Cleveland 275.0 false
76 49.0 Cleveland 166.0 false
77 64.0 Cleveland 211.0 false
78 58.0 Cleveland 224.0 true
79 58.0 Cleveland 284.0 false
80 58.0 Cleveland 224.0 false
81 60.0 Cleveland 206.0 false
82 50.0 Cleveland 219.0 false
83 58.0 Cleveland 226.0 false
84 66.0 Cleveland 226.0 false
85 63.0 Cleveland 233.0 true
86 67.0 Cleveland 286.0 false
87 67.0 Cleveland 229.0 false
88 37.0 Cleveland 250.0 false
89 41.0 Cleveland 204.0 false
90 36.0 Cleveland 236.0 false
91 62.0 Cleveland 268.0 false
92 57.0 Cleveland 354.0 false
93 62.0 Cleveland 254.0 false
94 53.0 Cleveland 203.0 true
95 57.0 Cleveland 192.0 false
96 56.0 Cleveland 284.0 false
97 56.0 Cleveland 256.0 true
98 44.0 Cleveland 263.0 false
99 52.0 Cleveland 199.0 true
100 57.0 Cleveland 168.0 false
101 48.0 Cleveland 229.0 true
102 54.0 Cleveland 139.0 false
103 48.0 Cleveland 275.0 false
104 49.0 Cleveland 166.0 false
105 64.0 Cleveland 211.0 false
106 58.0 Cleveland 224.0 true
107 58.0 Cleveland 284.0 false
108 58.0 Cleveland 224.0 false
109 60.0 Cleveland 206.0 false
110 50.0 Cleveland 219.0 false
111 58.0 Cleveland 226.0 false
112 66.0 Cleveland 226.0 false
113 63.0 Cleveland 233.0 true
114 67.0 Cleveland 286.0 false
115 67.0 Cleveland 229.0 false
116 37.0 Cleveland 250.0 false
117 41.0 Cleveland 204.0 false
118 36.0 Cleveland 236.0 false
119 62.0 Cleveland 268.0 false
120 57.0 Cleveland 354.0 false
121 62.0 Cleveland 254.0 false
122 53.0 Cleveland 203.0 true
123 57.0 Cleveland 192.0 false
124 56.0 Cleveland 284.0 false
125 56.0 Cleveland 256.0 true
126 44.0 Cleveland 263.0 false
127 52.0 Cleveland 199.0 true
128 57.0 Cleveland 168.0 false
129 48.0 Cleveland 229.0 true
130 54.0 Cleveland 139.0 false
131 48.0 Cleveland 275.0 false
132 49.0 Cleveland 166.0 false
133 64.0 Cleveland 211.0 false
134 58.0 Cleveland 224.0 true
135 58.0 Cleveland 284.0 false
136 58.0 Cleveland 224.0 false
137 60.0 Cleveland 206.0 false
138 50.0 Cleveland 219.0 false
139 58.0 Cleveland 226.0 false
140 66.0 Cleveland 226.0 false
141 63.0 Cleveland 233.0 true
142 67.0 Cleveland 286.0 false
143 67.0 Cleveland 229.0 false
144 37.0 Cleveland 250.0 false
145 41.0 Cleveland 204.0 false
146 36.0 Cleveland 236.0 false
147 62.0 Cleveland 268.0 false
148 57.0 Cleveland 354.0 false
149 62.0 Cleveland 254.0 false
150 53.0 Cleveland 203.0 true
151 57.0 Cleveland 192.0 false
152 56.0 Cleveland 284.0 false
153 56.0 Cleveland 256.0 true
154 44.0 Cleveland 263.0 false
155 52.0 Cleveland 199.0 true
156 57.0 Cleveland 168.0 false
157 48.0 Cleveland 229.0 true
158 54.0 Cleveland 139.0 false
159 48.0 Cleveland 275.0 false
160 49.0 Cleveland 166.0 false
161 64.0 Cleveland 211.0 false
162 58.0 Cleveland 224.0 true
163 58.0 Cleveland 284.0 false
164 58.0 Cleveland 224.0 false
165 60.0 Cleveland 206.0 false
166 50.0 Cleveland 219.0 false
167 58.0 Cleveland 226.0 false
168 66.0 Cleveland 226.0 false
169 63.0 Cleveland 233.0 true
170 67.0 Cleveland 286.0 false
171 67.0 Cleveland 229.0 false
172 37.0 Cleveland 250.0 false
173 41.0 Cleveland 204.0 false
174 36.0 Cleveland 236.0 false
175 62.0 Cleveland 268.0 false
176 57.0 Cleveland 354.0 false
177 62.0 Cleveland 254.0 false
178 53.0 Cleveland 203.0 true
179 57.0 Cleveland 192.0 false
180 56.0 Cleveland 284.0 false
181 56.0 Cleveland 256.0 true
182 44.0 Cleveland 263.0 false
183 52.0 Cleveland 199.0 true
184 57.0 Cleveland 168.0 false
185 48.0 Cleveland 229.0 true
186 54.0 Cleveland 139.0 false
187 48.0 Cleveland 275.0 false
188 49.0 Cleveland 166.0 false
189 64.0 Cleveland 211.0 false
190 58.0 Cleveland 224.0 true
191 58.0 Cleveland 284.0 false
192 58.0 Cleveland 224.0 false
193 60.0 Cleveland 206.0 false
194 50.0 Cleveland 219.0 false
195 58.0 Cleveland 226.0 false
196 66.0 Cleveland 226.0 false
197 63.0 Cleveland 233.0 true
198 67.0 Cleveland 286.0 false
199 67.0 Cleveland 229.0 false
200 37.0 Cleveland 250.0 false
201 41.0 Cleveland 204.0 false
202 36.0 Cleveland 236.0 false
203 62.0 Cleveland 268.0 false
204 57.0 Cleveland 354.0 false
205 62.0 Cleveland 254.0 false
206 53.0 Cleveland 203.0 true
207 57.0 Cleveland 192.0 false
208 56.0 Cleveland 284.0 false
209 56.0 Cleveland 256.0 true
210 44.0 Cleveland 263.0 false
211 52.0 Cleveland 199.0 true
212 57.0 Cleveland 168.0 false
213 48.0 Cleveland 229.0 true
214 54.0 Cleveland 139.0 false
215 48.0 Cleveland 275.0 false
216 49.0 Cleveland 166.0 false
217 64.0 Cleveland 211.0 false
218 58.0 Cleveland 224.0 true
219 58.0 Cleveland 284.0 false
220 58.0 Cleveland 224.0 false
221 60.0 Cleveland 206.0 false
222 50.0 Cleveland 219.0 false
223 58.0 Cleveland 226.0 false
224 66.0 Cleveland 226.0 false
225 63.0 Cleveland 233.0 true
226 67.0 Cleveland 286.0 false
227 67.0 Cleveland 229.0 false
228 37.0 Cleveland 250.0 false
229 41.0 Cleveland 204.0 false
230 36.0 Cleveland 236.0 false
231 62.0 Cleveland 268.0 false
232 57.0 Cleveland 354.0 false
233 62.0 Cleveland 254.0 false
234 53.0 Cleveland 203.0 true
235 57.0 Cleveland 192.0 false
236 56.0 Cleveland 284.0 false
237 56.0 Cleveland 256.0 true
238 44.0 Cleveland 263.0 false
239 52.0 Cleveland 199.0 true
240 57.0 Cleveland 168.0 false
241 48.0 Cleveland 229.0 true
242 54.0 Cleveland 139.0 false
243 48.0 Cleveland 275.0 false
244 49.0 Cleveland 166.0 false
245 64.0 Cleveland 211.0 false
246 58.0 Cleveland 224.0 true
247 58.0 Cleveland 284.0 false
248 58.0 Cleveland 224.0 false
249 60.0 Cleveland 206.0 false
250 50.0 Cleveland 219.0 false
251 58.0 Cleveland 226.0 false
252 66.0 Cleveland 226.0 false
253 63.0 Cleveland 233.0 true
254 67.0 Cleveland 286.0 false
255 67.0 Cleveland 229.0 false
256 37.0 Cleveland 250.0 false
257 41.0 Cleveland 204.0 false
258 36.0 Cleveland 236.0 false
259 62.0 Cleveland 268.0 false
260 57.0 Cleveland 354.0 false
261 62.0 Cleveland 254.0 false
262 53.0 Cleveland 203.0 true
263 57.0 Cleveland 192.0 false
264 56.0 Cleveland 284.0 false
265 56.0 Cleveland 256.0 true
266 44.0 Cleveland 263.0 false
267 52.0 Cleveland 199.0 true
268 57.0 Cleveland 168.0 false
269 48.0 Cleveland 229.0 true
270 54.0 Cleveland 139.0 false
271 48.0 Cleveland 275.0 false
272 49.0 Cleveland 166.0 false
273 64.0 Cleveland 211.0 false
274 58.0 Cleveland 224.0 true
275 58.0 Cleveland 284.0 false
276 58.0 Cleveland 224.0 false
277 60.0 Cleveland 206.0 false
278 50.0 Cleveland 219.0 false
279 58.0 Cleveland 226.0 false
280 66.0 Cleveland 226.0 false
281 63.0 Cleveland 233.0 true
282 67.0 Cleveland 286.0 false
283 67.0 Cleveland 229.0 false
284 37.0 Cleveland 250.0 false
285 41.0 Cleveland 204.0 false
286 36.0 Cleveland 236.0 false
287 62.0 Cleveland 268.0 false
288 57.0 Cleveland 354.0 false
289 62.0 Cleveland 254.0 false
290 53.0 Cleveland 203.0 true
291 57.0 Cleveland 192.0 false
292 56.0 Cleveland 284.0 false
293 56.0 Cleveland 256.0 true
294 44.0 Cleveland 263.0 false
295 52.0 Cleveland 199.0 true
296 57.0 Cleveland 168.0 false
297 48.0 Cleveland 229.0 true
298 54.0 Cleveland 139.0 false
299 48.0 Cleveland 275.0 false
300 49.0 Cleveland 166.0 false
301 64.0 Cleveland 211.0 false
302 58.0 Cleveland 224.0 true
303 58.0 Cleveland 284.0 false
304 58.0 Cleveland 224.0 false
305 60.0 Cleveland 206.0 false
306 50.0 Cleveland 219.0 false
307 58.0 Cleveland 226.0 false
308 66.0 Cleveland 226.0 false
309 63.0 Cleveland 233.0 true
310 67.0 Cleveland 286.0 false
311 67.0 Cleveland 229.0 false
312 37.0 Cleveland 250.0 false
313 41.0 Cleveland 204.0 false
314 36.0 Cleveland 236.0 false
315 62.0 Cleveland 268.0 false
316 57.0 Cleveland 354.0 false
317 62.0 Cleveland 254.0 false
318 53.0 Cleveland 203.0 true
319 57.0 Cleveland 192.0 false
320 56.0 Cleveland 284.0 false
321 56.0 Cleveland 256.0 true
322 44.0 Cleveland 263.0 false
323 52.0 Cleveland 199.0 true
324 57.0 Cleveland 168.0 false
325 48.0 Cleveland 229.0 true
326 54.0 Cleveland 139.0 false
327 48.0 Cleveland 275.0 false
328 49.0 Cleveland 166.0 false
329 64.0 Cleveland 211.0 false
330 58.0 Cleveland 224.0 true
331 58.0 Cleveland 284.0 false
332 58.0 Cleveland 224.0 false
333 60.0 Cleveland 206.0 false
334 50.0 Cleveland 219.0 false
335 58.0 Cleveland 226.0 false
336 66.0 Cleveland 226.0 false
337 63.0 Cleveland 233.0 true
338 67.0 Cleveland 286.0 false
339 67.0 Cleveland 229.0 false
340 37.0 Cleveland 250.0 false
341 41.0 Cleveland 204.0 false
342 36.0 Cleveland 236.0 false
343 62.0 Cleveland 268.0 false
344 57.0 Cleveland 354.0 false
345 62.0 Cleveland 254.0 false
346 53.0 Cleveland 203.0 true
347 57.0 Cleveland 192.0 false
348 56.0 Cleveland 284.0 false
349 56.0 Cleveland 256.0 true
350 44.0 Cleveland 263.0 false

```

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\_detail p JOIN health\_details h
ON (p.id = h.id) CLUSTER BY p.id;"

To arrange the output data in ascending or descending order, use the sort by and order by functions. The data is distributed to the reducers using the clustering and distribution methods, respectively.

The data is only sorted in the reducers when using sort by; they are not sorted globally. Therefore, each file in the output files from the reducers will be sorted, but the two files combined won't be sorted. Additionally, there can be overlaps in the data sets. The same names may appear in both reducers when we sort by name, for instance.

Data is sorted globally by order by, which sends all the data to a single reducer. The high file size could have a negative effect on performance.

When we distribute by name, all the same names are sent to the same reducer. Distribute by distributes the data to all reducers in a way that prevents data range overlap. For instance, reducer 2 will not have the name Regina if it receives it from reducer 1, reducer 2, etc. Distribute without sorting the data globally or within each reducer.

Cluster by does not sort the output data globally; instead, it sorts the data in each reducer and ensures that no data ranges overlap.

The screenshot shows a Cloud Shell terminal window with a Hadoop MapReduce job. The job progress table is as follows:

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 3	.....	container	SUCCEEDED	1	1	0	0	0
Map 1	.....	container	SUCCEEDED	1	1	0	0	0
Reducer 2	.....	container	SUCCEEDED	1	1	0	0	0

Below the progress table, the terminal shows the command execution and its output:

```
INFO : Completed executing command(queryId=hive_20221013185028_dfad8923-f8ae-484e-9e80-e1ecdb878e34); Time taken: 16.638 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
```

The output is a table with 5 columns: p.id, p.age, p.dataset, h.chol, and h.fbs. It contains 30 rows of data, all from the 'Cleveland' dataset. The 'h.fbs' column contains boolean values (true/false).

p.id	p.age	p.dataset	h.chol	h.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
20	49.0	Cleveland	266.0	false
21	64.0	Cleveland	211.0	false
22	58.0	Cleveland	283.0	true
23	58.0	Cleveland	284.0	false
24	58.0	Cleveland	224.0	false
25	60.0	Cleveland	206.0	false
26	50.0	Cleveland	219.0	false
27	58.0	Cleveland	340.0	false
28	66.0	Cleveland	226.0	false
29	43.0	Cleveland	247.0	false
30	40.0	Cleveland	167.0	false

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_detail p JOIN health_details h JOIN diet d
ON ((p.id = h.id) AND (h.weight = d.weight)) ORDER BY p.id;"
```



- Scalable. Hive is easy to distribute, and scale based on your needs.

Cons: -

- Deploy
- Maintenance
- Support