

# 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';

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103) X +

INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20221010050049_1992e919-bb6c-42f4-95da-a6f474ef0124): SELECT id,age,dataset
FROM personal_details
Where dataset='Hungary'
INFO : Query ID = hive_20221010050049_1992e919-bb6c-42f4-95da-a6f474ef0124
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: {} for queryId: hive_20221010050049_1992e919-bb6c-42f4-95da-a6f474ef0124
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: SELECT id,age,dataset
FR...dataset='Hungary' (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1665377184034_0002)

-----
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: 8.09 s
-----
INFO : Completed executing command(queryId=hive_20221010050049_1992e919-bb6c-42f4-95da-a6f474ef0124); Time taken: 15.456 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 |
+-----+
```

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103) X +

+-----+
| 478 | 58.0 | Hungary |
| 479 | 58.0 | Hungary |
| 578 | 58.0 | Hungary |
| 480 | 58.0 | Hungary |
| 524 | 58.0 | Hungary |
| 553 | 58.0 | Hungary |
| 525 | 59.0 | Hungary |
| 482 | 59.0 | Hungary |
| 486 | 59.0 | Hungary |
| 579 | 59.0 | Hungary |
| 580 | 59.0 | Hungary |
| 483 | 59.0 | Hungary |
| 484 | 59.0 | Hungary |
| 485 | 59.0 | Hungary |
| 526 | 60.0 | Hungary |
| 487 | 60.0 | Hungary |
| 488 | 61.0 | Hungary |
| 489 | 61.0 | Hungary |
| 490 | 62.0 | Hungary |
| 491 | 62.0 | Hungary |
| 527 | 63.0 | Hungary |
| 528 | 65.0 | Hungary |
| 581 | 65.0 | Hungary |
| 597 | 65.0 | Hungary |
| 582 | 66.0 | Hungary |
+-----+

293 rows selected (16.561 seconds)
jobUuid: f657aa71-476f-36c7-bdb4-04290b1b4a06
placement:
  clusterName: hive-cluster
  clusterUuid: 3ce64e06-ac66-4f50-91c4-0e15d02e146e
reference:
  jobId: 43a541a050964dc4b0b8f6812d90a108
  projectId: lab-3-assignment-365103
status:
  state: DONE
```

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;

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103) X +

INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20221010051246_8cecdaf0f-5a3f-4822-8da9-8f03e9916a27
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: SELECT a.id,a.age,a.dataset,b.chol,b.fb...id (Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1665377184034_0006)

-----
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: 8.37 s
-----

INFO : Completed executing command(queryId=hive_20221010051246_8cecdaf0f-5a3f-4822-8da9-8f03e9916a27); Time taken: 16.123 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

+-----+-----+-----+-----+
| a.id | a.age | a.dataset | b.chol | b.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 |
```

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103) X +

| 905 | 57.0 | VA Long Beach | 207.0 | false |
| 906 | 61.0 | VA Long Beach | 284.0 | false |
| 907 | 61.0 | VA Long Beach | 337.0 | false |
| 908 | 58.0 | VA Long Beach | 219.0 | false |
| 909 | 74.0 | VA Long Beach | 310.0 | false |
| 910 | 68.0 | VA Long Beach | 254.0 | true |
| 911 | 51.0 | VA Long Beach | 258.0 | true |
| 912 | 62.0 | VA Long Beach | 254.0 | true |
| 913 | 53.0 | VA Long Beach | 300.0 | true |
| 914 | 62.0 | VA Long Beach | 170.0 | false |
| 915 | 46.0 | VA Long Beach | 310.0 | false |
| 916 | 54.0 | VA Long Beach | 333.0 | true |
| 917 | 62.0 | VA Long Beach | 139.0 | false |
| 918 | 55.0 | VA Long Beach | 223.0 | true |
| 919 | 58.0 | VA Long Beach | 385.0 | true |
| 920 | 62.0 | VA Long Beach | 254.0 | false |
+-----+-----+-----+-----+

920 rows selected (17.045 seconds)
Beeline version 3.1.2 by Apache Hive
Closing: 0: jdbc:hive2://hive-cluster-m:10000
Job [2fd0b595a5d946aeade0c594d0620ad] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66-4f50-91
driverOutputResourceUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66-4f50-
hiveJob:
  queryList:
    queries:
      - |2-

      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;
jobUuid: 73fd273e-2601-3ec2-932d-26149a798edd
placement:
  clusterName: hive-cluster
  clusterUuid: 3ce64e06-ac66-4f50-91c4-0e15d02e146e
reference:
  jobId: 2fd0b595a5d946aeade0c594d0620ad
```

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.

```
Google Cloud Lab 3 Assignment
Search Products, resources, docs (/)

CLOUD SHELL
Terminal (lab-3-assignment-365103) X +

INFO : Query ID = hive_20221010054958_9db2bb8c-369f-4c48-9f10-d14a44f8eac6
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20221010054958_9db2bb8c-369f-4c48-9f10-d14a44f8eac6
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: SELECT a.id,a.age,a.dataset,b.chol,b.fb...id (Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1665377184034_0016)

-----
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: 8.17 s
-----
INFO : Completed executing command(queryId=hive_20221010054958_9db2bb8c-369f-4c48-9f10-d14a44f8eac6); Time taken: 16.727 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

+-----+-----+-----+-----+-----+
| a.id | a.age | a.dataset | b.chol | b.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 |
```

```
Google Cloud Lab 3 Assignment
Search Products, resources, docs (/)

CLOUD SHELL
Terminal (lab-3-assignment-365103) X +

| 263 | 60.0 | Cleveland | 240.0 | false |
| 757 | 60.0 | VA Long Beach | 281.0 | false |
| 761 | 61.0 | VA Long Beach | 0.0 | false |
| 93 | 62.0 | Cleveland | 231.0 | false |
| 856 | 62.0 | VA Long Beach | 220.0 | false |
| 692 | 62.0 | Switzerland | 0.0 | NULL |
| 818 | 62.0 | VA Long Beach | 258.0 | false |
| 753 | 63.0 | VA Long Beach | 230.0 | true |
| 581 | 65.0 | Hungary | 263.0 | true |
| 214 | 66.0 | Cleveland | 228.0 | true |
| 257 | 67.0 | Cleveland | 223.0 | false |
| 884 | 69.0 | VA Long Beach | 289.0 | true |
| 197 | 69.0 | Cleveland | 234.0 | true |
| 171 | 70.0 | Cleveland | 269.0 | false |
| 104 | 71.0 | Cleveland | 265.0 | true |
| 234 | 74.0 | Cleveland | 269.0 | false |
+-----+-----+-----+-----+-----+
920 rows selected (17.064 seconds)
Beeline version 3.1.2 by Apache Hive
Closing: 0: jdbc:hive2://hive-cluster-m:10000
Job [fe84f4a8cca84a85ad9b60b6c3931f4f] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66
driverOutputResourceUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66
hiveJob:
  queryList:
    queries:
      - |2-

      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;
jobUuid: 191ef288-e53e-3517-aa10-193aa28706f3
placement:
  clusterName: hive-cluster
  clusterUuid: 3ce64e06-ac66-4f50-91c4-0e15d02e146e
reference:
  jobId: fe84f4a8cca84a85ad9b60b6c3931f4f
```

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.

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103)

INFO : Query ID = hive_20221010055747_adb71b16-5a7d-4ae9-a02d-2a4e7f4759d0
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: {} for queryId: hive_20221010055747_adb71b16-5a7d-4ae9-a02d-2a4e7f4759d0
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: SELECT a.id,a.age,a.dataset,b.chol,b.fb...id (Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1665377184034_0017)

-----
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: 8.56 s

INFO : Completed executing command(queryId=hive_20221010055747_adb71b16-5a7d-4ae9-a02d-2a4e7f4759d0); Time taken: 16.404 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

+-----+
| a.id | a.age | a.dataset | b.chol | b.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 |
+-----+
```

```
Google Cloud Lab 3 Assignment Search Products, resources, docs (/)

CLOUD SHELL Terminal (lab-3-assignment-365103) Open Ec

+-----+
| 901 | 57.0 | VA Long Beach | 264.0 | false |
| 902 | 55.0 | VA Long Beach | NULL | false |
| 903 | 55.0 | VA Long Beach | 226.0 | false |
| 904 | 56.0 | VA Long Beach | 203.0 | true |
| 905 | 57.0 | VA Long Beach | 207.0 | false |
| 906 | 61.0 | VA Long Beach | 284.0 | false |
| 907 | 61.0 | VA Long Beach | 337.0 | false |
| 908 | 58.0 | VA Long Beach | 219.0 | false |
| 909 | 74.0 | VA Long Beach | 310.0 | false |
| 910 | 68.0 | VA Long Beach | 254.0 | true |
| 911 | 51.0 | VA Long Beach | 258.0 | true |
| 912 | 62.0 | VA Long Beach | 254.0 | true |
| 913 | 53.0 | VA Long Beach | 300.0 | true |
| 914 | 62.0 | VA Long Beach | 170.0 | false |
| 915 | 46.0 | VA Long Beach | 310.0 | false |
| 916 | 54.0 | VA Long Beach | 333.0 | true |
| 917 | 62.0 | VA Long Beach | 139.0 | false |
| 918 | 55.0 | VA Long Beach | 223.0 | true |
| 919 | 58.0 | VA Long Beach | 385.0 | true |
| 920 | 62.0 | VA Long Beach | 254.0 | false |
+-----+

920 rows selected (16.716 seconds)
Beeline version 3.1.2 by Apache Hive
Closing: 0: jdbc:hive2://hive-cluster-m:10000
Job [f993cd5ffld741ff9663c17f53157e42] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66-4f50-91c4-0e15d
driverOutputResourceUri: gs://dataproc-staging-us-central1-676827437038-h8bjyza6/google-cloud-dataproc-metainfo/3ce64e06-ac66-4f50-91c4-0e15d
hiveJob:
  queryList:
    queries:
      - |2-

      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;
jobUuid: 284c8e35-1a26-33ca-88ee-ddf8ba3a4a4a
```

```
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;
```

```

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 10
```

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.