

Monica Santana Week 3 -Exercise

Grades.csv Results

```
santanamonica@dc6bigdata: ~/dc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
SLF4J: Found binding in [jar:file:/usr/program/tez/1.10.4-SNAPSHOT/SLF4JLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = f092cbf1-dabf-4542-92ee-3d59d0980668

Logging initialized using configuration in file:/usr/program/hive/conf/hive-log4j2.properties Async: true
2024-06-22 04:02:04,159 INFO [Tez session start thread] client.RMProxy: Connect
ing to ResourceManager at master/172.28.1.1:8032
Hive Session ID = 9c31a424-39c0-442c-9990-079e91e90c1d
2024-06-22 04:02:05,175 INFO [pool-7-thread-1] client.RMProxy: Connecting to Re
sourceManager at master/172.28.1.1:8032
hive> CREATE TABLE grades(
>   'Last name' STRING,
>   'First name' STRING,
>   'SEX' STRING,
>   'Test1' DOUBLE,
>   'Test2' INT,
>   'Test3' DOUBLE,
>   'Test4' DOUBLE,
>   'Final' DOUBLE,
>   'Grade' STRING)
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> STORED AS TEXTFILE
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.351 seconds
hive> LOAD DATA INPATH '/grades.csv' INTO TABLE grades;
Loading data to table default.grades
OK
Time taken: 1.365 seconds
hive> SELECT * FROM grades;
OK
Alfalfa Aloyalus      123-45-6789      40.0      90      100.0      83.0      49.0      D
+
Alfred University    123-12-1234      41.0      97      96.0      97.0      48.0      D
+
Gerty Grams 367-89-0123      41.0      80      60.0      48.0      44.0      C
+
Android Electric     087-65-4321      42.0      23      36.0      45.0      47.0      B
+
Bumpkin Fred         456-78-9012      43.0      78      88.0      77.0      45.0      A-
+
Rubble Betty         234-56-7890      44.0      90      60.0      90.0      46.0      C-
+
Noshov Cecil         345-67-8901      45.0      11      -1.0      4.0      43.0      F
+
Buff Hif             632-79-9939      46.0      20      30.0      40.0      50.0      B+
+
Airpump Andrew       223-45-6789      49.0      1      90.0      100.0      83.0      A
+
Rastus Jim           143-12-1234      48.0      1      97.0      96.0      97.0      A+
+
Cannivore Art        565-99-0123      44.0      1      80.0      60.0      40.0      D
+
Dandy Jim            087-75-4321      47.0      1      23.0      36.0      45.0      C+
+
Elephant Ima         456-71-9012      45.0      1      78.0      88.0      77.0      B
+
Franklin Benny       234-56-2090      50.0      1      90.0      80.0      90.0      B
+
George Roy           345-67-3501      40.0      1      11.0      -1.0      4.0      H
+
Hoffalump Harvey     632-79-9439      30.0      1      20.0      30.0      40.0      C
Time taken: 3.617 seconds, Fetched: 16 row(s)
hive>
```

SQL Command #1

```
santanamonica@dc6bigdata: ~/dc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
hive> SELECT COUNT(*) FROM grades;
2024-06-22 04:54:57,739 INFO [f092cbf1-dabf-4542-92ee-3d59d0980668 main] reduce
sink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator.construc
tor: VectorReduceSinkInfo org.apache.hadoop.hive.q1.plan.VectorReduceSinkInfo@5b2
ff4d5
Query ID = root_20240622045456_84e0d41b-662a-48a6-a3d7-7f6e8c2f7cd4
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-06-22 04:54:58,716 INFO [f092cbf1-dabf-4542-92ee-3d59d0980668 main] client
.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1719028677711
0004)

-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
Map 1         container  INITED      1          0          0          0          1
-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
--2 [>>>-----] 0% ELAPSED TIME: 0.03 s
Map 1         container  INITED      1          0          0          0          1
-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
--2 [>>>-----] 0% ELAPSED TIME: 0.34 s
Map 1         container  INITED      1          0          0          0          1
-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
--2 [>>>-----] 0% ELAPSED TIME: 1.05 s
Map 1         container  INITED      1          0          0          0          1
-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
--2 [>>>-----] 0% ELAPSED TIME: 1.56 s
Map 1         container  INITED      1          0          0          0          1
-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FA
ILED_KILLED
-----
Map 1 ..... container SUCCEEDED 1          1          0          0          0          0
Reducer 2 ..... container SUCCEEDED 1          1          0          0          0          0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 6.51 s
```

SQL Command #2

```
santanamonica@dscbigdata: ~/dsc650 infra/bellevue bigdata/hadoop hive spark hbase
Tez session was closed. Reopening...
2024-06-22 05:29:39,04/ INFO [d6505330-d208-46a4-932f-67299f08f1ad main] client.RMPProxy: Connecting to ResourceManager at master/172.20.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1719028677711_0008)

-----
VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICKS: 02/02 [=====] 100% ELAPSED TIME: 6.41 s
-----
OK
4.0 1
40.0 2
43.0 1
44.0 1
45.0 2
46.0 1
47.0 1
48.0 1
49.0 1
50.0 1
77.0 1
83.0 1
90.0 1
97.0 1
Time taken: 15.382 seconds, Fetched: 14 row(s)
hive> SELECT grade, COUNT(*) FROM grades GROUP BY grade;
Query ID = root_20240622053159_04775250-7c3f-477d-b154-7a7a75134e53
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719028677711_0008)

-----
VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICKS: 02/02 [=====] 100% ELAPSED TIME: 6.74 s
-----
OK
A 1
A+ 1
A- 1
B 1
B+ 1
B- 3
C 2
C+ 1
C- 1
D+ 2
D- 1
E 1
Time taken: 7.749 seconds, Fetched: 12 row(s)
hive>
```

SQL Command #3 and #4

```
santanamonica@dscbigdata: ~/dsc650 infra/bellevue bigdata/hadoop hive spark hbase
hive> SELECT ssn FROM grades;
OK
123-45-6789
123-12-1234
567-89-0123
087-65-4321
456-78-9012
234-56-7890
345-67-8901
632-79-9939
223-45-6789
143-12-1234
565-89-0123
087-75-4321
456-71-9012
234-56-2890
345-67-3901
632-79-9439
Time taken: 0.208 seconds, Fetched: 16 row(s)
hive> SELECT ssn FROM grades ORDER BY ssn;
2024-06-22 05:43:02,412 INFO [d6505330-d208-46a4-932f-67299f08f1ad main] reduce
sink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObjectHashOperator.com
tractor.VectorReduceSinkInfo org.apache.hadoop.hive ql.plan.VectorReduceSinkInfo
810f192d8
Query ID = root_20240622054302_442d2482-3a40-4757-a11f-13d73b761785
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-06-22 05:43:02,590 INFO [d6505330-d208-46a4-932f-67299f08f1ad main] client
.RMPProxy: Connecting to ResourceManager at master/172.20.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1719028677711_0009)

-----
VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICKS: 02/02 [=====] 100% ELAPSED TIME: 5.70 s
-----
OK
087-65-4321
087-75-4321
123-12-1234
123-45-6789
143-12-1234
223-45-6789
234-56-2890
234-56-7890
345-67-3901
345-67-8901
456-71-9012
456-78-9012
565-89-0123
567-89-0123
632-79-9439
632-79-9939
```

The dataset I chose for my SQL query is on sleep, and what features are used to predict whether someone has a sleep disorder of insomnia, sleep apnea or no sleep disorder at all. I hope to gain insights such as what gender is most apparent in the dataset, maximum age, most common age, and how many people are in the dataset total.

```

santanamonica@dscbigdata: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
hive> LOAD DATA INPATH '/sleep_data.csv' INTO TABLE sleep;
Loading data to table default.sleep
OK
Time taken: 0.681 seconds
hive> SELECT * FROM sleep;
OK
1      Male      27      Software Engineer      6      6      42      6      0
verweight 126/83 77      4200      None
2      Male      28      Doctor      6      6      60      8      Normal 125/80 7
5      10000      None
3      Male      28      Doctor      6      6      60      8      Normal 125/80 7
5      10000      None
4      Male      28      Sales Representative 5      4      30      8      0
basee 140/90 85      3000      Sleep Apnea
5      Male      28      Sales Representative 5      4      30      8      0
basee 140/90 85      3000      Sleep Apnea
6      Male      28      Software Engineer      5      4      30      8      0
basee 140/90 85      3000      Insomnia
7      Male      29      Teacher      6      6      40      7      Obese 140/90 8
2      3500      Insomnia
8      Male      29      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
9      Male      29      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
10     Male      29      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
11     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      None
12     Male      29      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
13     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      None
14     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      None
15     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      None
16     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      None
17     Female 29      Nurse      6      5      40      7      Normal Weight 1
32/87 80      4000      Sleep Apnea
18     Male      29      Doctor      6      6      30      8      Normal 120/80 7
0      8000      Sleep Apnea
19     Female 29      Nurse      6      5      40      7      Normal Weight 1
32/87 80      4000      Insomnia
20     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
21     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
22     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
23     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
24     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
25     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
26     Male      30      Doctor      7      7      75      6      Normal 120/80 7
0      8000      None
27     Male      30      Doctor      7      7      75      6      Normal 120/80 7

```

```

CREATE TABLE sleep(
  `Person ID` INT,
  `Gender` STRING,
  `Age` INT,
  `Occupation` STRING,
  `Sleep Duration` INT,
  `Quality of Sleep` INT,
  `Physical Activity Level` INT,
  `Stress Level` INT,
  `BMI Category` STRING,
  `Blood Pressure` STRING,
  `Heart Rate` INT,
  `Daily Steps` INT,
  `Sleep Disorder` STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
tblproperties("skip.header.line.count"="1");

```

SQL Command #1 – to see what age is most common in the dataset which in this case is age 43

```

hives@sc16ct-aga:~/COINTELL-1$ FROM sleep GROUP BY age;
Query ID = root_20240624035307_7f5259be-e1ff-403e-9716-d6a78acd805b
Total jobs = 1
Launching job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719199137166_0004)

-----
VERTICES    MODE    STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED
-----
Map 1 ----- container INITED 1 0 0 1
-----
VERTICES    MODE    STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED
-----
Map 1 -----2 [>-----] 0% ELAPSED TIME: 0.00 s
----- container INITED 1 0 0 1
-----
VERTICES    MODE    STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED
-----
Map 1 -----2 [>-----] 0% ELAPSED TIME: 0.51 s
----- container INITED 1 0 0 1
-----
VERTICES    MODE    STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED
-----
Map 1 -----2 [>-----] 0% ELAPSED TIME: 1.02 s
----- container INITED 1 0 0 1
-----
VERTICES    MODE    STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====>] 100% ELAPSED TIME: 5.53 s

OK
27 1
28 5
29 13
30 13
31 18
32 17
33 13
34 2
35 12
36 12
37 20
38 20
39 15
40 4
41 12

```

```

santananonica@dscbigdata: ~/dsc650infra/bellevue.bigdata/hadoophive-spark-hbase
Map 1 .....-2 [==>-----] 0% ELAPSED TIME: 0.00 s
container INITED 1 0 0 1
container INITED 1 0 0 1
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED -----
Map 1 .....-2 [==>-----] 0% ELAPSED TIME: 0.51 s
container INITED 1 0 0 1
container INITED 1 0 0 1
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED -----
Map 1 .....-2 [==>-----] 0% ELAPSED TIME: 1.02 s
container INITED 1 0 0 1
container INITED 1 0 0 1
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 5.53 s
OK
27 1
28 5
29 13
30 13
31 18
32 17
33 13
34 2
35 12
36 12
37 20
38 20
39 15
40 4
41 12
42 9
43 34
44 30
45 14
48 3
49 11
50 20
51 8
52 9
53 17
54 7
55 2
56 2
57 9
58 6
59 15
Time taken: 6.433 seconds, Fetched: 31 row(s)
hive>

```

SQL Command #2 – separating the occupations from the sleep dataset

```
santanamonica@dcdigdata ~/dir650 infra/believe bigdata/hadoop hive spark hbase  
Time taken: 0.173 seconds, Fetched: 3/4 row(s)  
hive> SELECT Occupation FROM sleep;  
OK  
Software Engineer  
Doctor  
Doctor  
Sales Representative  
Sales Representative  
Software Engineer  
Teacher  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Nurse  
Doctor  
Nurse  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Nurse  
Nurse  
Nurse  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Doctor  
Engineer  
Engineer  
Doctor  
Doctor  
Doctor  
Doctor
```

SQL Command #3 – sorting the occupations in alphabetical order

[illegible]

SQL Command #4 – Counting how many people are in the dataset – which is 374

[illegible]

SQL Command #5 – seeing the distinct types of occupations and ages

```

$ santanamonica@dc3bjgdatz:~/dsc650 infra/belueve bigdata/hadoop-hive spark-hbase
Query ID = root_20240624040240_elcaca09-f339-45c6-8add-fbc2a7049b3c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719199137166_00005)

-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FAILED      KILLED
-----
Map 1 ..... container      SUCCEEDED      1            1            0            0            0            0
Reducer 2 ..... container      SUCCEEDED      1            1            0            0            0            0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 5.45 s
-----
OK
Accountant
Doctor
Engineer
Lawyer
Manager
Nurse
Sales Representative
Salesperson
Scientist
Software Engineer
Teacher
Time taken: 6.366 seconds, Fetched: 11 row(s)
hive> SELECT DISTINCT Age FROM alscsp;
Query ID = root_20240624040321_86f8cd34-ae45-4ac6-a8164-8f1e043daffe
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719199137166_00005)

-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FAILED      KILLED
-----
Map 1 ..... container      SUCCEEDED      1            1            0            0            0            0
Reducer 2 ..... container      SUCCEEDED      1            1            0            0            0            0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 7.71 s
-----
OK
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

```

SQL Command #6 – seeing the maximum value in occupations in the dataset, which is a teacher, seeing the maximum value in age which is age 59 (the oldest person in the data), and the maximum value of gender is male

```
santanamonica@dscbigdata: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
hive> SELECT MAX(Occupation) AS label FROM sleep;
2024-06-24 04:13:12,013 INFO [d49e6e27-e19b-480b-8892-79fbc6262211 main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceS
inkInfo org.apache.hadoop.hive.q1.plan.VectorReduceSinkInfo085f0ff4f1
Query ID = root_20240624041311_55e77d58-66ef-4e0c-aabf-f93121fb8070
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-06-24 04:13:12,172 INFO [d49e6e27-e19b-480b-8892-79fbc6262211 main] client.RMPProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1719199137166_0006)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 02/02 [=====] >>> 100% ELAPSED TIME: 6.34 s
-----
OK
Teacher
Time taken: 13.919 seconds, Fetched: 1 row(s)
hive> SELECT MAX(Age) AS label FROM sleep;
2024-06-24 04:13:44,631 INFO [d49e6e27-e19b-480b-8892-79fbc6262211 main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceS
inkInfo org.apache.hadoop.hive.q1.plan.VectorReduceSinkInfo0466add8d
Query ID = root_20240624041344_3bea2387-d4df-4a31-819d-35529b1792a8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719199137166_0006)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 02/02 [=====] >>> 100% ELAPSED TIME: 4.81 s
-----
OK
59
Time taken: 5.681 seconds, Fetched: 1 row(s)
hive> SELECT MAX(Gender) AS label FROM sleep;
2024-06-24 04:14:09,912 INFO [d49e6e27-e19b-480b-8892-79fbc6262211 main] reducesink.VectorReduceSinkEmptyKeyOperator: VectorReduceSinkEmptyKeyOperator constructor vectorReduceS
inkInfo org.apache.hadoop.hive.q1.plan.VectorReduceSinkInfo0665c79a2
Query ID = root_20240624041409_eba70def-da5c-40e5-b3e7-29bd19c1bdda
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1719199137166_0006)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 02/02 [=====] >>> 100% ELAPSED TIME: 5.14 s
-----
OK
```