Final Project

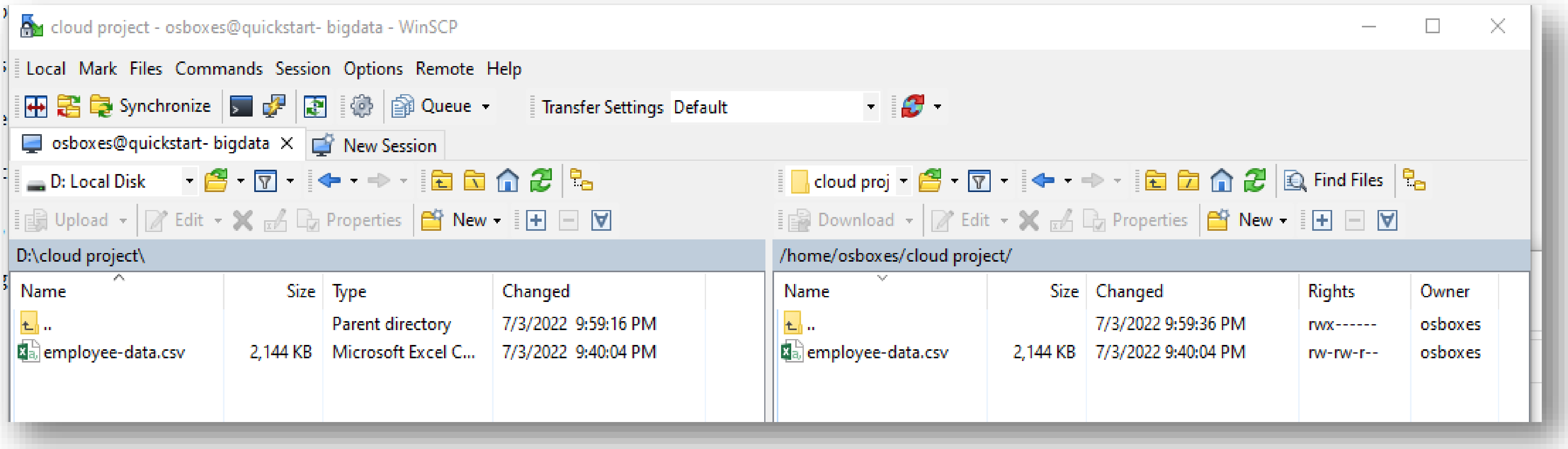
Data Description:

We have dataset that contains information about Chicago employee data.
The file name: employee-data.csv
It consists of 32928 record involved (Name, Job Titles, Department, Full or Part-Time, Typical Hours, Annual Salary, Hourly Rate) as columns names.

Requirements:

1) Create a Hive table named employee-data-hive based on the given dataset.

We first move the data file (employee-data.csv) from local on the disk to local location on the virtual machine under 'cloud project' directory.


cloud project - osboxes@quickstart- bigdata - WinSCP
Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
osboxes@quickstart- bigdata
New Session
D: Local Disk
cloud proj
Find Files
Upload
Edit
Properties
New
Download
Edit
Properties
New

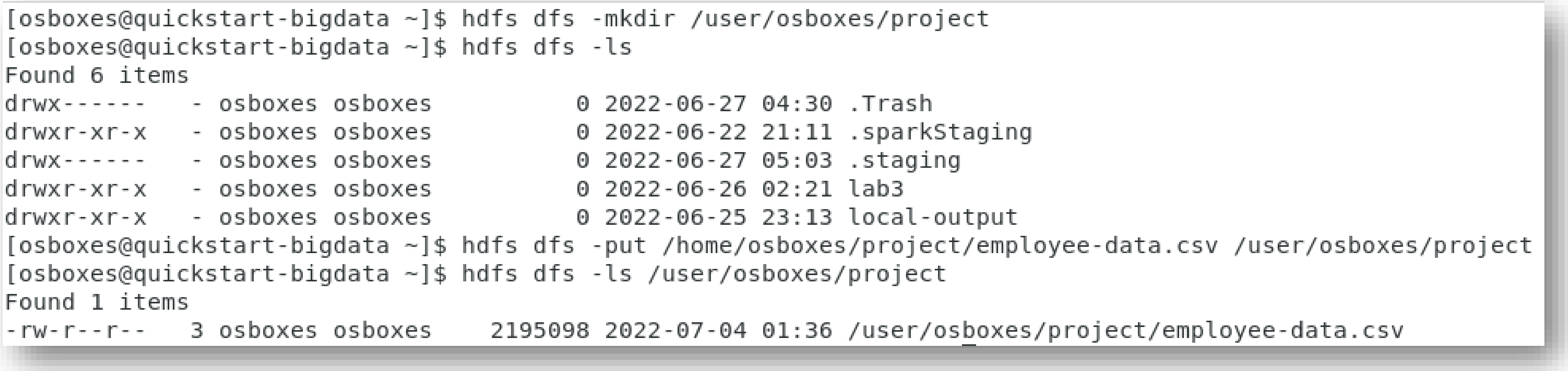
D:\cloud project\

Name	Size	Type	Changed
..		Parent directory	7/3/2022 9:59:16 PM
employee-data.csv	2,144 KB	Microsoft Excel C...	7/3/2022 9:40:04 PM

/home/osboxes/cloud project/

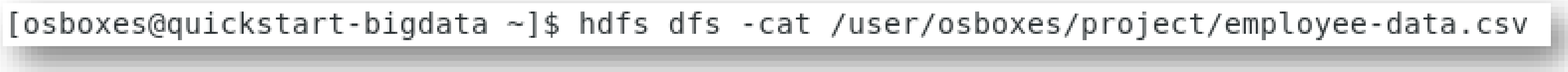
Name	Size	Changed	Rights	Owner
..		7/3/2022 9:59:36 PM	rwx------	osboxes
employee-data.csv	2,144 KB	7/3/2022 9:40:04 PM	rw-rw-r--	osboxes

Then we created directory called 'project' on hdfs location & copied the data file under its path.



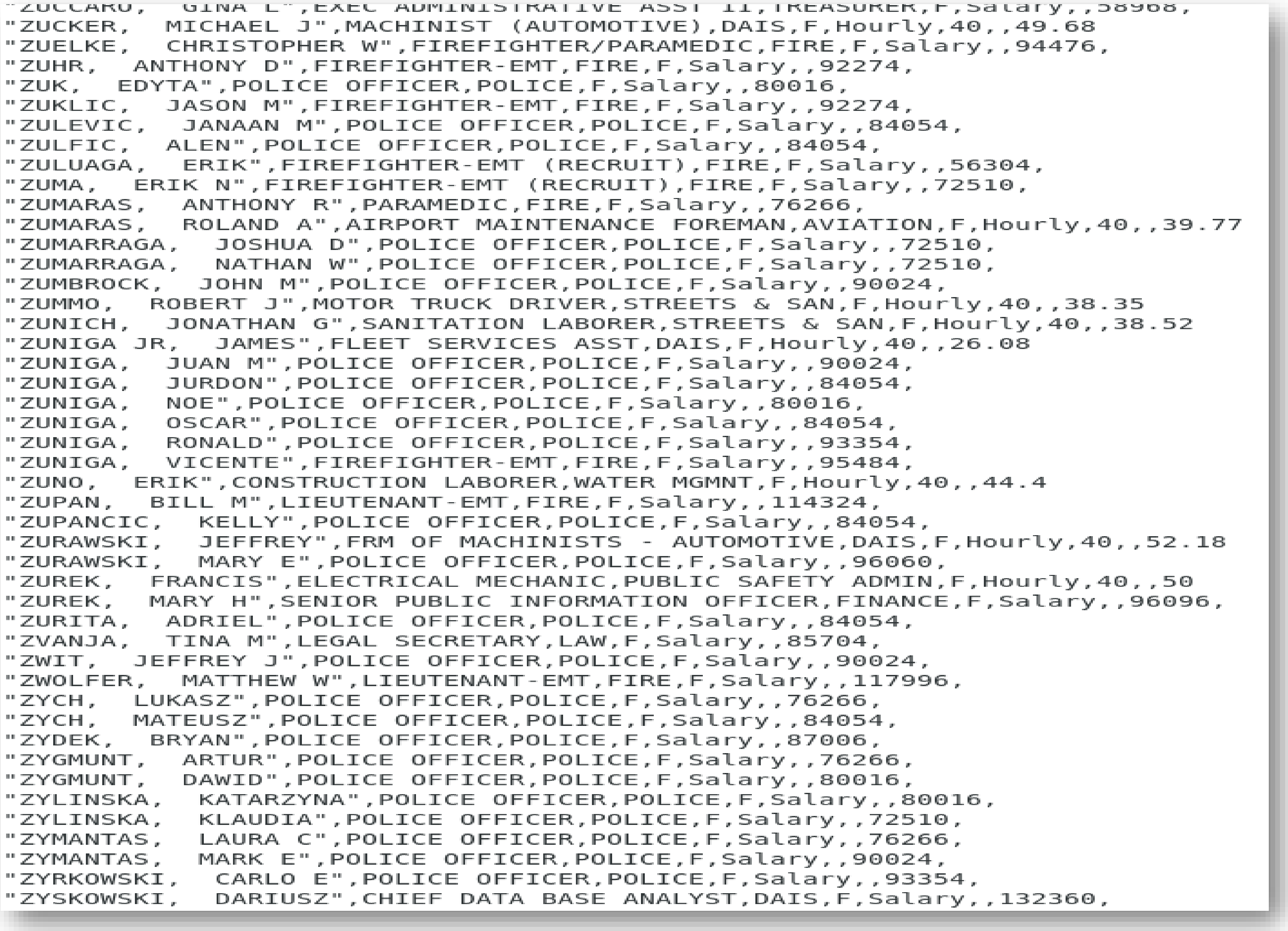
```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -mkdir /user/osboxes/project
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls
Found 6 items
drwx------   - osboxes osboxes          0 2022-06-27 04:30 .Trash
drwxr-xr-x   - osboxes osboxes          0 2022-06-22 21:11 .sparkStaging
drwx------   - osboxes osboxes          0 2022-06-27 05:03 .staging
drwxr-xr-x   - osboxes osboxes          0 2022-06-26 02:21 lab3
drwxr-xr-x   - osboxes osboxes          0 2022-06-25 23:13 local-output
[osboxes@quickstart-bigdata ~]$ hdfs dfs -put /home/osboxes/project/employee-data.csv /user/osboxes/project
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes/project
Found 1 items
-rw-r--r--   3 osboxes osboxes    2195098 2022-07-04 01:36 /user/osboxes/project/employee-data.csv
```

Then we read the file data content



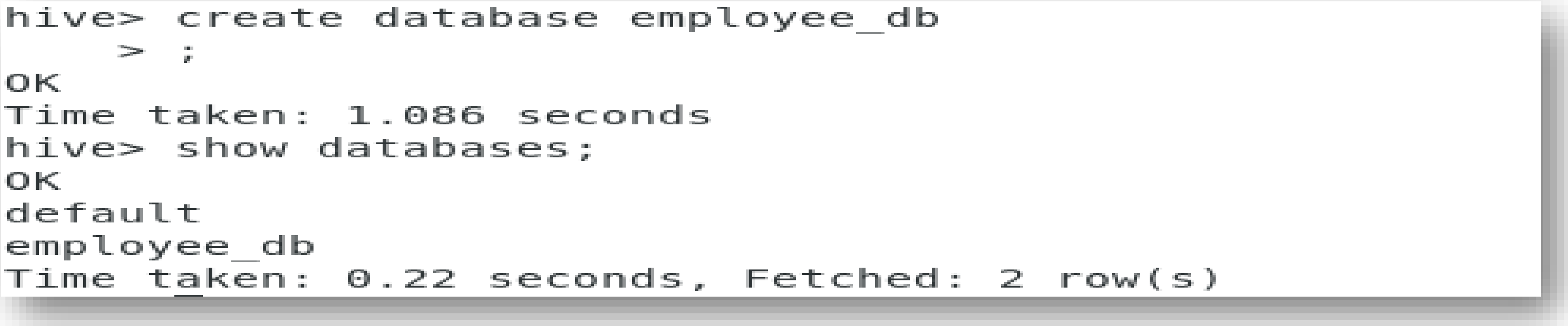
```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/osboxes/project/employee-data.csv
```

And here was the output sample.



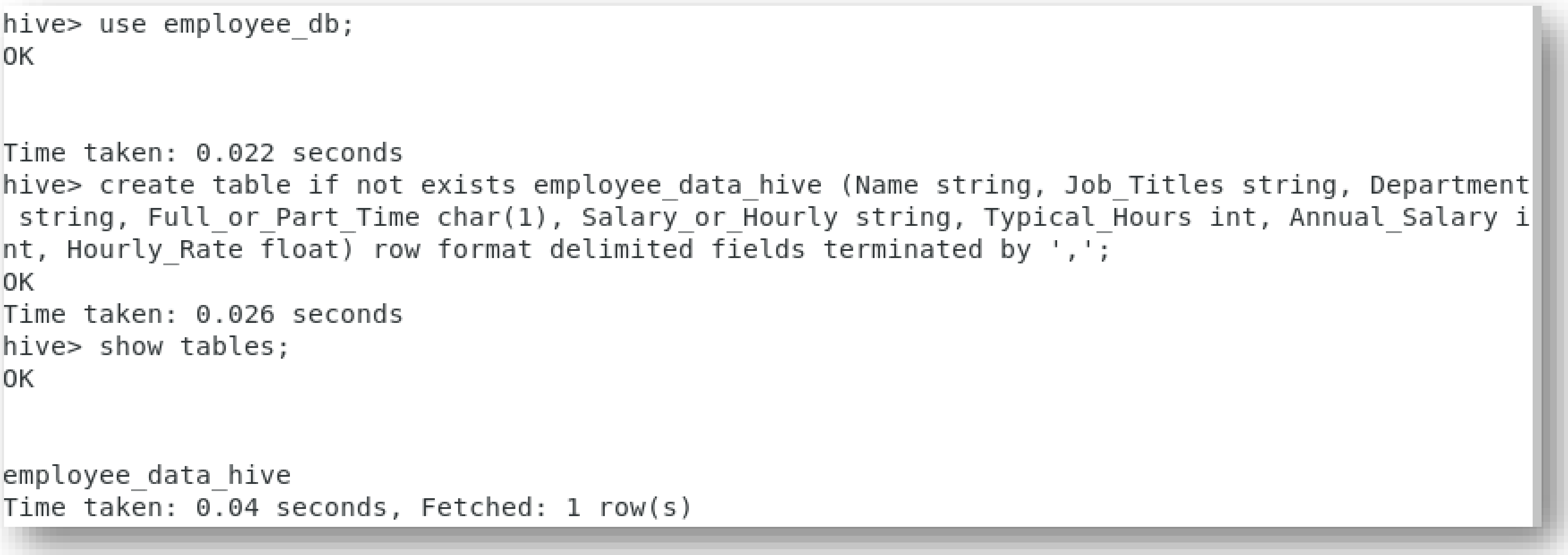
```
"ZUCCARO,  GINA L",EXEC ADMINISTRATIVE ASST II,TREASURER,F,Salary,,58968,
"ZUCKER,  MICHAEL J",MACHINIST (AUTOMOTIVE),DAIS,F,Hourly,40,,49.68
"ZUELKE,  CHRISTOPHER W",FIREFIGHTER/PARAMEDIC,FIRE,F,Salary,,94476,
"ZUHR,  ANTHONY D",FIREFIGHTER-EMT,FIRE,F,Salary,,92274,
"ZUK,  EDYTA",POLICE OFFICER,POLICE,F,Salary,,80016,
"ZUKLIC,  JASON M",FIREFIGHTER-EMT,FIRE,F,Salary,,92274,
"ZULEVIC,  JANAAN M",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZULFIC,  ALEN",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZULUAGA,  ERIK",FIREFIGHTER-EMT (RECRUIT),FIRE,F,Salary,,56304,
"ZUMA,  ERIK N",FIREFIGHTER-EMT (RECRUIT),FIRE,F,Salary,,72510,
"ZUMARAS,  ANTHONY R",PARAMEDIC,FIRE,F,Salary,,76266,
"ZUMARAS,  ROLAND A",AIRPORT MAINTENANCE FOREMAN,AVIATION,F,Hourly,40,,39.77
"ZUMARRAGA,  JOSHUA D",POLICE OFFICER,POLICE,F,Salary,,72510,
"ZUMARRAGA,  NATHAN W",POLICE OFFICER,POLICE,F,Salary,,72510,
"ZUMBROCK,  JOHN M",POLICE OFFICER,POLICE,F,Salary,,90024,
"ZUMMO,  ROBERT J",MOTOR TRUCK DRIVER,STREETS & SAN,F,Hourly,40,,38.35
"ZUNICH,  JONATHAN G",SANITATION LABORER,STREETS & SAN,F,Hourly,40,,38.52
"ZUNIGA JR,  JAMES",FLEET SERVICES ASST,DAIS,F,Hourly,40,,26.08
"ZUNIGA,  JUAN M",POLICE OFFICER,POLICE,F,Salary,,90024,
"ZUNIGA,  JURDON",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZUNIGA,  NOE",POLICE OFFICER,POLICE,F,Salary,,80016,
"ZUNIGA,  OSCAR",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZUNIGA,  RONALD",POLICE OFFICER,POLICE,F,Salary,,93354,
"ZUNIGA,  VICENTE",FIREFIGHTER-EMT,FIRE,F,Salary,,95484,
"ZUNO,  ERIK",CONSTRUCTION LABORER,WATER MGMNT,F,Hourly,40,,44.4
"ZUPAN,  BILL M",LIEUTENANT-EMT,FIRE,F,Salary,,114324,
"ZUPANCIC,  KELLY",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZURAWSKI,  JEFFREY",FRM OF MACHINISTS - AUTOMOTIVE,DAIS,F,Hourly,40,,52.18
"ZURAWSKI,  MARY E",POLICE OFFICER,POLICE,F,Salary,,96060,
"ZUREK,  FRANCIS",ELECTRICAL MECHANIC,PUBLIC SAFETY ADMIN,F,Hourly,40,,50
"ZUREK,  MARY H",SENIOR PUBLIC INFORMATION OFFICER,FINANCE,F,Salary,,96096,
"ZURITA,  ADRIEL",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZVANJA,  TINA M",LEGAL SECRETARY,LAW,F,Salary,,85704,
"ZWIT,  JEFFREY J",POLICE OFFICER,POLICE,F,Salary,,90024,
"ZWOLFER,  MATTHEW W",LIEUTENANT-EMT,FIRE,F,Salary,,117996,
"ZYCH,  LUKASZ",POLICE OFFICER,POLICE,F,Salary,,76266,
"ZYCH,  MATEUSZ",POLICE OFFICER,POLICE,F,Salary,,84054,
"ZYDEK,  BRYAN",POLICE OFFICER,POLICE,F,Salary,,87006,
"ZYGMUNT,  ARTUR",POLICE OFFICER,POLICE,F,Salary,,76266,
"ZYGMUNT,  DAWID",POLICE OFFICER,POLICE,F,Salary,,80016,
"ZYLINSKA,  KATARZYNA",POLICE OFFICER,POLICE,F,Salary,,80016,
"ZYLINSKA,  KLAUDIA",POLICE OFFICER,POLICE,F,Salary,,72510,
"ZYMANTAS,  LAURA C",POLICE OFFICER,POLICE,F,Salary,,76266,
"ZYMANTAS,  MARK E",POLICE OFFICER,POLICE,F,Salary,,90024,
"ZYRKOWSKI,  CARLO E",POLICE OFFICER,POLICE,F,Salary,,93354,
"ZYSKOWSKI,  DARIUSZ",CHIEF DATA BASE ANALYST,DAIS,F,Salary,,132360,
```

We entered the hive shell to start creating our database that was called ‘employee_db’



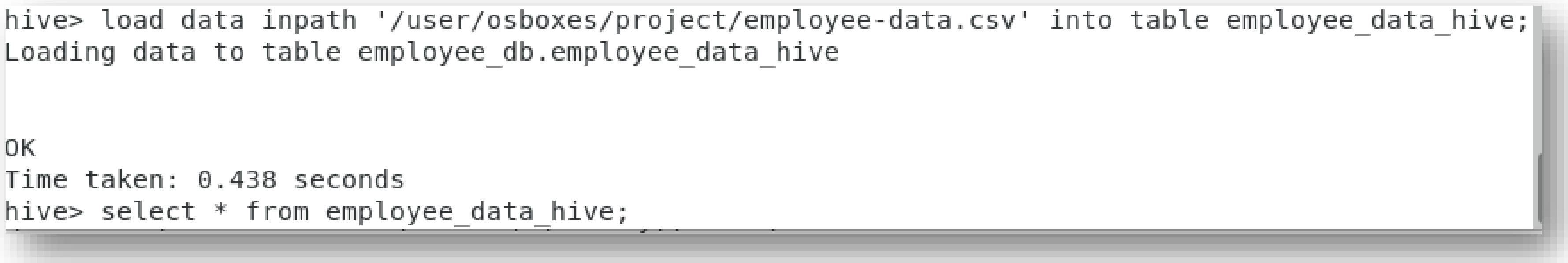
```
hive> create database employee_db
    > ;
OK
Time taken: 1.086 seconds
hive> show databases;
OK
default
employee_db
Time taken: 0.22 seconds, Fetched: 2 row(s)
```

We opened the ‘employee_db’ database by using ‘use employee_db’ command to create a new table under its directory called ‘employee_data_hive’

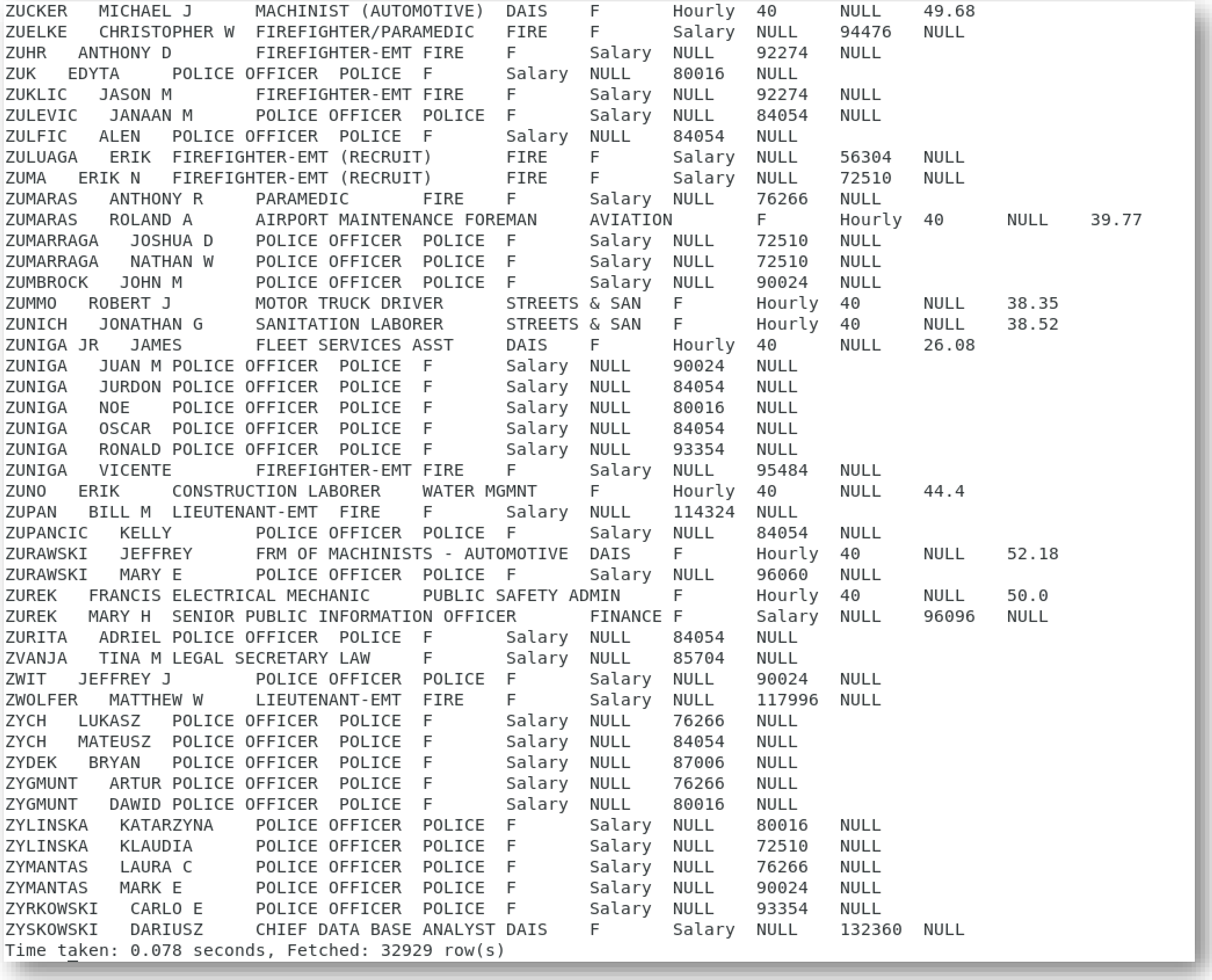


```
hive> use employee_db;
OK


Time taken: 0.022 seconds
hive> create table if not exists employee_data_hive (Name string, Job_Titles string, Department
 string, Full_or_Part_Time char(1), Salary_or_Hourly string, Typical_Hours int, Annual_Salary i
nt, Hourly_Rate float) row format delimited fields terminated by ',';
OK
Time taken: 0.026 seconds
hive> show tables;
OK


employee_data_hive
Time taken: 0.04 seconds, Fetched: 1 row(s)
```

Then we loaded the data from hdfs location to the hive created table ‘employee_data_hive’ and showed its content



```
hive> load data inpath '/user/osboxes/project/employee-data.csv' into table employee_data_hive;
Loading data to table employee_db.employee_data_hive


OK
Time taken: 0.438 seconds
hive> select * from employee_data_hive;
```



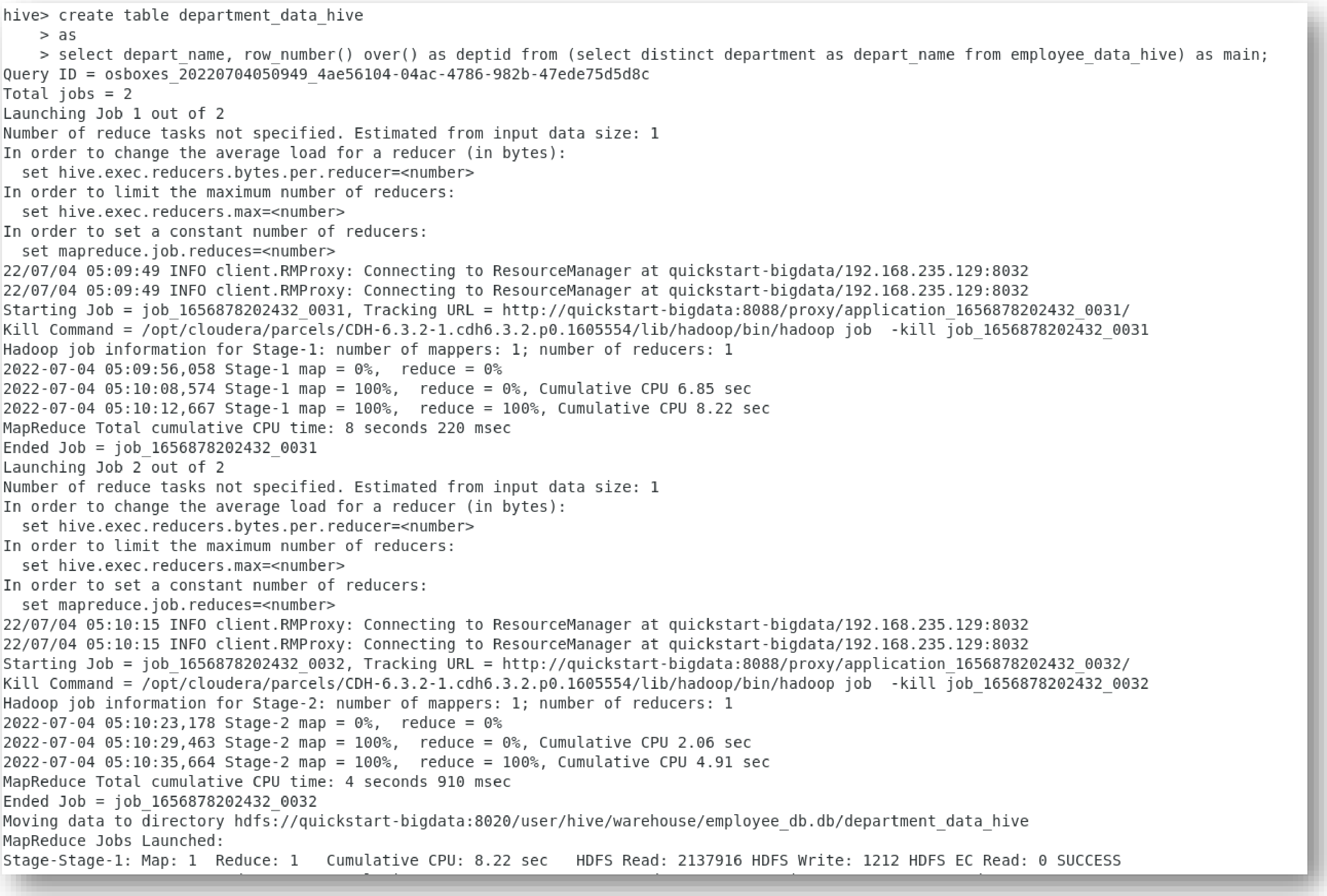
```
ZUCKER   MICHAEL J      MACHINIST (AUTOMOTIVE)  DAIS    F       Hourly  40      NULL    49.68
ZUELKE   CHRISTOPHER W  FIREFIGHTER/PARAMEDIC   FIRE    F       Salary  NULL    94476   NULL
ZUHR   ANTHONY D        FIREFIGHTER-EMT FIRE    F       Salary  NULL    92274   NULL
ZUK   EDYTA     POLICE OFFICER  POLICE  F       Salary  NULL    80016   NULL
ZUKLIC   JASON M        FIREFIGHTER-EMT FIRE    F       Salary  NULL    92274   NULL
ZULEVIC   JANAAN M      POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZULFIC   ALEN   POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZULUAGA   ERIK  FIREFIGHTER-EMT (RECRUIT)       FIRE    F       Salary  NULL    56304   NULL
ZUMA   ERIK N   FIREFIGHTER-EMT (RECRUIT)       FIRE    F       Salary  NULL    72510   NULL
ZUMARAS   ANTHONY R     PARAMEDIC       FIRE    F       Salary  NULL    76266   NULL
ZUMARAS   ROLAND A      AIRPORT MAINTENANCE FOREMAN     AVIATION        F       Hourly  40      NULL    39.77
ZUMARRAGA   JOSHUA D    POLICE OFFICER  POLICE  F       Salary  NULL    72510   NULL
ZUMARRAGA   NATHAN W    POLICE OFFICER  POLICE  F       Salary  NULL    72510   NULL
ZUMBROCK   JOHN M       POLICE OFFICER  POLICE  F       Salary  NULL    90024   NULL
ZUMMO   ROBERT J        MOTOR TRUCK DRIVER      STREETS & SAN   F       Hourly  40      NULL    38.35
ZUNICH   JONATHAN G     SANITATION LABORER      STREETS & SAN   F       Hourly  40      NULL    38.52
ZUNIGA JR   JAMES       FLEET SERVICES ASST     DAIS    F       Hourly  40      NULL    26.08
ZUNIGA   JUAN M POLICE OFFICER  POLICE  F       Salary  NULL    90024   NULL
ZUNIGA   JURDON POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZUNIGA   NOE    POLICE OFFICER  POLICE  F       Salary  NULL    80016   NULL
ZUNIGA   OSCAR  POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZUNIGA   RONALD POLICE OFFICER  POLICE  F       Salary  NULL    93354   NULL
ZUNIGA   VICENTE        FIREFIGHTER-EMT FIRE    F       Salary  NULL    95484   NULL
ZUNO   ERIK     CONSTRUCTION LABORER    WATER MGMNT     F       Hourly  40      NULL    44.4
ZUPAN   BILL M  LIEUTENANT-EMT  FIRE    F       Salary  NULL    114324  NULL
ZUPANCIC   KELLY        POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZURAWSKI   JEFFREY      FRM OF MACHINISTS - AUTOMOTIVE  DAIS    F       Hourly  40      NULL    52.18
ZURAWSKI   MARY E       POLICE OFFICER  POLICE  F       Salary  NULL    96060   NULL
ZUREK   FRANCIS ELECTRICAL MECHANIC     PUBLIC SAFETY ADMIN     F       Hourly  40      NULL    50.0
ZUREK   MARY H  SENIOR PUBLIC INFORMATION OFFICER       FINANCE F       Salary  NULL    96096   NULL
ZURITA   ADRIEL POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZVANJA   TINA M LEGAL SECRETARY LAW     F       Salary  NULL    85704   NULL
ZWIT   JEFFREY J        POLICE OFFICER  POLICE  F       Salary  NULL    90024   NULL
ZWOLFER   MATTHEW W     LIEUTENANT-EMT  FIRE    F       Salary  NULL    117996  NULL
ZYCH   LUKASZ   POLICE OFFICER  POLICE  F       Salary  NULL    76266   NULL
ZYCH   MATEUSZ  POLICE OFFICER  POLICE  F       Salary  NULL    84054   NULL
ZYDEK   BRYAN   POLICE OFFICER  POLICE  F       Salary  NULL    87006   NULL
ZYGMUNT   ARTUR POLICE OFFICER  POLICE  F       Salary  NULL    76266   NULL
ZYGMUNT   DAWID POLICE OFFICER  POLICE  F       Salary  NULL    80016   NULL
ZYLINSKA   KATARZYNA    POLICE OFFICER  POLICE  F       Salary  NULL    80016   NULL
ZYLINSKA   KLAUDIA      POLICE OFFICER  POLICE  F       Salary  NULL    72510   NULL
ZYMANTAS   LAURA C      POLICE OFFICER  POLICE  F       Salary  NULL    76266   NULL
ZYMANTAS   MARK E       POLICE OFFICER  POLICE  F       Salary  NULL    90024   NULL
ZYRKOWSKI   CARLO E     POLICE OFFICER  POLICE  F       Salary  NULL    93354   NULL
ZYSKOWSKI   DARIUSZ     CHIEF DATA BASE ANALYST DAIS    F       Salary  NULL    132360  NULL
Time taken: 0.078 seconds, Fetched: 32929 row(s)
```

2) Create a department-data-hive table by selecting unique department names from the employee-data-hive,
and adding a column named deptID in the new department-data-hive table,
and put unique values in the deptID column.

Alternatively, you can pre-process the employee-data
and select the unique department names,
add DeptID column

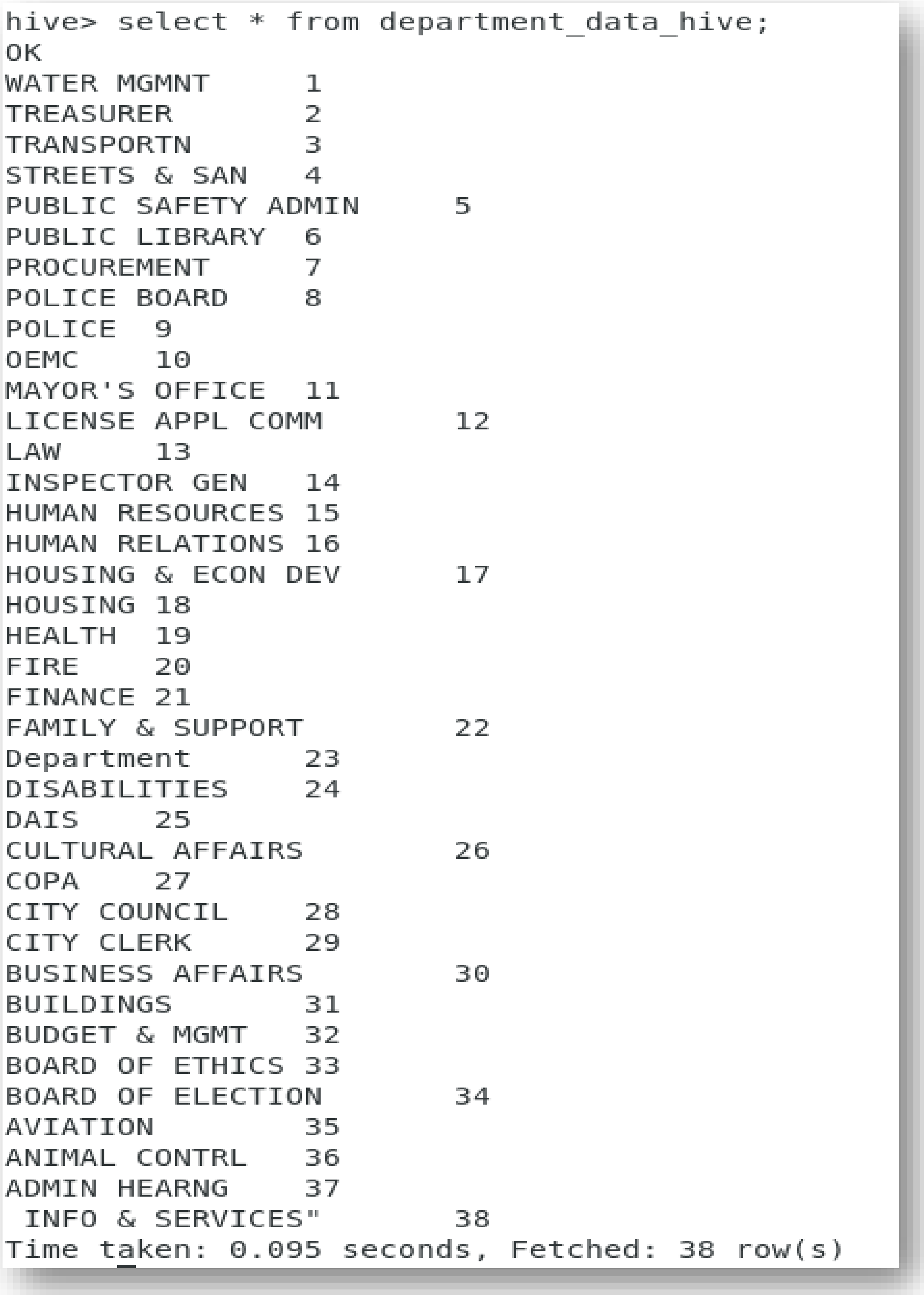
and assign unique value in the new Colum using excel or mySQL database separately, and then consider this structure (depart-name, DeptID) to create the department-data-hive table.

At first we created ‘department_data_hive’ table with selecting unique department names from the ‘employee-data-hive’ as its ‘depart-name’ column & adding the ‘deptID’ column.



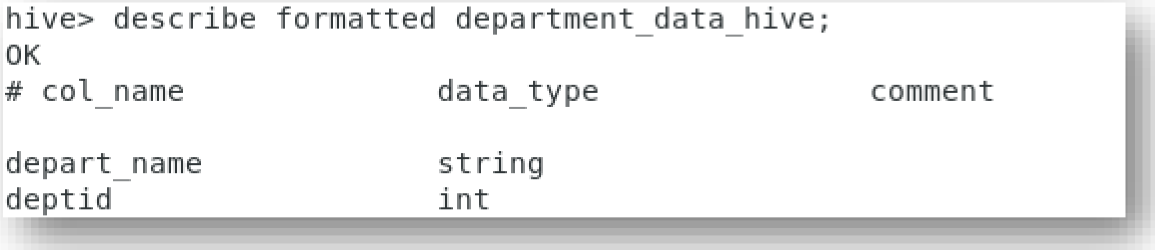
```
hive> create table department_data_hive
    > as
    > select depart_name, row_number() over() as deptid from (select distinct department as depart_name from employee_data_hive) as main;
Query ID = osboxes_20220704050949_4ae56104-04ac-4786-982b-47ede75d5d8c
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
22/07/04 05:09:49 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
22/07/04 05:09:49 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
Starting Job = job_1656878202432_0031, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1656878202432_0031/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1656878202432_0031
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-07-04 05:09:56,058 Stage-1 map = 0%,  reduce = 0%
2022-07-04 05:10:08,574 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.85 sec
2022-07-04 05:10:12,667 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.22 sec
MapReduce Total cumulative CPU time: 8 seconds 220 msec
Ended Job = job_1656878202432_0031
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
22/07/04 05:10:15 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
22/07/04 05:10:15 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
Starting Job = job_1656878202432_0032, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1656878202432_0032/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1656878202432_0032
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-07-04 05:10:23,178 Stage-2 map = 0%,  reduce = 0%
2022-07-04 05:10:29,463 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.06 sec
2022-07-04 05:10:35,664 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.91 sec
MapReduce Total cumulative CPU time: 4 seconds 910 msec
Ended Job = job_1656878202432_0032
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employee_db.db/department_data_hive
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.22 sec   HDFS Read: 2137916 HDFS Write: 1212 HDFS EC Read: 0 SUCCESS
```

Then we displayed the table content as following **(There were 38 unique department name fetched value)**.



```
hive> select * from department_data_hive;
OK
WATER MGMNT     1
TREASURER       2
TRANSPORTN      3
STREETS & SAN   4
PUBLIC SAFETY ADMIN     5
PUBLIC LIBRARY  6
PROCUREMENT     7
POLICE BOARD    8
POLICE  9
OEMC    10
MAYOR'S OFFICE  11
LICENSE APPL COMM       12
LAW     13
INSPECTOR GEN   14
HUMAN RESOURCES 15
HUMAN RELATIONS 16
HOUSING & ECON DEV      17
HOUSING 18
HEALTH  19
FIRE    20
FINANCE 21
FAMILY & SUPPORT        22
Department      23
DISABILITIES    24
DAIS    25
CULTURAL AFFAIRS        26
COPA    27
CITY COUNCIL    28
CITY CLERK      29
BUSINESS AFFAIRS        30
BUILDINGS       31
BUDGET & MGMT   32
BOARD OF ETHICS 33
BOARD OF ELECTION       34
AVIATION        35
ANIMAL CONTRL   36
ADMIN HEARNG    37
 INFO & SERVICES"       38
Time taken: 0.095 seconds, Fetched: 38 row(s)
```

To made sure from table structure, we used ‘describe’ command.

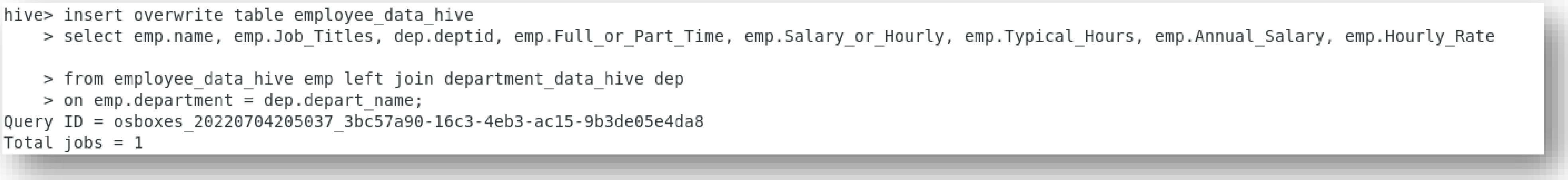


```
hive> describe formatted department_data_hive;
OK
# col_name              data_type               comment

depart_name             string
deptid                  int
```

3)Update the employee-data-hive table by replacing the department field data with the deptID values as created in the department-data-hive table.

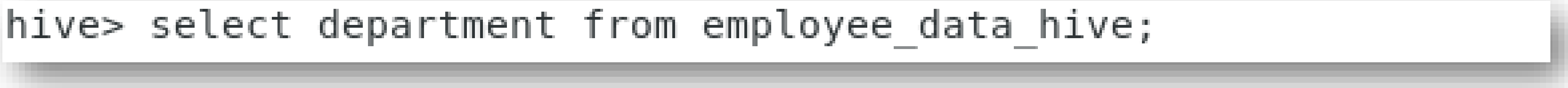
Also update the employee-data-hive table 'annual salary' field based on the 'Typical Hours' * 'Hourly Rate' * 52 if the annual salary field is empty.

At first we updated the ‘employee-data-hive’ table by replacing the department field data with the deptID values as created in the department-data-hive table.

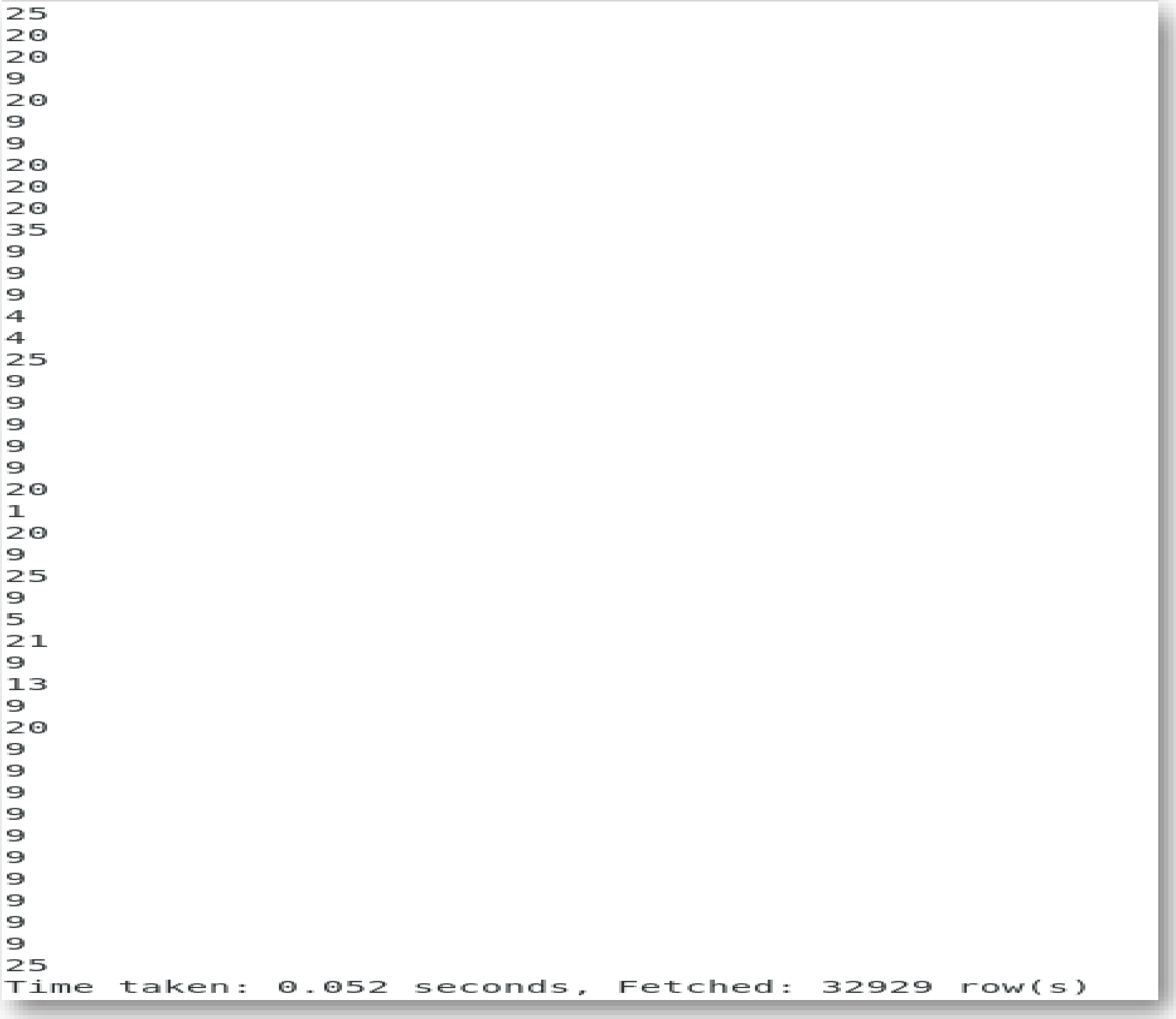


```
hive> insert overwrite table employee_data_hive
    > select emp.name, emp.Job_Titles, dep.deptid, emp.Full_or_Part_Time, emp.Salary_or_Hourly, emp.Typical_Hours, emp.Annual_Salary, emp.Hourly_Rate

    > from employee_data_hive emp left join department_data_hive dep
    > on emp.department = dep.depart_name;
Query ID = osboxes_20220704205037_3bc57a90-16c3-4eb3-ac15-9b3de05e4da8
Total jobs = 1
```

To made sure from the above command, we displayed the ‘employee-data-hive’ table updated column ‘department’.

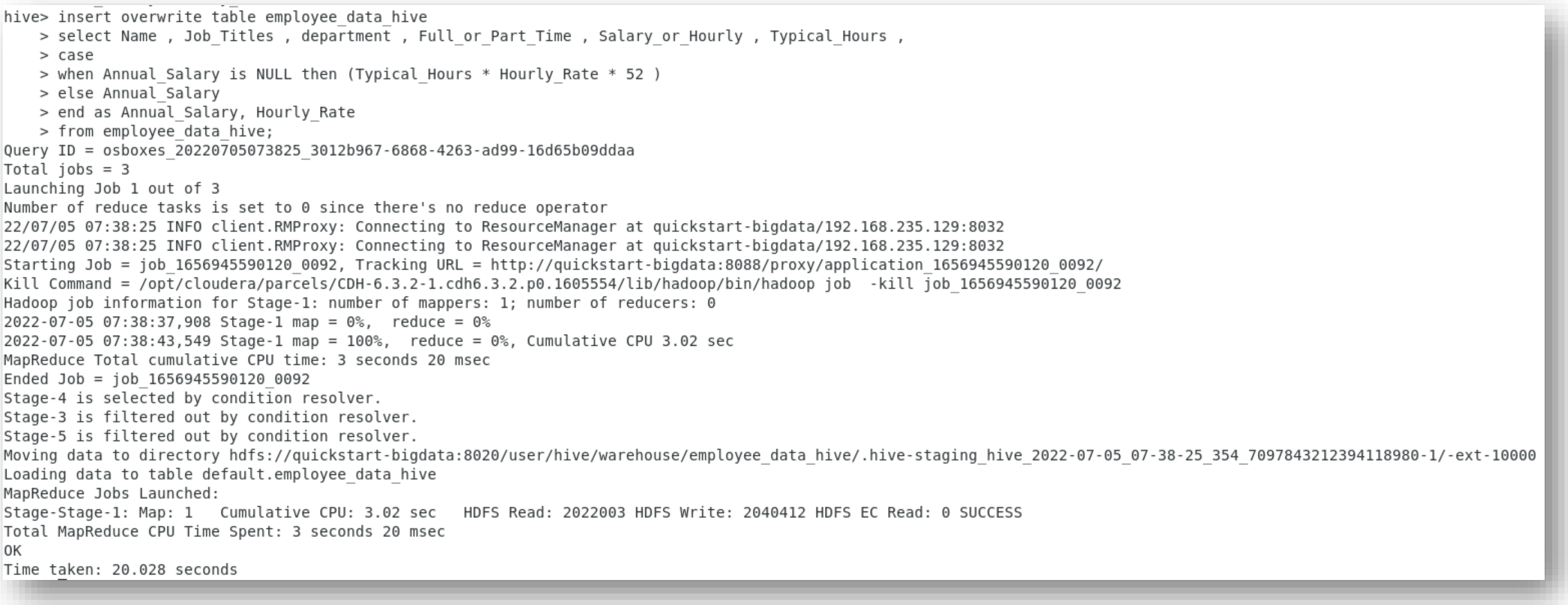


```
hive> select department from employee_data_hive;
```



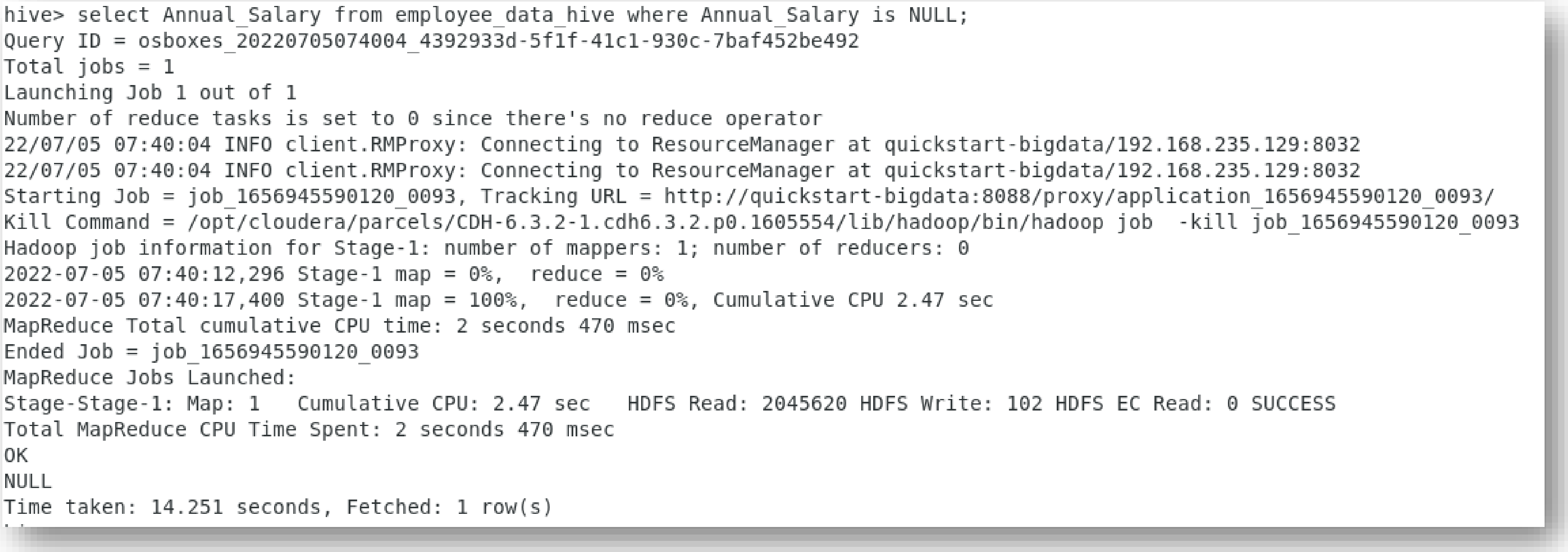
```
25
20
20
9
20
9
9
20
20
20
35
9
9
9
4
4
25
9
9
9
9
9
20
1
20
9
25
9
5
21
9
13
9
20
9
9
9
9
9
9
9
9
9
9
25
Time taken: 0.052 seconds, Fetched: 32929 row(s)
```

Then we updated the ‘employee-data-hive’ table 'annual salary' field based on the 'Typical Hours' * 'Hourly Rate' * 52 if the annual salary field is empty using the following commands.



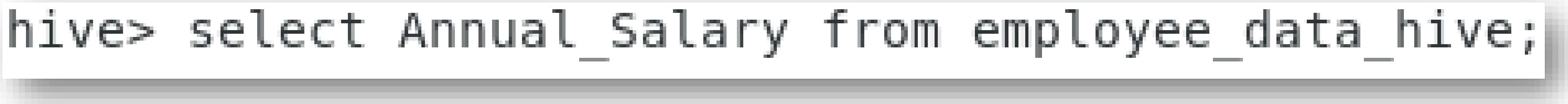
```
hive> insert overwrite table employee_data_hive
    > select Name , Job_Titles , department , Full_or_Part_Time , Salary_or_Hourly , Typical_Hours ,
    > case
    > when Annual_Salary is NULL then (Typical_Hours * Hourly_Rate * 52 )
    > else Annual_Salary
    > end as Annual_Salary, Hourly_Rate
    > from employee_data_hive;
Query ID = osboxes_20220705073825_3012b967-6868-4263-ad99-16d65b09ddaa
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
22/07/05 07:38:25 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
22/07/05 07:38:25 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
Starting Job = job_1656945590120_0092, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1656945590120_0092/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1656945590120_0092
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-07-05 07:38:37,908 Stage-1 map = 0%,  reduce = 0%
2022-07-05 07:38:43,549 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.02 sec
MapReduce Total cumulative CPU time: 3 seconds 20 msec
Ended Job = job_1656945590120_0092
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employee_data_hive/.hive-staging_hive_2022-07-05_07-38-25_354_7097843212394118980-1/-ext-10000
Loading data to table default.employee_data_hive
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.02 sec   HDFS Read: 2022003 HDFS Write: 2040412 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 20 msec
OK
Time taken: 20.028 seconds
```

Check if there exist any null values (it did not find any null values).

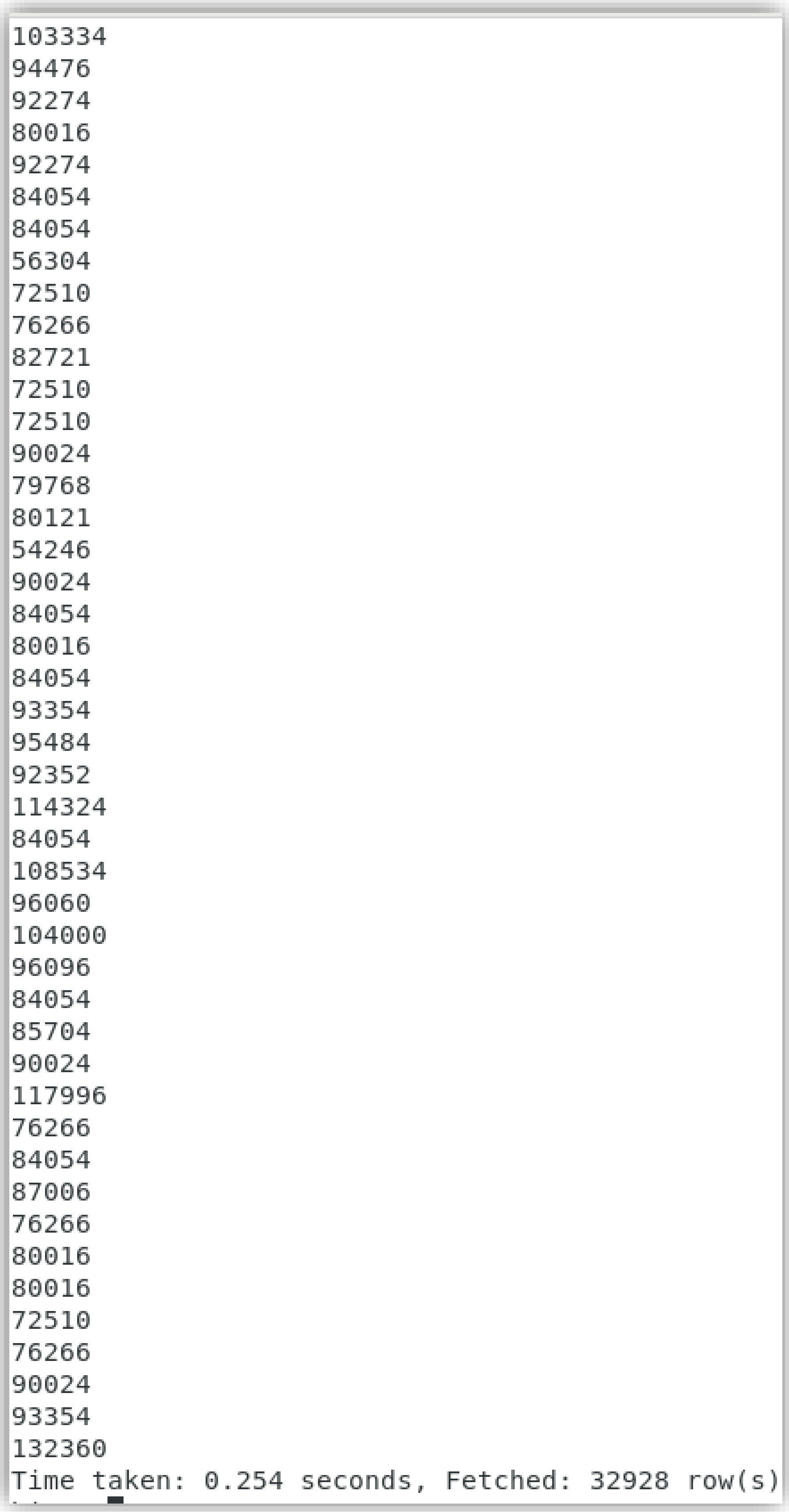


```
hive> select Annual_Salary from employee_data_hive where Annual_Salary is NULL;
Query ID = osboxes_20220705074004_4392933d-5f1f-41c1-930c-7baf452be492
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
22/07/05 07:40:04 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
22/07/05 07:40:04 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
Starting Job = job_1656945590120_0093, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1656945590120_0093/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1656945590120_0093
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-07-05 07:40:12,296 Stage-1 map = 0%,  reduce = 0%
2022-07-05 07:40:17,400 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.47 sec
MapReduce Total cumulative CPU time: 2 seconds 470 msec
Ended Job = job_1656945590120_0093
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.47 sec   HDFS Read: 2045620 HDFS Write: 102 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 470 msec
OK
NULL
Time taken: 14.251 seconds, Fetched: 1 row(s)
```

Check the success of updating the 'annual salary' null fields and displaying the column.



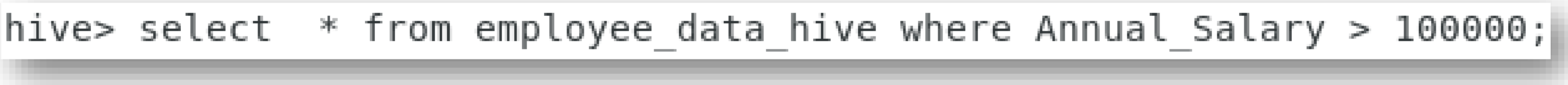
```
hive> select Annual_Salary from employee_data_hive;
```



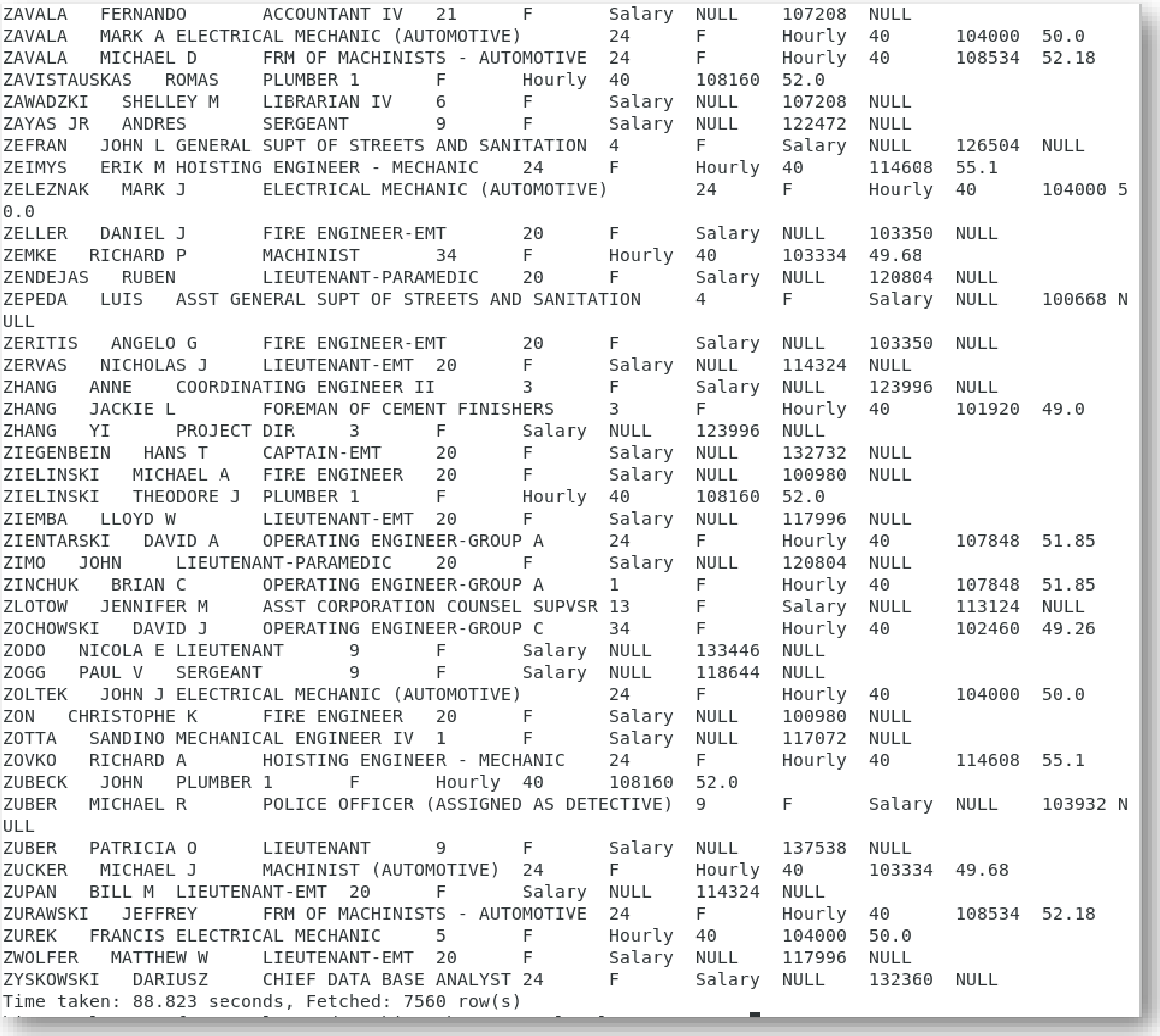
```
103334
94476
92274
80016
92274
84054
84054
56304
72510
76266
82721
72510
72510
90024
79768
80121
54246
90024
84054
80016
84054
93354
95484
92352
114324
84054
108534
96060
104000
96096
84054
85704
90024
117996
76266
84054
87006
76266
80016
80016
72510
76266
90024
93354
132360
Time taken: 0.254 seconds, Fetched: 32928 row(s)
```

4) Display all employees list with salary more than $100,000 based on employee-data-hive table. Also join the ‘employee-data-hive’ and ‘department-data-hive’ table to show the average salary of employees by department name.

At first we displayed all employees list with salary more than $100,000 based on ‘employee-data-hive’ table **(There were 7560 fetched value)**.

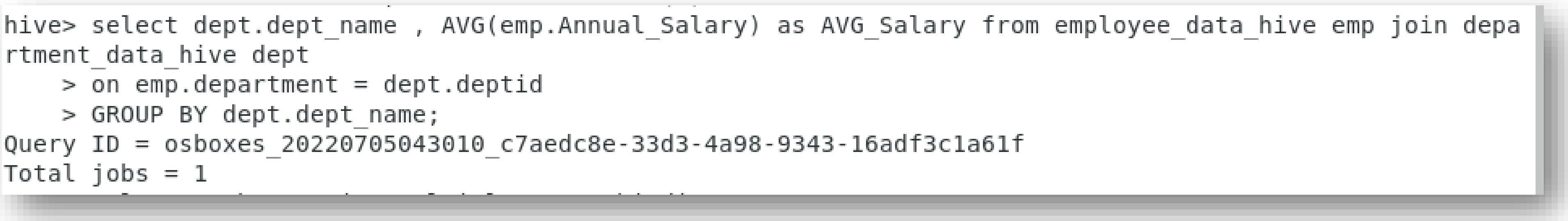


```
hive> select  * from employee_data_hive where Annual_Salary > 100000;
```

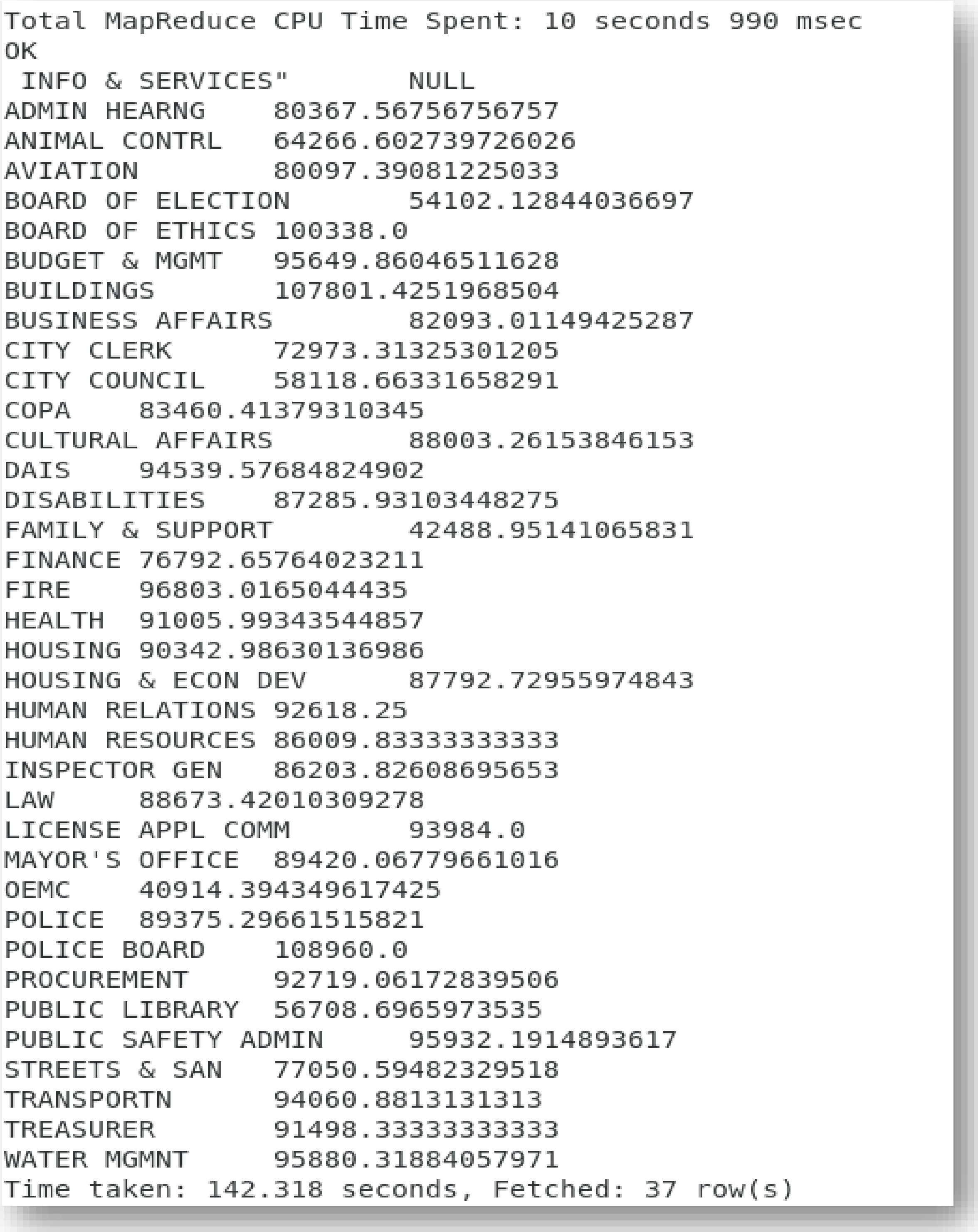


```
ZAVALA   FERNANDO       ACCOUNTANT IV   21      F       Salary  NULL    107208  NULL
ZAVALA   MARK A ELECTRICAL MECHANIC (AUTOMOTIVE)        24      F       Hourly  40      104000  50.0
ZAVALA   MICHAEL D      FRM OF MACHINISTS - AUTOMOTIVE  24      F       Hourly  40      108534  52.18
ZAVISTAUSKAS   ROMAS    PLUMBER 1       F       Hourly  40      108160  52.0
ZAWADZKI   SHELLEY M    LIBRARIAN IV    6       F       Salary  NULL    107208  NULL
ZAYAS JR   ANDRES       SERGEANT        9       F       Salary  NULL    122472  NULL
ZEFRAN   JOHN L GENERAL SUPT OF STREETS AND SANITATION  4       F       Salary  NULL    126504  NULL
ZEIMYS   ERIK M HOISTING ENGINEER - MECHANIC    24      F       Hourly  40      114608  55.1
ZELEZNAK   MARK J       ELECTRICAL MECHANIC (AUTOMOTIVE)        24      F       Hourly  40      104000 5
0.0
ZELLER   DANIEL J       FIRE ENGINEER-EMT       20      F       Salary  NULL    103350  NULL
ZEMKE   RICHARD P       MACHINIST       34      F       Hourly  40      103334  49.68
ZENDEJAS   RUBEN        LIEUTENANT-PARAMEDIC    20      F       Salary  NULL    120804  NULL
ZEPEDA   LUIS   ASST GENERAL SUPT OF STREETS AND SANITATION     4       F       Salary  NULL    100668 N
ULL
ZERITIS   ANGELO G      FIRE ENGINEER-EMT       20      F       Salary  NULL    103350  NULL
ZERVAS   NICHOLAS J     LIEUTENANT-EMT  20      F       Salary  NULL    114324  NULL
ZHANG   ANNE    COORDINATING ENGINEER II        3       F       Salary  NULL    123996  NULL
ZHANG   JACKIE L        FOREMAN OF CEMENT FINISHERS     3       F       Hourly  40      101920  49.0
ZHANG   YI      PROJECT DIR     3       F       Salary  NULL    123996  NULL
ZIEGENBEIN   HANS T     CAPTAIN-EMT     20      F       Salary  NULL    132732  NULL
ZIELINSKI   MICHAEL A   FIRE ENGINEER   20      F       Salary  NULL    100980  NULL
ZIELINSKI   THEODORE J  PLUMBER 1       F       Hourly  40      108160  52.0
ZIEMBA   LLOYD W        LIEUTENANT-EMT  20      F       Salary  NULL    117996  NULL
ZIENTARSKI   DAVID A    OPERATING ENGINEER-GROUP A      24      F       Hourly  40      107848  51.85
ZIMO   JOHN     LIEUTENANT-PARAMEDIC    20      F       Salary  NULL    120804  NULL
ZINCHUK   BRIAN C       OPERATING ENGINEER-GROUP A      1       F       Hourly  40      107848  51.85
ZLOTOW   JENNIFER M     ASST CORPORATION COUNSEL SUPVSR 13      F       Salary  NULL    113124  NULL
ZOCHOWSKI   DAVID J     OPERATING ENGINEER-GROUP C      34      F       Hourly  40      102460  49.26
ZODO   NICOLA E LIEUTENANT      9       F       Salary  NULL    133446  NULL
ZOGG   PAUL V   SERGEANT        9       F       Salary  NULL    118644  NULL
ZOLTEK   JOHN J ELECTRICAL MECHANIC (AUTOMOTIVE)        24      F       Hourly  40      104000  50.0
ZON   CHRISTOPHE K      FIRE ENGINEER   20      F       Salary  NULL    100980  NULL
ZOTTA   SANDINO MECHANICAL ENGINEER IV  1       F       Salary  NULL    117072  NULL
ZOVKO   RICHARD A       HOISTING ENGINEER - MECHANIC    24      F       Hourly  40      114608  55.1
ZUBECK   JOHN   PLUMBER 1       F       Hourly  40      108160  52.0
ZUBER   MICHAEL R       POLICE OFFICER (ASSIGNED AS DETECTIVE)  9       F       Salary  NULL    103932 N
ULL
ZUBER   PATRICIA O      LIEUTENANT      9       F       Salary  NULL    137538  NULL
ZUCKER   MICHAEL J      MACHINIST (AUTOMOTIVE)  24      F       Hourly  40      103334  49.68
ZUPAN   BILL M  LIEUTENANT-EMT  20      F       Salary  NULL    114324  NULL
ZURAWSKI   JEFFREY      FRM OF MACHINISTS - AUTOMOTIVE  24      F       Hourly  40      108534  52.18
ZUREK   FRANCIS ELECTRICAL MECHANIC     5       F       Hourly  40      104000  50.0
ZWOLFER   MATTHEW W     LIEUTENANT-EMT  20      F       Salary  NULL    117996  NULL
ZYSKOWSKI   DARIUSZ     CHIEF DATA BASE ANALYST 24      F       Salary  NULL    132360  NULL
Time taken: 88.823 seconds, Fetched: 7560 row(s)
```

Then we joined the ‘employee-data-hive’ and ‘department-data-hive’ table to show the average salary of employees by department name.



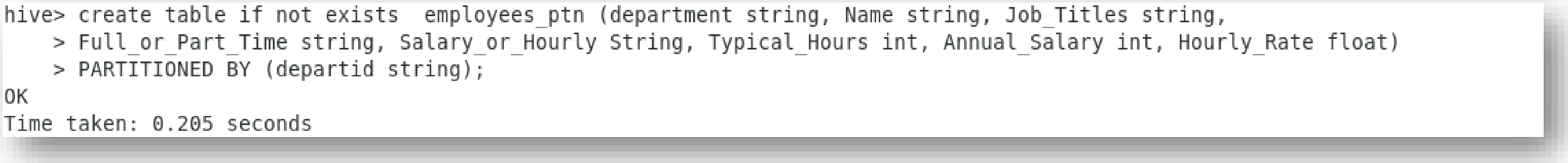
```
hive> select dept.dept_name , AVG(emp.Annual_Salary) as AVG_Salary from employee_data_hive emp join depa
rtment_data_hive dept
    > on emp.department = dept.deptid
    > GROUP BY dept.dept_name;
Query ID = osboxes_20220705043010_c7aedc8e-33d3-4a98-9343-16adf3c1a61f
Total jobs = 1
```



```
Total MapReduce CPU Time Spent: 10 seconds 990 msec
OK
 INFO & SERVICES"       NULL
ADMIN HEARNG    80367.56756756757
ANIMAL CONTRL   64266.602739726026
AVIATION        80097.39081225033
BOARD OF ELECTION       54102.12844036697
BOARD OF ETHICS 100338.0
BUDGET & MGMT   95649.86046511628
BUILDINGS       107801.4251968504
BUSINESS AFFAIRS        82093.01149425287
CITY CLERK      72973.31325301205
CITY COUNCIL    58118.66331658291
COPA    83460.41379310345
CULTURAL AFFAIRS        88003.26153846153
DAIS    94539.57684824902
DISABILITIES    87285.93103448275
FAMILY & SUPPORT        42488.95141065831
FINANCE 76792.65764023211
FIRE    96803.0165044435
HEALTH  91005.99343544857
HOUSING 90342.98630136986
HOUSING & ECON DEV      87792.72955974843
HUMAN RELATIONS 92618.25
HUMAN RESOURCES 86009.83333333333
INSPECTOR GEN   86203.82608695653
LAW     88673.42010309278
LICENSE APPL COMM       93984.0
MAYOR'S OFFICE  89420.06779661016
OEMC    40914.394349617425
POLICE  89375.29661515821
POLICE BOARD    108960.0
PROCUREMENT     92719.06172839506
PUBLIC LIBRARY  56708.6965973535
PUBLIC SAFETY ADMIN     95932.1914893617
STREETS & SAN   77050.59482329518
TRANSPORTN      94060.8813131313
TREASURER       91498.33333333333
WATER MGMNT     95880.31884057971
Time taken: 142.318 seconds, Fetched: 37 row(s)
```

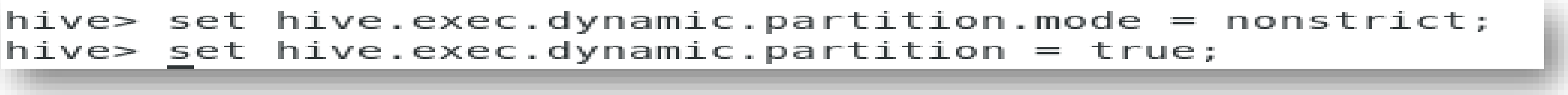
5) Create 5 partitions in an ‘employees_ptn’ table to store 5 departments in the appropriate partition. Display the partition structure.

At first we created the ‘employee-ptn’ table.



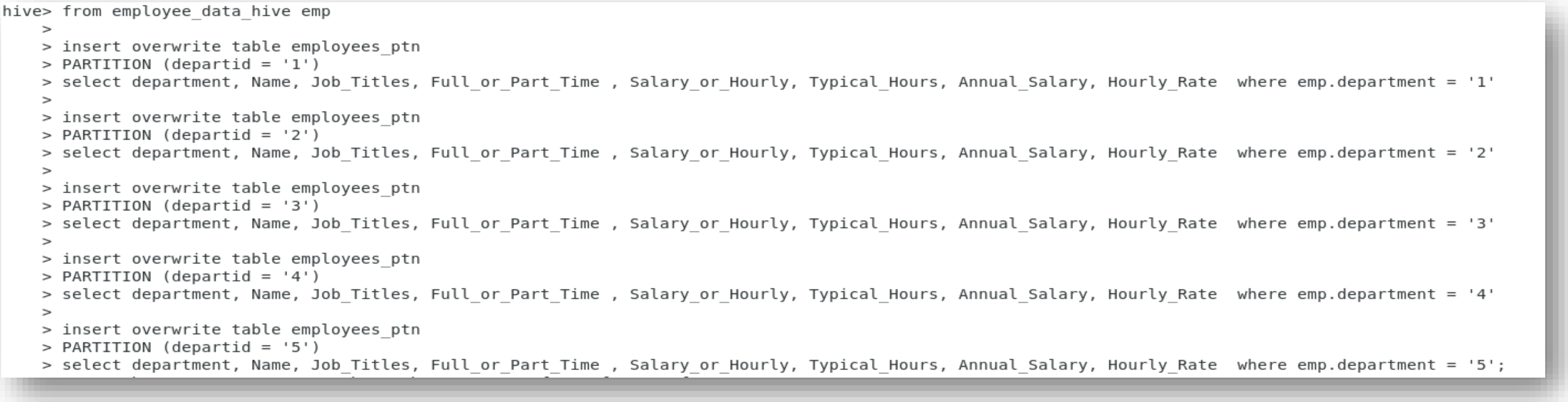
```
hive> create table if not exists  employees_ptn (department string, Name string, Job_Titles string,
    > Full_or_Part_Time string, Salary_or_Hourly String, Typical_Hours int, Annual_Salary int, Hourly_Rate float)
    > PARTITIONED BY (departid string);
OK
Time taken: 0.205 seconds
```

Then we set the mode of dynamic partition to ‘nonstrict’ mode.



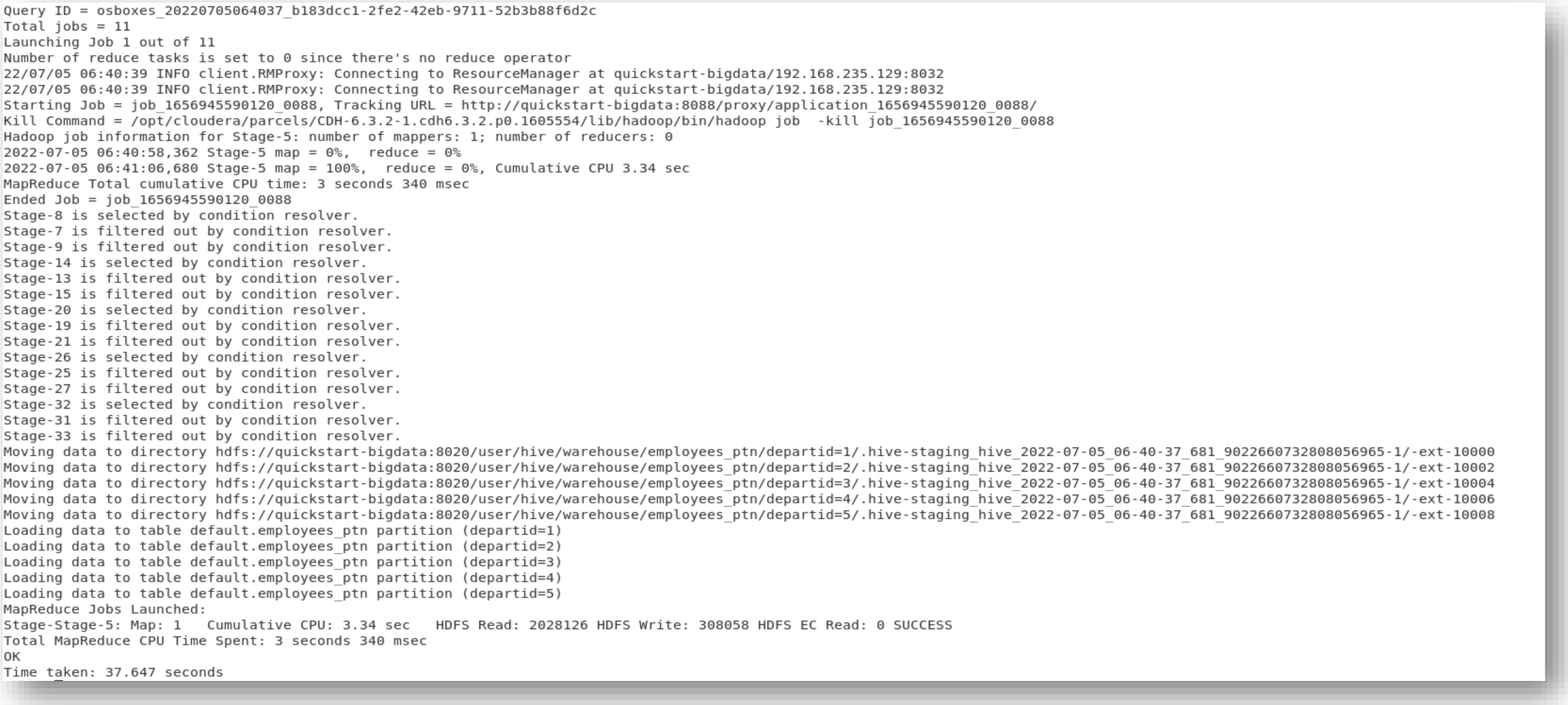
```
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> set hive.exec.dynamic.partition = true;
```

Then the step of creating 5 partitions in an ‘employees_ptn’ table to store 5 departments**.**



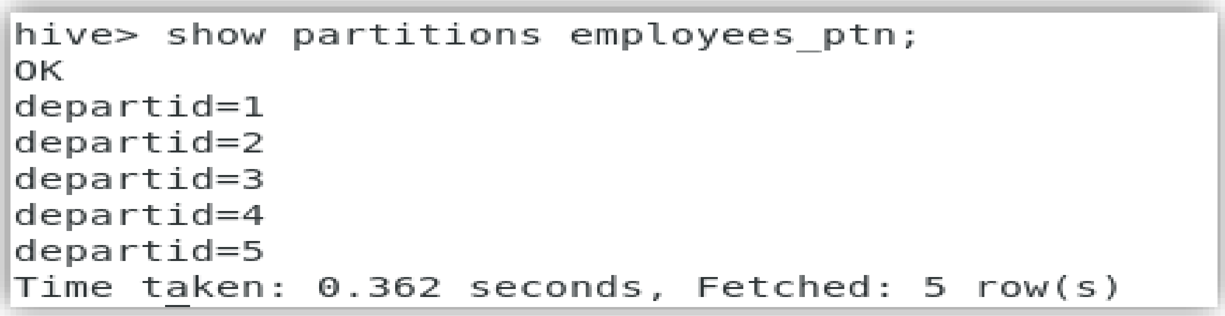
```
hive> from employee_data_hive emp
    >
    > insert overwrite table employees_ptn
    > PARTITION (departid = '1')
    > select department, Name, Job_Titles, Full_or_Part_Time , Salary_or_Hourly, Typical_Hours, Annual_Salary, Hourly_Rate  where emp.department = '1'
    >
    > insert overwrite table employees_ptn
    > PARTITION (departid = '2')
    > select department, Name, Job_Titles, Full_or_Part_Time , Salary_or_Hourly, Typical_Hours, Annual_Salary, Hourly_Rate  where emp.department = '2'
    >
    > insert overwrite table employees_ptn
    > PARTITION (departid = '3')
    > select department, Name, Job_Titles, Full_or_Part_Time , Salary_or_Hourly, Typical_Hours, Annual_Salary, Hourly_Rate  where emp.department = '3'
    >
    > insert overwrite table employees_ptn
    > PARTITION (departid = '4')
    > select department, Name, Job_Titles, Full_or_Part_Time , Salary_or_Hourly, Typical_Hours, Annual_Salary, Hourly_Rate  where emp.department = '4'
    >
    > insert overwrite table employees_ptn
    > PARTITION (departid = '5')
    > select department, Name, Job_Titles, Full_or_Part_Time , Salary_or_Hourly, Typical_Hours, Annual_Salary, Hourly_Rate  where emp.department = '5';
```

Here was the output.



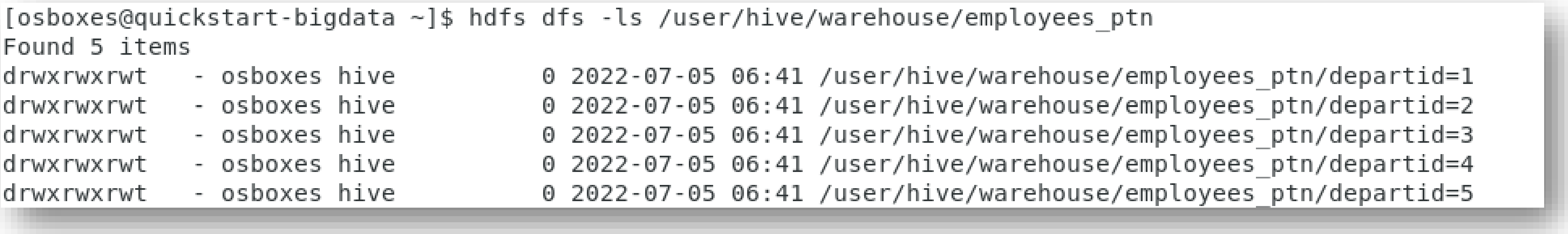
```
Query ID = osboxes_20220705064037_b183dcc1-2fe2-42eb-9711-52b3b88f6d2c
Total jobs = 11
Launching Job 1 out of 11
Number of reduce tasks is set to 0 since there's no reduce operator
22/07/05 06:40:39 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
22/07/05 06:40:39 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.235.129:8032
Starting Job = job_1656945590120_0088, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1656945590120_0088/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job  -kill job_1656945590120_0088
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 0
2022-07-05 06:40:58,362 Stage-5 map = 0%,  reduce = 0%
2022-07-05 06:41:06,680 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 3.34 sec
MapReduce Total cumulative CPU time: 3 seconds 340 msec
Ended Job = job_1656945590120_0088
Stage-8 is selected by condition resolver.
Stage-7 is filtered out by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-14 is selected by condition resolver.
Stage-13 is filtered out by condition resolver.
Stage-15 is filtered out by condition resolver.
Stage-20 is selected by condition resolver.
Stage-19 is filtered out by condition resolver.
Stage-21 is filtered out by condition resolver.
Stage-26 is selected by condition resolver.
Stage-25 is filtered out by condition resolver.
Stage-27 is filtered out by condition resolver.
Stage-32 is selected by condition resolver.
Stage-31 is filtered out by condition resolver.
Stage-33 is filtered out by condition resolver.
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employees_ptn/departid=1/.hive-staging_hive_2022-07-05_06-40-37_681_9022660732808056965-1/-ext-10000
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employees_ptn/departid=2/.hive-staging_hive_2022-07-05_06-40-37_681_9022660732808056965-1/-ext-10002
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employees_ptn/departid=3/.hive-staging_hive_2022-07-05_06-40-37_681_9022660732808056965-1/-ext-10004
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employees_ptn/departid=4/.hive-staging_hive_2022-07-05_06-40-37_681_9022660732808056965-1/-ext-10006
Moving data to directory hdfs://quickstart-bigdata:8020/user/hive/warehouse/employees_ptn/departid=5/.hive-staging_hive_2022-07-05_06-40-37_681_9022660732808056965-1/-ext-10008
Loading data to table default.employees_ptn partition (departid=1)
Loading data to table default.employees_ptn partition (departid=2)
Loading data to table default.employees_ptn partition (departid=3)
Loading data to table default.employees_ptn partition (departid=4)
Loading data to table default.employees_ptn partition (departid=5)
MapReduce Jobs Launched:
Stage-Stage-5: Map: 1   Cumulative CPU: 3.34 sec   HDFS Read: 2028126 HDFS Write: 308058 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 340 msec
OK
Time taken: 37.647 seconds
```

Here the list of 5 partitions.

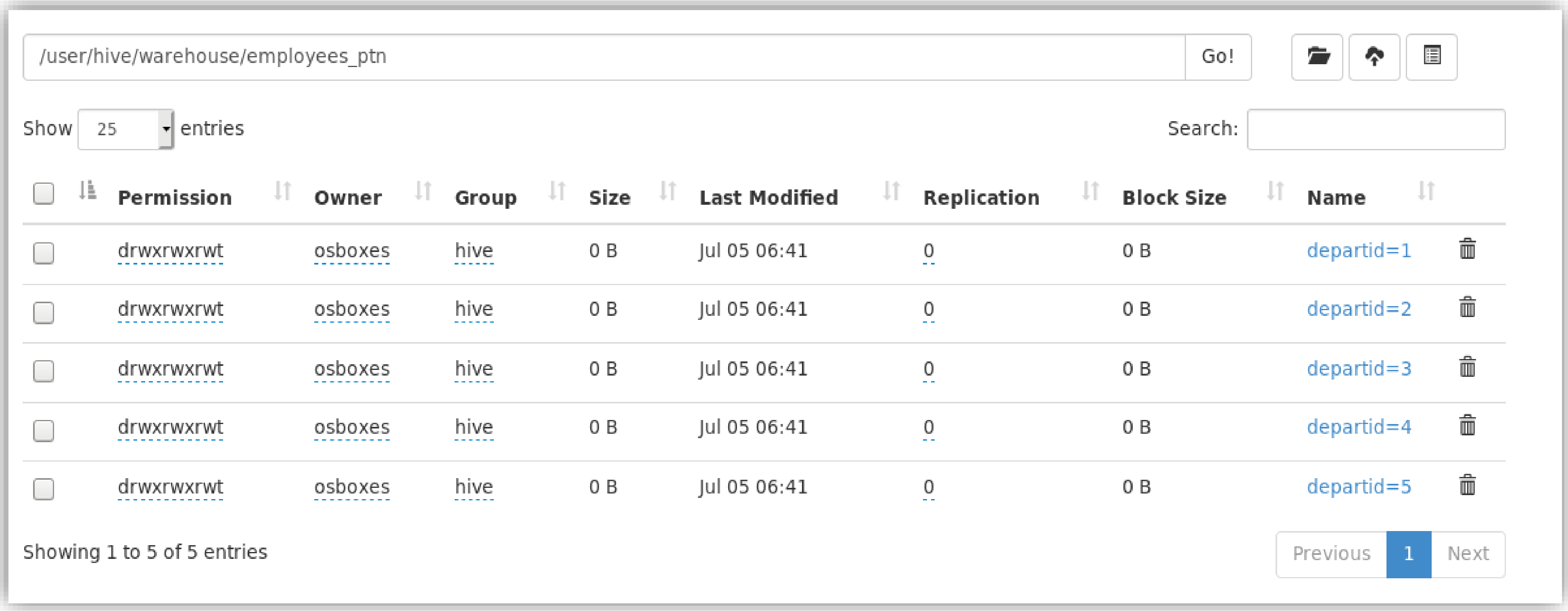


```
hive> show partitions employees_ptn;
OK
departid=1
departid=2
departid=3
departid=4
departid=5
Time taken: 0.362 seconds, Fetched: 5 row(s)
```

The list of 5 partitions in their path in hdfs.



```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/hive/warehouse/employees_ptn
Found 5 items
drwxrwxrwt   - osboxes hive          0 2022-07-05 06:41 /user/hive/warehouse/employees_ptn/departid=1
drwxrwxrwt   - osboxes hive          0 2022-07-05 06:41 /user/hive/warehouse/employees_ptn/departid=2
drwxrwxrwt   - osboxes hive          0 2022-07-05 06:41 /user/hive/warehouse/employees_ptn/departid=3
drwxrwxrwt   - osboxes hive          0 2022-07-05 06:41 /user/hive/warehouse/employees_ptn/departid=4
drwxrwxrwt   - osboxes hive          0 2022-07-05 06:41 /user/hive/warehouse/employees_ptn/departid=5
```



/user/hive/warehouse/employees_ptn Go!

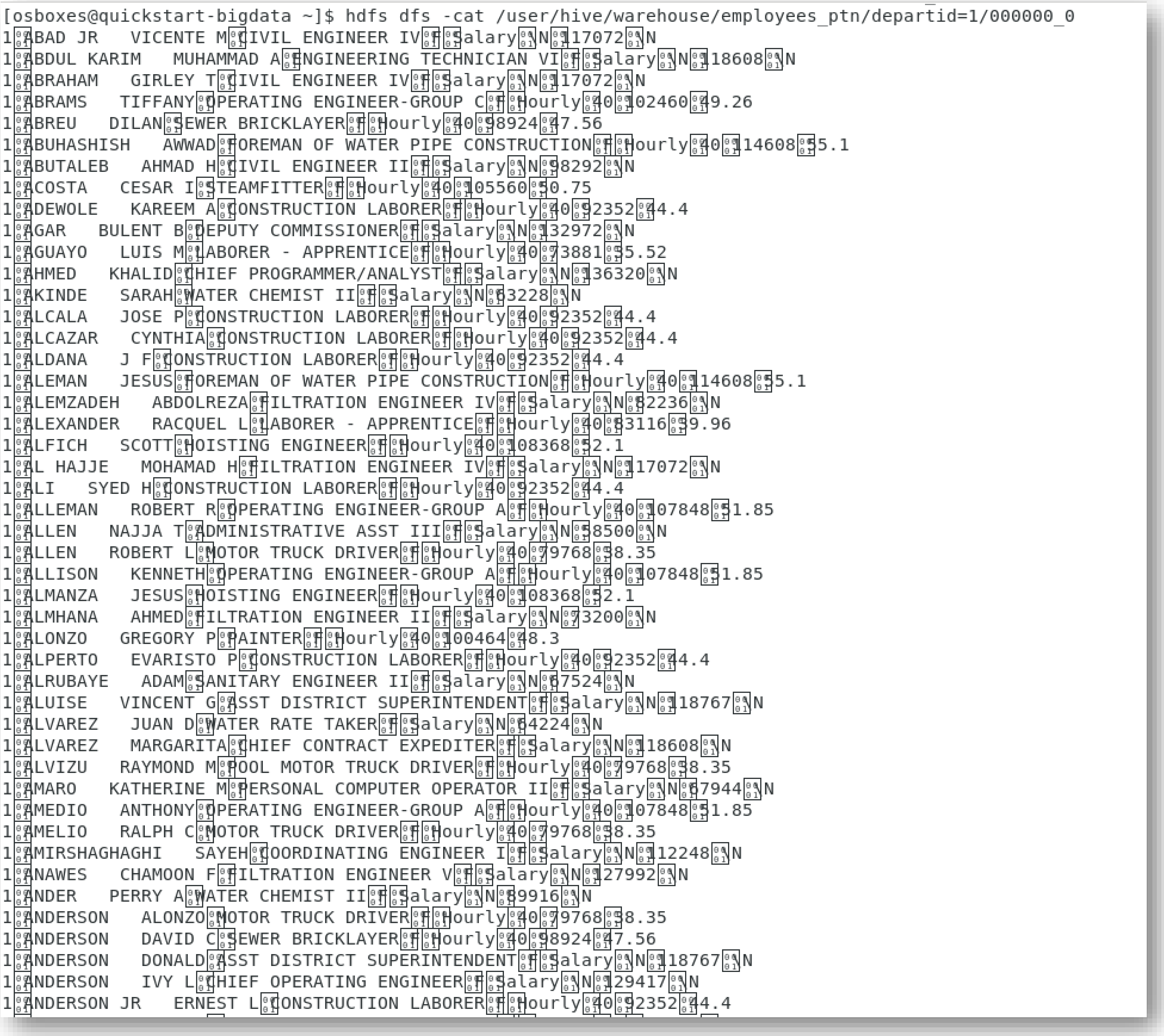
Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxrwxrwt	osboxes	hive	0 B	Jul 05 06:41	0	0 B	departid=1
□	drwxrwxrwt	osboxes	hive	0 B	Jul 05 06:41	0	0 B	departid=2
□	drwxrwxrwt	osboxes	hive	0 B	Jul 05 06:41	0	0 B	departid=3
□	drwxrwxrwt	osboxes	hive	0 B	Jul 05 06:41	0	0 B	departid=4
□	drwxrwxrwt	osboxes	hive	0 B	Jul 05 06:41	0	0 B	departid=5

Showing 1 to 5 of 5 entries

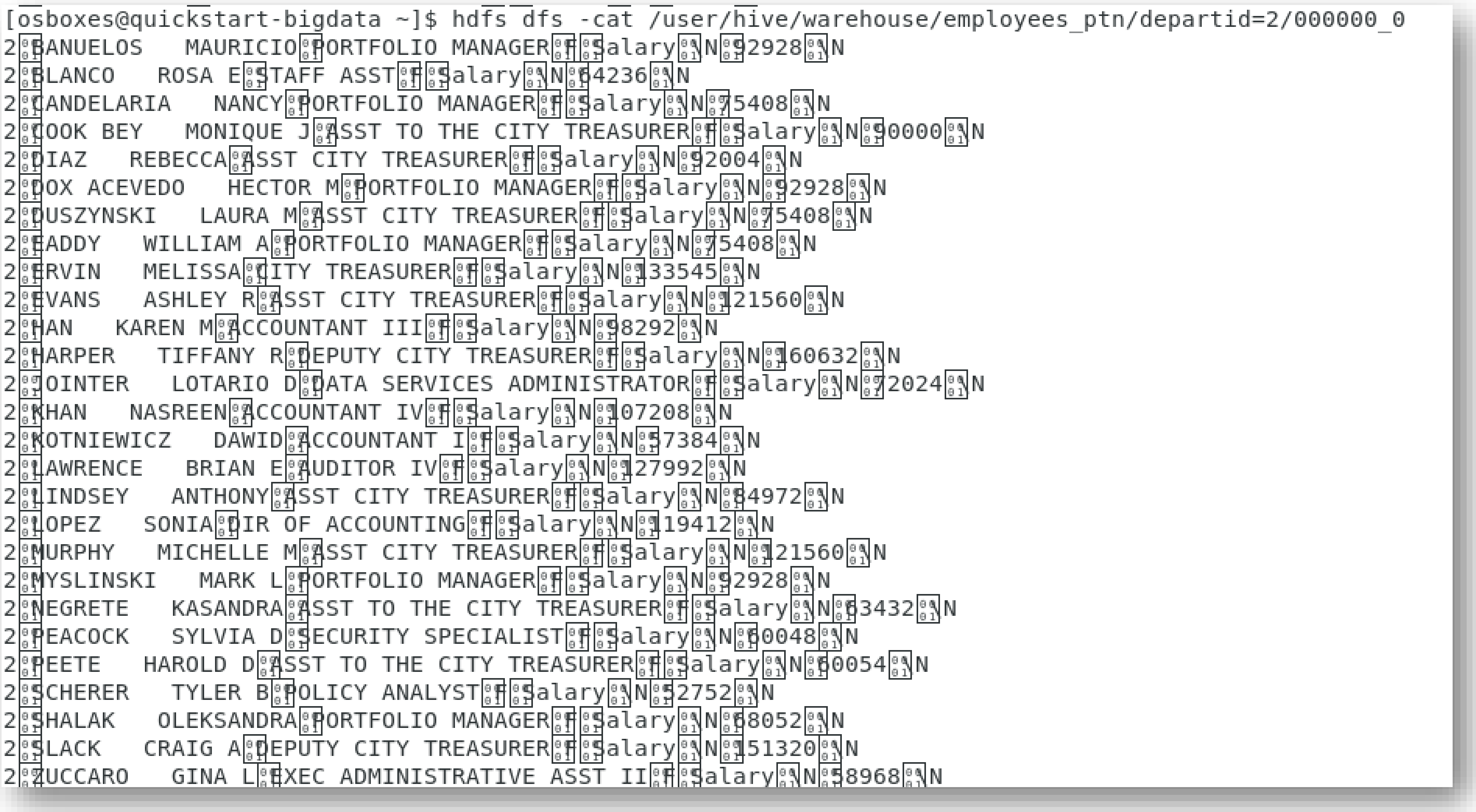
Previous 1 Next

The first partition ‘s output structure.



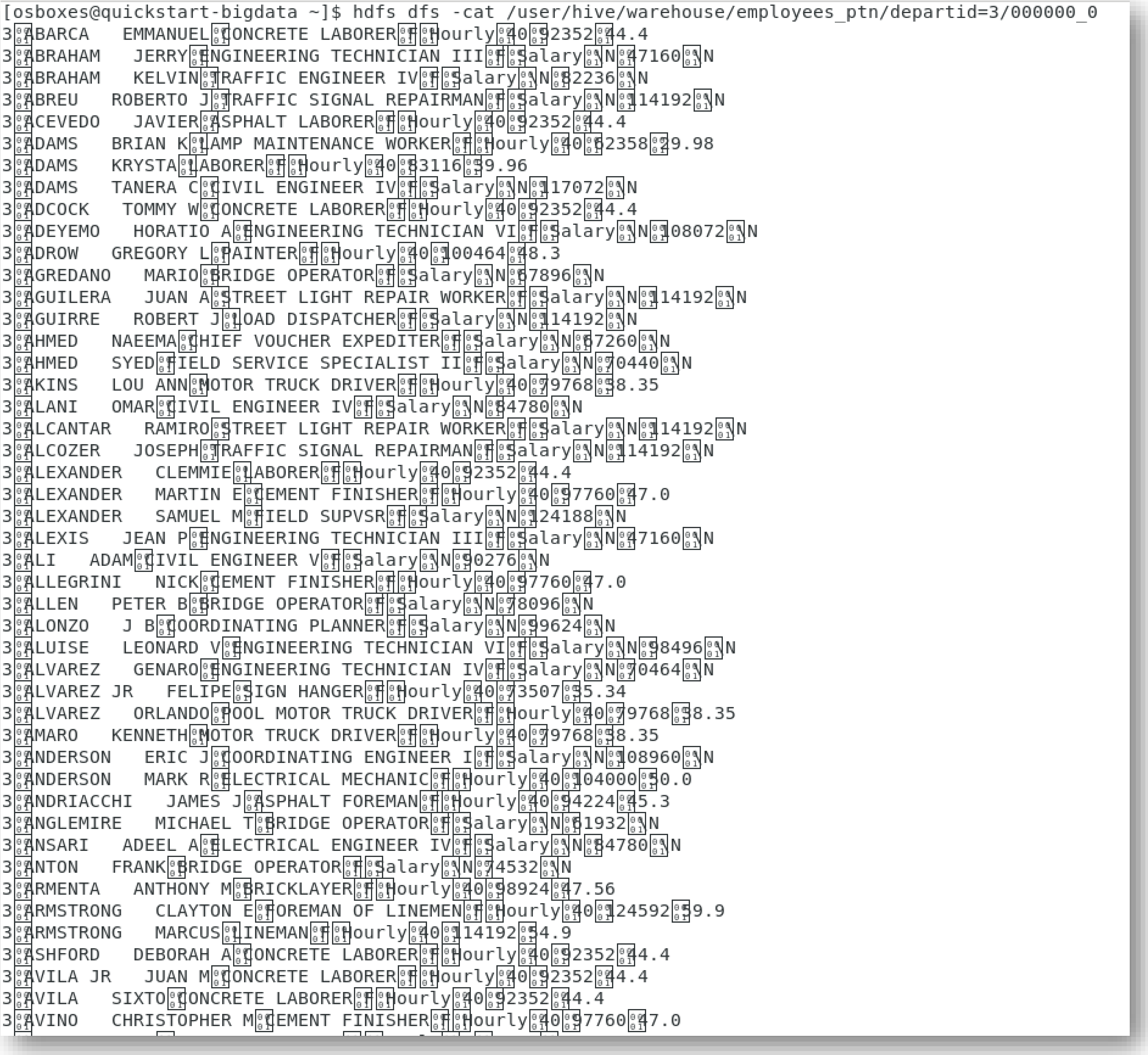
```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/hive/warehouse/employees_ptn/departid=1/000000_0
1▯ABAD JR   VICENTE M▯CIVIL ENGINEER IV▯F▯Salary▯\N▯117072▯\N
1▯ABDUL KARIM   MUHAMMAD A▯ENGINEERING TECHNICIAN VI▯F▯Salary▯\N▯118608▯\N
1▯ABRAHAM   GIRLEY T▯CIVIL ENGINEER IV▯F▯Salary▯\N▯117072▯\N
1▯ABRAMS   TIFFANY▯OPERATING ENGINEER-GROUP C▯F▯Hourly▯40▯102460▯49.26
1▯ABREU   DILAN▯SEWER BRICKLAYER▯F▯Hourly▯40▯98924▯47.56
1▯ABUHASHISH   AWWAD▯FOREMAN OF WATER PIPE CONSTRUCTION▯F▯Hourly▯40▯114608▯55.1
1▯ABUTALEB   AHMAD H▯CIVIL ENGINEER II▯F▯Salary▯\N▯98292▯\N
1▯ACOSTA   CESAR I▯STEAMFITTER▯F▯Hourly▯40▯105560▯50.75
1▯ADEWOLE   KAREEM A▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯AGAR   BULENT B▯DEPUTY COMMISSIONER▯F▯Salary▯\N▯132972▯\N
1▯AGUAYO   LUIS M▯LABORER - APPRENTICE▯F▯Hourly▯40▯73881▯35.52
1▯AHMED   KHALID▯CHIEF PROGRAMMER/ANALYST▯F▯Salary▯\N▯136320▯\N
1▯AKINDE   SARAH▯WATER CHEMIST II▯F▯Salary▯\N▯63228▯\N
1▯ALCALA   JOSE P▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯ALCAZAR   CYNTHIA▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯ALDANA   J F▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯ALEMAN   JESUS▯FOREMAN OF WATER PIPE CONSTRUCTION▯F▯Hourly▯40▯114608▯55.1
1▯ALEMZADEH   ABDOLREZA▯FILTRATION ENGINEER IV▯F▯Salary▯\N▯82236▯\N
1▯ALEXANDER   RACQUEL L▯LABORER - APPRENTICE▯F▯Hourly▯40▯83116▯39.96
1▯ALFICH   SCOTT▯HOISTING ENGINEER▯F▯Hourly▯40▯108368▯52.1
1▯AL HAJJE   MOHAMAD H▯FILTRATION ENGINEER IV▯F▯Salary▯\N▯117072▯\N
1▯ALI   SYED H▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯ALLEMAN   ROBERT R▯OPERATING ENGINEER-GROUP A▯F▯Hourly▯40▯107848▯51.85
1▯ALLEN   NAJJA T▯ADMINISTRATIVE ASST III▯F▯Salary▯\N▯58500▯\N
1▯ALLEN   ROBERT L▯MOTOR TRUCK DRIVER▯F▯Hourly▯40▯79768▯38.35
1▯ALLISON   KENNETH▯OPERATING ENGINEER-GROUP A▯F▯Hourly▯40▯107848▯51.85
1▯ALMANZA   JESUS▯HOISTING ENGINEER▯F▯Hourly▯40▯108368▯52.1
1▯ALMHANA   AHMED▯FILTRATION ENGINEER II▯F▯Salary▯\N▯73200▯\N
1▯ALONZO   GREGORY P▯PAINTER▯F▯Hourly▯40▯100464▯48.3
1▯ALPERTO   EVARISTO P▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
1▯ALRUBAYE   ADAM▯SANITARY ENGINEER II▯F▯Salary▯\N▯67524▯\N
1▯ALUISE   VINCENT G▯ASST DISTRICT SUPERINTENDENT▯F▯Salary▯\N▯118767▯\N
1▯ALVAREZ   JUAN D▯WATER RATE TAKER▯F▯Salary▯\N▯64224▯\N
1▯ALVAREZ   MARGARITA▯CHIEF CONTRACT EXPEDITER▯F▯Salary▯\N▯118608▯\N
1▯ALVIZU   RAYMOND M▯POOL MOTOR TRUCK DRIVER▯F▯Hourly▯40▯79768▯38.35
1▯AMARO   KATHERINE M▯PERSONAL COMPUTER OPERATOR II▯F▯Salary▯\N▯67944▯\N
1▯AMEDIO   ANTHONY▯OPERATING ENGINEER-GROUP A▯F▯Hourly▯40▯107848▯51.85
1▯AMELIO   RALPH C▯MOTOR TRUCK DRIVER▯F▯Hourly▯40▯79768▯38.35
1▯AMIRSHAGHAGHI   SAYEH▯COORDINATING ENGINEER I▯F▯Salary▯\N▯112248▯\N
1▯ANAWES   CHAMOON F▯FILTRATION ENGINEER V▯F▯Salary▯\N▯127992▯\N
1▯ANDER   PERRY A▯WATER CHEMIST II▯F▯Salary▯\N▯89916▯\N
1▯ANDERSON   ALONZO▯MOTOR TRUCK DRIVER▯F▯Hourly▯40▯79768▯38.35
1▯ANDERSON   DAVID C▯SEWER BRICKLAYER▯F▯Hourly▯40▯98924▯47.56
1▯ANDERSON   DONALD▯ASST DISTRICT SUPERINTENDENT▯F▯Salary▯\N▯118767▯\N
1▯ANDERSON   IVY L▯CHIEF OPERATING ENGINEER▯F▯Salary▯\N▯129417▯\N
1▯ANDERSON JR   ERNEST L▯CONSTRUCTION LABORER▯F▯Hourly▯40▯92352▯44.4
```

The second partition ‘s output structure.



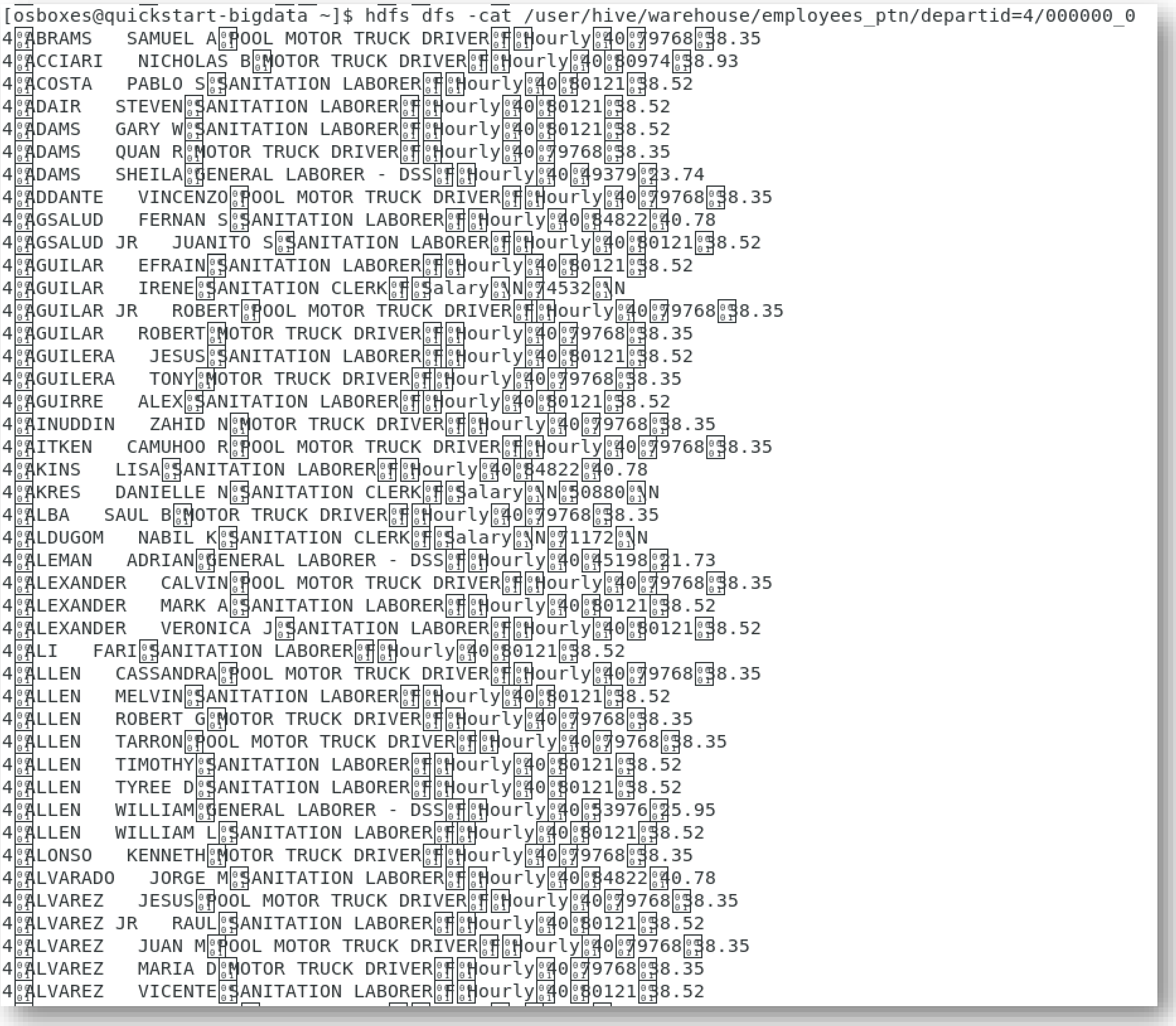
```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/hive/warehouse/employees_ptn/departid=2/000000_0
2␁BANUELOS   MAURICIO␁PORTFOLIO MANAGER␁F␁Salary␁\N␁92928␁\N
2␁BLANCO   ROSA E␁STAFF ASST␁F␁Salary␁\N␁64236␁\N
2␁CANDELARIA   NANCY␁PORTFOLIO MANAGER␁F␁Salary␁\N␁75408␁\N
2␁COOK BEY   MONIQUE J␁ASST TO THE CITY TREASURER␁F␁Salary␁\N␁90000␁\N
2␁DIAZ   REBECCA␁ASST CITY TREASURER␁F␁Salary␁\N␁92004␁\N
2␁DOX ACEVEDO   HECTOR M␁PORTFOLIO MANAGER␁F␁Salary␁\N␁92928␁\N
2␁DUSZYNSKI   LAURA M␁ASST CITY TREASURER␁F␁Salary␁\N␁75408␁\N
2␁EADDY   WILLIAM A␁PORTFOLIO MANAGER␁F␁Salary␁\N␁75408␁\N
2␁ERVIN   MELISSA␁CITY TREASURER␁F␁Salary␁\N␁133545␁\N
2␁EVANS   ASHLEY R␁ASST CITY TREASURER␁F␁Salary␁\N␁121560␁\N
2␁HAN   KAREN M␁ACCOUNTANT III␁F␁Salary␁\N␁98292␁\N
2␁HARPER   TIFFANY R␁DEPUTY CITY TREASURER␁F␁Salary␁\N␁160632␁\N
2␁JOINTER   LOTARIO D␁DATA SERVICES ADMINISTRATOR␁F␁Salary␁\N␁72024␁\N
2␁KHAN   NASREEN␁ACCOUNTANT IV␁F␁Salary␁\N␁107208␁\N
2␁KOTNIEWICZ   DAWID␁ACCOUNTANT I␁F␁Salary␁\N␁57384␁\N
2␁LAWRENCE   BRIAN E␁AUDITOR IV␁F␁Salary␁\N␁127992␁\N
2␁LINDSEY   ANTHONY␁ASST CITY TREASURER␁F␁Salary␁\N␁84972␁\N
2␁LOPEZ   SONIA␁DIR OF ACCOUNTING␁F␁Salary␁\N␁119412␁\N
2␁MURPHY   MICHELLE M␁ASST CITY TREASURER␁F␁Salary␁\N␁121560␁\N
2␁MYSLINSKI   MARK L␁PORTFOLIO MANAGER␁F␁Salary␁\N␁92928␁\N
2␁NEGRETE   KASANDRA␁ASST TO THE CITY TREASURER␁F␁Salary␁\N␁63432␁\N
2␁PEACOCK   SYLVIA D␁SECURITY SPECIALIST␁F␁Salary␁\N␁60048␁\N
2␁PEETE   HAROLD D␁ASST TO THE CITY TREASURER␁F␁Salary␁\N␁60054␁\N
2␁SCHERER   TYLER B␁POLICY ANALYST␁F␁Salary␁\N␁52752␁\N
2␁SHALAK   OLEKSANDRA␁PORTFOLIO MANAGER␁F␁Salary␁\N␁68052␁\N
2␁SLACK   CRAIG A␁DEPUTY CITY TREASURER␁F␁Salary␁\N␁151320␁\N
2␁ZUCCARO   GINA L␁EXEC ADMINISTRATIVE ASST II␁F␁Salary␁\N␁58968␁\N
```

The third partition ‘s output structure.



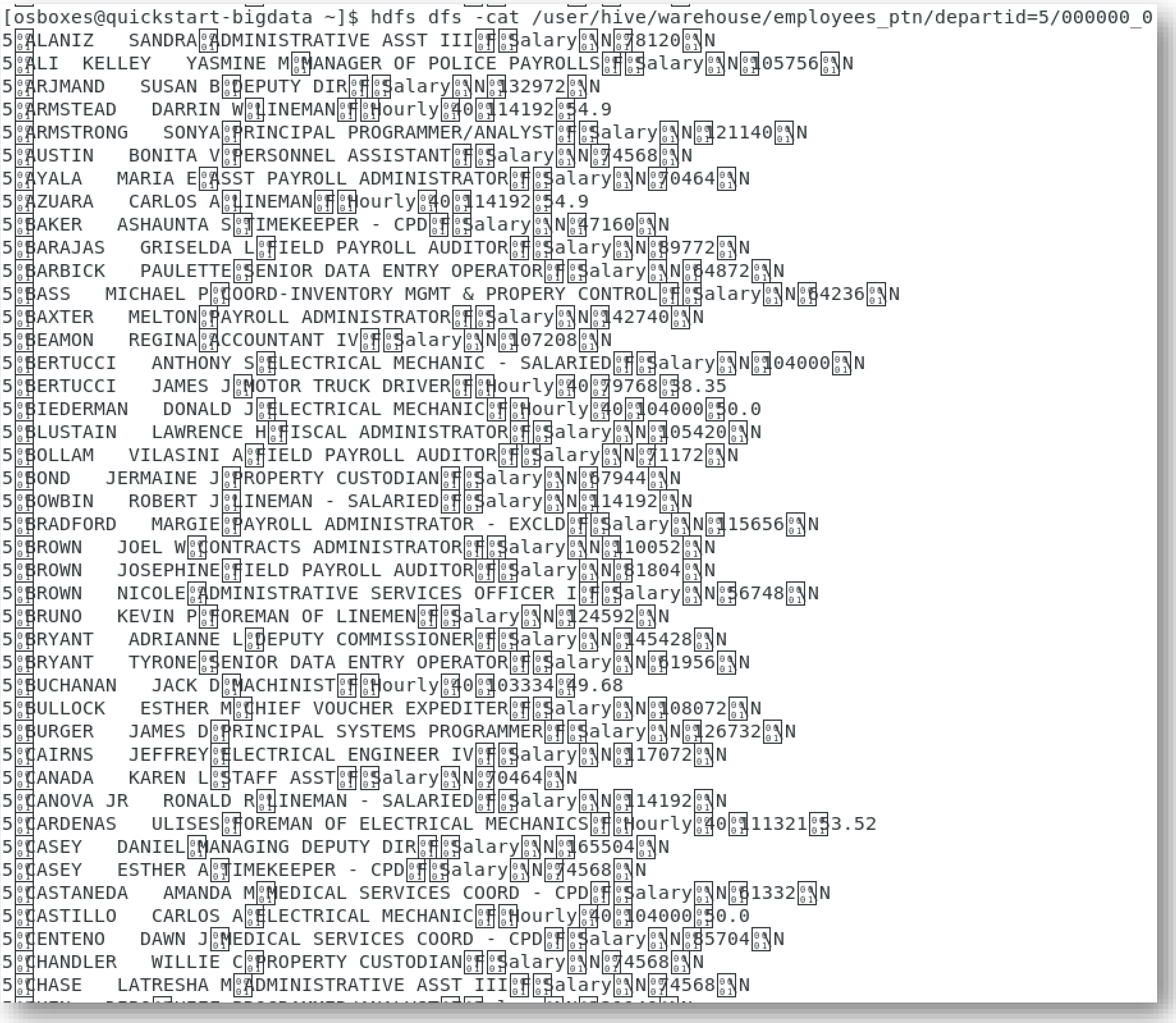
```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/hive/warehouse/employees_ptn/departid=3/000000_0
3␁ABARCA   EMMANUEL␁CONCRETE LABORER␁F␁Hourly␁40␁92352␁44.4
3␁ABRAHAM   JERRY␁ENGINEERING TECHNICIAN III␁F␁Salary␁\N␁47160␁\N
3␁ABRAHAM   KELVIN␁TRAFFIC ENGINEER IV␁F␁Salary␁\N␁82236␁\N
3␁ABREU   ROBERTO J␁TRAFFIC SIGNAL REPAIRMAN␁F␁Salary␁\N␁114192␁\N
3␁ACEVEDO   JAVIER␁ASPHALT LABORER␁F␁Hourly␁40␁92352␁44.4
3␁ADAMS   BRIAN K␁LAMP MAINTENANCE WORKER␁F␁Hourly␁40␁62358␁29.98
3␁ADAMS   KRYSTA␁LABORER␁F␁Hourly␁40␁83116␁39.96
3␁ADAMS   TANERA C␁CIVIL ENGINEER IV␁F␁Salary␁\N␁117072␁\N
3␁ADCOCK   TOMMY W␁CONCRETE LABORER␁F␁Hourly␁40␁92352␁44.4
3␁ADEYEMO   HORATIO A␁ENGINEERING TECHNICIAN VI␁F␁Salary␁\N␁108072␁\N
3␁ADROW   GREGORY L␁PAINTER␁F␁Hourly␁40␁100464␁48.3
3␁AGREDANO   MARIO␁BRIDGE OPERATOR␁F␁Salary␁\N␁67896␁\N
3␁AGUILERA   JUAN A␁STREET LIGHT REPAIR WORKER␁F␁Salary␁\N␁114192␁\N
3␁AGUIRRE   ROBERT J␁LOAD DISPATCHER␁F␁Salary␁\N␁114192␁\N
3␁AHMED   NAEEMA␁CHIEF VOUCHER EXPEDITER␁F␁Salary␁\N␁67260␁\N
3␁AHMED   SYED␁FIELD SERVICE SPECIALIST II␁F␁Salary␁\N␁70440␁\N
3␁AKINS   LOU ANN␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
3␁ALANI   OMAR␁CIVIL ENGINEER IV␁F␁Salary␁\N␁84780␁\N
3␁ALCANTAR   RAMIRO␁STREET LIGHT REPAIR WORKER␁F␁Salary␁\N␁114192␁\N
3␁ALCOZER   JOSEPH␁TRAFFIC SIGNAL REPAIRMAN␁F␁Salary␁\N␁114192␁\N
3␁ALEXANDER   CLEMMIE␁LABORER␁F␁Hourly␁40␁92352␁44.4
3␁ALEXANDER   MARTIN E␁CEMENT FINISHER␁F␁Hourly␁40␁97760␁47.0
3␁ALEXANDER   SAMUEL M␁FIELD SUPVSR␁F␁Salary␁\N␁124188␁\N
3␁ALEXIS   JEAN P␁ENGINEERING TECHNICIAN III␁F␁Salary␁\N␁47160␁\N
3␁ALI   ADAM␁CIVIL ENGINEER V␁F␁Salary␁\N␁90276␁\N
3␁ALLEGRINI   NICK␁CEMENT FINISHER␁F␁Hourly␁40␁97760␁47.0
3␁ALLEN   PETER B␁BRIDGE OPERATOR␁F␁Salary␁\N␁78096␁\N
3␁ALONZO   J B␁COORDINATING PLANNER␁F␁Salary␁\N␁99624␁\N
3␁ALUISE   LEONARD V␁ENGINEERING TECHNICIAN VI␁F␁Salary␁\N␁98496␁\N
3␁ALVAREZ   GENARO␁ENGINEERING TECHNICIAN IV␁F␁Salary␁\N␁70464␁\N
3␁ALVAREZ JR   FELIPE␁SIGN HANGER␁F␁Hourly␁40␁73507␁35.34
3␁ALVAREZ   ORLANDO␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
3␁AMARO   KENNETH␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
3␁ANDERSON   ERIC J␁COORDINATING ENGINEER I␁F␁Salary␁\N␁108960␁\N
3␁ANDERSON   MARK R␁ELECTRICAL MECHANIC␁F␁Hourly␁40␁104000␁50.0
3␁ANDRIACCHI   JAMES J␁ASPHALT FOREMAN␁F␁Hourly␁40␁94224␁45.3
3␁ANGLEMIRE   MICHAEL T␁BRIDGE OPERATOR␁F␁Salary␁\N␁61932␁\N
3␁ANSARI   ADEEL A␁ELECTRICAL ENGINEER IV␁F␁Salary␁\N␁84780␁\N
3␁ANTON   FRANK␁BRIDGE OPERATOR␁F␁Salary␁\N␁74532␁\N
3␁ARMENTA   ANTHONY M␁BRICKLAYER␁F␁Hourly␁40␁98924␁47.56
3␁ARMSTRONG   CLAYTON E␁FOREMAN OF LINEMEN␁F␁Hourly␁40␁124592␁59.9
3␁ARMSTRONG   MARCUS␁LINEMAN␁F␁Hourly␁40␁114192␁54.9
3␁ASHFORD   DEBORAH A␁CONCRETE LABORER␁F␁Hourly␁40␁92352␁44.4
3␁AVILA JR   JUAN M␁CONCRETE LABORER␁F␁Hourly␁40␁92352␁44.4
3␁AVILA   SIXTO␁CONCRETE LABORER␁F␁Hourly␁40␁92352␁44.4
3␁AVINO   CHRISTOPHER M␁CEMENT FINISHER␁F␁Hourly␁40␁97760␁47.0
```

The fourth partition ‘s output structure.



```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/hive/warehouse/employees_ptn/departid=4/000000_0
4␁ABRAMS   SAMUEL A␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ACCIARI   NICHOLAS B␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁80974␁38.93
4␁ACOSTA   PABLO S␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ADAIR   STEVEN␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ADAMS   GARY W␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ADAMS   QUAN R␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ADAMS   SHEILA␁GENERAL LABORER - DSS␁F␁Hourly␁40␁49379␁23.74
4␁ADDANTE   VINCENZO␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AGSALUD   FERNAN S␁SANITATION LABORER␁F␁Hourly␁40␁84822␁40.78
4␁AGSALUD JR   JUANITO S␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁AGUILAR   EFRAIN␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁AGUILAR   IRENE␁SANITATION CLERK␁F␁Salary␁\N␁74532␁\N
4␁AGUILAR JR   ROBERT␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AGUILAR   ROBERT␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AGUILERA   JESUS␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁AGUILERA   TONY␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AGUIRRE   ALEX␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁AINUDDIN   ZAHID N␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AITKEN   CAMUHOO R␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁AKINS   LISA␁SANITATION LABORER␁F␁Hourly␁40␁84822␁40.78
4␁AKRES   DANIELLE N␁SANITATION CLERK␁F␁Salary␁\N␁50880␁\N
4␁ALBA   SAUL B␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALDUGOM   NABIL K␁SANITATION CLERK␁F␁Salary␁\N␁71172␁\N
4␁ALEMAN   ADRIAN␁GENERAL LABORER - DSS␁F␁Hourly␁40␁45198␁21.73
4␁ALEXANDER   CALVIN␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALEXANDER   MARK A␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALEXANDER   VERONICA J␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALI   FARI␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALLEN   CASSANDRA␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALLEN   MELVIN␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALLEN   ROBERT G␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALLEN   TARRON␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALLEN   TIMOTHY␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALLEN   TYREE D␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALLEN   WILLIAM␁GENERAL LABORER - DSS␁F␁Hourly␁40␁53976␁25.95
4␁ALLEN   WILLIAM L␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALONSO   KENNETH␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALVARADO   JORGE M␁SANITATION LABORER␁F␁Hourly␁40␁84822␁40.78
4␁ALVAREZ   JESUS␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALVAREZ JR   RAUL␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
4␁ALVAREZ   JUAN M␁POOL MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALVAREZ   MARIA D␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
4␁ALVAREZ   VICENTE␁SANITATION LABORER␁F␁Hourly␁40␁80121␁38.52
```

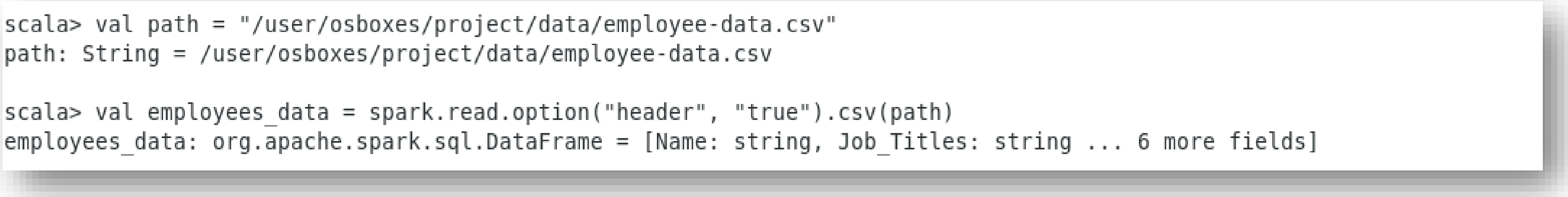
The fifth partition 's output structure.



```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/hive/warehouse/employees_ptn/departid=5/000000_0
5␁ALANIZ   SANDRA␁ADMINISTRATIVE ASST III␁F␁Salary␁N␁78120␁N
5␁ALI  KELLEY   YASMINE M␁MANAGER OF POLICE PAYROLLS␁F␁Salary␁N␁105756␁N
5␁ARJMAND   SUSAN B␁DEPUTY DIR␁F␁Salary␁N␁132972␁N
5␁ARMSTEAD   DARRIN W␁LINEMAN␁F␁Hourly␁40␁114192␁54.9
5␁ARMSTRONG   SONYA␁PRINCIPAL PROGRAMMER/ANALYST␁F␁Salary␁N␁121140␁N
5␁AUSTIN   BONITA V␁PERSONNEL ASSISTANT␁F␁Salary␁N␁74568␁N
5␁AYALA   MARIA E␁ASST PAYROLL ADMINISTRATOR␁F␁Salary␁N␁70464␁N
5␁AZUARA   CARLOS A␁LINEMAN␁F␁Hourly␁40␁114192␁54.9
5␁BAKER   ASHAUNTA S␁TIMEKEEPER - CPD␁F␁Salary␁N␁47160␁N
5␁BARAJAS   GRISELDA L␁FIELD PAYROLL AUDITOR␁F␁Salary␁N␁89772␁N
5␁BARBICK   PAULETTE␁SENIOR DATA ENTRY OPERATOR␁F␁Salary␁N␁64872␁N
5␁BASS   MICHAEL P␁COORD-INVENTORY MGMT & PROPERY CONTROL␁F␁Salary␁N␁64236␁N
5␁BAXTER   MELTON␁PAYROLL ADMINISTRATOR␁F␁Salary␁N␁142740␁N
5␁BEAMON   REGINA␁ACCOUNTANT IV␁F␁Salary␁N␁107208␁N
5␁BERTUCCI   ANTHONY S␁ELECTRICAL MECHANIC - SALARIED␁F␁Salary␁N␁104000␁N
5␁BERTUCCI   JAMES J␁MOTOR TRUCK DRIVER␁F␁Hourly␁40␁79768␁38.35
5␁BIEDERMAN   DONALD J␁ELECTRICAL MECHANIC␁F␁Hourly␁40␁104000␁50.0
5␁BLUSTAIN   LAWRENCE H␁FISCAL ADMINISTRATOR␁F␁Salary␁N␁105420␁N
5␁BOLLAM   VILASINI A␁FIELD PAYROLL AUDITOR␁F␁Salary␁N␁71172␁N
5␁BOND   JERMAINE J␁PROPERTY CUSTODIAN␁F␁Salary␁N␁67944␁N
5␁BOWBIN   ROBERT J␁LINEMAN - SALARIED␁F␁Salary␁N␁114192␁N
5␁BRADFORD   MARGIE␁PAYROLL ADMINISTRATOR - EXCLD␁F␁Salary␁N␁115656␁N
5␁BROWN   JOEL W␁CONTRACTS ADMINISTRATOR␁F␁Salary␁N␁110052␁N
5␁BROWN   JOSEPHINE␁FIELD PAYROLL AUDITOR␁F␁Salary␁N␁81804␁N
5␁BROWN   NICOLE␁ADMINISTRATIVE SERVICES OFFICER I␁F␁Salary␁N␁56748␁N
5␁BRUNO   KEVIN P␁FOREMAN OF LINEMEN␁F␁Salary␁N␁124592␁N
5␁BRYANT   ADRIANNE L␁DEPUTY COMMISSIONER␁F␁Salary␁N␁145428␁N
5␁BRYANT   TYRONE␁SENIOR DATA ENTRY OPERATOR␁F␁Salary␁N␁61956␁N
5␁BUCHANAN   JACK D␁MACHINIST␁F␁Hourly␁40␁103334␁49.68
5␁BULLOCK   ESTHER M␁CHIEF VOUCHER EXPEDITER␁F␁Salary␁N␁108072␁N
5␁BURGER   JAMES D␁PRINCIPAL SYSTEMS PROGRAMMER␁F␁Salary␁N␁126732␁N
5␁CAIRNS   JEFFREY␁ELECTRICAL ENGINEER IV␁F␁Salary␁N␁117072␁N
5␁CANADA   KAREN L␁STAFF ASST␁F␁Salary␁N␁70464␁N
5␁CANOVA JR   RONALD R␁LINEMAN - SALARIED␁F␁Salary␁N␁114192␁N
5␁CARDENAS   ULISES␁FOREMAN OF ELECTRICAL MECHANICS␁F␁Hourly␁40␁111321␁53.52
5␁CASEY   DANIEL␁MANAGING DEPUTY DIR␁F␁Salary␁N␁165504␁N
5␁CASEY   ESTHER A␁TIMEKEEPER - CPD␁F␁Salary␁N␁74568␁N
5␁CASTANEDA   AMANDA M␁MEDICAL SERVICES COORD - CPD␁F␁Salary␁N␁61332␁N
5␁CASTILLO   CARLOS A␁ELECTRICAL MECHANIC␁F␁Hourly␁40␁104000␁50.0
5␁CENTENO   DAWN J␁MEDICAL SERVICES COORD - CPD␁F␁Salary␁N␁85704␁N
5␁CHANDLER   WILLIE C␁PROPERTY CUSTODIAN␁F␁Salary␁N␁74568␁N
5␁CHASE   LATRESHA M␁ADMINISTRATIVE ASST III␁F␁Salary␁N␁74568␁N
```

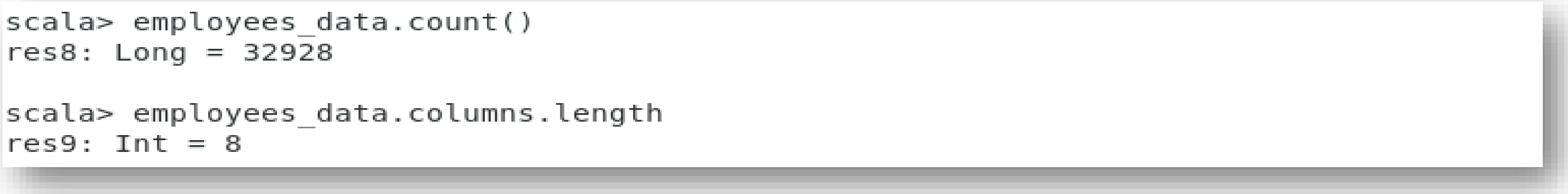
6) Create spark DataFrame based on the given dataset. Identify # of records in the DataFrame and show top 10 records.

At first we read the csv data file



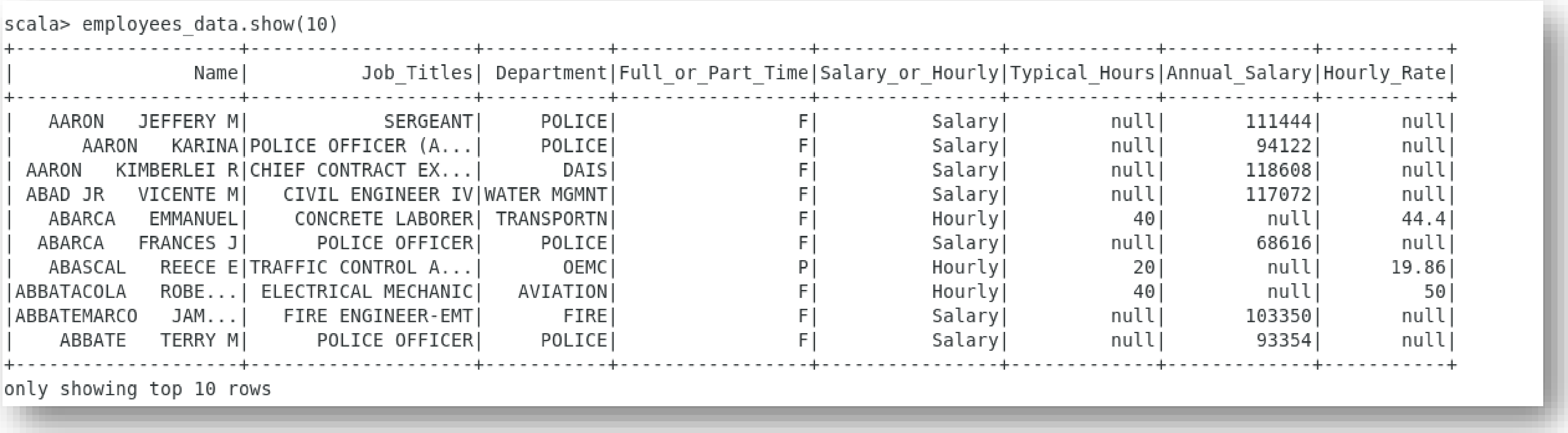
```
scala> val path = "/user/osboxes/project/data/employee-data.csv"
path: String = /user/osboxes/project/data/employee-data.csv

scala> val employees_data = spark.read.option("header", "true").csv(path)
employees_data: org.apache.spark.sql.DataFrame = [Name: string, Job_Titles: string ... 6 more fields]
```

Then we Identified # of records in the DataFrame using 'count' command.



```
scala> employees_data.count()
res8: Long = 32928

scala> employees_data.columns.length
res9: Int = 8
```

Then we displayed the top 10 records.



```
scala> employees_data.show(10)
+--------------------+--------------------+-----------+-----------------+----------------+-------------+-------------+-----------+
|                Name|          Job_Titles| Department|Full_or_Part_Time|Salary_or_Hourly|Typical_Hours|Annual_Salary|Hourly_Rate|
+--------------------+--------------------+-----------+-----------------+----------------+-------------+-------------+-----------+
|   AARON   JEFFERY M|            SERGEANT|     POLICE|                F|          Salary|         null|       111444|       null|
|      AARON   KARINA|POLICE OFFICER (A...|     POLICE|                F|          Salary|         null|        94122|       null|
| AARON   KIMBERLEI R|CHIEF CONTRACT EX...|       DAIS|                F|          Salary|         null|       118608|       null|
| ABAD JR   VICENTE M|   CIVIL ENGINEER IV|WATER MGMNT|                F|          Salary|         null|       117072|       null|
|   ABARCA   EMMANUEL|    CONCRETE LABORER| TRANSPORTN|                F|          Hourly|           40|         null|       44.4|
|  ABARCA   FRANCES J|      POLICE OFFICER|     POLICE|                F|          Salary|         null|        68616|       null|
|   ABASCAL   REECE E|TRAFFIC CONTROL A...|       OEMC|                P|          Hourly|           20|         null|      19.86|
|ABBATACOLA   ROBE...| ELECTRICAL MECHANIC|   AVIATION|                F|          Hourly|           40|         null|         50|
|ABBATEMARCO   JAM...|   FIRE ENGINEER-EMT|       FIRE|                F|          Salary|         null|       103350|       null|
|    ABBATE   TERRY M|      POLICE OFFICER|     POLICE|                F|          Salary|         null|        93354|       null|
+--------------------+--------------------+-----------+-----------------+----------------+-------------+-------------+-----------+
only showing top 10 rows
```

DONE BY:

- **Ahmed Ibrahim Salem (21aisa)**