Practical 07 Hadoop and Hive

Task 1:

First, I downloaded the Apache hive. Then I loaded up docker Hadoop folder I used in practical 6 in docker.

I then moved the compressed apache hive file into docker Hadoop folder. I then copied the apache hive file into namenode container.

```
C:\Users\waleed\Desktop\PRACTICAL7_Waleed_Wazir_19396951\ex1\docker-hadoop>docker cp apache-hive-3.1.3-bin.tar.gz namenode:/
```

I then went into the namenode container using the command:

```
C:\Users\waleed\Desktop\PRACTICAL7_Waleed_Wazir_19396951\ex1\docker-hadoop>docker exec -it namenode bash
root@a5286aaf2748:~#
```

uncompressed the Apache-hive file inside the namenode container using the command in the image below:

```
root@006134cf364a:/# tar -xvzf apache-hive-3.1.3-bin.tar.gz
apache-hive-3.1.3-bin/LICENSE
apache-hive-3.1.3-bin/RELEASE_NOTES.txt
apache-hive-3.1.3-bin/NOTICE
apache-hive-3.1.3-bin/binary-package-licenses/com.thoughtworks.paranamer-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.codehaus.janino-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.jamon.jamon-runtime-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.mozilla.rhino-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.jruby-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/jline-LICENSE
```

I then moved the the apache-hive file into bin directory inside namenode container so we can use Hive.

```
root@006134cf364a:/# mv apache-hive-3.1.3-bin bin
```

Checking if hive was moved properly.

```
root@006134cf364a:/bin# cd apache-hive-3.1.3-bin
root@006134cf364a:/bin/apache-hive-3.1.3-bin#
```

I then went into bin and removed the files using the command in the image below:

apache-hive-3.1.3-bin/jdbc/hive-jdbc-3.1.3-standalone.jar,
apache-hive-3.1.3-bin/lib/log4j-slf4j-impl-2.17.1.jar, and
apache-hive-3.1.3-bin/lib/guava-19.0.jar files

```
root@006134cf364a:/bin/apache-hive-3.1.3-bin# cd ..
root@006134cf364a:/bin# rm apache-hive-3.1.3-bin/jdbc/hive-jdbc-3.1.3-standalone.jar
root@006134cf364a:/bin# rm apache-hive-3.1.3-bin/lib/log4j-slf4j-impl-2.17.1.jar
root@006134cf364a:/bin# rm apache-hive-3.1.3-bin/lib/guava-19.0.jar
root@006134cf364a:/bin#
```

I then copied the file below into apache-hive-3.1.3-bin/lib.

/opt/hadoop-3.3.1/share/hadoop/hdfs/lib/guava-27.0-jre.jar file

```
root@006134cf364a:/# cp opt/hadoop-3.3.1/share/hadoop/hdfs/lib/guava-27.0-jre.jar /bin/apache-hive-3.1.3-bin/lib
```

I then cd into apache-hive-3.1.3-bin/bin and ran the command schematool -dbType derby -initSchema.

```
root@a5286aaf2748:/# cd bin
root@a5286aaf2748:/bin# cd apache-hive-3.1.3-bin
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin# cd bin
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin#
```

```
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin# ./schematool -dbType derby -initSchema

Metastore connection URL:        jdbc:derby:;databaseName=metastore_db;create=true
Metastore Connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:       APP
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.derby.sql
```

```
Initialization script completed
schemaTool completed
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin#
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin#
```

I then logged into hive using ./hive and created a new database called P7.

```
root@006134cf364a:/# ./hive
bash: ./hive: No such file or directory
root@006134cf364a:/# cd ..
root@006134cf364a:/# cd bin
root@006134cf364a:/bin# cd apache-hive-3.1.3-bin
root@006134cf364a:/bin/apache-hive-3.1.3-bin# cd bin
root@006134cf364a:/bin/apache-hive-3.1.3-bin/bin# ./hive
./hive: line 351: ps: command not found
./hive: line 351: ps: command not found
Hive Session ID = 812f4b61-0b45-405a-8273-50a8e5f33924

Logging initialized using configuration in jar:file:/bin/apache-hive-3.1.3-bin/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: trueHive-on-MR is deprecated i
d may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

```
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin# ./hive
./hive: line 351: ps: command not found
./hive: line 351: ps: command not found
Hive Session ID = 518da79a-1bd3-42f0-85d1-649043a0742a

Logging initialized using configuration in jar:file:/bin/apache-hive-3
perties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the fut
engine (i.e. spark, tez) or using Hive 1.X releases.Hive Session ID =

hive> CREATE DATABASE P7;
OK
Time taken: 0.497 seconds
```

When I used the command below we can see the database that I created is shown.

```
hive> SHOW DATABASES;
OK
default
p7
Time taken: 0.571 seconds, Fetched: 2 row(s)
hive>
```

Task 2:

I created two tab-delimited files from the given employee_data.xlsx as follows so that a relation can be established via CountyID to normalise the data into two tables.

I changed the county ID from 0 to 15, divided the original files into two, and saved the files in.txt format as required. This file is called "county" and represents 5 features, as shown below, that correspond with the county ID. This file contains both Dublin and Kildare counties.

File Edit Format View Help

County ID	County Name	Province	Area	Population
0	Dublin	Leinster	922	1.451
1	Dublin	Leinster	922	1.451
2	Kildare	Leinster	1695	0.247
3	Dublin	Leinster	922	1.451
4	Kildare	Leinster	1695	0.247
5	Kildare	Leinster	1695	0.247
6	Dublin	Leinster	922	1.451
7	Dublin	Leinster	922	1.451
8	Dublin	Leinster	922	1.451
9	Dublin	Leinster	922	1.451
10	Kildare	Leinster	1695	0.247
11	Kildare	Leinster	1695	0.247
12	Dublin	Leinster	922	1.451
13	Dublin	Leinster	922	1.451
14	Dublin	Leinster	922	1.451
15	Kildare	Leinster	1695	0.247

The second file, called the employee file, contains everything that is represented by the features shown below.

File Edit Format View Help

Employee ID	Name	Date of Birth	Gender	Job Title	Salary	Date Joined	Date Left	Address	County ID
100	Smith	27041	M	Director	100000	01-Aug-01		"12, Green Avenue, Howth, Co. Dublin"	0
125	Jones	32604	F	Technician	30000	01-May-01	31-Aug-02	"43, School Road, Malahide, Co. Dublin"	1
167	Davis	29970	F	Senior Technician	50000	01-Dec-02		"10, Main Street, Naas, Co. Kildare"	2
200	O'Bien	35553	M	Technician	25000	01-May-02	30-Nov-02	"Apt 02, High Court, Condalkin. Co. Dublin"	3
205	Edward	35019	M	Technician	33000	01-Jan-01		"33, Barake Street, Clane, Co. Kildare"	4
216	Evans	34780	F	S' Technician	44000	01-Aug-01	31-Mar-02	"143, High Street, Niwbridge, Co. Kildare"	5
220	Moore	35244	F	Jnr. Technician	22000	01-Jan-02		"Apt 01, Shreedon Court, Rathcoole, Co. Dublin"	6
301	Rogers	27712	M	Deputy Director	60000	01-May-02		"Manor House, Naas Road, Inchico, Co. Dublin"	7
303	Phillip	28047	F	HR Manager	70000	01-Jan-02		"44, Dublin Road, Finglas, Co. Dublin"	8
344	Shane	31599	M	"D"" Director"	50000	01-Jan-01	30-Apr-02	"65, Waterway, Killiney, Co. Dublin"	9
351	Alan	26806	M	Dep. Director	80000	01-Apr-01		"43, Shandon Court, Clane, Co. Kildare"	10
364	Gary	29901	M	Eng` Manager	85000	01-Jan-01	31-Mar-02	"22, Earls, Newbridge, Co. Kildare"	11
371	Robert	33996	M	J Technician	27000	01-Jan-01	31-Oct-01	"61, Robert Street, Dublin 04, Co. Dublin"	12
380	Jason	29645	M	Lead Technician	45000	02-Feb-01		"56, Alex Street, Dublin 01, Co. Dublin"	13
393	Marry	27898	F	Director	70000	01-Jan-01	30-Nov-02	"59, Sea Forth, Sword, Co. Dublin"	14
409	Hilary	32309	F	Marketing Director	78000	01-Mar-01		"87, Bray Street, Kildare, Co. Kildare"	15

I then saved the above files and move them into docker-hadoop/jobs/data folder. I then used docker cp to copy them into the namenode container.

```
C:\Users\waleed\Desktop\PRACTICAL7_Waleed_Wazir_19396951\ex1\docker-hadoop>docker cp Employee.txt namenode:/app/data

C:\Users\waleed\Desktop\PRACTICAL7_Waleed_Wazir_19396951\ex1\docker-hadoop>docker cp County.txt namenode:/app/data
```

I then created a directory called /P7 in namenode container.

```
root@a5286aaf2748:~# hdfs dfs -mkdir /P7
```

```
root@a5286aaf2748:~# hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - root supergroup          0 2022-12-07 23:31 /P7
drwxr-xr-x   - root supergroup          0 2022-12-05 18:06 /rmstate
drwx-wx-wx   - root supergroup          0 2022-12-05 18:18 /tmp
drwxr-xr-x   - root supergroup          0 2022-12-05 18:18 /user
```

I then copied both tab delimited files called County.txt and Employee.txt into directory /P7 and Checking to see if it was moved:

```
root@a5286aaf2748:/# hadoop fs -copyFromLocal -f /app/data/County.txt /P7/
root@a5286aaf2748:/# hadoop fs -copyFromLocal -f /app/data/Employee.txt /P7/
root@a5286aaf2748:/# hdfs dfs -ls /P7/
Found 2 items
-rw-r--r--   3 root supergroup        513 2022-12-08 16:39 /P7/County.txt
-rw-r--r--   3 root supergroup       1696 2022-12-08 16:39 /P7/Employee.txt
```

I then went into the hive shell using ./hive and used the P7 database we created.

```
root@a5286aaf2748:/bin/apache-hive-3.1.3-bin/bin# ./hive
Hive Session ID = 77ea559b-4249-480a-8706-6e45a2481da6

Logging initialized using configuration in jar:file:/bin/a
ies Async: true
Hive Session ID = f60e5578-1783-414f-869f-95317ba663e4
Hive-on-MR is deprecated in Hive 2 and may not be availabl
ne (i.e. spark, tez) or using Hive 1.X releases.
hive> USE P7;
OK
Time taken: 0.45 seconds
hive>
```

We create a table called employee, giving it features that are the same as the file we had before. We pass the same names into the employee table because it will be easy to have the data loaded there when we load it. The commands below allow us to delimit by \t, which represents a tab. We also use another command, which allows us to skip the header.

```
hive> create table employee (`id` int , name string , dob string , gender string , job_title string , salary int , date_joined date , date_left date , address string
`county_id` int )
    > row format delimited fields terminated by '\t'
    > tblproperties ("skip.header.line.count"="1");
OK
```

I then load data into the table by using the command in the image below:

```
hive> load data inpath '/P7/Employee.txt' overwrite into table employee;
```

As you can see in the screenshot below, when we use the select * from employee, it will show us everything associated with the file we had earlier. We have now successfully moved it into Apache hive.

```
hive> select * from employee;
OK
100     Smith   1974-01-12      M       Director        100000  2001-08-01      NULL    12, Green Avenue, Howth, Co. Dublin     0
125     Jones   1989-04-06      F       Technician      30000   2001-05-01      2002-08-31      43, School Road, Malahide, Co. Dublin   1
167     Davis   1982-01-19      F       Senior Technician       50000   2002-12-01      NULL    10, Main Street, Naas, Co. Kildare      2
200     O'Bien  1997-05-03      M       Technician      25000   2002-05-01      2002-11-30      Apt 02, High Court, Condalkin. Co. Dublin       3
205     Edward  1995-11-16      M       Technician      33000   2001-01-01      NULL    33, Barake Street, Clane, Co. Kildare   4
216     Evans   1995-03-22      F       S' Technician   44000   2001-08-01      2002-03-31      143, High Street, Niwbridge, Co. Kildare        5
220     Moore   1996-06-28      F       Jnr. Technician 22000   2002-01-01      NULL    Apt 01, Shreedon Court, Rathcoole, Co. Dublin   6
301     Rogers  1975-11-14      M       Deputy Director 60000   2002-05-01      NULL    Manor House, Naas Road, Inchico, Co. Dublin     7
303     Phillip 1976-10-14      F       HR Manager      70000   2002-01-01      NULL    44, Dublin Road, Finglas, Co. Dublin    8
344     Shane   1986-07-06      M       "D"" Director"  50000   2001-01-01      2002-04-30      65, Waterway, Killiney, Co. Dublin      9
351     Alan    1973-05-22      M       Dep. Director   80000   2001-04-01      NULL    43, Shandon Court, Clane, Co. Kildare   10
364     Gary    1981-11-11      M       Eng` Manager    85000   2001-01-01      2002-03-31      22, Earls, Newbridge, Co. Kildare       11
371     Robert  1993-01-27      M       J Technician    27000   2001-01-01      2001-10-31      61, Robert Street, Dublin 04, Co. Dublin        12
380     Jason   1981-02-28      M       Lead Technician 45000   2001-02-02      NULL    56, Alex Street, Dublin 01, Co. Dublin  13
393     Marry   1976-05-18      F       Director        70000   2001-01-01      2002-11-30      59, Sea Forth, Sword, Co. Dublin        14
409     Hilary  1988-06-15      F       Marketing Director      78000   2001-03-01      NULL    87, Bray Street, Kildare, Co. Kildare   15
Time taken: 1.1 seconds, Fetched: 16 row(s)
hive>
```

Using the command “desc” we can see all the data types for the table employee.

```
hive> desc employee;
OK
id                      int
name                    string
dob                     string
gender                  string
job_title               string
salary                  int
date_joined             date
date_left               date
address                 string
county_id               int
```

I then created a table called county. giving it features that are the same as the file we had before. We pass the same names into county, because when we load them, it will be easy to have the data loaded there. The commands below allow us to delimit by \t, which represents a tab. We also use another command, which allows us to skip the headers.

```
hive> create table county (`id` int , county string , province string , area string , population double )
    > row format delimited fields terminated by '\t'
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.49 seconds
```

I then load data into the table by using the command in the image below:

```
hive> load data inpath '/P7/County.txt' overwrite into table county;
Loading data to table default.county
OK
Time taken: 0.172 seconds
hive>
```

As you can see in the screenshot below, when we use the select * from county, it will show us everything associated with the file we had earlier. We have now successfully moved it into Apache hive

```
hive> select * from county;
OK
0       Dublin  Leinster        922     1.451
1       Dublin  Leinster        922     1.451
2       Kildare Leinster        1695    0.247
3       Dublin  Leinster        922     1.451
4       Kildare Leinster        1695    0.247
5       Kildare Leinster        1695    0.247
6       Dublin  Leinster        922     1.451
7       Dublin  Leinster        922     1.451
8       Dublin  Leinster        922     1.451
9       Dublin  Leinster        922     1.451
10      Kildare Leinster        1695    0.247
11      Kildare Leinster        1695    0.247
12      Dublin  Leinster        922     1.451
13      Dublin  Leinster        922     1.451
14      Dublin  Leinster        922     1.451
15      Kildare Leinster        1695    0.247
Time taken: 0.089 seconds, Fetched: 16 row(s)
hive>
```

Using the command “desc” we can see all the data types for the table county

```
hive> desc county;
OK
id                      int
county                  string
province                string
area                    string
population              double
Time taken: 0.141 seconds, Fetched: 5 row(s)
```

Task 3:

I calculated the average salary of employees in two counties ‘Dublin’ and ‘Kildare’ separately. First, I selected county names and calculated employee salaries and salary totals. We divided these totals by the number of employees. Then join the Employees table (renamed to e) and the Counties table (renamed to a). Then, please link to the county ID value. Finally, county name groups the average scores. The query and results are displayed in the screenshot above. The average salary of employees in the two districts of "Dublin" and "Kildare" is 49900.0 and 61666.66666

The script is named: “script task 3” inside the ex1 folder.

```
hive> select a.county , sum(b.salary)/count(distinct b.id) as average_salary from employee b join county a on (b.county_id = a.id) group by a.county;
Query ID = root_20221209124744_7b462325-aa94-40f5-9047-36bfdffe5f79
Total jobs = 1
2022-12-09 12:47:50     Dump the side-table for tag: 1 with group count: 16 into file: file:/tmp/root/bc8ab761-6ba9-45bb-ad2e-dbcb8196142f/hive_2022-1
-44_861_8424207496902593974-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile01--.hashtable
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
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>
Starting Job = job_1670589414881_0001, Tracking URL = http://resourcemanager:8088/proxy/application_1670589414881_0001/
Kill Command = /opt/hadoop-3.3.1/bin/mapred job  -kill job_1670589414881_0001
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-12-09 12:48:00,418 Stage-2 map = 0%,  reduce = 0%
2022-12-09 12:48:05,539 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.25 sec
2022-12-09 12:48:10,649 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.57 sec
MapReduce Total cumulative CPU time: 4 seconds 570 msec
Ended Job = job_1670589414881_0001
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.57 sec   HDFS Read: 16746 HDFS Write: 153 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 570 msec
OK
Dublin  49900.0
Kildare 61666.666666666664
Time taken: 26.894 seconds, Fetched: 2 row(s)
```

Task 4:

In here I manually copied one file into docker-hadoop/jobs/data folder. This file is called employee_data which I turned into a txt file. I then copied it to namenode container using docker cp.

```
C:\Users\waleed\Desktop\PRACTICAL7_Waleed_Wazir_19396951\ex1\docker-hadoop>docker cp employee_data.txt namenode:/app/data
```

I then copy the folder into the P7 directory in namenode.

```
root@a5286aaf2748:/# hadoop fs -copyFromLocal -f /app/data/employee_data.txt /P7/
root@a5286aaf2748:/# hdfs dfs -ls /P7/
Found 1 items
-rw-r--r--   3 root supergroup       2121 2022-12-09 13:13 /P7/employee_data.txt
root@a5286aaf2748:/#
```

I then created a new table called employee_data with the same features as the file we had before. We pass the same names into employee_data because it will be easy to have the data loaded there when we load it. The commands below allow us to delimit by \t, which represents a tab. The command below can also be used to skip the header, in the image below:

```
hive> create table employee_data (`id` int , name string , DOB string , gender string , job_title string , salary int ,
date_joined date , date_left date , address string , county string , province string , area string , population double ,
 `country_ID` int )
    > row format delimited fields terminated by '\t'
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.5 seconds
```

I then load the data employee_data into the table.

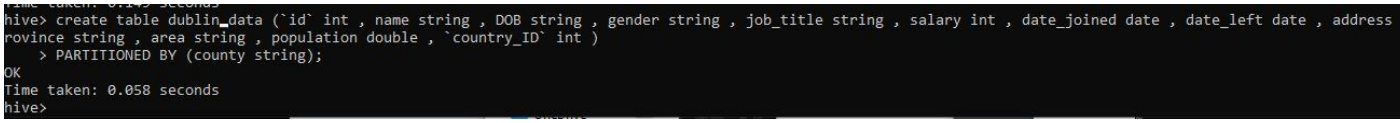
```
hive> load data inpath '/P7/employee_data.txt' overwrite into table employee_data;
Loading data to table p7.employee_data
OK
Time taken: 0.785 seconds
hive>
```

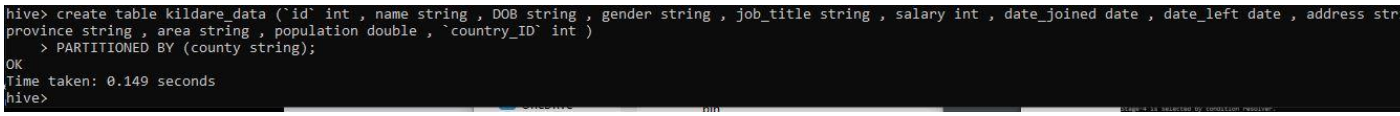
```
hive> select * from employee;
OK
100     Smith   1974-01-12      M       Director        100000  2001-08-01              12, Green Avenue, Howth, Co. Dublin     0
125     Jones   1989-04-06      F       Technician      30000   2001-05-01      2002-08-31      43, School Road, Malahide, Co. Dublin   1
167     Davis   1982-01-19      F       Senior Technician       50000   2002-12-01              10, Main Street, Naas, Co. Kildare      2
200     O'Bien  1997-05-03      M       Technician      25000   2002-05-01      2002-11-30      Apt 02, High Court, Condalkin. Co. Dublin       3
205     Edward  1995-11-16      M       Technician      33000   2001-01-01              33, Barake Street, Clane, Co. Kildare   4
216     Evans   1996-03-22      F       S' Technician   44000   2001-08-01      2002-03-31      143, High Street, Niwbridge, Co. Kildare        5
220     Moore   1996-06-28      F       Jnr. Technician 22000   2002-01-01              Apt 01, Shreedon Court, Rathcoole, Co. Dublin   6
301     Rogers  1975-11-14      M       Deputy Director 60000   2002-05-01              Manor House, Naas Road, Inchico, Co. Dublin     7
303     Phillip 1976-10-14      F       HR Manager      70000   2002-01-01              44, Dublin Road, Finglas, Co. Dublin    8
344     Shane   1986-07-06      M       "D"" Director"  50000   2001-01-01      2002-04-30      65, Waterway, Killiney, Co. Dublin      9
351     Alan    1973-05-22      M       Dep. Director   80000   2001-04-01              43, Shandon Court, Clane, Co. Kildare   10
364     Gary    1981-11-11      M       Eng' Manager    85000   2001-01-01      2002-03-31      22, Earls, Newbridge, Co. Kildare       11
371     Robert  1993-01-27      M       J Technician    27000   2001-01-01      2001-10-31      61, Robert Street, Dublin 04, Co. Dublin        12
380     Jason   1981-02-28      M       Lead Technician 45000   2001-02-02              56, Alex Street, Dublin 01, Co. Dublin  13
393     Marry   1976-05-18      F       Director        70000   2001-01-01      2002-11-30      59, Sea Forth, Sword, Co. Dublin        14
409     Hilary  1988-06-15      F       Marketing Director      78000   2001-03-01              87, Bray Street, Kildare, Co. Kildare   15
Time taken: 0.15 seconds, Fetched: 16 row(s)
```

I then checked the data types using the command desc, which allows us to see all types of features, such as string, int, which is helpful in showing the types, if we're not sure if they were exactly right or not.:

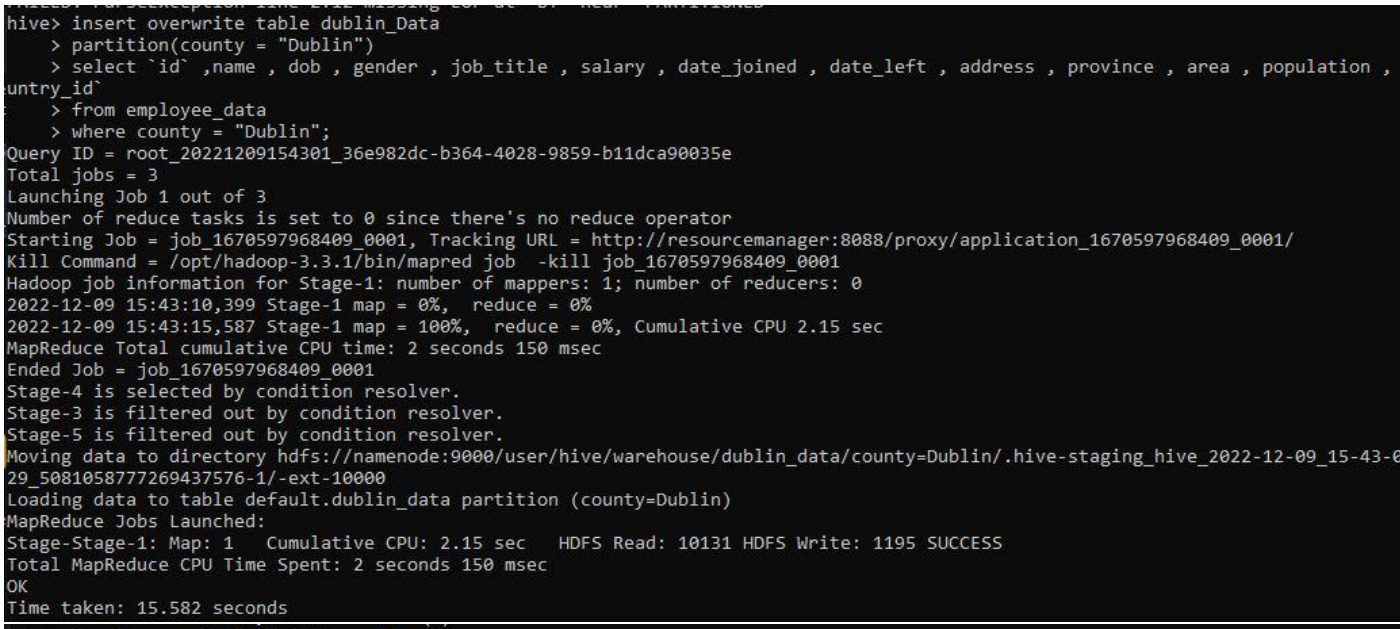
```
hive> desc employee_data;
OK
id                      int
name                    string
dob                     string
gender                  string
job_title               string
salary                  int
date_joined             date
date_left               date
address                 string
county                  string
province                string
area                    string
population              double
country_id              int
Time taken: 0.061 seconds, Fetched: 14 row(s)
```

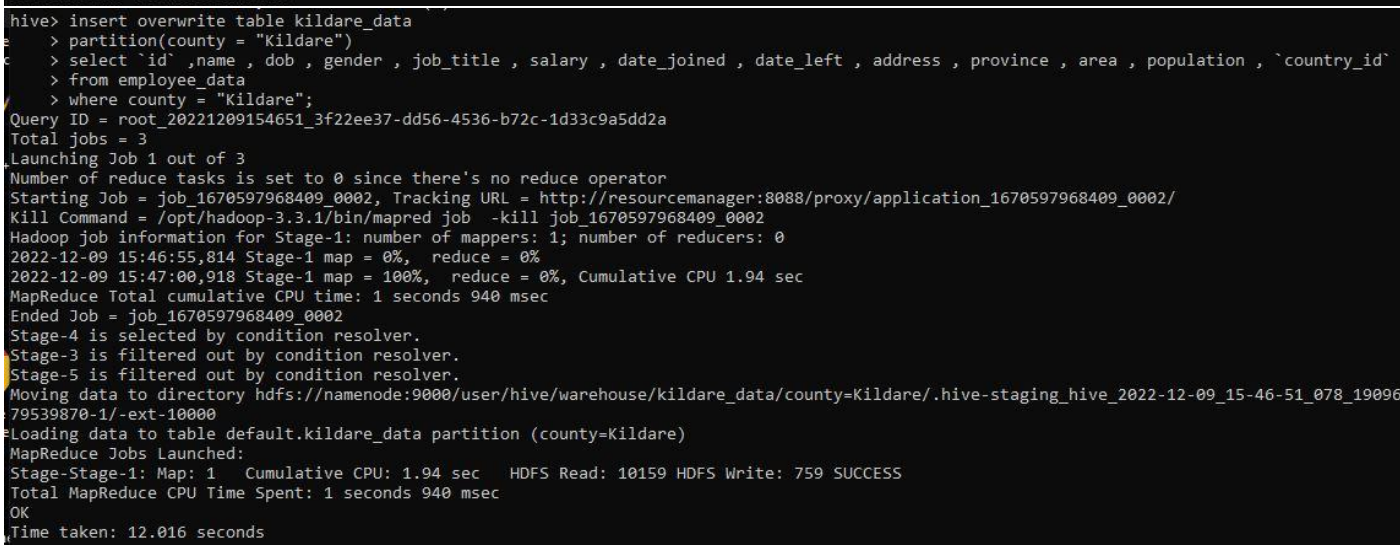
I then create a static partitioning using 'partitioned by' to store Kildare and Dublin workers separately. The command above will allow us to separate both counties and passing features only for Dublin, and for Kildare as shown in the image below.





Then for reading the partitioned files from the Hadoop dashboard, we use insert/partition method showing in the screenshot for both files.





Then we used ‘select *’ to show the datasets we get to check if it correct.

```
hive> select * from dublin_data;
OK
100     Smith   12-Jan-74       M       Director        100000  NULL    NULL    "12, Green Avenue, How
125     Jones   06-Apr-89       F       Technician      30000   NULL    NULL    "43, School Road, Malal
200     O'Bien  03-May-97       M       Technician      25000   NULL    NULL    "Apt 02, High Court, C
220     Moore   28-Jun-96       F       Jnr. Technician 22000   NULL    NULL    "Apt 01, Shreedon Court
301     Rogers  14-Nov-75       M       Deputy Director 60000   NULL    NULL    "Manor House, Naas Roa
303     Phillip 14-Oct-76       F       HR Manager      70000   NULL    NULL    "44, Dublin Road, Fing
344     Shane   06-Jul-86       M       "D"" Director"  50000   NULL    NULL    "65, Waterway, Killine
371     Robert  27-Jan-93       M       J Technician    27000   NULL    NULL    "61, Robert Street, Du
380     Jason   28-Feb-81       M       Lead Technician 45000   NULL    NULL    "56, Alex Street, Dubl
393     Marry   18-May-76       F       Director        70000   NULL    NULL    "59, Sea Forth, Sword,
Time taken: 0.107 seconds, Fetched: 10 row(s)
```

```
hive> select * from kildare_data;
OK
167     Davis   19-Jan-82       F       Senior Technician       50000   NULL    NULL
"10, Main Street, Naas, Co. Kildare"    Leinster        1695    0.247   NULL    Kildare
205     Edward  16-Nov-95       M       Technician      33000   NULL    NULL    "33, Barake Street, Clar
216     Evans   22-Mar-95       F       S' Technician   44000   NULL    NULL    "143, High Street, Niwbr
351     Alan    22-May-73       M       Dep. Director   80000   NULL    NULL    "43, Shandon Court, Clar
364     Gary    11-Nov-81       M       Eng` Manager    85000   NULL    NULL    "22, Earls, Newbridge, C
409     Hilary  15-Jun-88       F       Marketing Director      78000   NULL    NULL
"87, Bray Street, Kildare, Co. Kildare" Leinster        1695    0.247   NULL    Kildare
Time taken: 0.09 seconds, Fetched: 6 row(s)
```

Partitioning can greatly enhance performance. Hive tables and materialised views partitions can be configured to correspond to actual file system folders. A table partitioned by date-time, for example, might arrange data loaded into Hive sequentially each day. There can be found tens of thousands of partitions in large installations.

Task 5:

In this task we select 3 columns that will be viewed, the difference of the employee when he joined and left. I have named the variables “days_is_joined” and the other variables, such as “days_is_left”. We also use another method, which will help us and benefit us if the employee didn’t leave. We also add an strategy if the employee has never left, the last column is filled with 0, with select * we can see al the data types for employee_dates.

```
hive> create view employee_dates as select name , datediff(to_date('2002-12-31'), date_joined) as days_is_joined , if (isnotnull(date_left)
, datediff(to_date('2002-12-31') , date_left), 0) as days_is_left from employee_data;
OK
Time taken: 0.107 seconds
```

```
hive> select * from employee_dates;
OK
Smith   517     0
Jones   609     122
Davis   30      0
O'Bien  244     31
Edward  729     0
Evans   517     275
Moore   364     0
Rogers  244     0
Phillip 364     0
Shane   729     245
Alan    639     0
Gary    729     275
Robert  729     426
Jason   697     0
Marry   729     31
Hilary  670     0
Time taken: 0.128 seconds, Fetched: 16 row(s)
```

I then use the select methods to see the employee dates that give us the answer.

```
hive> select name , sum(days_is_joined-days_is_left) as total_number_of_days_worked from employee_dates
    > group by name;
Query ID = root_20221204145520_ccfcb0bc-9afe-4a94-a9c2-a81964ea3843
Total jobs = 1
Launching Job 1 out of 1
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>
Starting Job = job_1670162575812_0003, Tracking URL = http://resourcemanager:8088/proxy/application_167016
Kill Command = /opt/hadoop-3.3.1/bin/mapred job  -kill job_1670162575812_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-12-04 14:55:33,531 Stage-1 map = 0%,  reduce = 0%
2022-12-04 14:55:44,081 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.13 sec
2022-12-04 14:55:49,265 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.3 sec
MapReduce Total cumulative CPU time: 6 seconds 300 msec
Ended Job = job_1670162575812_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.3 sec   HDFS Read: 20677 HDFS Write: 443 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 300 msec
```

Then, for both workers, I enter the number of working days for all employees, People who are actively working and those who have retired using the sum method and grouping by name, total the total Working days for each employee as we can see in the image below.

```
Total MapReduce CPU Time Spent: 6 seconds 300 msec
OK
Alan    639
Davis   30
Edward  729
Evans   242
Gary    454
Hilary  670
Jason   697
Jones   487
Marry   698
Moore   364
O'Bien  213
Phillip 364
Robert  303
Rogers  244
Shane   484
Smith   517
Time taken: 31.107 seconds, Fetched: 16 row(s)
hive>
```

As you can see in the screenshot above, the query was successful, displaying the names and numbers required for this task.