

#Basic Operations: Hive#

1. To get started with the hive shell

hive

```
komal@komal: ~/hadoop-2.7.3
konal@komal:~/hadoop-2.7.3$ jps
2776 DataNode
3283 NodeManager
5573 Jps
2998 SecondaryNameNode
4888 JobHistoryServer
2634 NameNode
3150 ResourceManager
konal@komal:~/hadoop-2.7.3$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/komal/apache-hive-2.1.1-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/komal/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.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]
Logging initialized using configuration in jar:file:/home/komal/apache-hive-2.1.1-bin/lib/hive-common-2.1.1.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and 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>
```

2. To create a new database, in the hive shell

create database hive_projects;

```
hive> create database hive_projects;
OK
Time taken: 0.083 seconds
hive>
```

3. To see existing databases,

show databases;

```
hive> show databases;
OK
default
hive_projects
Time taken: 0.035 seconds, Fetched: 2 row(s)
hive>
```

4. To start using the database which is already created,
(say, if 'hive_projects' named database is already created)

use 'hive_projects';

```
hive> use hive_projects;
OK
Time taken: 0.039 seconds
hive>
```

5. To check the existing tables present in the database,

show tables;

```
hive> show tables;  
OK  
Time taken: 0.112 seconds  
hive> █
```

6. To know the schema of the table,

describe airports;

```
hive> describe airports;  
OK  
airport_id          int  
airport_name        string  
airport_city        string  
airport_country     string  
airport_faa         string  
airport_icao         string  
airport_lat         double  
airport_long        double  
airport_alt         double  
airport_timezone    double  
airport_dst         string  
airport_tz          string  
Time taken: 0.106 seconds, Fetched: 12 row(s)  
hive> █
```

7. To drop a table,

drop table airports;

```
hive> drop table airports;  
OK  
Time taken: 0.203 seconds  
hive> █
```

8. To drop all tables inside a database,

drop database airports cascade; (if database contains any tables then by using cascade, database will be deleted)

```
hive> drop database hive_projects cascade;  
OK  
Time taken: 0.88 seconds  
hive> █
```

9. To query the table.

(to check if the data is loaded into the hive table)

select * from airports; (to see entire data) OR

select * from airports limit 10;

```
hive> select * from airports limit 10;
OK
1  Goroka Goroka Papua New Guinea GKA AYGA -6.081689 145.391881 5282.0 10.0 U Pacific/Port_Moresby
2  Madang Madang Papua New Guinea MAG AYMD -5.207083 145.7887 20.0 10.0 U Pacific/Port_Moresby
3  Mount Hagen Mount Hagen Papua New Guinea HGU AYMH -5.826789 144.295861 5388.0 10.0 U Pacific/Port_Moresby
4  Nadzab Nadzab Papua New Guinea LAE AYNZ -6.569828 146.726242 239.0 10.0 U Pacific/Port_Moresby
5  Port Moresby Jacksons Intl Port Moresby Papua New Guinea POM AYPY -9.443383 147.22005 146.0 10.0 U Pacific/Port_Moresby
6  Wewak Intl Wewak Papua New Guinea WWK AYWK -3.583828 143.669186 19.0 10.0 U Pacific/Port_Moresby
7  Narsarsuaq Narsarsuaq Greenland UAK BGBW 61.160517 -45.425978 112.0 -3.0 E America/Godthab
8  Nuuk Godthaab Greenland GOH BGGH 64.190922 -51.678864 283.0 -3.0 E America/Godthab
9  Sondre Stromfjord Sondrestrom Greenland SFJ BGSF 67.016969 -50.089325 165.0 -3.0 E America/Godthab
10 Thule Air Base Thule Greenland THU BGTL 76.531203 -68.703161 251.0 -4.0 E America/Thule
Time taken: 0.239 seconds, Fetched: 10 row(s)
hive>
```

10. To come out of the hive shell,

quit; or **ctrl+c**

```
hive> quit;
komal@komal:~/hadoop-2.7.3$
```

#Store Files on HDFS To perform operations (skip if files are stored in local storage)

1.Create directory

bin/hdfs dfs mkdir /hive

bin/hdfs dfs mkdir /hive/hive

2. Move files to hdfs

bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/Airports_mod.dat /hive/Hive1

bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/routes.dat /hive/Hive1

bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/Final_airlines /hive/Hive1

```
komal@komal:~/hadoop-2.7.3$ bin/hdfs dfs -mkdir /hive
komal@komal:~/hadoop-2.7.3$ bin/hdfs dfs -mkdir /hive/Hive1
komal@komal:~/hadoop-2.7.3$ bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/airports_mod.dat /hive/Hive1
komal@komal:~/hadoop-2.7.3$ bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/routes.dat /hive/Hive1
komal@komal:~/hadoop-2.7.3$ bin/hdfs dfs -put /home/komal/Downloads/Air-datasets/Final_airlines /hive/Hive1
komal@komal:~/hadoop-2.7.3$
```

Problem Solution: (Part 1)

Analysis of Airport Data using Hadoop-Hive

1. Creating table airport for airports_mod.dat:

create table airports (airport_id int,airport_name string,airport_city string,airport_country string,airport_faa string,airport_icao string,airport_lat double,airport_long double,airport_alt double,airport_timezone double,airport_dst string,airport_tz string) row format delimited fields terminated by ',';

```
hive> create table airports (airport_id int,airport_name string,airport_city string,airport_country string,airport_faa string,airport_icao string,airport_lat double,airport_long double,airport_alt double,airport_timezone double,airport_dst string,airport_tz string) row format delimited fields terminated by ',';
OK
Time taken: 0.209 seconds
hive> |
```

2. Creating table finalairlines for Final_airlines :

create table final_airlines (airlineID string,airline_name string, airline_alias string, airline_iata string, airline_icao string,callsign string,territory string, active string) row format delimited fields terminated by ',';

```
hive> create table final_airlines (airlineID string,airline_name string, airline_alias string, airline_iata string, airline_icao string,callsign string,territory string, active string) row format delimited fields terminated by ',';
OK
Time taken: 0.181 seconds
hive>
```

3. Creating table route for routes.dat:

create table routes (route_iata string,route_airid int,route_source_iata string,route_source_airid int,route_des_iata string,route_des_airid int,route_codeshare string,route_stops int,route equip string) row format delimited fields terminated by ',';

```
hive> create table routes (route_iata string,route_airid int,route_source_iata string,route_source_airid int,route_des_iata string,route_des_airid int,route_codeshare string,route_stops int,route equip string) row format delimited fields terminated by ',';
OK
Time taken: 0.178 seconds
hive>
```

4. Loading data into airport table

- If the file is stored in hdfs then
load data inpath '/hive/Hive1/airports_mod.dat' into table airports;

```
hive> load data inpath '/hive/Hive1/airports_mod.dat' into table airports;
Loading data to table hive_projects.airports
OK
Time taken: 0.743 seconds
hive> |
```

OR

- If the file is stored in local storage then
load data local inpath '/home/komal/Downloads/Air-datasets/airports_mod.dat' into table airports;

```
hive> load data local inpath '/home/komal/Downloads/Air-datasets/airports_mod.dat' into table airports;
Loading data to table hive_projects.airports
OK
Time taken: 0.379 seconds
hive> |
```

5. Loading data into final airlines table

load data inpath '/hive/Hive1/Final_airlines' into table final_airlines;

```
hive> load data inpath '/hive/Hive1/Final_airlines' into table final_airlines;
Loading data to table hive_projects.final_airlines
OK
Time taken: 0.466 seconds
hive> |
```

6. Loading data into route table

load data inpath '/hive/Hive1/routes.dat' into table routes;

```
hive> load data inpath '/hive/Hive1/routes.dat' into table routes;
Loading data to table hive_projects.routes
OK
Time taken: 0.687 seconds
hive> |
```

Problem solution: (Part 2)

a) Find list of Airports operating in the Country India;

create table india_opert_airport as select * from airports where airport_country LIKE '%India%';

```
hive> create table india_opert_airport as select * from airports where airport_country LIKE '%India%';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using
Hive 1.X releases.
Query ID = komal_20200622173244_36958156-dd93-4c70-acef-cc0ff4ca5f25
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1588147351813_0021, Tracking URL = http://komal:8088/proxy/application_1588147351813_0021/
Kill Command = /home/komal/hadoop-2.7.3/bin/hadoop job -kill job_1588147351813_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-06-22 17:32:58,082 Stage-1 map = 0%, reduce = 0%
2020-06-22 17:33:04,986 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.69 sec
MapReduce Total cumulative CPU time: 1 seconds 690 msec
Ended Job = job_1588147351813_0021
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://localhost:9000/user/hive/warehouse/hive_projects.db/.hive-staging_hive_2020-06-22_17-32-44_804_2059376122631739209-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/hive_projects.db/india_opert_airport
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 1.69 sec HDFS Read: 745691 HDFS Write: 11946 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 690 msec
OK
Time taken: 22.828 seconds
```

select * from india_opert_airport limit 10; (to show only first 10 values)

```
hive> select * from india_opert_airport limit 10;
OK
895   Diego Garcia Nsf      Diego Garcia Island    British Indian Ocean Territory    FJDG   -7.313267    72.411089    9.0    6.0    U    Ind
ian/Chagos
2994  Ahmedabad      Ahmedabad      India  AMD   VAAH   23.877242    72.63465    189.0   5.5    N    Asia/Calcutta
2995  Akola      Akola      India  AKD   VAAK   20.699006    77.058628    999.0   5.5    N    Asia/Calcutta
2996  Aurangabad  Aurangabad    India  IXU   VAAU   19.862728    75.398114    1911.0  5.5    N    Asia/Calcutta
2997  Chhatrapati Shivaji Intl  Mumbai  India  BOM   VABB   19.088686    72.867919    37.0    5.5    N    Asia/Calcutta
2998  Bilaspur    Bilaspur      India  PAB   VABI   21.9884 82.110983    899.0   5.5    N    Asia/Calcutta
2999  Bhuj        Bhuj          India  BHJ   VABJ   23.287828    69.670147    268.0   5.5    N    Asia/Calcutta
3000  Belgaum    Belgaum      India  IXG   VABM   15.859286    74.618292    2487.0  5.5    N    Asia/Calcutta
3001  Vadodara   Baroda       India  BDQ   VABO   22.336164    73.226289    129.0   5.5    N    Asia/Calcutta
3002  Bhopal     Bhopal       India  BHO   VABP   23.287467    77.337375    1719.0  5.5    N    Asia/Calcutta
Time taken: 0.249 seconds, Fetched: 10 row(s)
hive>
```

b) Find the list of Airlines having zero stops

create table stop as select * from routes where route_stops LIKE '%0';

```
hive> create table stop as select * from routes where route_stops LIKE '%0';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using
Hive 1.X releases.
Query ID = komal_20200622175407_870e720c-b67a-4088-a625-819fa68671e6
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1588147351813_0022, Tracking URL = http://komal:8088/proxy/application_1588147351813_0022/
Kill Command = /home/komal/hadoop-2.7.3/bin/hadoop job -kill job_1588147351813_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-06-22 17:54:17,770 Stage-1 map = 0%, reduce = 0%
2020-06-22 17:54:26,478 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.76 sec
MapReduce Total cumulative CPU time: 2 seconds 760 msec
Ended Job = job_1588147351813_0022
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://localhost:9000/user/hive/warehouse/hive_projects.db/.hive-staging_hive_2020-06-22_17-54-07_307_6130414784412129547-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/hive_projects.db/stop
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 2.76 sec HDFS Read: 2380962 HDFS Write: 2307569 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 760 msec
OK
Time taken: 21.12 seconds
hive>
```

select * from stop limit 10;

```
hive> select * from stop limit 10;
OK
2B      410      AER      2965      KZN      2990      0      CR2
2B      410      ASF      2966      KZN      2990      0      CR2
2B      410      ASF      2966      MRV      2962      0      CR2
2B      410      CEK      2968      KZN      2990      0      CR2
2B      410      CEK      2968      OVB      4078      0      CR2
2B      410      DME      4029      KZN      2990      0      CR2
2B      410      DME      4029      NBC      6969      0      CR2
2B      410      DME      4029      TGK      NULL      0      CR2
2B      410      DME      4029      UUA      6160      0      CR2
2B      410      EGO      6156      KGD      2952      0      CR2
Time taken: 0.208 seconds, Fetched: 10 row(s)
hive>
```

c) List of Airlines operating with code share

create table codeshare_1 as select * from routes where route_codeshare LIKE '%Y%';

```
hive> create table codeshare_1 as select * from routes where route_codeshare LIKE '%Y%';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using
Hive 1.X releases.
Query ID = komal_20200622175717_7c06d76c-0587-463a-bc12-900ed77cc69e
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1588147351813_0023, Tracking URL = http://komal:8088/proxy/application_1588147351813_0023/
Kill Command = /home/komal/hadoop-2.7.3/bin/hadoop job -kill job_1588147351813_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-06-22 17:57:26,767 Stage-1 map = 0%, reduce = 0%
2020-06-22 17:57:34,330 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.87 sec
MapReduce Total cumulative CPU time: 1 seconds 870 msec
Ended Job = job_1588147351813_0023
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://localhost:9000/user/hive/warehouse/hive_projects.db/.hive-staging_hive_2020-06-22_17-57-17_131_4166500346485711494-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/hive_projects.db/codeshare_1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 1.87 sec HDFS Read: 2380969 HDFS Write: 511487 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 870 msec
OK
Time taken: 18.777 seconds
hive>
```

select * from codeshare_1 limit 10;

```
hive> select * from codeshare_1 limit 10;
OK
2P      897      GES      6011     MNL      2397      Y        0        320
2P      897      MNL      2397     GES      6011      Y        0        320
4M      3201     DFW      3670     EZE      3988      Y        0        777
4M      3201     EZE      3988     DFW      3670      Y        0        777
4M      3201     EZE      3988     JFK      3797      Y        0        777
4M      3201     JFK      3797     EZE      3988      Y        0        777
5N      503      ARH      4362     CSH      6110      Y        0        AN4
5N      503      ARH      4362     MMK      2949      Y        0        AN4
5N      503      ARH      4362     USK      4369      Y        0        AN4
5N      503      CSH      6110     ARH      4362      Y        0        AN4
Time taken: 0.191 seconds, Fetched: 10 row(s)
hive>
```


d) Which country (or) territory having highest Airports

select airport_country,count(*) as cnt from airports group by airport_country ORDER BY cnt DESC;

```
hive> select airport_country,count(*) as cnt from airports group by airport_country ORDER BY cnt DESC;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using
Hive 1.X releases.
Query ID = komal_20200622175958_fcb546c6-4f12-420d-a969-7cf4f074fe6d
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>
Starting Job = job_1588147351813_0024, Tracking URL = http://komal:8088/proxy/application_1588147351813_0024/
Kill Command = /home/komal/hadoop-2.7.3/bin/hadoop job -kill job_1588147351813_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-22 18:00:08,362 Stage-1 map = 0%, reduce = 0%
2020-06-22 18:00:15,934 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.51 sec
2020-06-22 18:00:24,556 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.86 sec
MapReduce Total cumulative CPU time: 2 seconds 860 msec
Ended Job = job_1588147351813_0024
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_1588147351813_0025, Tracking URL = http://komal:8088/proxy/application_1588147351813_0025/
Kill Command = /home/komal/hadoop-2.7.3/bin/hadoop job -kill job_1588147351813_0025
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-06-22 18:00:38,855 Stage-2 map = 0%, reduce = 0%
2020-06-22 18:00:46,331 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.13 sec
2020-06-22 18:00:53,915 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.45 sec
MapReduce Total cumulative CPU time: 2 seconds 450 msec
Ended Job = job_1588147351813_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.86 sec HDFS Read: 748740 HDFS Write: 7092 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.45 sec HDFS Read: 12340 HDFS Write: 6226 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 310 msec
OK
United States 1697
```

e) Find the list of Active Airlines in United state

select * from final_airlines where territory like '%United States%' AND active like '%Y%';

```
hive> select * from final_airlines where territory like '%United States%' AND active like '%Y%';
OK
10 40-Mile Air NULL Q5 MLA MILE-AIR United States Y
22 Aloha Airlines NULL A0 AAH ALOHA United States Y
24 American Airlines NULL AA AAL AMERICAN United States Y
35 Allegiant Air NULL C4 AAY ALLEGiant United States Y
109 Alaska Central Express NULL KO AER ACE AIR United States Y
149 Air Cargo Carriers NULL ZQ SNC NIGHT CARGO United States Y
210 Airlift International NULL AIR AIRLIFT United States Y
281 America West Airlines NULL HP AWE CACTUS United States Y
282 Air Wisconsin NULL ZW AWI AIR WISCONSIN United States Y
287 Allegheny Commuter Airlines NULL ALO ALLEGHENY United States Y
295 Air Sunshine NULL RSI AIR SUNSHINE United States Y
315 ATA Airlines NULL AMT AMTRAN United States Y
397 Arrow Air NULL JW APW BIG A United States Y
452 Atlantic Southeast Airlines NULL EV ASQ ACEY United States Y
659 American Eagle Airlines NULL MQ EGF EAGLE FLIGHT United States Y
792 Access Air NULL ZA CYD CYCLONE United States Y
882 Air Florida NULL OH FLZ AIR FLORIDA United States Y
928 Atlas Air NULL SY GTI GIANT United States Y
1316 AirTran Airways NULL FL TRS CITRUS United States Y
1442 Benidji Airlines NULL CH BMJ BEMIDJII United States Y
1472 Bering Air NULL 8E BRG BERING AIR United States Y
1629 Cape Air NULL 9K KAP CAIR United States Y
1739 Chautauqua Airlines NULL RP CHQ CHAUTAUQUA United States Y
1814 Coastal Air NULL DQ U.S. Virgin Islands United States Y
1821 Colgan Air NULL 9L CJC COLGAN United States Y
1828 Comair NULL OH COM COMAIR United States Y
1843 CommutAir NULL CS UCA COMMUTAIR United States Y
1860 Compass Airlines NULL CP CPZ Compass Rose United States Y
1881 Continental Airlines NULL CO COA CONTINENTAL United States Y
1883 Continental Express NULL CO JETLINK United States Y
1884 Continental Micronesia NULL CS CMI AIR MIKE United States Y
1931 Crown Airways NULL CRO CROWN AIRWAYS United States Y
2009 Delta Air Lines NULL DL DAL DELTA United States Y
2261 Evergreen International Airlines NULL EZ EIA EVERGREEN United States Y
2293 Express One International NULL EO LHN LONGHORN United States Y
2295 ExpressJet NULL XE BTA JET LINK United States Y
2404 Florida West International Airways NULL RF FWL FLO WEST United States Y
2454 Freedom Air NULL FP FRE FREEDOM United States Y
```