

BDA EXPERIMENT-3

Name: Sahil Dinesh Patil	Class: BE3
Roll no: 38	Batch: B

- hive
- show databases
- create database testing;
- use testing;
- create table posts (user string, post string, words bigint)
row format delimited
fields terminated by ',';
- show tables;
- load data local inpath '/home/cloudera/Desktop/DB/posts.txt' into table posts;
- select * from posts;

```
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 0.283 seconds, Fetched: 1 row(s)
hive> create database testing;
OK
Time taken: 0.054 seconds
hive> use testing;
OK
Time taken: 0.024 seconds
hive> create table posts(user string,post string,words bigint)
> row format delimited
> fields terminated by ',';
OK
Time taken: 0.116 seconds
hive> show tables;
OK
posts
Time taken: 0.021 seconds, Fetched: 1 row(s)
hive> load data local inpath '/Desktop/DB/posts.txt' into table posts;
FAILED: SemanticException Line 1:23 Invalid path ''/Desktop/DB/posts.txt': No f
iles matching path file:/Desktop/DB/posts.txt
hive> load data local inpath '/home/cloudera/Desktop/DB/posts.txt' into table posts;
Loading data to table testing.posts
Table testing.posts stats: [numFiles=1, totalSize=72]
OK
Time taken: 0.383 seconds
hive> select * from posts;
OK
user1    funny_story    1550
user2    funny_story1   1551
user3    funny_story2   1552
         NULL      NULL
Time taken: 0.179 seconds, Fetched: 4 row(s)
hive> select * from posts;
OK
user1    funny_story    1550
user2    funny_story1   1551
user3    funny_story2   1552
         NULL      NULL
Time taken: 0.041 seconds, Fetched: 4 row(s)
hive> █
```

- select count (1) from posts;
- drop table posts;

```
hive> select count(1) from posts;
Query ID = cloudera_20240728232828_55b8bade-ea6f-4a4a-8dec-1b210c253c17
Total jobs = 1
Launching Job 1 out of 1
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_1722229410265_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1722229410265_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1722229410265_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-07-28 23:28:18,100 Stage-1 map = 0%, reduce = 0%
2024-07-28 23:28:23,397 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.61 sec
2024-07-28 23:28:29,678 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.42 sec
MapReduce Total cumulative CPU time: 1 seconds 420 msec
Ended Job = job_1722229410265_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.42 sec HDFS Read: 7508 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 420 msec
OK
4
Time taken: 19.751 seconds, Fetched: 1 row(s)
hive> drop table posts;
OK
Time taken: 0.138 seconds
hive>
```

- CREATE EXTERNAL TABLE posts
(user STRING, post STRING, time BIGINT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
- CREATE TABLE posts1 (user STRING, post STRING, time BIGINT)
PARTITIONED BY (country STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
- desc post1;
- show partitions post1;
- load data local inpath '/home/cloudera/Desktop/hdfs/userp-AUS'
overwrite into table post1 partition (country='AUS')

```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System

[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE EXTERNAL TABLE posts
  > (user STRING, post STRING, time BIGINT)
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE
  > LOCATION '/home/cloudera/Desktop/hdfs/';
OK
Time taken: 1.1 seconds
hive> CREATE TABLE post1 (user STRING, post STRING, time BIGINT)
  > PARTITIONED BY(country STRING)
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE;
OK
Time taken: 0.072 seconds
hive> desc post1;
OK
user                string
post                string
time                bigint
country             string

# Partition Information
# col_name          data_type          comment
country             string
Time taken: 0.163 seconds, Fetched: 9 row(s)
hive> show partitions post1;
OK
Time taken: 0.063 seconds
hive> load data local inpath '/home/cloudera/Desktop/hdfs/userp-AUS'
  > overwrite into table post1 partition (country='AUS');
Loading data to table default.post1 partition (country=AUS)
Partition default.post1{country=AUS} stats: [numFiles=1, numRows=0, totalSize=0,
rawDataSize=0]
OK
Time taken: 0.369 seconds
hive> S

```

- load data local inpath '/home/cloudera/Desktop/hdfs/userp-US' overwrite into table post1 partition (country='US')

```
hive> load data local inpath '/home/cloudera/Desktop/hdfs/userp-US'
> overwrite into table post1 partition (country='US');
Loading data to table default.post1 partition (country=US)
Partition default.post1{country=US} stats: [numFiles=1, numRows=0, totalSize=66, rawDataSize=0]
OK
Time taken: 0.235 seconds
```

- **BUCKETING**
- CREATE TABLE post_count (user STRING, count INT) CLUSTERED BY (user) INTO 5 BUCKETS;
set hive.enforce.bucketing=true;
insert overwrite table post_cpunt
select user, count(post) from posts group by user;

```
hive> CREATE TABLE post_count (user STRING, count INT)
> CLUSTERED BY (user) INTO 5 BUCKETS;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive ql.exec.DDLTask. AlreadyExistsException(message:Table post_count already exists)
hive> set hive.enforce.bucketing = true;
hive> insert overwrite table post_count
```

```
hive> select user, count(post) from posts group by user;
Query ID = cloudera_20240804232626_6459d7c3-4bd3-49e5-9c7f-ad5ab1be13f5
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_1722834154938_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1722834154938_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1722834154938_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-08-04 23:26:32,616 Stage-1 map = 0%, reduce = 0%
2024-08-04 23:26:39,020 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.66 sec
2024-08-04 23:26:46,336 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.46 sec
MapReduce Total cumulative CPU time: 1 seconds 460 msec
Ended Job = job_1722834154938_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.46 sec HDFS Read: 7776 HDFS Write: 24 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 460 msec
OK
user1 1
user2 1
user3 1
Time taken: 22.617 seconds, Fetched: 3 row(s)
hive> █
```

- exit

```
hive> exit;
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
[cloudera@quickstart ~]$ hdfs dfs -ls -R /user/hive/warehouse/post_count/
-rwxrwxrwx 1 cloudera supergroup 0 2024-08-04 23:34 /user/hive/warehouse/post_count/000000_0
-rwxrwxrwx 1 cloudera supergroup 8 2024-08-04 23:34 /user/hive/warehouse/post_count/000001_0
-rwxrwxrwx 1 cloudera supergroup 8 2024-08-04 23:34 /user/hive/warehouse/post_count/000002_0
-rwxrwxrwx 1 cloudera supergroup 8 2024-08-04 23:34 /user/hive/warehouse/post_count/000003_0
-rwxrwxrwx 1 cloudera supergroup 0 2024-08-04 23:34 /user/hive/warehouse/post_count/000004_0
[cloudera@quickstart ~]$ █
```

- JOINS

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table customers(id INT, name STRING, age INT, address STRING, salary INT)
  > row format delimited
  > fields terminated by ',';
OK
Time taken: 0.457 seconds
hive> create table orders(oid INT, c_id INT, amount INT)
  > row format delimited
  > fields terminated by ',';
OK
Time taken: 0.086 seconds
hive> █
```