

Apache Hive

Apache Hive provides a SQL-like interface to data stored in HDP(Hortonworks Data Platform) whose component is Hive.

Steps:

1] I have also created a Amazon EMR Hadoop cluster and I am going to use HIVE on it

2] Uploaded my dataset on Amazon S3 bucket and then into hdfs:

The screenshot displays a Linux terminal window with a dark theme. The terminal title bar indicates the date and time as 'Dec 13 19:25'. The terminal content shows the following sequence of commands and outputs:

```
[hadoop@ip-172-31-7-155 ~]$ aws s3 ls
2019-12-14 03:09:13 amazonreviewssayali
2019-12-13 06:27:24 aws-logs-642781628235-us-east-2
[hadoop@ip-172-31-7-155 ~]$ hadoop distcp s3://amazonreviewssayali/AmazonReviews.tsv /user/hadoop
19/12/14 03:24:42 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile=null, copyStrategy='uniformsize', preserveStatus=true, preserveRawXattr=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://amazonreviewssayali/AmazonReviews.tsv], targetPath=/user/hadoop, targetPathExists=true, filtersFile=null}
19/12/14 03:24:42 INFO client.RMPProxy: Connecting to ResourceManager at ip-172-31-7-155.us-east-2.compute.internal/172.31.7.155:8032
[hadoop@ip-172-31-7-155 ~]$
```

Below this, a second terminal window is shown, titled 'Activities Terminal', with the date and time 'Dec 13 19:25'. It displays the command to list files in the Hadoop file system:

```
[hadoop@ip-172-31-7-155 ~]$ hadoop fs -ls
Found 3 items
drwxr-xr-x   - hadoop hadoop          0 2019-12-14 03:16 .hiveJars
-rw-r--r--   1 hadoop hadoop 423823337 2019-12-14 03:25 AmazonReviews.tsv
-rw-r--r--   1 hadoop hadoop          0 2019-12-14 03:14 fp24FCiT
[hadoop@ip-172-31-7-155 ~]$
```

Amazon Review Data Analysis HIVE:

Create a table into Hive:

```
CREATE TABLE IF NOT EXISTS AmazonReviews (  
  marketplace String,  
  customer_id int,  
  review_id String,  
  product_id String,  
  product_parent String,  
  product_title String,  
  product_category String,  
  star_rating float,  
  helpful_votes float,  
  total_votes int,  
  Vine String,  
  verified_purchase String,  
  review_headline String,  
  review_body String,  
  review_date String)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\\t' LINES TERMINATED BY '\\n' tblproperties("skip.header.line.count"="1");
```

A screenshot of a terminal window with a dark background. The window title is 'hadoop@ip-172-31-7-155'. The terminal shows the following sequence of commands and output:
hive> show tables;
OK
Time taken: 0.503 seconds
hive> CREATE TABLE IF NOT EXISTS AmazonReviews (
 > marketplace String,
 > customer_id int,
 > review_id String,
 > product_id String,
 > product_parent String,
 > product_title String,
 > product_category String,
 > star_rating float,
 > helpful_votes float,
 > total_votes int,
 > Vine String,
 > verified_purchase String,
 > review_headline String,
 > review_body String,
 > review_date String)
 > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\\t' LINES TERMINATED BY '\\n' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.519 seconds
hive> show tables;
OK
amazonreviews
Time taken: 0.022 seconds, Fetched: 1 row(s)
hive>

Load the dataset into hive table:

```
hadoop@ip-172-31-7-155:~$ hive> show tables;
OK
amazonreviews
amazonreviews1
averagerating
Time taken: 0.504 seconds, Fetched: 3 row(s)
hadoop@ip-172-31-7-155:~$ hive> select * from amazonreviews1 limit 5;
OK
US      1797882 R3I2DHQBR577SS B001AN000E 2102612 The Naked Bee Vitmin C Moisturizing Sunscreen SPF 30, 5.5 oz (163 ml.) Bea
uty 5.0 0.0 0 N Y Five Stars Love this, excellent sun block!! 2015-08-31
US      18381298 R1QNE9N0FJC2Y4 B0016J22EQ 106393691 Alba Botanica Sunless Tanning Lotion, 4 Ounce Beauty 5.0
0.0 0 N Y Thank you Alba Bontanica! The great thing about this cream is that it doesn't smell weird lik
e all those chemical laden ones. I get a nice healthy un-fake looking tan that isn't orange and it makes my skin soft too. 201
5-08-31
US      19242472 R3LIDG204LJBAO B00HU6UOAG 375449471 Elysee Infusion Skin Therapy Elixir, 2oz. Beauty 5.0
0.0 0 N Y Five Stars Great Product, I'm 65 years old and this is all it claims to be! 2015-08-31
US      19551372 R3KSZHPAEVPEAL B002HWS7RM 255651889 Diane D722 Color, Perm And Conditioner Processing Caps - 10
0-Pack - Clear Beauty 5.0 0.0 0 N Y GOOD DEAL! I use them as shower caps & conditioning caps. I li
ke that they're in bulk. It saves a lot of money. 2015-08-31
US      14802407 RAI20IG50KZ43 B00SM99KWU 116158747 Biore UV Aqua Rich Watery Essence SPF50+/PA++++ (pack of 2)
Beauty 5.0 0.0 0 N Y this soaks in quick and provides a nice base for makeup This is my go-to daily sunb
lock. It leaves no white cast at all and has a clean, pleasant scent. If you're a makeup wearer, this soaks in quick and provides a
nice base for makeup. I've been using this brand for over a year. With daily use, this tube will last you a couple months. 201
5-08-31
Time taken: 1.345 seconds, Fetched: 5 row(s)
hadoop@ip-172-31-7-155:~$
```

Top 10 Products based on average rating

I have calculated average rating for each product. I have demonstrated similar script on map reduce however hive gave the fastest performance.

>Select product_id, avg(star_rating) as AvgRating

>From AmazonReviews1

>group by Product_ID

>order by AvgRating DESC limit 10;

```
hadoop@ip-172-31-7-155:~$ hive> > From AmazonReviews1
> group by Product ID
> order by AvgRating desc limit 10;
Query ID = hadoop_20191214045748_78af6204-edf0-4047-98d5-f310546448d2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1576258964838_0015)

-----
VERTICES    MODE          STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 12.99 s
-----
OK
189112966X 5.0
B01C2EFBZU 5.0
5357956227 5.0
3254895630 5.0
3222000565 5.0
4186001189 5.0
4967318290 5.0
B01GUDYS20 5.0
5214785224 5.0
4937826484 5.0
Time taken: 13.467 seconds, Fetched: 10 row(s)
hadoop@ip-172-31-7-155:~$
```

Most valuable Customer

I have counted no of products ordered by each customer and Customer who ordered highest number of products is the most valuable customer.

```
>select customer_id, count(product_id) as totalProducts
> from AmazonReviews1
> group by customer_id
> order by totalProducts desc limit 5;
```

```
hive> select customer_id, count(product_id) as totalProducts
> from AmazonReviews1
> group by customer_id
> order by totalProducts desc limit 5;
Query ID = hadoop_20191214051432_dc50de44-eb74-4a40-8992-d54d7f5bb6d1
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1576258964838_0017)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	1	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	1	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 13.46 s
OK
12201275      289
10942711      268
37446839      154
4090438 141
51812418      129
Time taken: 18.702 seconds, Fetched: 5 row(s)
hive>
```

Most Popular Product

I have counted how many times each product was ordered.

```
>select product_id, count(product_id) as mostordered
> from AmazonReviews1
> group by product_id
> order by mostordered desc limit 5;
```

```
hive> select product_id, count(product_id) as mostordered
> from AmazonReviews1
> group by product_id
> order by mostordered desc limit 5;
Query ID = hadoop_20191214052439_5debf73-850b-4091-acb1-c27eb23041a8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1576258964838_0018)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	1	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	1	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 10.49 s
OK
B001MA0QY2    3548
B0049LUI90    1631
B00I3VHKVK    1351
B00DPE9EQ0    1266
B000JNOSIQ    1245
Time taken: 13.968 seconds, Fetched: 5 row(s)
hive>
```

Number of Products Per day

- > Select review_date , count(product_id) as noOfProducts
- > From AmazonReviews1
- > group by review_date
- > order by noOfProducts desc limit 5;

```
hive> Select review_date , count(product_id) as noOfProducts
> From AmazonReviews1
> group by review_date
> order by noOfProducts desc limit 5;
Query ID = hadoop_20191214055647_f30afe45-7ee4-4021-84b6-2a3bb2a5923e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1576258964838_0019)

VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  2      2      0      0      0      0
Reducer 3 ..... container  SUCCEEDED  1      1      0      0      0      0
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 7.45 s
OK
2015-06-03      10599
2015-03-10      8641
2015-07-22      8596
2015-05-11      8568
2015-06-23      8364
Time taken: 10.879 seconds, Fetched: 5 row(s)
hive>
```

Number of Products Per Rating

- >select star_rating, count(product_id) as noOfProducts
- > From AmazonReviews1
- > group by star_rating
- > order by noOfProducts desc limit 5;

```
hive> select star_rating, count(product_id) as noOfProducts
> From AmazonReviews1
> group by star_rating
> order by noOfProducts desc limit 5;
Query ID = hadoop_20191214060454_1e3a6d68-efe0-4538-aed2-b876f16f7a6e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1576258964838_0020)

VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  2      2      0      0      0      0
Reducer 3 ..... container  SUCCEEDED  1      1      0      0      0      0
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 7.65 s
OK
5.0      761569
4.0      157827
1.0      98995
3.0      85570
2.0      55917
Time taken: 11.125 seconds, Fetched: 5 row(s)
hive>
```