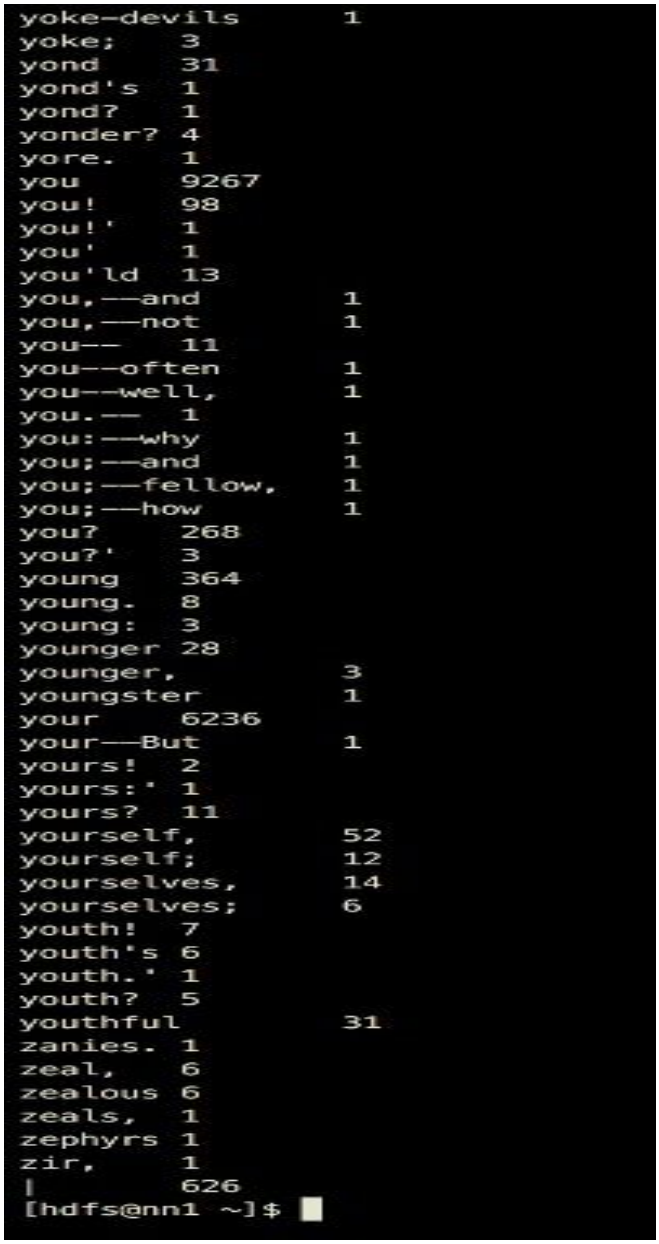
Map-Reduce using Cloudera:

Cloudera provides us with a convenient API to handle our hadoop environment and data clusters. The cloudera manager is set up with our hostname and ip address at port 7180 by first running the cloudera-installation.bin file. Before we set up the cloudera manager, we disable selinux in order to allow unlabeled files to be executed in the system. We add the port 7180 to our firewall in order to allow our system to send and receive information from that node, and hence facilitate communication with the cloudera API. We then add the cdh5 parcel files to setup the appropriate version of cloudera. Since cloudera API is run on a web browser, we place the cloudera.cdh5 files in a parcels directory that we create in the /var/www/html/ folder, which is our systems localhost address, making it possible to run cloudera using our local server. Lynx or w3m, which are command line browsers, is run on http://hostname/parcels to make sure that packages are hosted. Finally we have a single node cluster which comprises of everything- name node, data node, resource manager, secondary name node,etc.

Map-Reduce using Cloudera Multi-Node :
This task was performed using the linode services by having 1 name node and 3 data nodes but in this case the name node also acted as a data node and hence we used only 3 virtual machines.
Configuration of each node was 8GB RAM and 4 CPU cores.

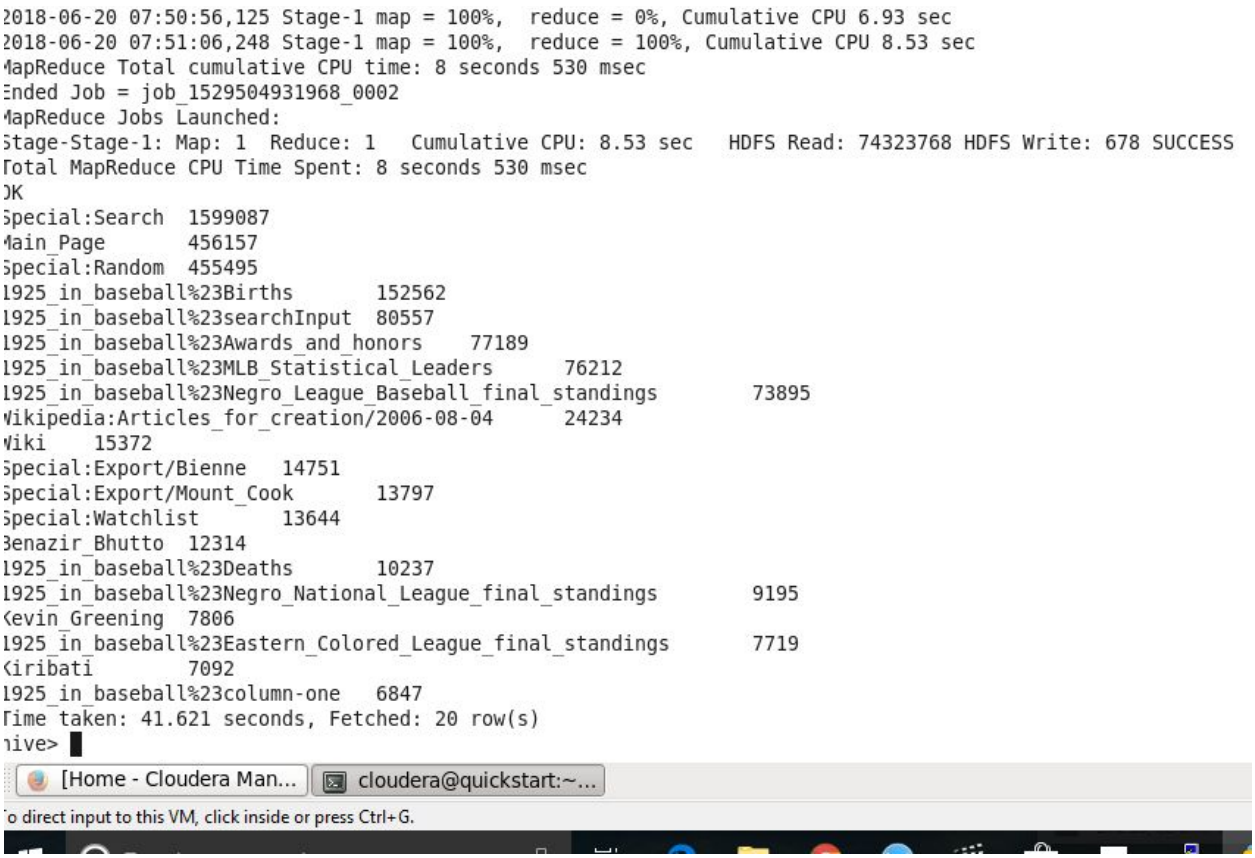
- In all the three nodes we add all the host names and its corresponding ip addresses(/etc/hosts) so that every node can communicate with every other node.
- We setup the cluster using cloudera manager by first creating the name node and we add the two other nodes as data nodes.
- We then apply a host template to each of the data nodes so that the workload can be distributed using the load balancer.
- Once we set this up, our cluster is ready and we ran the example wordcount program jar file that's already generated for the shakespeare dataset.



Hive database

Hive is used to create and manage databases, in hive format, where information can be updated and accessed efficiently using mysql queries. The databases are stored in the hdfs. The data to be loaded into the hive table is put inside hdfs. We then create a database with a table using hive and load the data on which word count is to be performed into the table that was created. We then use an sql query to create a new table with the words and their counts, thus performing wordcount.
The word count was performed for the shakespeare dataset.

Wikipedia dataset :-
The wikipedia dataset was used to query the number of visits to each of their different web pages.
We used the hive database for the querying.
The output is as follows :-



Movie Dataset :-
The movie dataset comprised of 3 tables,

1. Movies :- This table had the information about a movie name its name and its genre's.
2. Users :- This table had the information about every user, their id, age, occupation, etc.
3. Rating :- This table had the information of every user with an id watching a movie and rating it at a certain timestamp.

Query 1:-
Top Viewed Movies

select Title, count(*) as cnt from (select Title, UserID from movies, rating where movies.MovieID == rating.MovieID) q1 group by Title order by cnt desc limit 10;
Output :-

```
MapReduce Total cumulative CPU time: 3 seconds 430 msec
Ended Job = job_1529543818338_0020
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.26 sec   HDFS Read: 2160481
4 HDFS Write: 164690 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 3.43 sec   HDFS Read: 169670
HDFS Write: 379 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 690 msec
OK
American Beauty (1999)  3428
Star Wars: Episode IV - A New Hope (1977)       2991
Star Wars: Episode V - The Empire Strikes Back (1980)   2990
Star Wars: Episode VI - Return of the Jedi (1983)       2883
Jurassic Park (1993)    2672
Saving Private Ryan (1998)      2653
Terminator 2: Judgment Day (1991)       2649
Matrix, The (1999)      2590
Back to the Future (1985)       2583
Silence of the Lambs, The (1991)        2578
Time taken: 69.081 seconds, Fetched: 10 row(s)
```

Query 2:-
Highest Rated Movies
select Title, AVG(rating.Rating) as average from (select Title, UserID from movies, rating where movies.MovieID == rating.MovieID) q1 group by Title order by cnt desc limit 20

Output :-

```
Baby, The (1973)        5.0
Follow the Bitch (1998) 5.0
One Little Indian (1973)        5.0
Lured (1947)    5.0
Ulysses (Ulisse) (1954) 5.0
Smashing Time (1967)    5.0
Song of Freedom (1936)  5.0
Schlafes Bruder (Brother of Sleep) (1995)       5.0
Bittersweet Motel (2000)        5.0
Gate of Heavenly Peace, The (1995)      5.0
I Am Cuba (Soy Cuba/Ya Kuba) (1964)     4.8
Lamerica (1994) 4.75
Apple, The (Sib) (1998) 4.666666666666667
Sanjuro (1962)  4.608695652173913
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)     4.560509
554140127
Shawshank Redemption, The (1994)        4.554557700942973
Godfather, The (1972)   4.524966261808367
Close Shave, A (1995)   4.52054794520548
Usùal Suspects, The (1995)      4.517106001121705
Schindler's List (1993) 4.510416666666667
Time taken: 64.35 seconds, Fetched: 20 row(s)
```

Query 3 :-
Order by age group and occupation ranking of genres

```
CREATE TABLE  movie_genre_rank AS
select occupation,agegroup,genreName,dense_rank() over(partition by occupation order by
avgRating DESC) as generRank from
(
select occupation, agegroup, genreName, avg(rating) as avgRating from
(
select rating, genreName, agegroup, occupation from
(
select a.rating as rating, b.genre as genreList, c.agegroup as agegroup, c.occupation as
occupation
FROM ratings a
JOIN movies b
ON a.movieId=b.movieId
JOIN enriched_user_table c
ON a.userId=c.userId
)a
LATERAL VIEW explode(genreList)l as genreName
)b
group by occupation, agegroup, genreName
)c
group by  occupation,agegroup,genreName, avgRating;
```

```
writer  18-24   Action|Adventure|Children's|Fantasy     535
writer  18-24   Action|Children's       536
writer  35-44   Action|Adventure|Children's|Sci-Fi      537
writer  18-24   Action|Adventure|Crime|Thriller 537
writer  50-55   Adventure|Sci-Fi|Thriller       537
writer  25-34   Action|Children's       538
writer  45-49   Adventure|Drama|Romance 539
writer  25-34   Action|Adventure|Mystery|Sci-Fi 540
writer  45-49   Action|Sci-Fi|Western   541
writer  25-34   Comedy|Film-Noir|Thriller       541
writer  18-24   Action|Adventure|Comedy|War     541
writer  45-49   Animation|Children's|Comedy|Romance     541
writer  45-49   Animation|Children's|Fantasy|War        541
writer  45-49   Action|Comedy|Musical|Sci-Fi    541
writer  18-24   Children's|Sci-Fi       541
writer  18-24   Children's|Fantasy      541
writer  45-49   Adventure|Comedy|Musical        541
writer  56+     Comedy|Crime|Drama      541
writer  45-49   Adventure|Animation|Children's|Fantasy  541
writer  45-49   Action|Adventure|Comedy|War     541
writer  45-49   Action|Adventure|Children's|Sci-Fi      541
writer  25-34   Action|Adventure|Children's     541
writer  45-49   Adventure|Animation|Children's|Comedy|Fantasy   541
writer  50-55   Action|Adventure|Children's|Sci-Fi      541
writer  18-24   Adventure|Children's|Comedy|Fantasy|Romance     541
writer  35-44   Action|Adventure|Children's     541
writer  45-49   Adventure|Musical|Romance       541
Time taken: 0.105 seconds, Fetched: 28959 row(s)
hive>
```