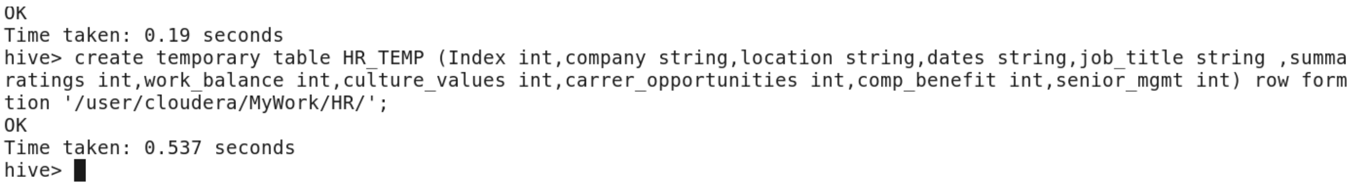
Project – 2: Employee Review of Comment Analysis

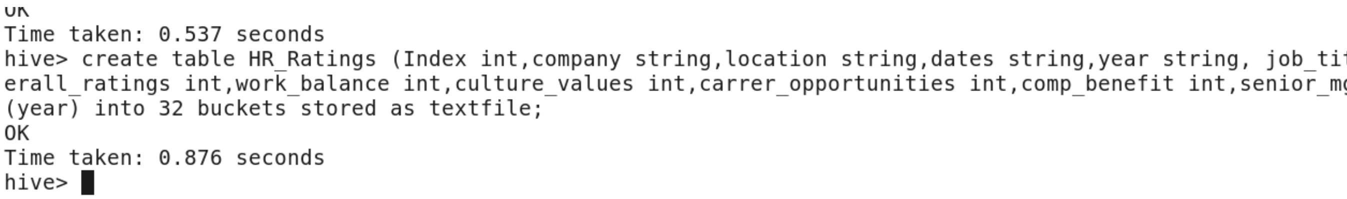
Candidate Name: Rahul Ramachandran

1. Copied the input file to the system and Created a temporary table using that.
   create temporary table HR_TEMP (Index int,company string,location string,dates string,job_title string ,summary string ,pros string,cons string,overall_ratings int,work_balance int,culture_values int,carrer_opportunities int,comp_benefit int,senior_mgmt int) row format delimited fields terminated by ',' location '/user/cloudera/MyWork/HR/';



```
OK
Time taken: 0.19 seconds
hive> create temporary table HR_TEMP (Index int,company string,location string,dates string,job_title string ,summa
ratings int,work_balance int,culture_values int,carrer_opportunities int,comp_benefit int,senior_mgmt int) row form
tion '/user/cloudera/MyWork/HR/';
OK
Time taken: 0.537 seconds
hive> █
```

2. Using the above temporary table, created the table **HR_Ratings** partitioned by company and bucketed by the year.
   create table HR_Ratings (Index int,company string,location string,dates string,year string, job_title string ,summary string ,pros string,cons string,overall_ratings int,work_balance int,culture_values int,carrer_opportunities int,comp_benefit int,senior_mgmt int) partitioned by(country string) clustered by (year) into 32 buckets stored as textfile;

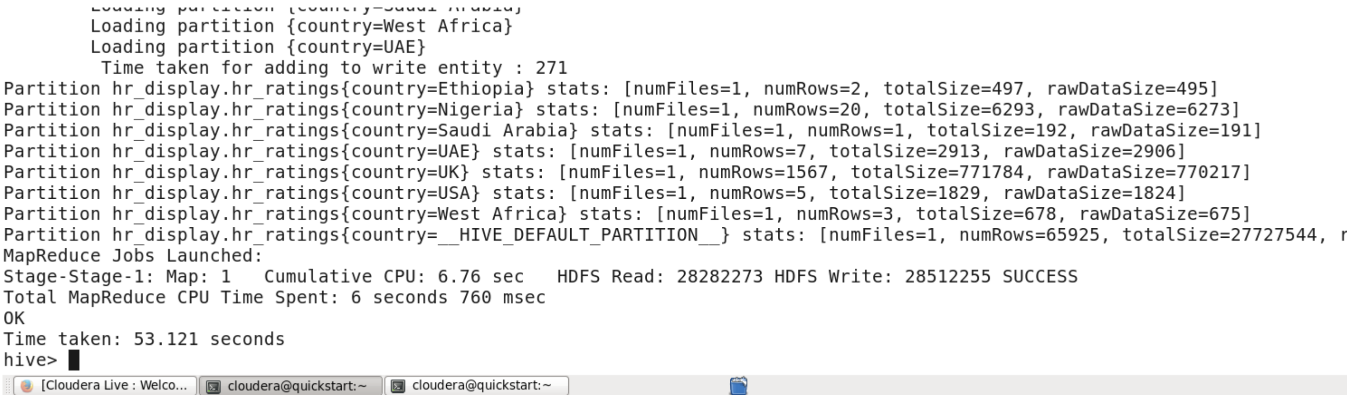


```
OK
Time taken: 0.537 seconds
hive> create table HR_Ratings (Index int,company string,location string,dates string,year string, job_ti
erall_ratings int,work_balance int,culture_values int,carrer_opportunities int,comp_benefit int,senior_mg
(year) into 32 buckets stored as textfile;
OK
Time taken: 0.876 seconds
hive> █
```

3. Now the metadata is created. Time to load the data. Set the right environment variable and then loaded the data into the table.
   set hive.exec.dynamic.partition.mode=nonstrict;
   set hive.exec.dynamic.partition = true;
   set hive.exec.dynamic.partitions.pernode = 1000;

   insert overwrite table hr_ratings partition (country) select Index,company,location,dates,substr(dates,length(dates)-4,length(dates)), job_title ,summary,pros,cons,overall_ratings,work_balance,culture_values,carrer_opportunities,comp_benefit,senior_mgmt,

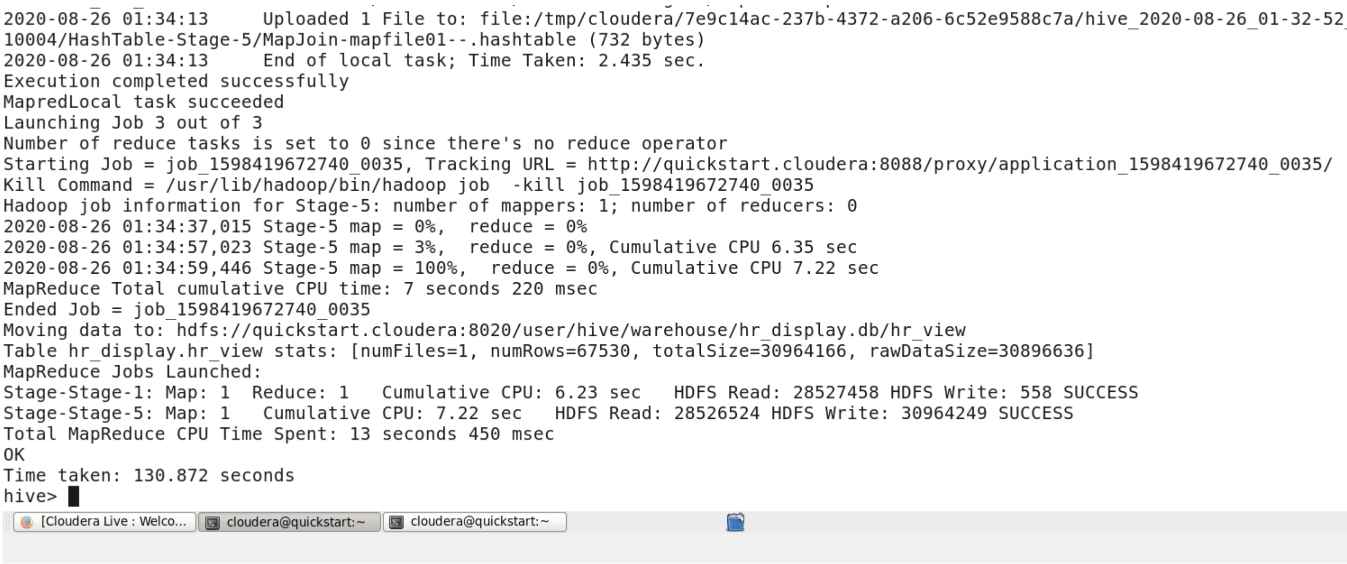
```
case when location like 'Aberdeen%' or location like 'Aberdeen%' or location like 'Acworth%' then
'USA'
when location like  '%UK%' then 'UK'
when location like 'Abha %' then 'Saudi Arabia'
when location like 'Abidjan%' then 'West Africa'
when location like 'Abu Dhabi%' then 'UAE'
when location like '%Nigeria%' then 'Nigeria'
when location like '%Ethiopia%' then 'Ethiopia' else null
end as country from hr_temp;
```



```
        Loading partition {country=Saudi Arabia}
        Loading partition {country=West Africa}
        Loading partition {country=UAE}
         Time taken for adding to write entity : 271
Partition hr_display.hr_ratings{country=Ethiopia} stats: [numFiles=1, numRows=2, totalSize=497, rawDataSize=495]
Partition hr_display.hr_ratings{country=Nigeria} stats: [numFiles=1, numRows=20, totalSize=6293, rawDataSize=6273]
Partition hr_display.hr_ratings{country=Saudi Arabia} stats: [numFiles=1, numRows=1, totalSize=192, rawDataSize=191]
Partition hr_display.hr_ratings{country=UAE} stats: [numFiles=1, numRows=7, totalSize=2913, rawDataSize=2906]
Partition hr_display.hr_ratings{country=UK} stats: [numFiles=1, numRows=1567, totalSize=771784, rawDataSize=770217]
Partition hr_display.hr_ratings{country=USA} stats: [numFiles=1, numRows=5, totalSize=1829, rawDataSize=1824]
Partition hr_display.hr_ratings{country=West Africa} stats: [numFiles=1, numRows=3, totalSize=678, rawDataSize=675]
Partition hr_display.hr_ratings{country=__HIVE_DEFAULT_PARTITION__} stats: [numFiles=1, numRows=65925, totalSize=27727544, r
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 6.76 sec   HDFS Read: 28282273 HDFS Write: 28512255 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 760 msec
OK
Time taken: 53.121 seconds
hive> █
```

4. Below step I performed below activities.
   a. For missing value imputation, calculated the median/average for each company and use the same to fill the null values.
   b. Split the job title column to Job Status and Job Destination for better evaluation
   c. Push the calculated record set to a new table to perform the data analysis

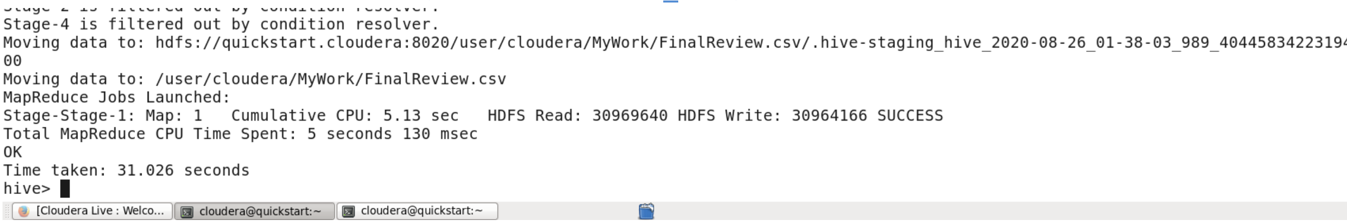
```
create table hr_view as
select index,hr.company,hr.country,location,dates, year,substr(job_title,0,instr(job_title,'-')-2)
as EMP_STATUS,substr(job_title,instr(job_title,'-')+2,length(job_title)) EMP_DESIG, summary,
pros, cons, nvl(hr.overall_ratings,c.avg_overall) as overall_ratings,
nvl(hr.work_balance,c.avg_balance) as work_bal,nvl(hr.culture_values,c.avg_culture) as
culture_val,nvl(hr.carrer_opportunities,c.avg_opport) as
career_opport,nvl(hr.comp_benefit,c.avg_benefit) as comp_ben,
nvl(hr.senior_mgmt,c.avg_mgmnt) as senior_mngnt  from hr_ratings hr left join (select
company,round(avg(overall_ratings),2) as avg_overall,round(avg(work_balance),2) as
avg_balance, round(avg(culture_values),2) as avg_culture,round(avg(carrer_opportunities),2)
as avg_opport,round(avg(comp_benefit),2) as avg_benefit, round(avg(senior_mgmt),2) as
avg_mgmnt from hr_ratings group by company) c  on(hr.company = c.company);
```



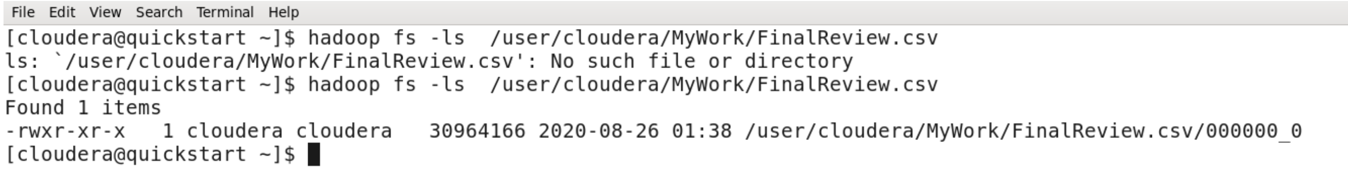
```
2020-08-26 01:34:13     Uploaded 1 File to: file:/tmp/cloudera/7e9c14ac-237b-4372-a206-6c52e9588c7a/hive_2020-08-26_01-32-52_
10004/HashTable-Stage-5/MapJoin-mapfile01--.hashtable (732 bytes)
2020-08-26 01:34:13     End of local task; Time Taken: 2.435 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1598419672740_0035, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1598419672740_0035/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1598419672740_0035
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 0
2020-08-26 01:34:37,015 Stage-5 map = 0%,  reduce = 0%
2020-08-26 01:34:57,023 Stage-5 map = 3%,  reduce = 0%, Cumulative CPU 6.35 sec
2020-08-26 01:34:59,446 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 7.22 sec
MapReduce Total cumulative CPU time: 7 seconds 220 msec
Ended Job = job_1598419672740_0035
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hr_display.db/hr_view
Table hr_display.hr_view stats: [numFiles=1, numRows=67530, totalSize=30964166, rawDataSize=30896636]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.23 sec   HDFS Read: 28527458 HDFS Write: 558 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 7.22 sec   HDFS Read: 28526524 HDFS Write: 30964249 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 450 msec
OK
Time taken: 130.872 seconds
hive>
```

5. Exported the final table to HDFS

*INSERT OVERWRITE DIRECTORY '/user/cloudera/MyWork/FinalReview.csv' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM hr_view;*



```
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/cloudera/MyWork/FinalReview.csv/.hive-staging_hive_2020-08-26_01-38-03_989_40445834223194
00
Moving data to: /user/cloudera/MyWork/FinalReview.csv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 5.13 sec   HDFS Read: 30969640 HDFS Write: 30964166 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 130 msec
OK
Time taken: 31.026 seconds
hive>
```

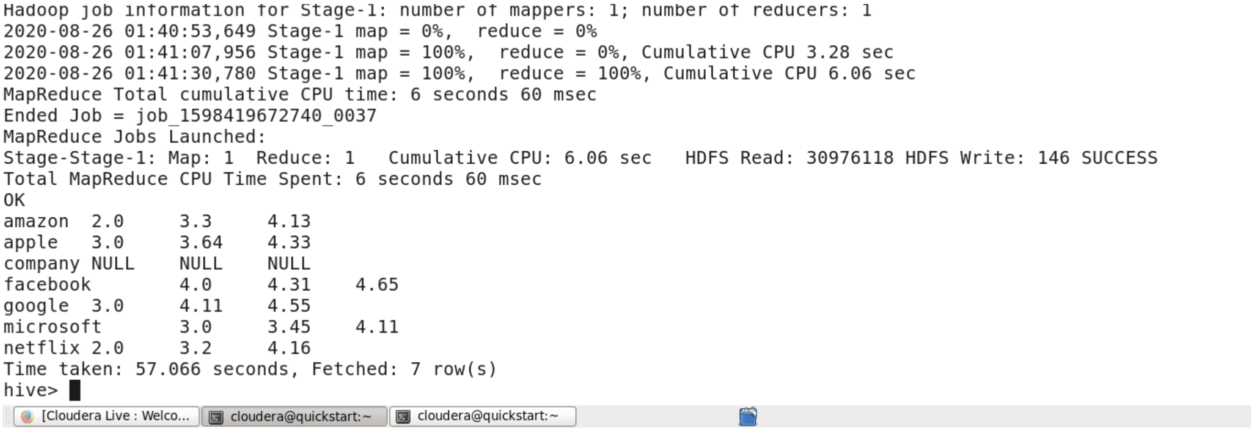


```
[cloudera@quickstart ~]$ hadoop fs -ls  /user/cloudera/MyWork/FinalReview.csv
ls: `/user/cloudera/MyWork/FinalReview.csv': No such file or directory
[cloudera@quickstart ~]$ hadoop fs -ls  /user/cloudera/MyWork/FinalReview.csv
Found 1 items
-rwxr-xr-x   1 cloudera cloudera   30964166 2020-08-26 01:38 /user/cloudera/MyWork/FinalReview.csv/000000_0
[cloudera@quickstart ~]$
```

6. **Trend Calculations below**

   a. Globally by company

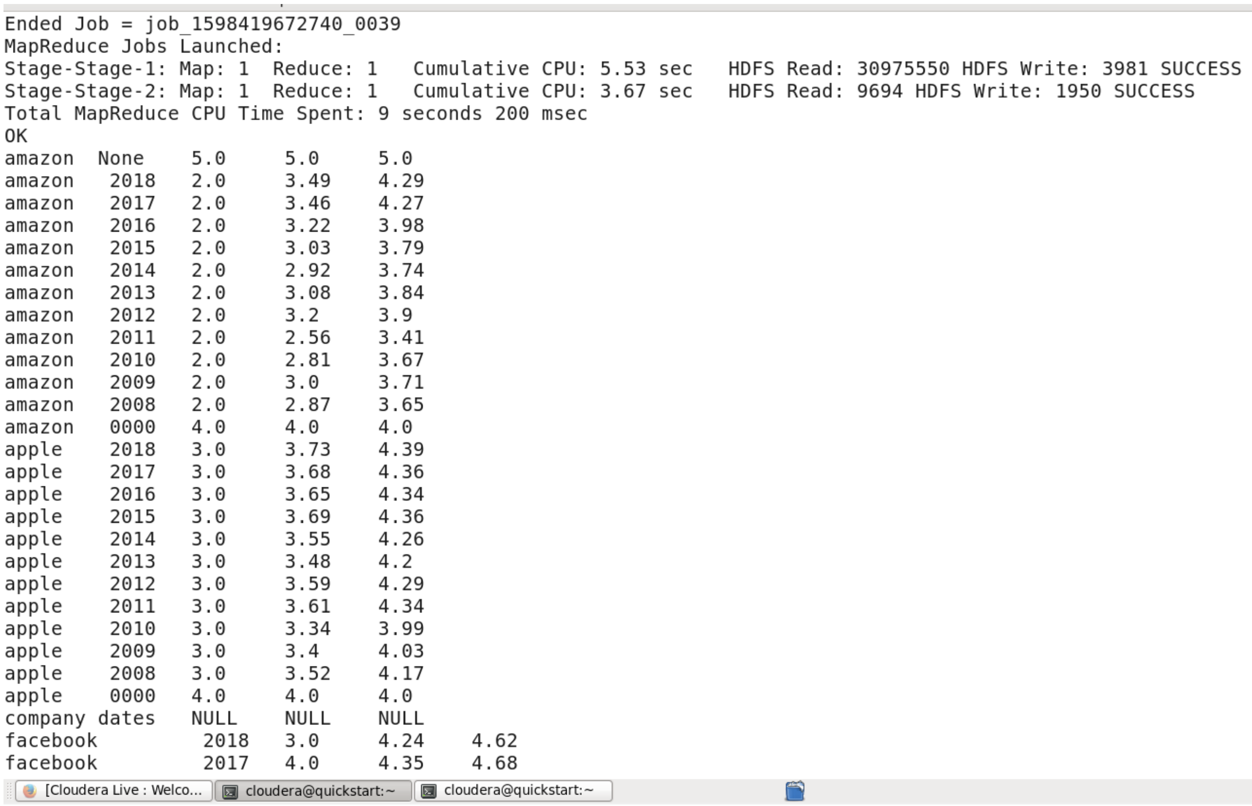
   select company,round(percentile_approx(overall_ratings,0.25),0) as Trend_25,round(percentile_approx(overall_ratings,0.5),2) as Trend_50, round(percentile_approx(overall_ratings,0.75),2) as Trend_75 from hr_view group by company;



```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-08-26 01:40:53,649 Stage-1 map = 0%,  reduce = 0%
2020-08-26 01:41:07,956 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.28 sec
2020-08-26 01:41:30,780 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.06 sec
MapReduce Total cumulative CPU time: 6 seconds 60 msec
Ended Job = job_1598419672740_0037
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.06 sec   HDFS Read: 30976118 HDFS Write: 146 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 60 msec
OK
amazon  2.0     3.3     4.13
apple   3.0     3.64    4.33
company NULL    NULL    NULL
facebook        4.0     4.31    4.65
google  3.0     4.11    4.55
microsoft       3.0     3.45    4.11
netflix 2.0     3.2     4.16
Time taken: 57.066 seconds, Fetched: 7 row(s)
hive> █
```

   b. Globally by company and year

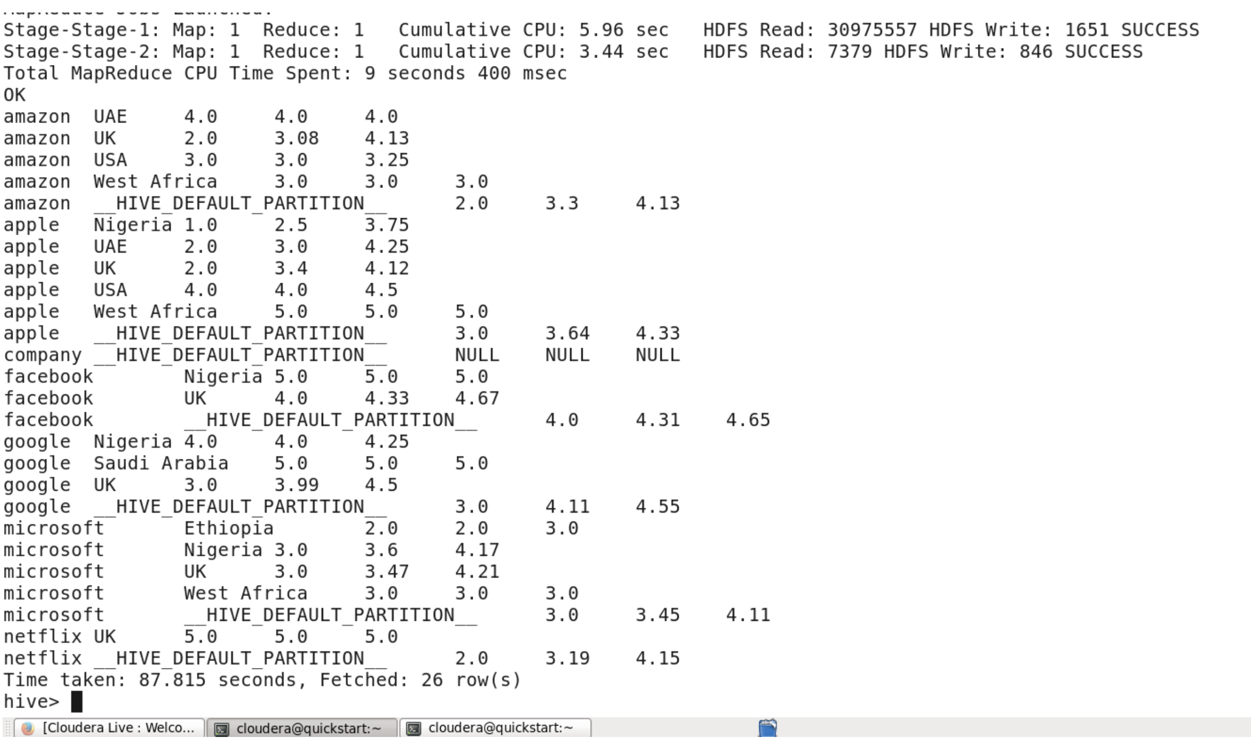
   select company,year,round(percentile_approx(overall_ratings,0.25),0) as Trend_25,round(percentile_approx(overall_ratings,0.5),2) as Trend_50, round(percentile_approx(overall_ratings,0.75),2) as Trend_75 from hr_view group by company,year order by company, year desc;



```
Ended Job = job_1598419672740_0039
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.53 sec   HDFS Read: 30975550 HDFS Write: 3981 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 9694 HDFS Write: 1950 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 200 msec
OK
amazon   None    5.0     5.0     5.0
amazon   2018    2.0     3.49    4.29
amazon   2017    2.0     3.46    4.27
amazon   2016    2.0     3.22    3.98
amazon   2015    2.0     3.03    3.79
amazon   2014    2.0     2.92    3.74
amazon   2013    2.0     3.08    3.84
amazon   2012    2.0     3.2     3.9
amazon   2011    2.0     2.56    3.41
amazon   2010    2.0     2.81    3.67
amazon   2009    2.0     3.0     3.71
amazon   2008    2.0     2.87    3.65
amazon   0000    4.0     4.0     4.0
apple    2018    3.0     3.73    4.39
apple    2017    3.0     3.68    4.36
apple    2016    3.0     3.65    4.34
apple    2015    3.0     3.69    4.36
apple    2014    3.0     3.55    4.26
apple    2013    3.0     3.48    4.2
apple    2012    3.0     3.59    4.29
apple    2011    3.0     3.61    4.34
apple    2010    3.0     3.34    3.99
apple    2009    3.0     3.4     4.03
apple    2008    3.0     3.52    4.17
apple    0000    4.0     4.0     4.0
company  dates   NULL    NULL    NULL
facebook         2018    3.0     4.24    4.62
facebook         2017    4.0     4.35    4.68
```

c. By company by country

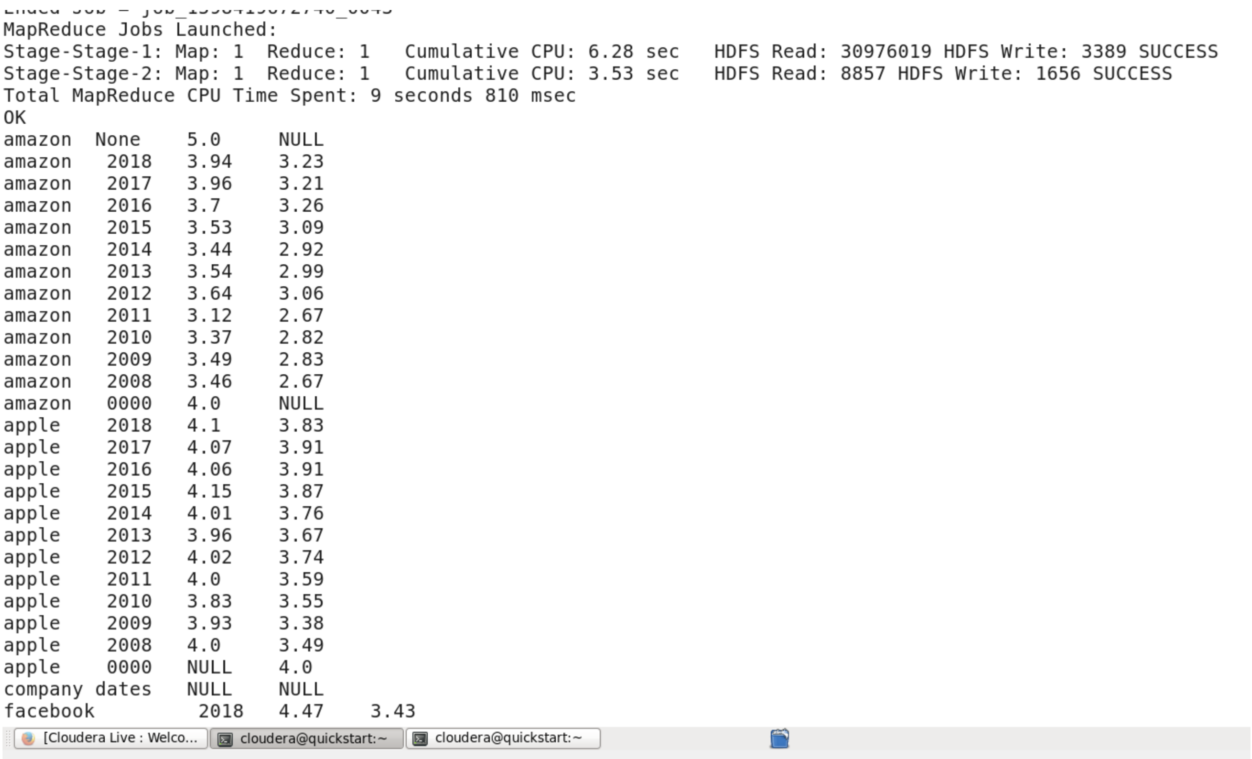
select company,country,round(percentile_approx(overall_ratings,0.25),0) as Trend_25,round(percentile_approx(overall_ratings,0.5),2) as Trend_50, round(percentile_approx(overall_ratings,0.75),2) as Trend_75 from hr_view group by company,country order by company,country;



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.96 sec   HDFS Read: 30975557 HDFS Write: 1651 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.44 sec   HDFS Read: 7379 HDFS Write: 846 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 400 msec
OK
amazon  UAE     4.0     4.0     4.0
amazon  UK      2.0     3.08    4.13
amazon  USA     3.0     3.0     3.25
amazon  West Africa     3.0     3.0     3.0
amazon  __HIVE_DEFAULT_PARTITION__      2.0     3.3     4.13
apple   Nigeria 1.0     2.5     3.75
apple   UAE     2.0     3.0     4.25
apple   UK      2.0     3.4     4.12
apple   USA     4.0     4.0     4.5
apple   West Africa     5.0     5.0     5.0
apple   __HIVE_DEFAULT_PARTITION__      3.0     3.64    4.33
company __HIVE_DEFAULT_PARTITION__      NULL    NULL    NULL
facebook        Nigeria 5.0     5.0     5.0
facebook        UK      4.0     4.33    4.67
facebook        __HIVE_DEFAULT_PARTITION__      4.0     4.31    4.65
google  Nigeria 4.0     4.0     4.25
google  Saudi Arabia    5.0     5.0     5.0
google  UK      3.0     3.99    4.5
google  __HIVE_DEFAULT_PARTITION__      3.0     4.11    4.55
microsoft       Ethiopia        2.0     2.0     3.0
microsoft       Nigeria 3.0     3.6     4.17
microsoft       UK      3.0     3.47    4.21
microsoft       West Africa     3.0     3.0     3.0
microsoft       __HIVE_DEFAULT_PARTITION__      3.0     3.45    4.11
netflix UK      5.0     5.0     5.0
netflix __HIVE_DEFAULT_PARTITION__      2.0     3.19    4.15
Time taken: 87.815 seconds, Fetched: 26 row(s)
hive> █
```

7. Display the impact of employee status on rating a company using the overall-ratings field by the company by year.

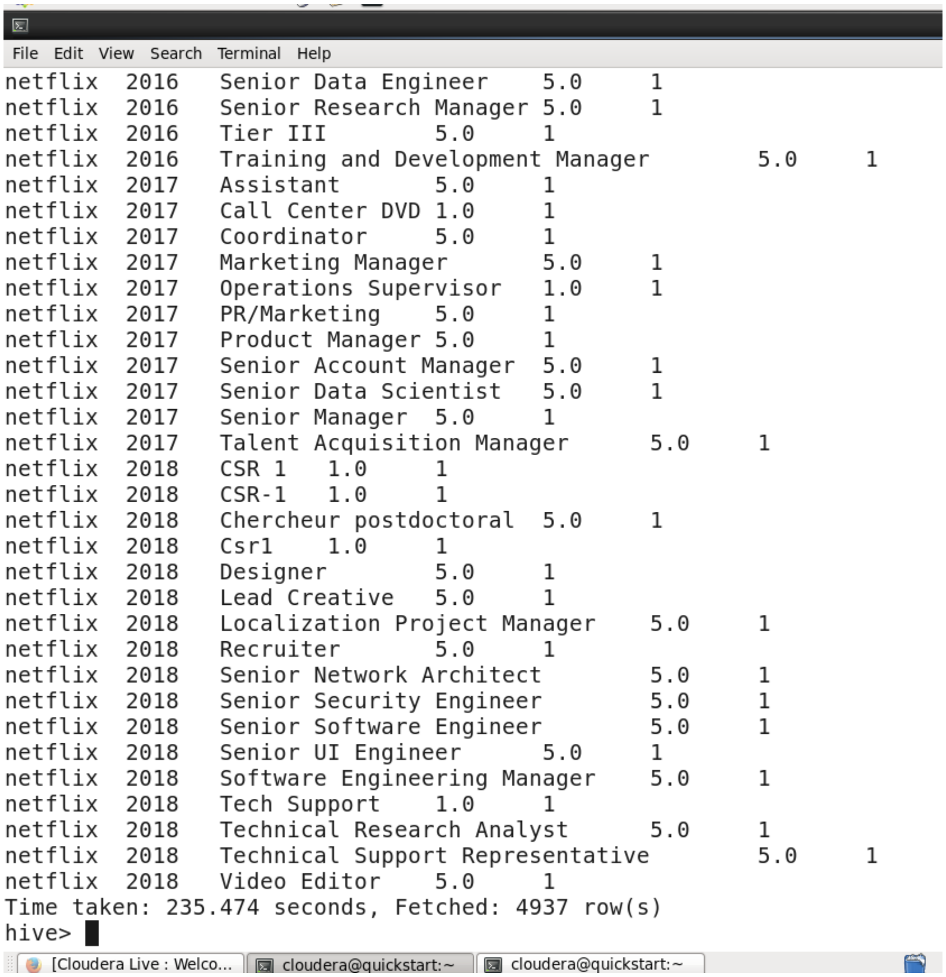
   select hr.company,hr.year,round(avg(hr.curr_emp),2) as Curr_Emp_Ratings,round(avg(hr.prev_emp),2) Former_Emp_Ratings from (select company,year,case when emp_status = 'Current Employee' then overall_ratings end as curr_emp, case when emp_status = 'Former Employee' then overall_ratings end as prev_emp from hr_view where year is not null and year != '0000') hr group by hr.company,hr.year order by hr.company,hr.year desc;



```
Ended Job = job_1538419872740_0045
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.28 sec   HDFS Read: 30976019 HDFS Write: 3389 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.53 sec   HDFS Read: 8857 HDFS Write: 1656 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 810 msec
OK
amazon  None    5.0     NULL
amazon   2018   3.94    3.23
amazon   2017   3.96    3.21
amazon   2016   3.7     3.26
amazon   2015   3.53    3.09
amazon   2014   3.44    2.92
amazon   2013   3.54    2.99
amazon   2012   3.64    3.06
amazon   2011   3.12    2.67
amazon   2010   3.37    2.82
amazon   2009   3.49    2.83
amazon   2008   3.46    2.67
amazon   0000   4.0     NULL
apple    2018   4.1     3.83
apple    2017   4.07    3.91
apple    2016   4.06    3.91
apple    2015   4.15    3.87
apple    2014   4.01    3.76
apple    2013   3.96    3.67
apple    2012   4.02    3.74
apple    2011   4.0     3.59
apple    2010   3.83    3.55
apple    2009   3.93    3.38
apple    2008   4.0     3.49
apple    0000   NULL    4.0
company dates   NULL    NULL
facebook         2018   4.47    3.43
```

8. Display the impact of job role on rating a company using the overall-ratings field by the company by year.

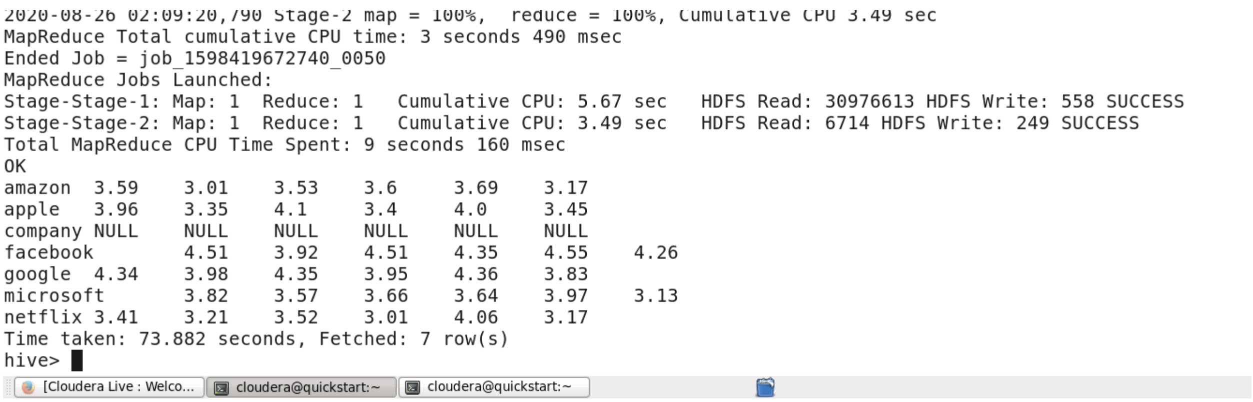
select company,year,emp_desig,ratings,rank_ratings from (
select company,year,emp_desig,round(ov,2) ratings,dense_rank() over (partition by company,year order by ov) rank_ratings from (
select company,year, emp_desig,avg(overall_ratings) as ov from hr_view where year is not null and year != '0000' group by company,year,emp_desig) hr_bad
union all
select company,year,emp_desig,round(ov,2) ratings,dense_rank() over (partition by company,year order by ov desc) rank_ratings from (
select company,year, emp_desig,avg(overall_ratings) as ov from hr_view where year is not null and year != '0000' group by company,year,emp_desig) hr_good) HR where rank_ratings < 2
order by company,year,emp_desig,ratings;


File Edit View Search Terminal Help
[Cloudera Live : Welco...
cloudera@quickstart:~
cloudera@quickstart:~

```
netflix  2016   Senior Data Engineer    5.0     1
netflix  2016   Senior Research Manager 5.0     1
netflix  2016   Tier III        5.0     1
netflix  2016   Training and Development Manager        5.0     1
netflix  2017   Assistant       5.0     1
netflix  2017   Call Center DVD 1.0     1
netflix  2017   Coordinator     5.0     1
netflix  2017   Marketing Manager       5.0     1
netflix  2017   Operations Supervisor   1.0     1
netflix  2017   PR/Marketing    5.0     1
netflix  2017   Product Manager 5.0     1
netflix  2017   Senior Account Manager  5.0     1
netflix  2017   Senior Data Scientist   5.0     1
netflix  2017   Senior Manager  5.0     1
netflix  2017   Talent Acquisition Manager      5.0     1
netflix  2018   CSR 1   1.0     1
netflix  2018   CSR-1   1.0     1
netflix  2018   Chercheur postdoctoral  5.0     1
netflix  2018   Csr1    1.0     1
netflix  2018   Designer        5.0     1
netflix  2018   Lead Creative   5.0     1
netflix  2018   Localization Project Manager    5.0     1
netflix  2018   Recruiter       5.0     1
netflix  2018   Senior Network Architect        5.0     1
netflix  2018   Senior Security Engineer        5.0     1
netflix  2018   Senior Software Engineer        5.0     1
netflix  2018   Senior UI Engineer      5.0     1
netflix  2018   Software Engineering Manager    5.0     1
netflix  2018   Tech Support    1.0     1
netflix  2018   Technical Research Analyst      5.0     1
netflix  2018   Technical Support Representative        5.0     1
netflix  2018   Video Editor    5.0     1
Time taken: 235.474 seconds, Fetched: 4937 row(s)
hive>
```

9. Display the relationship between the overall rating score vs. the rest of the rating field scores by company. Also, document your findings

```
select company,
round(avg(overall_ratings),2),
round(avg(work_bal),2),
round(avg(culture_val),2),
round(avg(career_opport),2),
round(avg(comp_ben),2),
round(avg(senior_mngnt),2) from hr_view group by company order by company;
```



```
2020-08-26 02:09:20,790 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.49 sec
MapReduce Total cumulative CPU time: 3 seconds 490 msec
Ended Job = job_1598419672740_0050
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.67 sec   HDFS Read: 30976613 HDFS Write: 558 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.49 sec   HDFS Read: 6714 HDFS Write: 249 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 160 msec
OK
amazon  3.59    3.01    3.53    3.6     3.69    3.17
apple   3.96    3.35    4.1     3.4     4.0     3.45
company NULL    NULL    NULL    NULL    NULL    NULL
facebook        4.51    3.92    4.51    4.35    4.55    4.26
google  4.34    3.98    4.35    3.95    4.36    3.83
microsoft       3.82    3.57    3.66    3.64    3.97    3.13
netflix 3.41    3.21    3.52    3.01    4.06    3.17
Time taken: 73.882 seconds, Fetched: 7 row(s)
hive> 
```

10. Document your findings for the following:
    a. Which corporation is worth working for
    Based on the Analysis done, **Facebook** has overall higher overall ratings. Also, Facebook is the second largest in Senior management ratings and third largest ratings in career opportunities.
    b. Classification of satisfied or unsatisfied employees
    First Finding: Based on the analysis performed, all the mentioned companies need to focus on company benefits and culture as there are significant change required. Ratings for these components are bad across all the companies.
    Second finding – Ratings from the senior management employees are relatively impressive than the junior members. This is a common scenario find across all the companies. Organizations should focus more on the Junior members welface which could make a remarkable change in the company ratings.