Big Data Hadoop –Social Media Project

Problem Statement

Analyze data set from Stack Exchange
As part of a recruiting exercise of the biggest social media company, they asked candidates to analyze data set from Stack Exchange. We will be using similar data set to arrive at certain key insights.
Download the data set from the following link:
http://www.ics.uci.edu/~duboisc/stackoverflow/answers.csv
The data set contains the following attributes:
qid: Unique question id
i: User id of questioner
qs: Score of the question
qt: Time of the question (in epoch time)
tags: a comma-separated list of the tags associated with the question. Examples of tags are ``html'', ``R'', ``mysql'', ``python'', and so on; often between two and six tags are used on each question.
qvc: Number of views of this question (at the time of the datadump)
qac: Number of answers for this question (at the time of the datadump)
aid: Unique answer id
j: User id of answerer
as: Score of the answer
at: Time of the answer (in epoch time)

We need to arrive at following results:

- Top 10 most commonly used tags in this data set.
- Average time to answer questions.
- Number of questions which got answered within 1 hour.
- Tags of questions which got answered within 1 hour.

Analysis Approach :

Data Import :

The Data is imported into a table in Hive . Serde is used since the Input Data is a comma separated file containing the Tags enclosed in Quotes separated by Commas .

QUERY:

```
CREATE TABLE SOWM_R.SOCIAL_MEDIA_PROJECT
( ROW_NUMBER BIGINT,QUESTION_ID BIGINT,USER_ID_Q BIGINT,QUESTION_SCORE INT,TIME_OF_QUESTION DOUBLE,TAGS STRING,
NO_OF_VIEWS BIGINT ,NO_OF_ANSWERS BIGINT, ANSWER_ID BIGINT,USER_ID BIGINT , ANSWER_SCORE INT,TIME_OF_ANSWER DOUBLE)
ROW FORMAT SERDE 'ORG.APACHE.HADOOP.HIVE.SERDE2.OPENCSVSERDE'  WITH SERDEPROPERTIES (
 "SEPARATORCHAR" = ",",
 "QUOTECHAR" = "\"",
  "SERIALIZATION.ENCODING" = 'UTF-8' );
```

OUTPUT:

```
Time taken: 0.671 seconds
hive> Create table sowm_r.social_media_project (
    > Row_Number bigint,Question_Id bigint,User_Id_Q bigint,Question_Score int,Time_of_Question double,Tags string,
    >   No_of_Views bigint ,No_of_Answers bigint, Answer_Id bigint,User_Id bigint , Answer_Score int,Time_of_Answer double)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > with serdeproperties (
    >    "separatorChar" = ",",
    >    "quoteChar"     = "\"",
    > "serialization.encoding" = 'UTF-8' );
OK
Time taken: 5.268 seconds
```

```
hive> Load Data LOCAL INPATH '/home/sowmyaradha_gmail/SMP_Data_Analysis.csv' INTO TABLE social_media_project ;
Loading data to table sowm_r.social_media_project
Table sowm_r.social_media_project stats: [numFiles=1, totalSize=24805760]
OK
Time taken: 3.748 seconds
```

RESULT :

It can be seen the Data is imported successfully using the Serde .

```
Time taken: 26.534 seconds
hive> describe social_media_project ;
OK
row_number              string                  from deserializer
question_id             string                  from deserializer
user_id_q               string                  from deserializer
question_score          string                  from deserializer
time_of_question        string                  from deserializer
tags                    string                  from deserializer
no_of_views             string                  from deserializer
no_of_answers           string                  from deserializer
answer_id               string                  from deserializer
user_id                 string                  from deserializer
answer_score            string                  from deserializer
time_of_answer          string                  from deserializer
Time taken: 0.384 seconds, Fetched: 12 row(s)
hive> select * from social_media_project where row_number = 1 ;
OK
1       563355  62701   0       1235000081      php,error,gd,image-processing   220     2       563372  67183   2       1235000501
Time taken: 0.19 seconds, Fetched: 1 row(s)
```

Data Import – Part 2 : Explode the Array of Tag Strings :

However , the tags have to be separated and stored in the Table to be used for Analysis .

Hence a new External table is created using the Lateral Explode feature to explode the Tag Array of Strings Column in the Social_Media_Project table . It can be seen that the single row is now split into the different tags and stored in the new external table .

```
CREATE EXTERNAL TABLE SOWM_R.SOCIAL_MEDIA_ET
( ROW_NUMBER BIGINT,QUESTION_ID BIGINT,USER_ID_Q BIGINT,QUESTION_SCORE INT,TIME_OF_QUESTION DOUBLE,TAGS STRING,
NO_OF_VIEWS BIGINT ,NO_OF_ANSWERS BIGINT, ANSWER_ID BIGINT,USER_ID BIGINT , ANSWER_SCORE INT,TIME_OF_ANSWER DOUBLE)
LOCATION 'HDFS:/USER/SOWMYARADHA_GMAIL/SMP_ET' ;

INSERT OVERWRITE TABLE SOCIAL_MEDIA_ET
 SELECT
ROW_NUMBER,QUESTION_ID,USER_ID_Q,QUESTION_SCORE,TIME_OF_QUESTION,TAGS,NO_OF_VIEWS,NO_OF_ANSWERS,ANSWER_ID,USER_I
D,ANSWER_SCORE,TIME_OF_ANSWER FROM SOCIAL_MEDIA_PROJECT LATERAL VIEW EXPLODE(SPLIT(TAGS,',')) SOCIAL_MEDIA_PROJECT AS
TAGS ;
```

OUTPUT:

```
hive> Insert overwrite Table Social_Media_ET
    > select row_number,question_id,user_id_q,question_score,time_of_question,tags,no_of_views,no_of_answers,answer_id,user_id,answer_score,time_of_
nswer from social_media_project  lateral view explode(split(tags,',')) social_media_project as tags ;
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_1488335516570_0733, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0733/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0733
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2017-03-04 05:22:36,502 Stage-1 map = 0%,  reduce = 0%
2017-03-04 05:22:52,003 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.83 sec
MapReduce Total cumulative CPU time: 14 seconds 830 msec
Ended Job = job_1488335516570_0733
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: hdfs://cloudlabns/user/sowmyaradha_gmail/SMP_ET/.hive-staging_hive_2017-03-04_05-22-27_305_781702246282507619 5-1/-ext-10000
Loading data to table sowm_r.social_media_et
Table sowm_r.social_media_et stats: [numFiles=1, numRows=814629, totalSize=8261958, rawDataSize=63440739]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 14.83 sec   HDFS Read: 24814569 HDFS Write: 8262046 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 830 msec
OK
Time taken: 26.534 seconds
```

RESULT :

```
hive> describe social_media_ET;
OK
row_number              bigint
question_id             bigint
user_id_q               bigint
question_score          int
time_of_question        double
tags                    string
no_of_views             bigint
no_of_answers           bigint
answer_id               bigint
user_id                 bigint
answer_score            int
time_of_answer          double
Time taken: 0.373 seconds, Fetched: 12 row(s)
hive> select * from social_media_ET where row_number = 1 ;
OK
1       563355  62701   0       1.235000081E9   php     220     2       563372  67183   2       1.235000501E9
1       563355  62701   0       1.235000081E9   error   220     2       563372  67183   2       1.235000501E9
1       563355  62701   0       1.235000081E9   gd      220     2       563372  67183   2       1.235000501E9
1       563355  62701   0       1.235000081E9   image-processing        220     2       563372  67183   2       1.235000501E9
Time taken: 0.085 seconds, Fetched: 4 row(s)
```

RESULT:

Now the table is ready to be used for analysis.

Problem Statement:

Top 10 most commonly used tags in this data set.

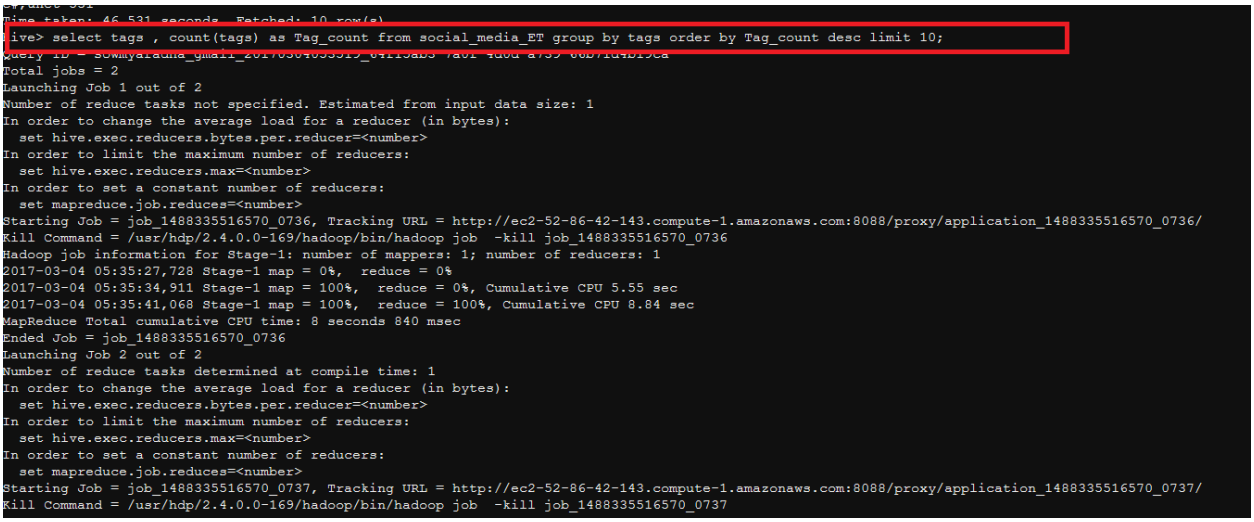
Analysis Approach :

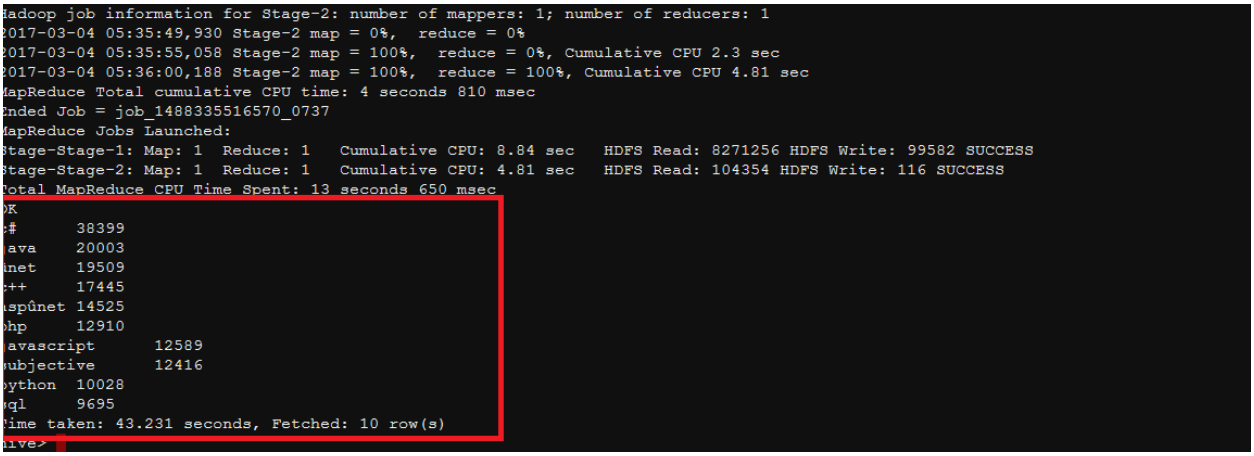
A Count of the Tags is performed by Grouping on the Tags and sorting them in descending order to obtain the Top 10 Tags .

QUERY:

SELECT TAGS , COUNT(TAGS) AS TAG_COUNT FROM SOCIAL_MEDIA_ET GROUP BY TAGS ORDER BY TAG_COUNT DESC LIMIT 10;

OUTPUT:





RESULT :

c# 38399
java 20003
ûnet 19509
c++ 17445
aspûnet 14525
php 12910
javascript 12589
subjective 12416
python 10028
sql 9695

Problem Statement:

Number of questions which got answered within 1 hour.

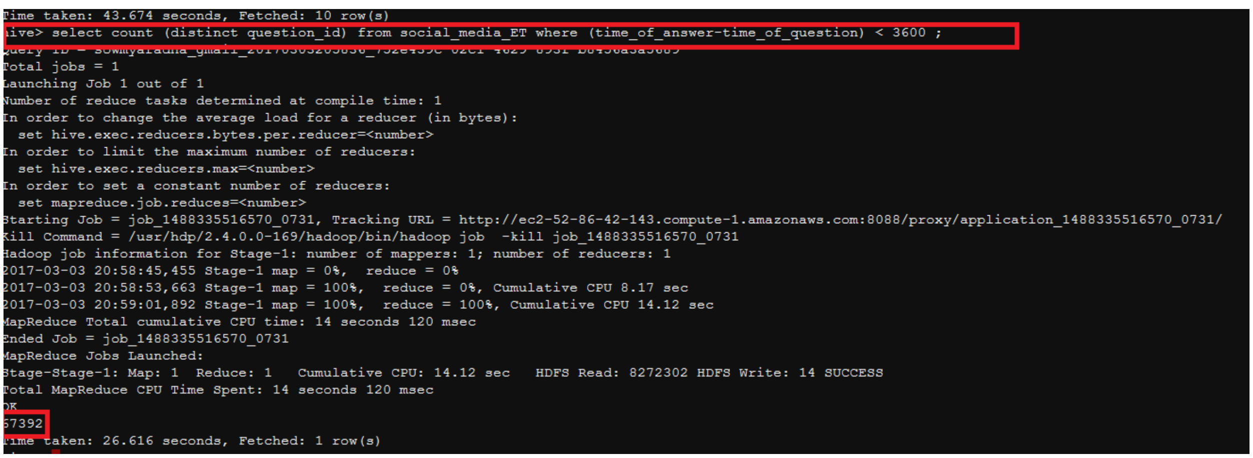
Analysis Approach:

A Count of Unique Question_Id is performed on the table whose response time is less than 60*60 =3600 seconds .

QUERY:

SELECT COUNT (DISTINCT QUESTION_ID) FROM SOCIAL_MEDIA_ET WHERE (TIME_OF_ANSWER-TIME_OF_QUESTION) < 3600 ;

OUTPUT:



```
Time taken: 43.674 seconds, Fetched: 10 row(s)
hive> select count (distinct question_id) from social_media_ET where (time_of_answer-time_of_question) < 3600 ;
[illegible]
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_1488335516570_0731, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0731/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0731
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-03 20:58:45,455 Stage-1 map = 0%,  reduce = 0%
2017-03-03 20:58:53,663 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.17 sec
2017-03-03 20:59:01,892 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.12 sec
MapReduce Total cumulative CPU time: 14 seconds 120 msec
Ended Job = job_1488335516570_0731
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 14.12 sec   HDFS Read: 8272302 HDFS Write: 14 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 120 msec
OK
67392
Time taken: 26.616 seconds, Fetched: 1 row(s)
```

RESULT :

67392

Problem Statement

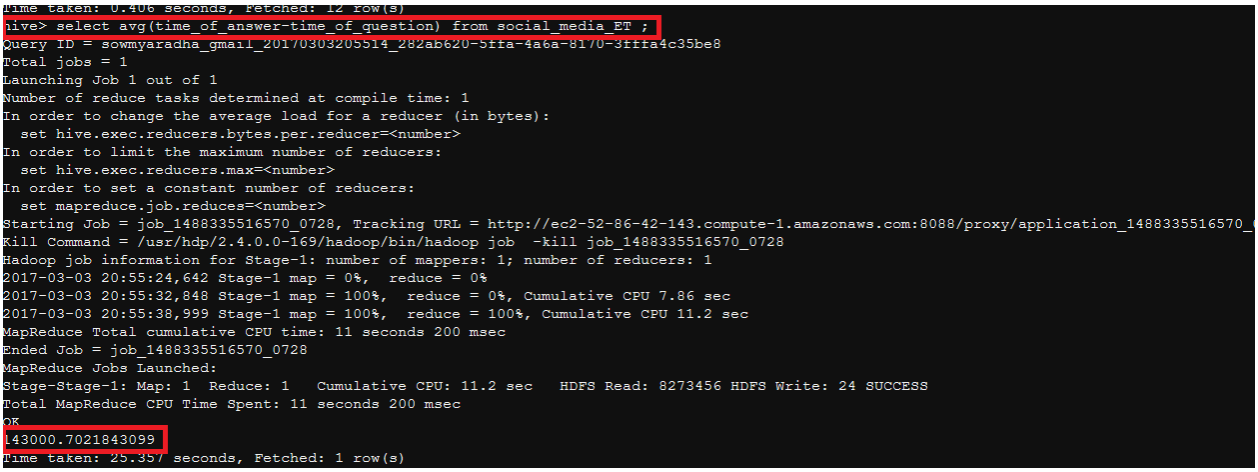
Average time to answer questions.

Analysis Approach :

An Aggregate Function Average is performed on the Response time calculated as the difference between the Time of question and Answer .

SELECT AVG(TIME_OF_ANSWER-TIME_OF_QUESTION) FROM SOCIAL_MEDIA_ET ;

OUTPUT:



RESULT:
143000.7

Problem Statement:
Tags of questions which got answered within 1 hour.

Analysis Approach:

The Social_Media Table is queried for the Tags of the questions where the difference between the time of the question and the time of the answer is less than 60*60 seconds.

The Output is written to an External Table for further analysis with the Question ID, Answer ID,Tags and Response Times .

OUTPUT:

CREATE EXTERNAL TABLE OUTPUT_RESULT_SMP (QUESTION_ID BIGINT , ANSWER_ID BIGINT, TAGS STRING, RESPONSE_TIME BIGINT) LOCATION 'HDFS:/USER/SOWMYARADHA_GMAIL/OUTPUT_SMP' ;
INSERT OVERWRITE TABLE OUTPUT_RESULT_SMP SELECT QUESTION_ID,ANSWER_ID ,TAGS,(TIME_OF_ANSWER - TIME_OF_QUESTION) RESPONSE_TIME FROM SOCIAL_MEDIA_ET WHERE (TIME_OF_ANSWER - TIME_OF_QUESTION) BETWEEN 0 AND 3600 ORDER BY RESPONSE_TIME ;

```
hive> create External table Output_Result_SMP (
    > Question_Id bigint , Answer_Id bigint, Tags string, Response_time bigint)
    > LOCATION 'hdfs:/user/sowmyaradha_gmail/Output_SMP' ;
OK
Time taken: 0.185 seconds
hive> Insert overwrite Table Output_Result_SMP
    > select question_id,answer_id ,tags,(time_of_answer - time_of_question ) response_time from social_media_ET where
    > (time_of_answer - time_of_question) between 0 and 3600 order by response_time ;
Query ID = sowmyaradha_gmail_20170304054117_19840df0-4b4a-4e3d-8910-ab9eeb314685
Total jobs = 3
Launching Job 1 out of 3
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_1488335516570_0739, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0739/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0739
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
```

```
2017-03-04 05:41:28,114 Stage-1 map = 0%,  reduce = 0%
2017-03-04 05:41:36,357 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.25 sec
2017-03-04 05:41:46,619 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.48 sec
MapReduce Total cumulative CPU time: 17 seconds 480 msec
Ended Job = job_1488335516570_0739
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: hdfs://cloudlabns/user/sowmyaradha_gmail/Output_SMP/.hive-staging_hive_2017-03-04_05-41-17_841_7885245136654915317-1/-ext-10000
Loading data to table sowm_r.output_result_smp
Moved: 'hdfs://cloudlabns/user/sowmyaradha_gmail/Output_SMP/000000_0.deflate' to trash at: hdfs://cloudlabns/user/sowmyaradha_gmail/.Trash/Current
Table sowm_r.output_result_smp stats: [numFiles=1, numRows=527467, totalSize=3886526, rawDataSize=13715069]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.48 sec   HDFS Read: 8274151 HDFS Write: 3886616 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 480 msec
OK
Time taken: 31.602 seconds
```

```
Time taken: 24.013 seconds, Fetched: 20 row(s)
hive> select * from output_result_SMP limit 20;
OK
913697  913698  vista   0
913697  913698  windows 0
913697  913699  vista   0
935697  935698  oracle  0
935697  935698  sqlserver       0
935697  935698  automation      0
935697  935698  scheduled-task  0
913697  913699  windows 0
851980  851984  books   1
851980  851983  operating-system        1
851980  851983  system-programming      1
851980  851984  subjective      1
851980  851983  linux   1
851980  851983  books   1
851980  851982  subjective      1
851980  851982  operating-system        1
851980  851982  system-programming      1
851980  851982  linux   1
851980  851982  books   1
851980  851981  subjective      1
Time taken: 0.105 seconds, Fetched: 20 row(s)
```

FURTHER ANALYSIS:

The Tags with maximum occurrence and average Response time for the questions with the tag have been obtained from the earlier output .

QUERY :

SELECT TAGS,COUNT(TAGS) TAG_COUNT ,AVG(RESPONSE_TIME) FROM OUTPUT_RESULT_SMP GROUP BY TAGS ORDER BY TAG_COUNT DESC LIMIT 20;

```
Time taken: 44.806 seconds, Fetched: 20 row(s)
hive> select tags,count(tags) tag_count ,avg(response_time) from output_result_SMP group by tags order by tag_count desc limit 20;
Query ID = sowmyaradha_gmail_20170304054633_3aee282d-2307-4e53-a894-219518729e47
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_1488335516570_0746, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0746/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0746
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-04 05:46:42,824 Stage-1 map = 0%,  reduce = 0%
2017-03-04 05:46:50,035 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.11 sec
2017-03-04 05:46:55,183 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.56 sec
MapReduce Total cumulative CPU time: 9 seconds 560 msec
Ended Job = job_1488335516570_0746
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_1488335516570_0747, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0747/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0747
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-04 05:47:05,850 Stage-2 map = 0%,  reduce = 0%
2017-03-04 05:47:12,011 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.35 sec
2017-03-04 05:47:17,142 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.71 sec
MapReduce Total cumulative CPU time: 4 seconds 710 msec
Ended Job = job_1488335516570_0747
```

```
Starting Job = job_1488335516570_0747, Tracking URL = http://ec2-52-86-42-143.compute-1.amazonaws.com:8088/proxy/application_1488335516570_0747/
Kill Command = /usr/hdp/2.4.0.0-169/hadoop/bin/hadoop job  -kill job_1488335516570_0747
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-03-04 05:47:05,850 Stage-2 map = 0%,  reduce = 0%
2017-03-04 05:47:12,011 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.35 sec
2017-03-04 05:47:17,142 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.71 sec
MapReduce Total cumulative CPU time: 4 seconds 710 msec
Ended Job = job_1488335516570_0747
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.56 sec   HDFS Read: 3895471 HDFS Write: 142353 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.71 sec   HDFS Read: 147729 HDFS Write: 410 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 270 msec
OK
c#      28965   826.8738132228551
inet    14065   839.2381798791326
ava     12784   961.707681476846
c++     12162   924.1094392369677
aspůnet 10226   887.8102875024448
avascript       9152    863.9641608391609
hp      8924    902.2889959659345
ubjective       7716    943.694142042509
ql      7134    860.7942248388001
ython   6375    1016.7074509803922
est-practices   6017    915.5281701844773
        5678    901.5540683339204
qlserver        5290    880.452741020794
query   5045    922.1302279484638
tml     5022    850.6326164874552
ysql    4074    852.980117820324
ss      3485    867.1190817790531
atabase         3230    919.3674922600619
indows 3177     950.4249291784703
ot-programming-related 2898     797.7097998619738
ime taken: 44.227 seconds, Fetched: 20 row(s)
```