Data Analysis and Visualization Using Hive and Tableau

Project II

Instructor:

Dr. Daisy Zhe Wang

Submitted By
Atul Garg - 94432505
Ishadutta Yadav- 54931916

Enron Dataset Analysis

Background:

- The Enron Corpus is a large database of emails of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.[1] A copy of the database was subsequently purchased for $10,000 by Andrew McCallum, a computer Scientist at the University of Massachusetts Amherst.[2]
- The Enron scandal, revealed in October 2001, eventually led to the bankruptcy of the Enron Corporation, an American energy company which was found in 1985 by Kenneth Lay. Several years later, when Jeffrey Skilling was hired, he developed a staff of executives that, by the use of accounting loopholes, special purpose entities, and poor financial reporting, were able to hide billions of dollars in debt from failed deals and projects.

Data processing pipeline & Hive queries

Person-Pairs who have the Top-40 Correspondences

Query:

```
Select * from (select mailto as email1, mailfrom as email2, count(*) as count
from emails where mailto like '%@%.com' and mailto > mailfrom group by
mailto ,mailfrom UNION ALL select mailfrom as email1, mailto as email2,
count(*) as count from emails where mailfrom like '%@%.com' and mailfrom >
mailto group by mailfrom,mailto) a order by a.count desc limit 40;
```

Hypothesis and Assumptions:

For evaluating top-k correspondence in the mail data set of Enron following hypothesis have been applied:

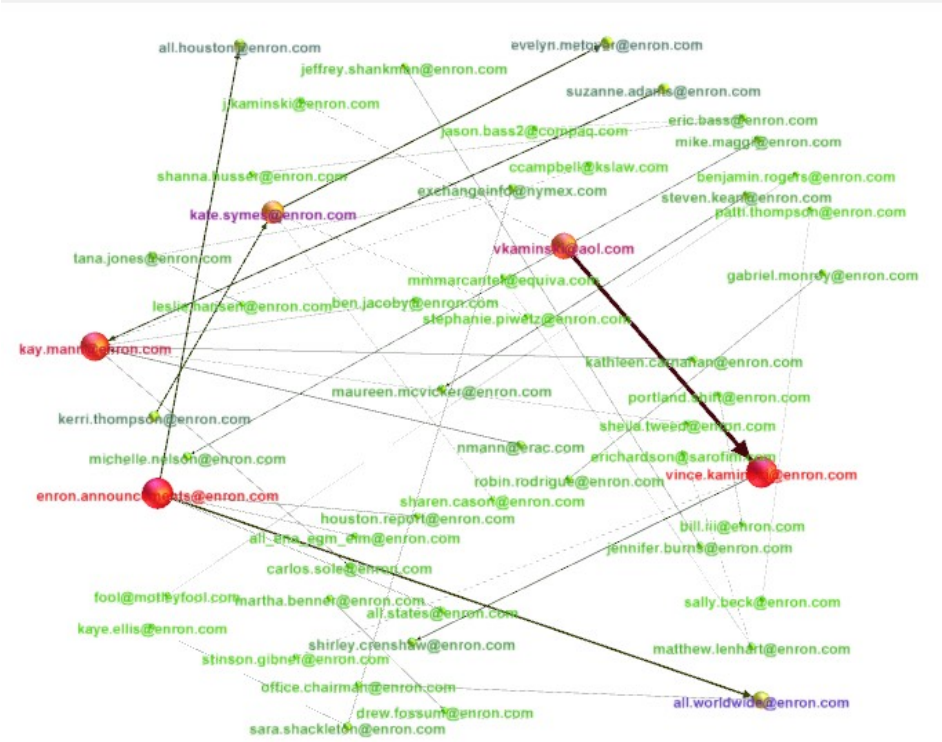
1. Mails sent to more than one set of recipients is not accounted for each of the recipients individually.
2. Number of mails exchanged decides importance for the person in company.

Analysis and Inference

By the Mail Exchange network following people are inferred to be of high significance:

1. Vincent Kaminski (Managing Director for research)
2. C Kay Mann (Assistant General Counsel)
3. Suzanne Adams (Legal Assistant)
4. Steven Kean (Vice President)
5. Kate Symes (Employee)

Visualization:



Most Important Topics in Year 2001

Query:

```
SELECT explode(ngrams(sentences(lower(a.subject)), 1, 10)) as abc FROM (select
subject from emails where timestamp like '%2001%') a;
```

Hypothesis:

Evaluation of Important topics in year 2001 are accounted on the basis of frequency of occurrence of each term in subject of emails exchanged within Enron network.

Analysis and Inference:

Many of the common stop words have been removed from the result set using list of stop words provided by MYSQL. Lot of conventional business language words are seen occurring frequently.

Visualization:

Total Messages per Month

Query:

```
select count(*),sentences(e.timestamp)[0][2],m.mid as
smid,sentences(e.timestamp)[0][3] as syear
from emails e JOIN monthid m
ON (m.mon=sentences(e.timestamp)[0][2])
sentences(e.timestamp)[0][2],m.mid,sentences(e.timestamp)[0][3] order by
syear,smid ;
```

Hypothesis:

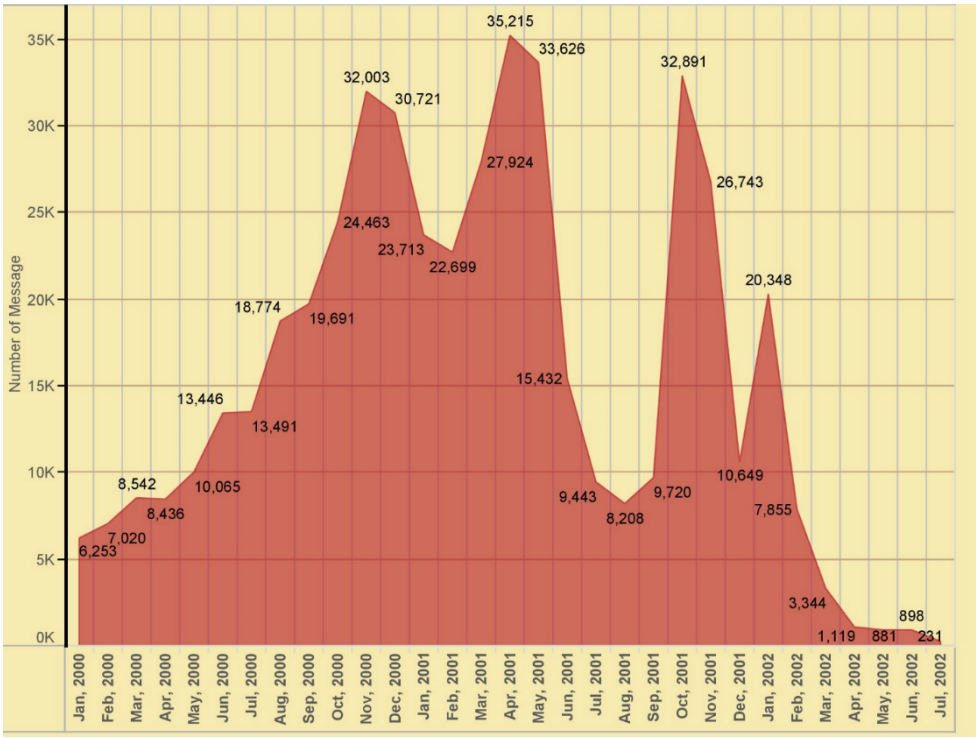
Each mail sent to more than one recipients is only considered as count of one and not counted individually for each of the recipients.

Analysis and Inference

We get 3 peak in total number of message around October 2000, April-May 2001 and October 20001. These 3 peak correspond to following 3 major event in Enron.

1. October-November 2000 - Enron Stock prices are all time high.
2. April-May 2001- Several Wall Street analyst accuses Enron for unusual accounting practice.
3. October- November 2001 - Restructuring losses and SEC investigation, credit rating downgrading and filed bankruptcy.

Visualization:



Top-40 Central Person in Enron Case

Query:

```
Select * from (select mailto as email1, mailfrom as email2, count(*) as count from emails where mailto like '%@%.com' and mailto > mailfrom group by mailto ,mailfrom UNION ALL select mailfrom as email1, mailto as email2, count(*) as count from emails where mailfrom like '%@%.com' and mailfrom > mailto group by mailfrom,mailto) a order by a.count desc limit 40;
```

Hypothesis:

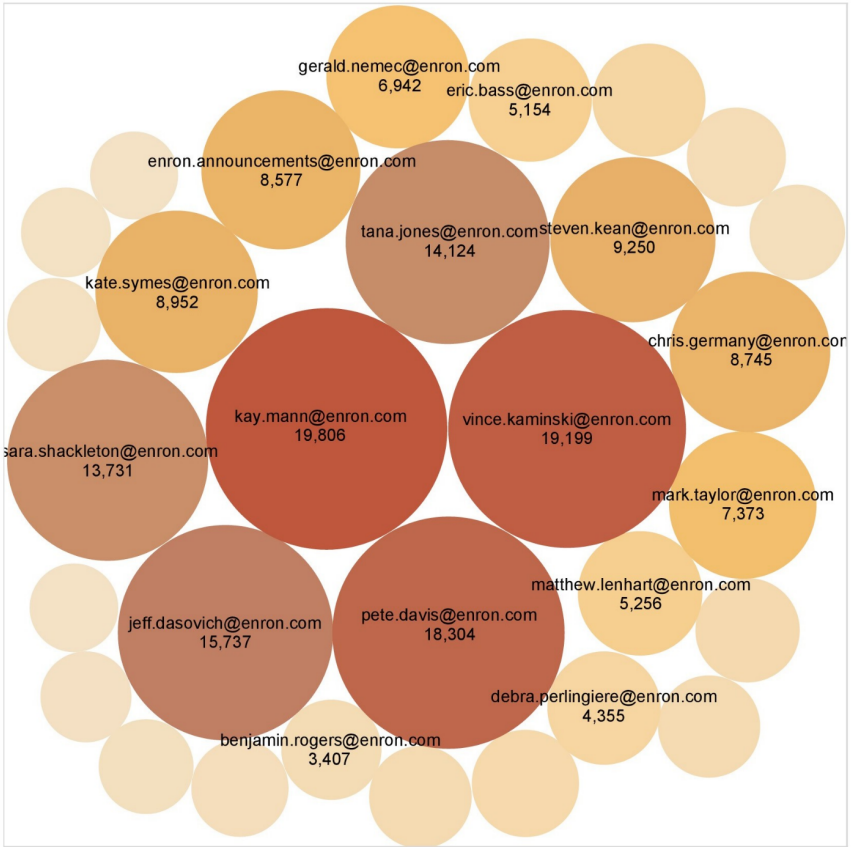
For Evaluation of Top – 40 people at Enron, hypothesis so used is number of mails exchanged by each person decides centrality of the person. Higher is the number of mails, more important is the person.

Analysis and Inference

As per the Results following people are central figures at Enron in order.

1. Vincent Kaminski
2. C Kay Mann
3. Pete Davis
4. Jeff Dasovich
5. Sara Shackleton
6. Steven Kean
7. Mark Taylor
8. Kate Syme

Visualization:



Other Hive Queries and Results

➢ 2001 Top Sender

Query:

```
select mailfrom, sentences(timestamp)[0][2] , count(*) as count from emails
where mailfrom like '%@%.com' and sentences(timestamp)[0][3] ='2001' group by
mailfrom,sentences(timestamp)[0][2] order by count desc limit 10;
```

Results:

Person	Month	Number of Mails Sent
pete.davis	Apr, 2001	4375
kay.mann	Apr, 2001	2132
kay.mann	Jan, 2001	1638
kay.mann	Mar, 2001	1544
kay.mann	May, 2001	1519
vince.kaminski	Apr, 2001	1358
vince.kaminski	Jan, 2001	1340
jeff.dasovich	Apr, 2001	1265
kate.symes	Mar, 2001	1205
kay.mann	Feb, 2001	1189

Netflix Dataset Analysis

Background:

- The movie rating files contain over 100 million ratings from 480 thousand randomly-chosen, anonymous Netflix customers over 17 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received during this period. The ratings are on a scale from 1 to 5 (integral) stars. To protect customer privacy, each customer id has been replaced with a randomly-assigned id. The date of each rating and the title and year of release for each movie id are also provided.[5]

Data processing pipeline & Hive queries

Total Number of Movies Watched Per Month

Query:

```
select count(1) as num_views, month(date) as month, year(date) as year
from movies_ratings group by year, month order by year, month;
```

Hypothesis:

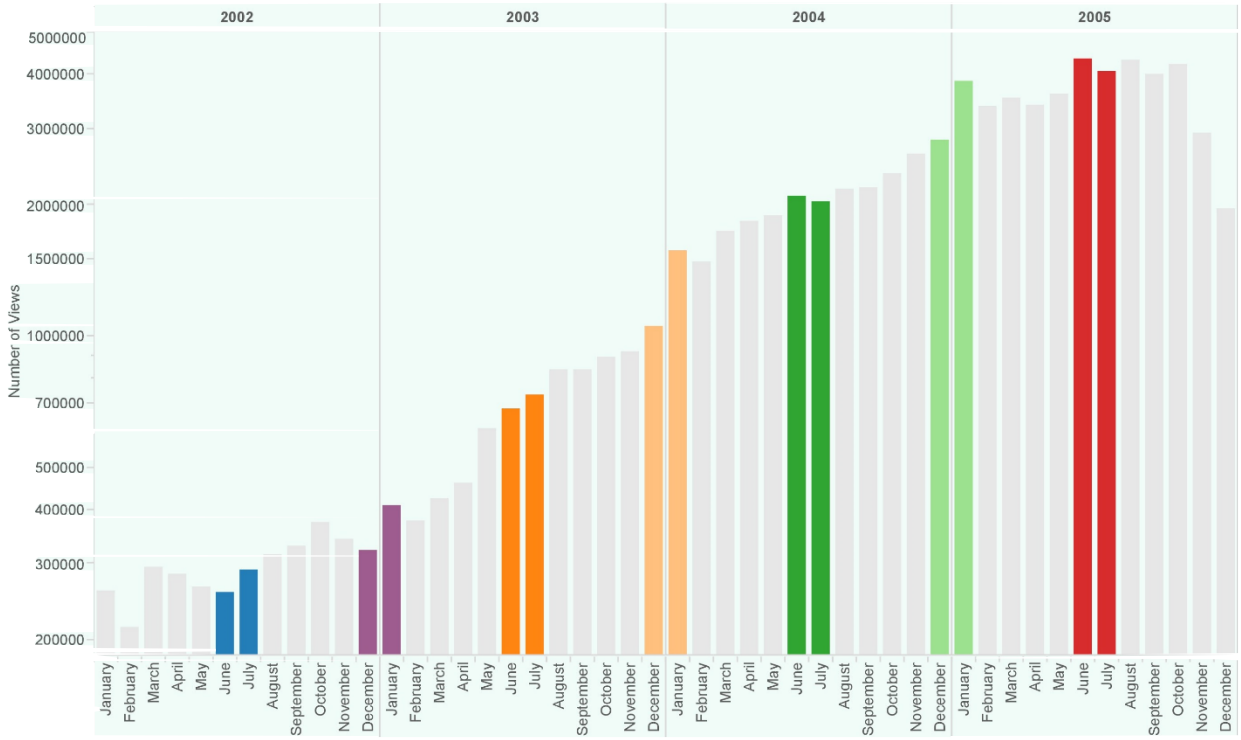
Movie view count is based on the assumption that every a time a movie is seen at Netflix, by any user rating for the movie is updated. Association rule so used here is if movie is seen, it will be rated.

Analysis and Inference:

Following Inferences can be concluded based on the representation.

1. Total number of views experience a spike every January which may be credited to either holiday season or scheduled release of movie/season.
2. Average Number of views during holiday season is moderately high during period of year (2002-2005).

Visualization:



Top 40 Movies of All Time

Query:

```
select t.title, a.rating from movie_titles t JOIN (select  avg(mr.rating)
as rating, mt.mid as mid from
movie_titles mt JOIN movie_ratings mr
ON mr.mid=mt.mid group by mt.mid) a
ON t.mid=a.mid
order by a.rating limit 20;
```

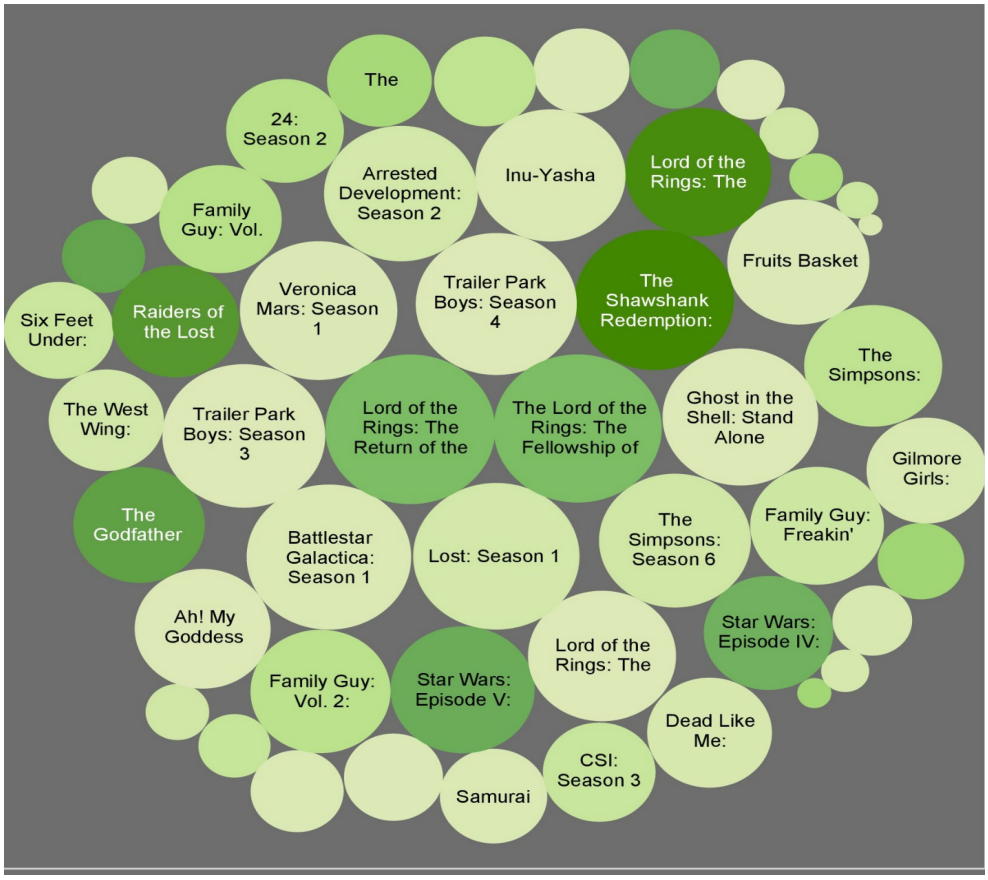
Hypothesis:

Top movies are evaluated using average rating of each movie as per the user, which is normalized by number of viewers of each movie.

Analysis and Inference

It can be inferred from the result set that number of views for the movie do not actually contribute for best rated movie. While if evaluated on both parameter a better rating for the movie can be evaluated. Movie like “The Shawshank Redemption: Special Edition” had maximum number of views as 1 39,660 while rating for it was 4.59, while movie “Lord of the Rings: The Return of the King: Extended Edition” which tops the chart with best rating of 4.72 has only 73,335 views.

Visualization:



Trends in Movie Title

Query:

```
SELECT explode(ngrams(sentences(lower(a.title)), 1, 300)) as abc FROM (select title
from movie_title) a;
```

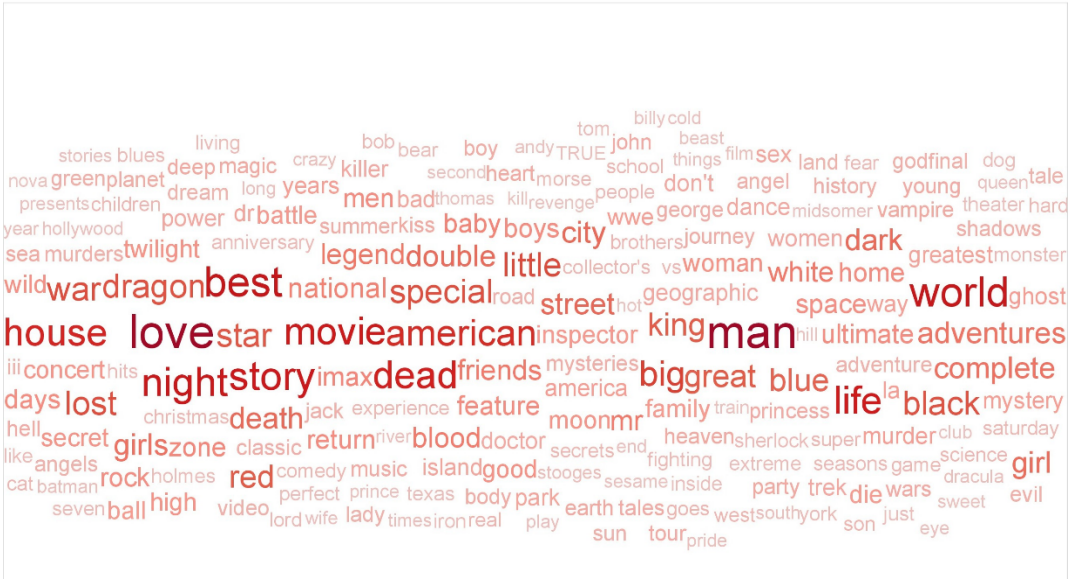
Hypothesis:

Evaluation of Important topics in movie title are accounted on the basis of frequency of occurrence of each word in movie title.

Analysis and Inference

Many of the common stop words have been removed from the result set using list of stop words provided by MYSQL. Lot of conventional movie topics such as love,man,life are seen occurring frequently.

Visualization:



Programing lessons and good resources

Hive Tutorials

1. Hive Manual (https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF)
2. http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/
3. http://www.orzota.com/hive-for-beginners/

Data Mining:

1. https://cwiki.apache.org/confluence/display/Hive/StatisticsAndDataMining
2. http://www.autonlab.org/tutorials/
3. http://blog.sqlauthority.com/2013/10/21/big-data-data-mining-with-hive-what-is-hive-what-is-hiveql-hql-day-15-of-21/

Enron Dataset

1. http://www.economist.com/node/940091
2. http://snap.stanford.edu/data/email-Enron.html

Netflix Dataset

1. http://www.rasch.org/rmt/rmt233d.htm
2. http://www.igvita.com/2006/10/29/dissecting-the-netflix-dataset/

Runtime experience for queries and visualizations

1. A Join Query over Hive database is a very heavy operation. E.g. self-Join over movie_ratings takes more than 1 Hour.
2. Tableau allows you to create a wide variety of interactive graphs, maps and tables and organize them. Example: We have easily created bar plot, treemaps, word clouds, and bubble charts by just drag and drop different dimensions.
3. It is not very convenient to draw graph related visualization in Tableau. We have used open source software “Gephi” for this purpose.

Difficulties faced and Learning_

1. Map join is faster than the common join, it's better to run the map join whenever possible.
2. We learn different techniques to write efficient hive query. E.g. Partition Hive table, bucketing hive table, Bucket sampling, block Sampling and parallel execution will decrease the latency of hive queries. While working on query to find similar movies we realized importance of using partition on attributes while table creation which will greatly enhance performance of query execution.
3. Better understanding of Hive which help in solving big data Challenge and keep us upfront with upcoming technology.
4. Sound understanding of distributed environment and use of distributed systems to solve computational problems.
5. Ability to visualize data effectively leads directly to better understanding, insight and better decisions.

Division of Work:

We realized not to divide the data set among the members since this will limit the scope of data-interpretation. Correspondence analysis was done by Atul Garg while context analysis was done by Ishadutta Yadav for Enron Data set. For Netflix dataset collaborative approach

REFERENCES

1. http://hive.apache.org/
2. https://cwiki.apache.org/confluence/display/Hive/Home
3. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-hive.html
4. http://hortonworks.com/products/hortonworks-sandbox/#tutorial_gallery
5. http://aws.amazon.com/redshift/partners/tableau/
6. http://www.tableausoftware.com/learn