Startup Data Analysis Using Hadoop

Roshik Ganesan, Utsav Malpani, Vignesh Srinivas
Department of Information Systems, California State University
Los Angeles
rganesa@caltstatela.edu
umalpan@calstatela.edu
vravish@calstatela.edu

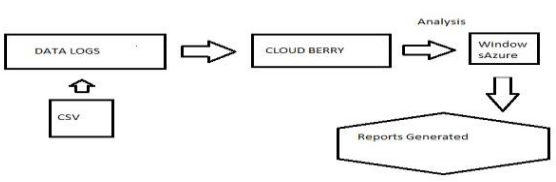
Abstract - The paper analyses over two decades of startup data from around the world. The data is analyzed using Hadoop cluster and Hive QL. The best market domain, location and the average funding that can be received by a startup have been inferred from this analysis, which can be used as a guide by future entrepreneurs.

1. Introduction

A startup is a company designed to grow fast. Being newly found does not in itself make a company a startup but it should explore an unknown or innovative business model in order to disrupt existing markets. Eg. Amazon and Uber.
In the past few years there has been a rapid increase in the number of startups that are starting but all of them do not succeed and close down due to factors like poor funding, slow market segment, lack of expertise, no buyers to take over the company etc.
Through this analysis we aim to provide a guide to future entrepreneurs which will guide them regarding the segments in which they should invest, the place where they should start, whom they should approach for funding and who can they contact if they want their company to be acquired.

2. Work Flow

Initially a data set comprising the details of startups was downloaded from a trusted source. This data set contains of the name of the company their funding details, the acquirement details along with the function market. As the data was in separate sheet in the same workbook each data was stored as a separate excel workbook. These data logs in CSV format were uploaded in Azure using cloudberry, a tool that helps users to visualize and manage files in Microsoft azure efficiently. This tool serves as a bridge between the local system and the cloud data storage. The "COPY" function from cloudberry is used in this case to put the data logs into the cloud storage. The data is analyzed using the hive queries and the output files are stored in cloud which is then downloaded and opened in excel format. The visualization from the output files are obtained using excel charts (Power View) and tableau.



2.1 Data Storage

The Azure cluster comprises of the following architecture, a master and slave. A master node also called as the Head node in Azure manages the queries in the cluster and the data node which is the also called the worker node executes the commands. HiveQL is used as a querying language to extract the data from the data logs in the cluster.

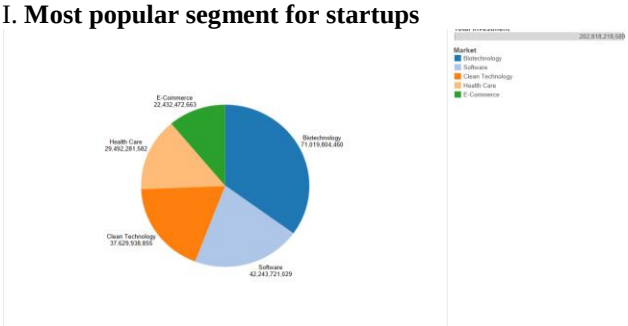
2.2 Data Analysis and Representation

In this Project, we have considered four main parameters to analyze the data.
1.The best market segments to enter into
2.Suitable place to start the venture
3.Average funding based on market segment
4.Prospective investors
These data were obtained by writing suitable queries in Hive QL.
The Following Hive Query was executed to obtain the net worth of the companies:

```
select distinct
c.name,c.funding_total_usd,a.acquirer_name,a.price_amou
nt,(a.price_amount - c.funding_total_usd) as
acqvalue,c.market from companies c join acquirer a on
(c.name = a.company_name) sort by acqvalue desc;
```

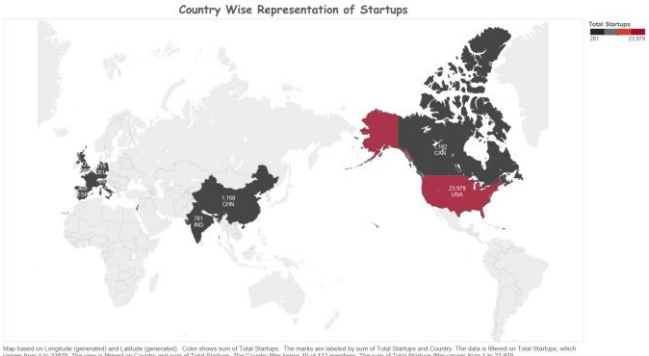
Below is the screenshot for the output of the above query -

I. **Most popular segment for startups**



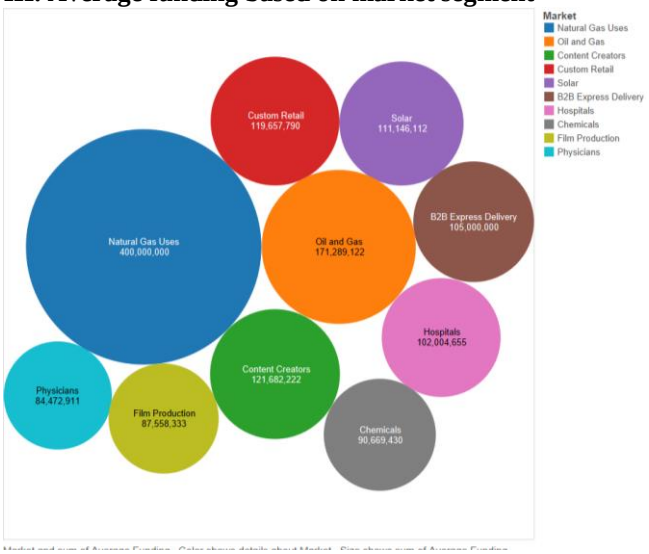
From this pie chart we see that the Biotechnology market is the most funded with comprising 35% of the entire funding followed by the software and clean technology with 21% and 19% respectively.

II. Countries with the maximum startups



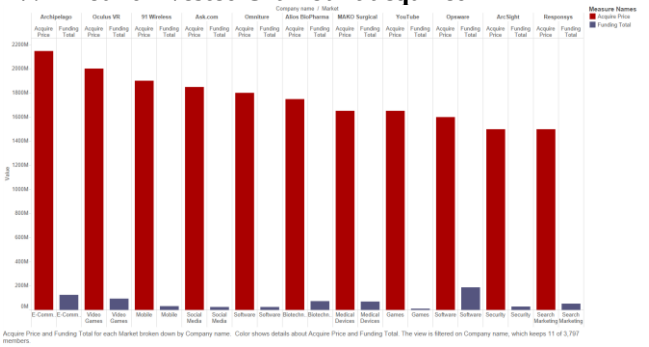
This analysis shows that USA leads with the most number of registered startups with 23k startups followed by London, Canada, China and India.

III. Average funding based on market segment



As we can see in this bubble chart, Natural Gas segment has received the highest funds i.e 400M followed by Oil and Gas with 171 M along with Content Creators having 121M of funds.

IV. Amount invested & amount acquired



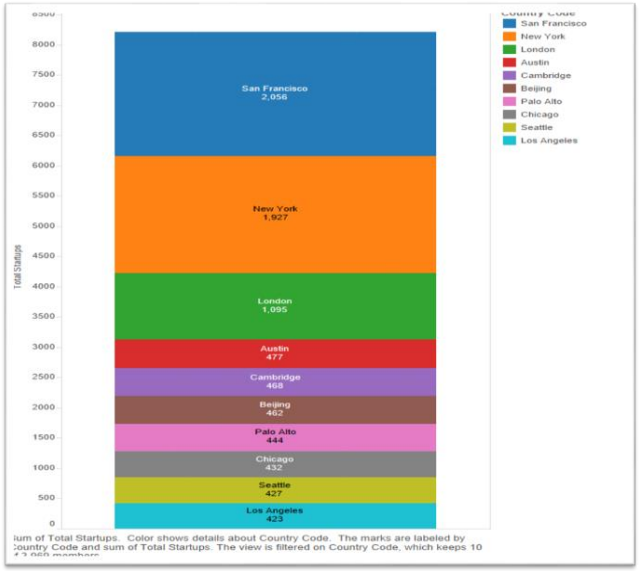
E-commerce companies had the highest increase in its net worth followed video games and mobile companies. It means that these companies had invested very little as compared to their acquisition prices.

It is worthwhile to mention that companies from the highest funded market domain and companies from the high average funding market domain are acquired the least and *7/10* companies with highest value are from the software or the entertainment market.

V. States in the U.S that made the highest acquisitions

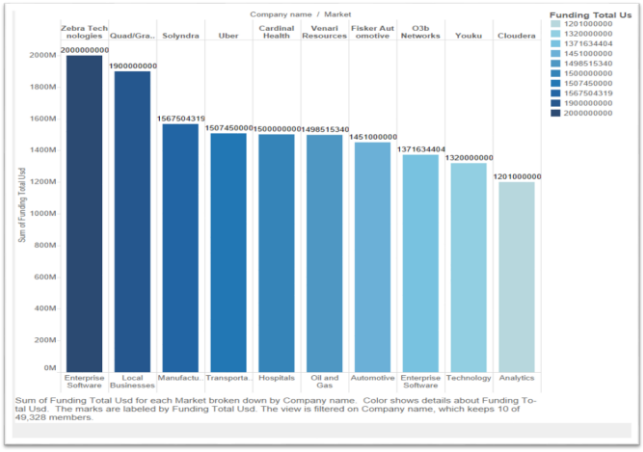
Since the maximum number of startups and acquisitions are in the United States, we analyzed which state has acquired the maximum number of startups and we found that 14.5% of the acquisitions are form the state of California followed by New York 5%.

VI. Popular City for Startups

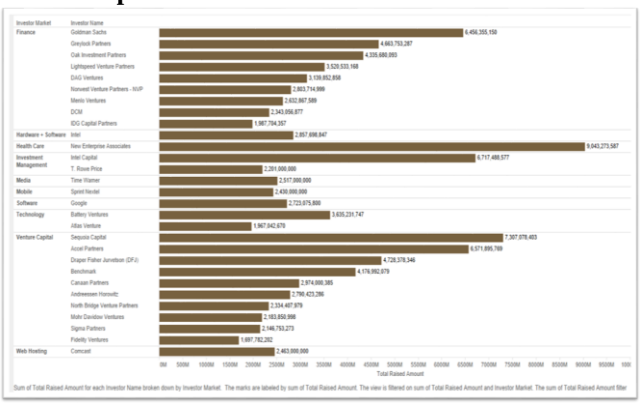


We analyzed that San Francisco tops the cities list with 2k followed by New York and London having 1.9k and 1k respectively

VII. Companies with Maximum Funding



VIII. Companies with Maximum Investment



The analysis gives an idea on the companies to target while startups are looking for funding.

3.Conclusion

From the above analysis, we can conclude that companies seeking funding from foreign companies could approach Goldman Sachs, Intel, Google based on their respective market segments.

Entrepreneurs looking only to make profit by selling their venture some years down the line for high net worth could venture into markets like software, entertainment, E-Commerce or Social Media preferably in states of CA or NY. The perks of doing so are low investments and high acquirement chances.

Start-ups looking for a long-term company ownership with high funding could target the following market domains -

- Bio-Technology
- Natural Gas
- Oil and Gas

4. Limitations and Future Scope

From the available data, holding few parameters, we were able to provide the solution that will cater the basic needs of a startup. Had the data been more detailed i.e. holding information regarding the scale of the company funding (Large or Small) and the amount it received in each round of, it would have been possible to analyze the funding that a particular segment receives on the basis of the size of the company.

The dataset did not provide the details on the investment the company has made apart from the funding, which would have been helpful in suggesting the amount, required to build a startup.

Github Link –
https://github.com/vigyr/Calstatela

Data Set Link –
https://www.dropbox.com/s/jr997tkty186apu/Crunchbase.xlsx?dl=0

References

Teach Asia
[1] https://www.techinasia.com/talk/27-striking-facts-startups-world-infographic

Hadoop Tutorials
[2] https://hadoop.apache.org/

Apache Hive TM
[3] https://hive.apache.org/

How To Process Data with Apache Hive
[4]http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/

Intro to Hive
[5]http://blog.cloudera.com/wp-content/uploads/2010/01/6-IntroToHive.pdf

Demo Analyzing data with hue and hive
[6]http://blog.cloudera.com/blog/2013/04/demo-analyzing-data-with-hue-and-hive/

HD Insights: Get Started Hadoop Tutorial
[7]https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-tutorial-get-started-windows/

Connecting excel with Hive
[8]https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/

HD Insights connect to excel power query
[9]https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-power-query/

Mapred Tutorial
[10]https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

10 Ways to query hadoop
[11]http://www.infoworld.com/article/2683729/hadoop/10-ways-to-query-hadoop-with-sql.html