**STARTUP DATA ANALYSIS USING AZURE HD INSIGHTS AND TABLEAU**

By Roshik Ganesan ([rganesa@calstatela.edu](mailto:rganesa@calstatela.edu)), Utsav Malpani ([umalpan@calstatela.edu](mailto:umalpan@calstatela.edu)) and Vignesh Srinivas ([vravish@calstatela.edu](mailto:vravish@calstatela.edu))

12/10/2016

**Objective:**

This tutorial is about the Startup data analysis using Hadoop cluster in Azure HD insights and visualizations using tableau. You will gain working knowledge about creating a Hadoop cluster in Azure HD insights, using Cloudberry explorer to import data into HDFS and visualizing the analyzed data in tableau.

The data set has a rich data on startup ventures from year 2000 – 2014. It is categorized by the Name of the company, market domain, country/region, current status, found year, investor details with amount invested, and also the acquisition details. The analyzation results will give us an idea about safest market segment, safest city, average funding one could receive for a startup etc., based on the past trend.

**Technologies Used:**

* Microsoft Azure HD Insight
* Cloudberry Explorer for Azure Blob Storage
* Tableau

**Pre-requisites:**

**Setup Tableau with the License:**

If you have the Tableau software with active license you are good to go. Else you can get the Tableau for students from the below link,

<http://www.tableau.com/academic/students>

You can use your institution ID to register and get free tableau desktop license for 1-year .

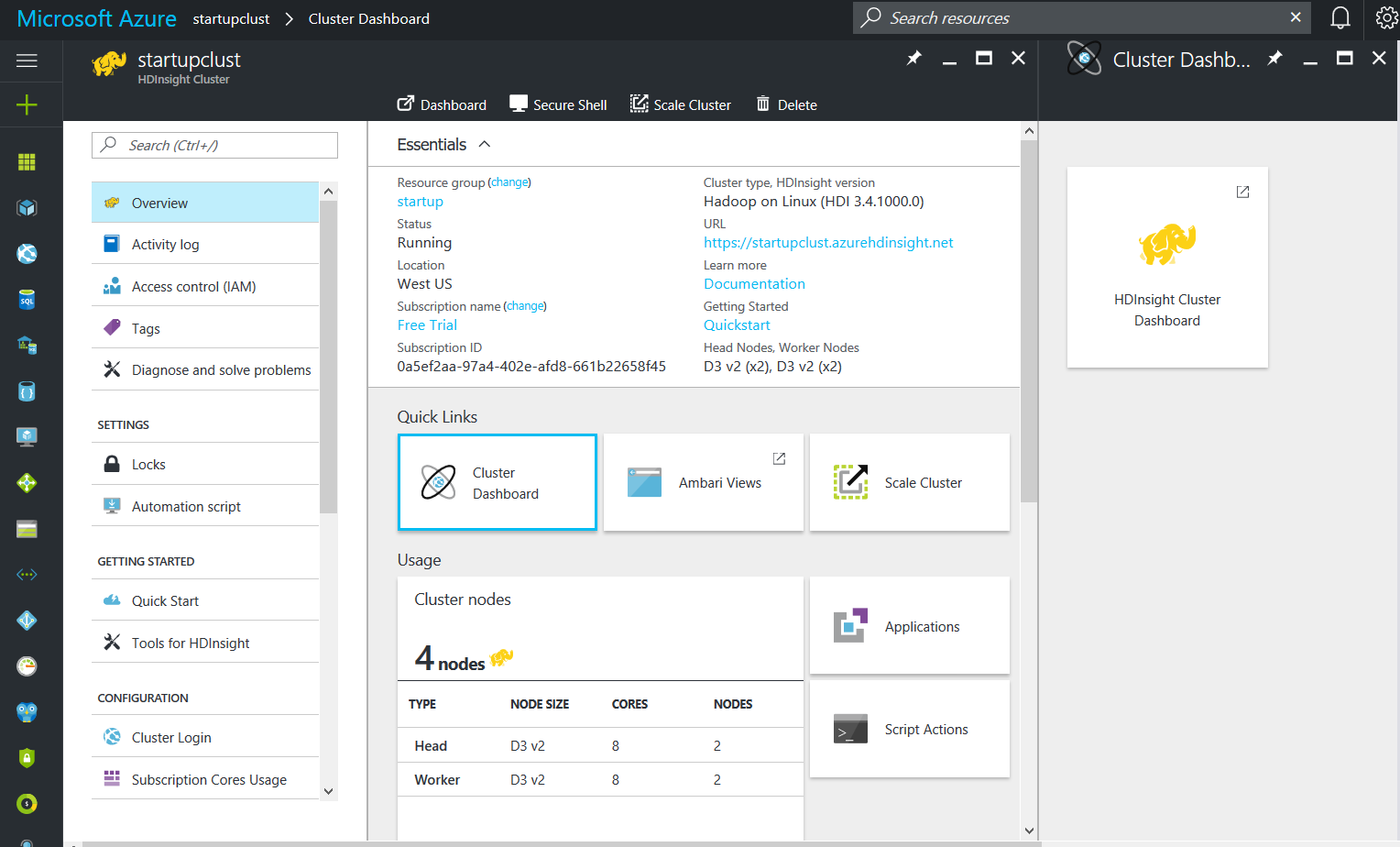
**Creating a Hadoop Cluster on Microsoft Azure HD insights:**

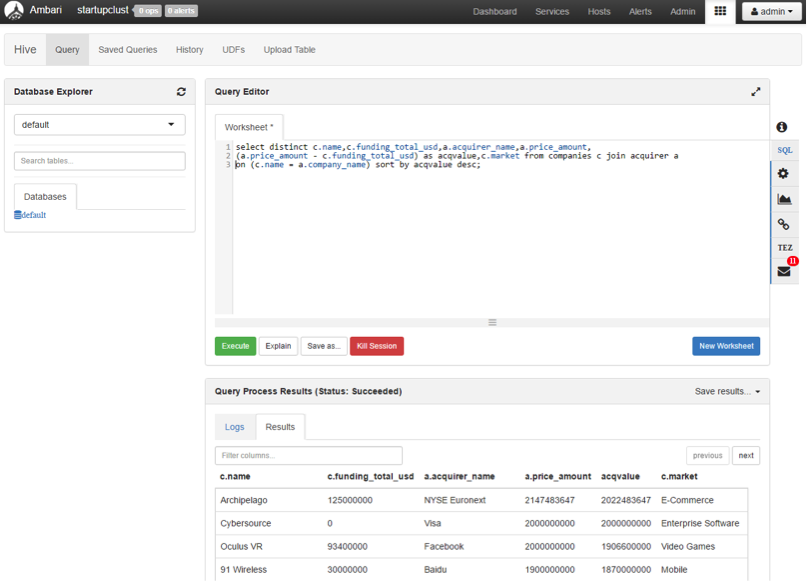
You have to sign up Microsoft Azure cloud computing account. You may go to https://azure.microsoft.com/en-us/free/ in order create a 30 days free trial with 200$ credit limit. (Make sure you sign up with free trail and **not with Pay-as-you-go subscription** which will charge you from your card, once the free credit is over)

* In Azure portal, Click New 🡪 Data + Analytics 🡪 HDInsight
* Choose the cluster type as Hadoop and give the authentication details.
* Choose the hardware configuration as below,

|  |
| --- |
| Number Of Head Nodes = 2 | CPU = 8 Cores |
| Number Of Worker Nodes = 2 |CPU = 8 Cores |
| Processor - 2.4 GHz Intel Xeon® E5-2673 v3 |
| Ram – 24 GB |
| Disks – 16 |
| Local SSD – 400 GB |

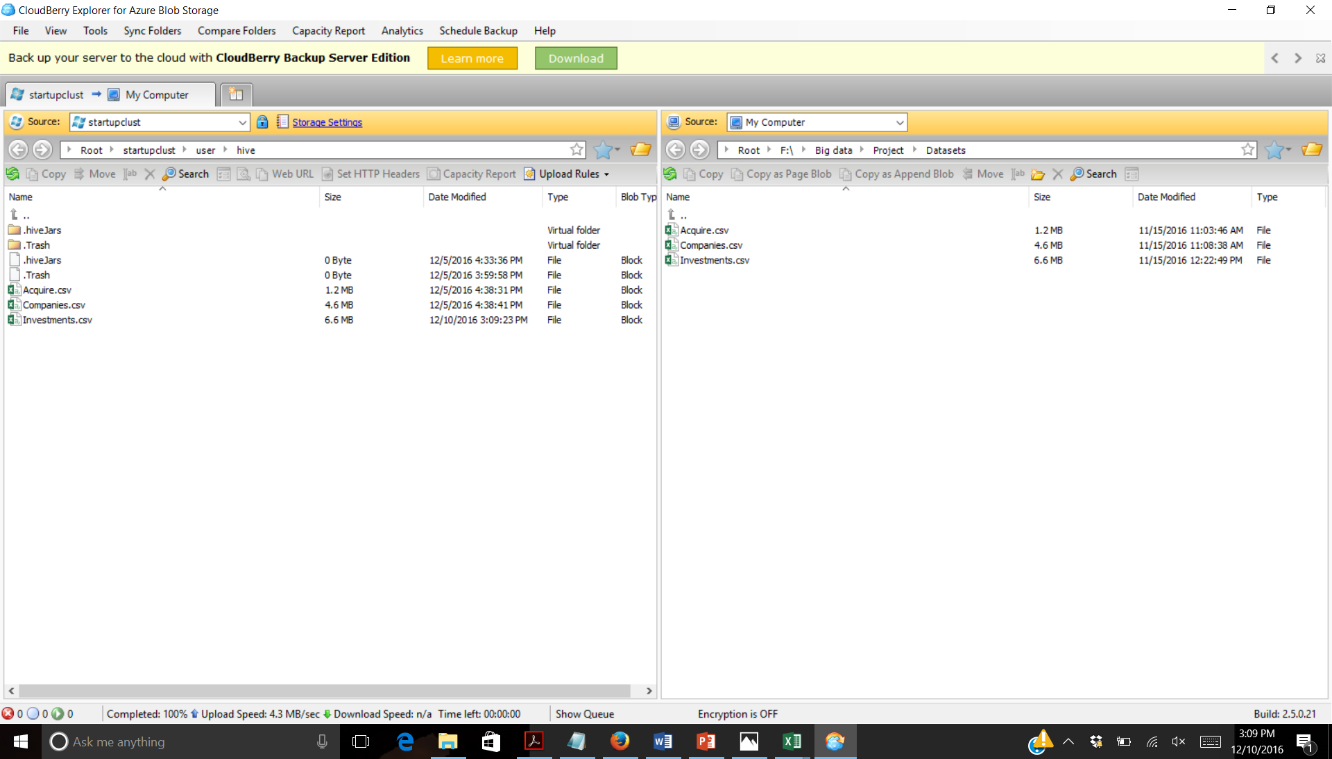
* The cluster creation may take a while.
* Once the cluster is created you can see the cluster dashboard as below,



* In the cluster dashboard, you will see a Hive query editor, where you will execute all the Hive commands. The hive editor looks like below,

**Importing Data into HDFS using Cloudberry:**

* We have 3 datasets for Startup Analysis. They are Companies, Investments and Acquirer. The download link for the dataset is given below,
* Companies: <https://www.dropbox.com/s/hskbwwdyzwm1qwa/Companies.xlsx?dl=0>
* Investments: https://www.dropbox.com/s/5r9uljbv3dls41i/Investments.xlsx?dl=0
* Acquirer: <https://www.dropbox.com/s/il206e92bwwy05k/Acquirer.xlsx?dl=0>
* Total Size : 200 MB
* Once the data is downloaded into local disk, use the below link to download the Cloudberry Explorer for Azure Blob Storage which is a freeware,
* <http://www.cloudberrylab.com/explorer/microsoft-azure.aspx>
* Once the cloudberry is downloaded, connect to the Azure storage blob with the credentials of the storage account you used while creating a HD insight Hadoop cluster.
* Once the Azure storage blob is connected, you can see local disks on one screen and your HDFS on the other side as below,



* The data now can be copied to the HDFS just by dragging and dropping the files from our local system or by Rightclick 🡪 Move option on the file.
* Once the data is moved to the HDFS, you are good to create the external tables for these CSV files.

**Analyze Startup data in Hive on Azure:**

In the Hive editor, execute the following queries to create an external table for Companies, Investors and Acquirer. This will create a schema for three tables and load the data from the CSV files we moved to HDFS using cloudberry.

|  |
| --- |
| DROP TABLE companies;  CREATE EXTERNAL TABLE IF NOT EXISTS  companies(name STRING,market STRING,funding\_total\_usd INT,status STRING,country\_code STRING,state\_code STRING,region STRING,city STRING,funding\_rounds INT,founded\_at DATE,founded\_year INT,first\_funding\_at DATE,last\_funding\_at DATE)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  tblproperties ("skip.header.line.count"="1");  LOAD DATA INPATH 'Companies.csv'  OVERWRITE INTO TABLE companies; |

|  |
| --- |
| DROP TABLE investments;  CREATE EXTERNAL TABLE IF NOT EXISTS  investments(company\_name STRING,company\_market STRING,company\_country STRING,company\_state STRING,company\_city STRING,investor\_name STRING,investor\_market STRING,investor\_country STRING,investor\_state STRING,investor\_city STRING,funding\_round\_type STRING,funded\_at DATE,funded\_year INT,raised\_amount\_usd INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  tblproperties ("skip.header.line.count"="1");  LOAD DATA INPATH 'Investments.csv'  OVERWRITE INTO TABLE investments; |

|  |
| --- |
| DROP TABLE acquirer;  CREATE EXTERNAL TABLE IF NOT EXISTS  acquirer(company\_name STRING,company\_market STRING,company\_country STRING,company\_state STRING,company\_city STRING,acquirer\_name STRING,acquirer\_market STRING,acquirer\_country STRING,acquirer\_state STRING,acquirer\_city STRING,acquired\_at DATE,acquired\_year INT,price\_amount INT,price\_currency\_code STRING)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  tblproperties ("skip.header.line.count"="1");  LOAD DATA INPATH 'Acquire.csv'  OVERWRITE INTO TABLE acquirer; |

Once the external table creation is successfully completed, execute the following HiveQL to analyze various factors from our dataset.

1. Most Popular and least popular market segment for starting a starup? (Based on investments)

|  |
| --- |
| DROP TABLE IF EXISTS PopMarket;  CREATE TABLE PopMarket  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn1/'  AS  select SUM(funding\_total\_usd) as TotalInvestment,market from companies where market IS NOT NULL group by market sort by TotalInvestment desc;  select \* from PopMarket limit 5; |

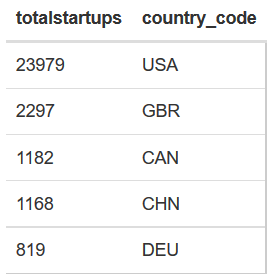
Output:



1. Which is the safest Country to setup a Startup?

|  |
| --- |
| DROP TABLE IF EXISTS SafestCountry;  CREATE TABLE SafestCountry  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn2\_1/'  AS  select count(country\_code) as TotalStartups,country\_code from companies where status='operating' and country\_code NOT LIKE '' group by country\_code sort by TotalStartups desc;  Select \* from SafestCountry limit 5; |

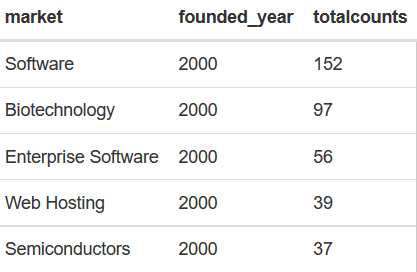
Output:



1. What are the total companies found by year and market segment?

|  |
| --- |
| DROP TABLE IF EXISTS TotComp;  CREATE TABLE TotComp  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn3/'  AS  select market,founded\_year,count(founded\_year) as Totalcounts from companies where founded\_year>=2000 and founded\_year<=2014 group by market,founded\_year order by founded\_year asc,Totalcounts desc;  Select \* from TotComp limit 5; |

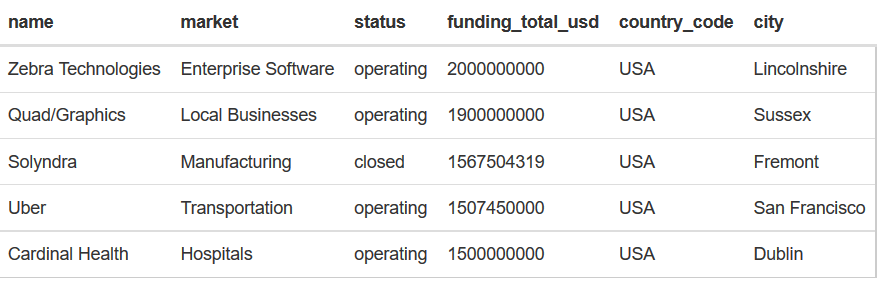
Output:



1. What are the popular companies based on highest funding?

|  |
| --- |
| CREATE TABLE HighFunding  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn4/'  AS  select name,market,status,funding\_total\_usd,country\_code,city from companies order by funding\_total\_usd desc;  select \* from HighFunding limit 5; |

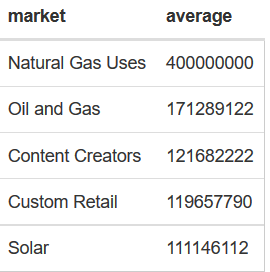
Output:



1. What is the Average funding for a company based on market segment?

|  |
| --- |
| DROP TABLE IF EXISTS AvgFunding;  CREATE TABLE AvgFunding  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn5/'  AS  select market,cast(avg(funding\_total\_usd)as bigint) as average from companies group by market order by average desc;  Select \* from AvgFunding limit 5; |

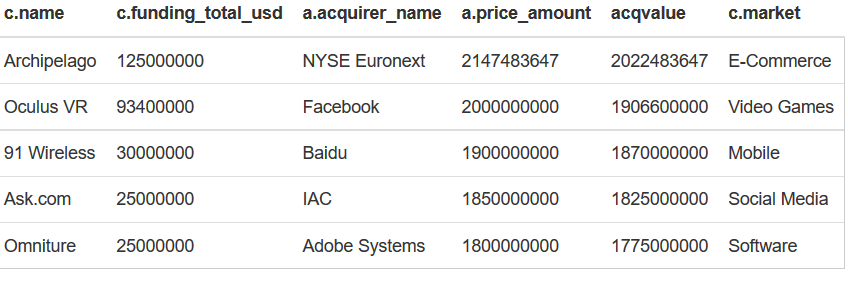
Output:



1. Which acquired company has the maximum networth?

|  |
| --- |
| DROP TABLE IF EXISTS Networth;  CREATE TABLE Networth  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn8/'  AS  select distinct c.name,c.funding\_total\_usd,a.acquirer\_name,a.price\_amount,(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;  select \* from Networth limit 5; |

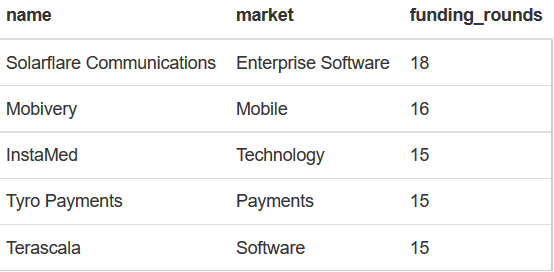
Output:



1. Which are the companies with most funding rounds?

|  |
| --- |
| DROP TABLE IF EXISTS Fundrounds;  CREATE TABLE Fundrounds  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn9/'  AS  select name,market,funding\_rounds from companies sort by funding\_rounds desc;  select \* from Fundrounds limit 5; |

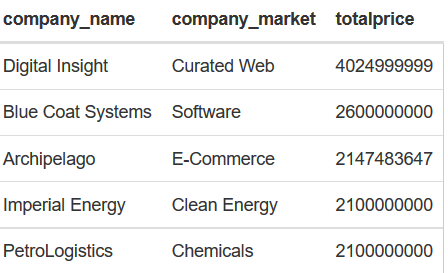
Output:



1. Which are the top 5 companies with highest acquisition value?

|  |
| --- |
| DROP TABLE IF EXISTS HighAcq;  CREATE TABLE HighAcq  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn10/'  AS  select company\_name, company\_market, sum(price\_amount) as Totalprice from acquirer group by company\_name,company\_market sort by Totalprice desc;  select \* from HighAcq limit 5; |

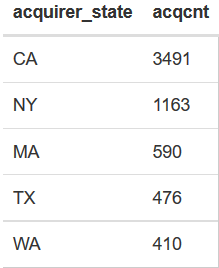
Output:



1. Which are the Cities with Highest acquisitions were made?

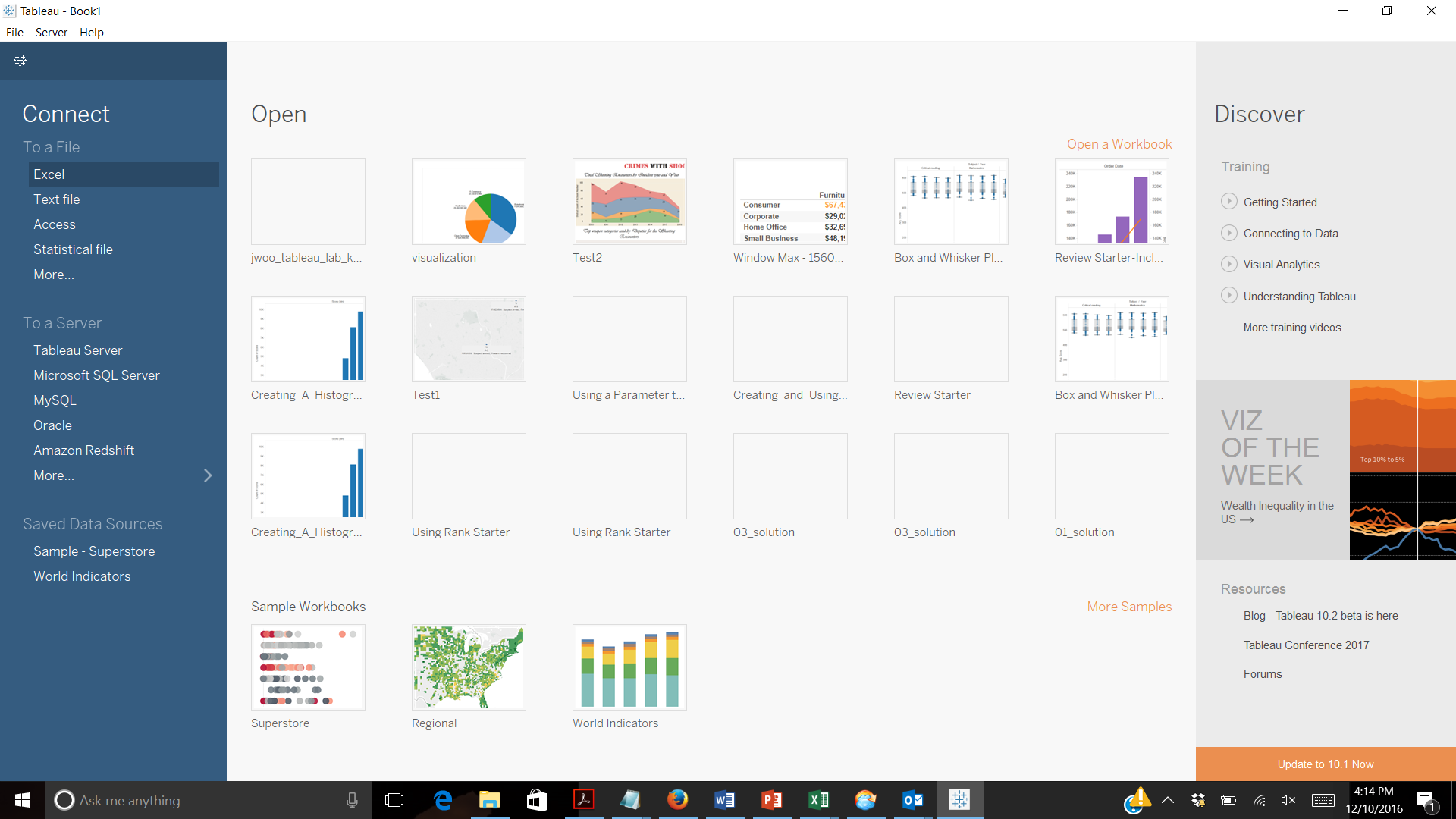
|  |
| --- |
| DROP TABLE IF EXISTS HighAcqCity;  CREATE TABLE HighAcqCity  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION 'wasbs://viggy@startupanalysis.blob.core.windows.net/Qn11/'  AS  select acquirer\_state, count(acquirer\_state) as acqcnt from acquirer where acquirer\_state != '' group by acquirer\_state sort by acqcnt desc;  select \* from HighAcqCity limit 5; |

Output:



**Visualizations using Tableau:**

* Download the analyzed files from HDFS to local system using Cloudberry.
* Save the files with .XLS (excel format)
* Open Tableau desktop and import the excel file as below,



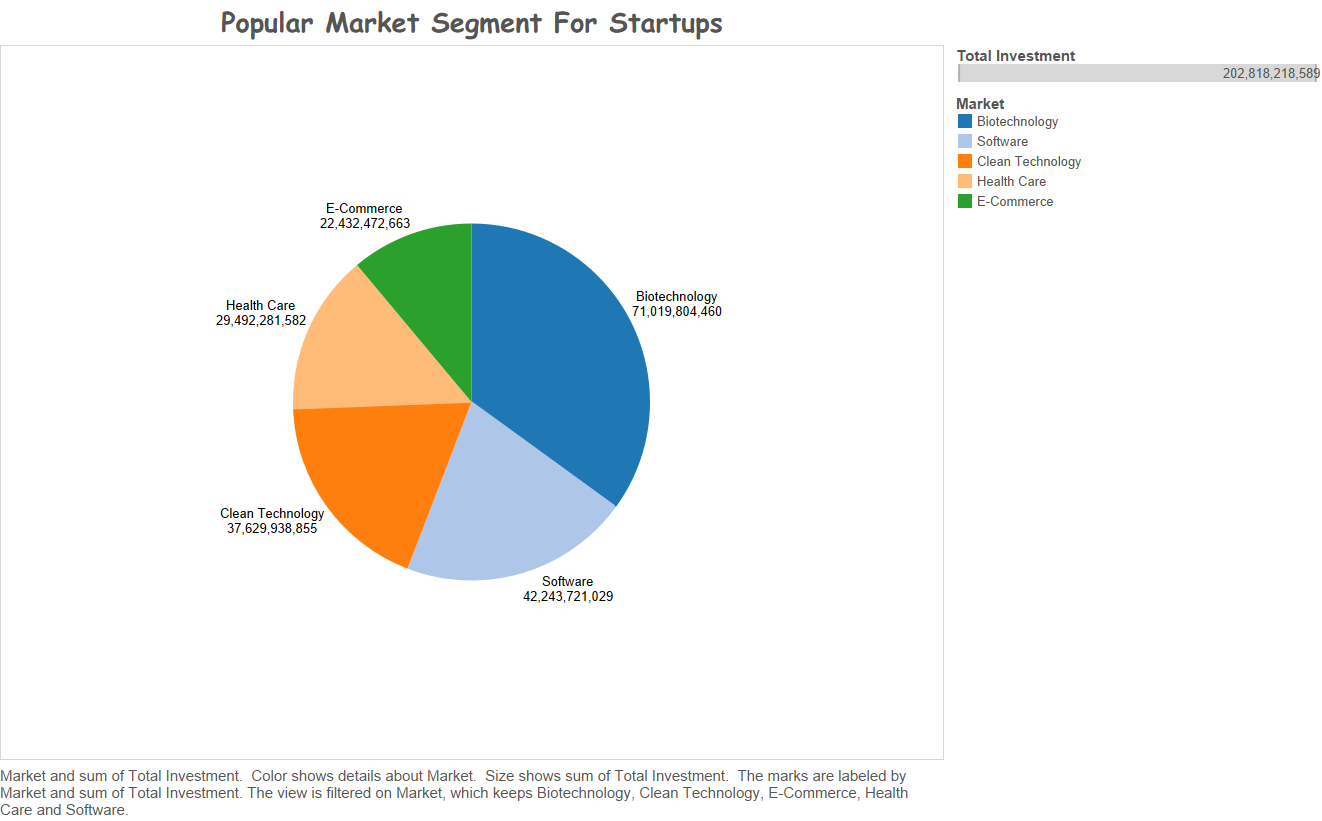
The visualization can be made in tableau by dropping appropriate measures and dimensions in rows, columns, filters and in marks. The following are the visualizations for the various factors that are analyzed,

1. Most Popular and least popular market segment for starting a starup? (Based on investments)

**Rows:** Market

**Column:** Total Investment

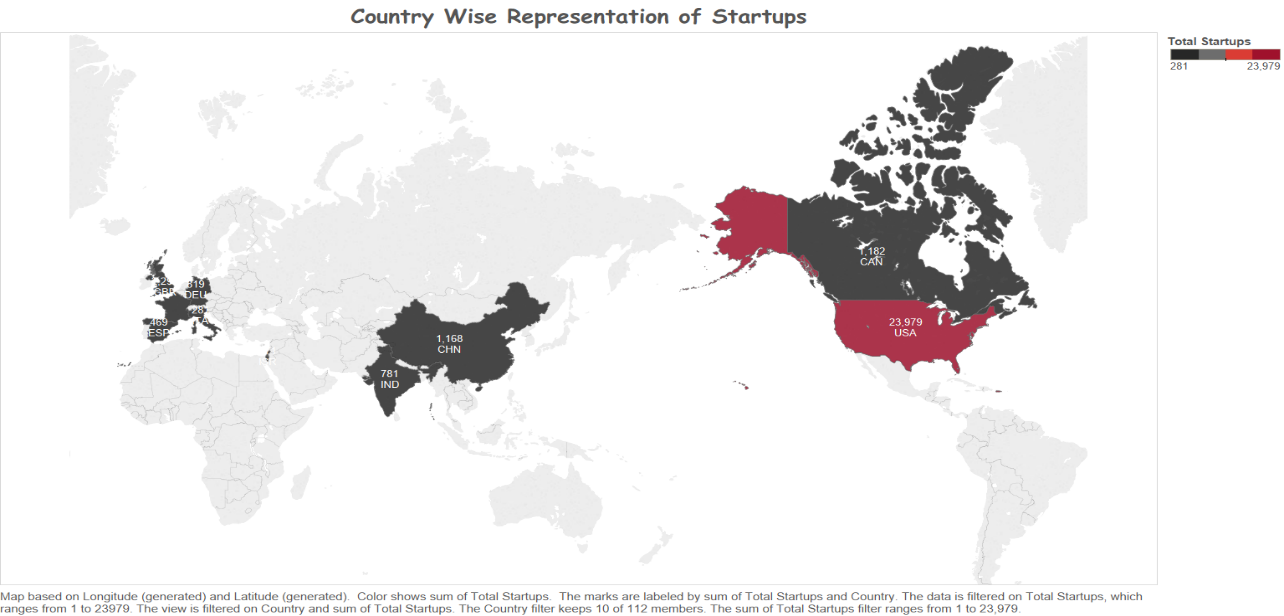
Selectt Pie chart from the type of charts



1. Which is the safest Country to setup a Startup?

Select Maps from Show Me on the right top

Add country and number of startups to the details card in the marksheet

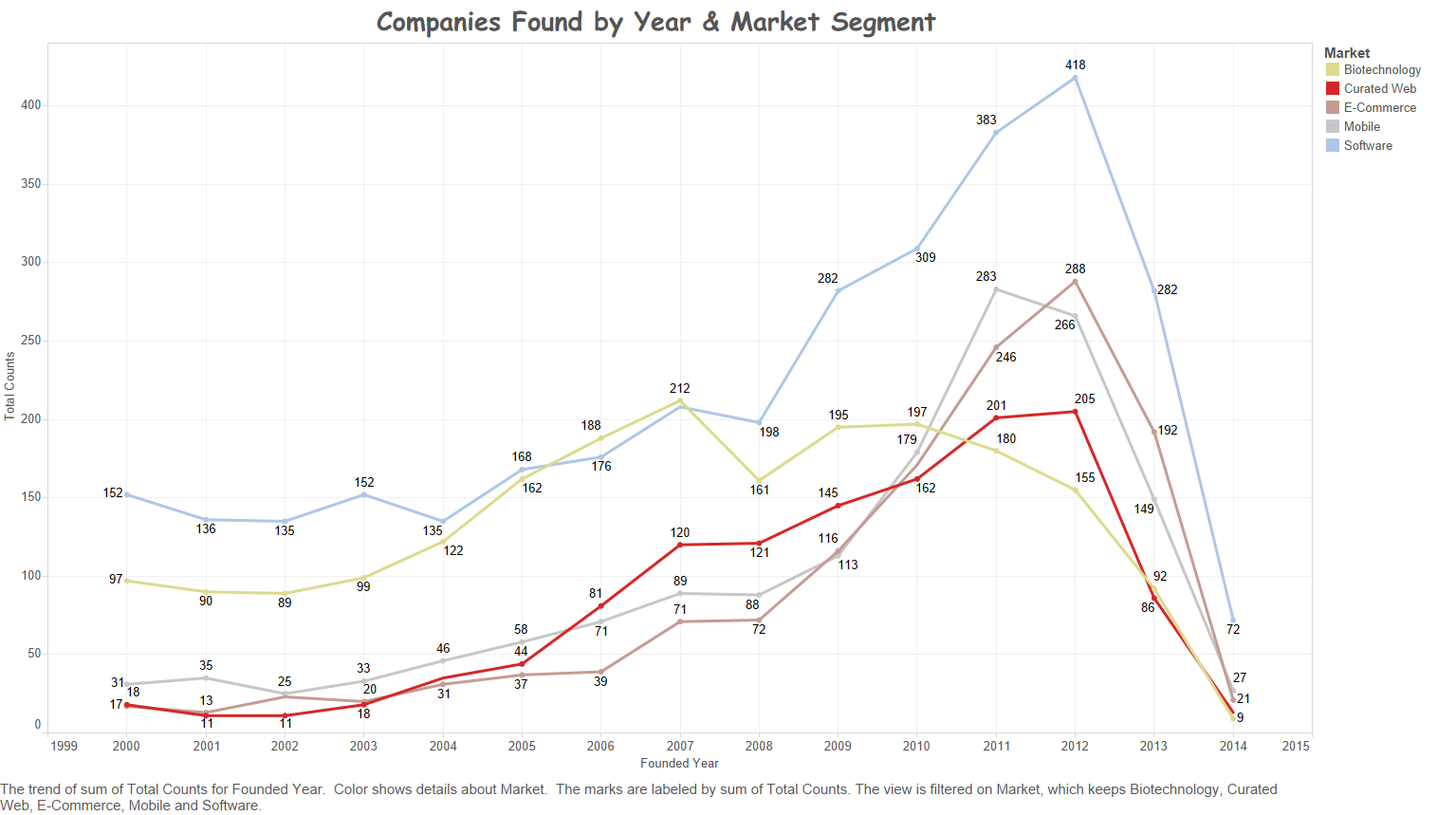


1. What are the total companies found by year and market segment?

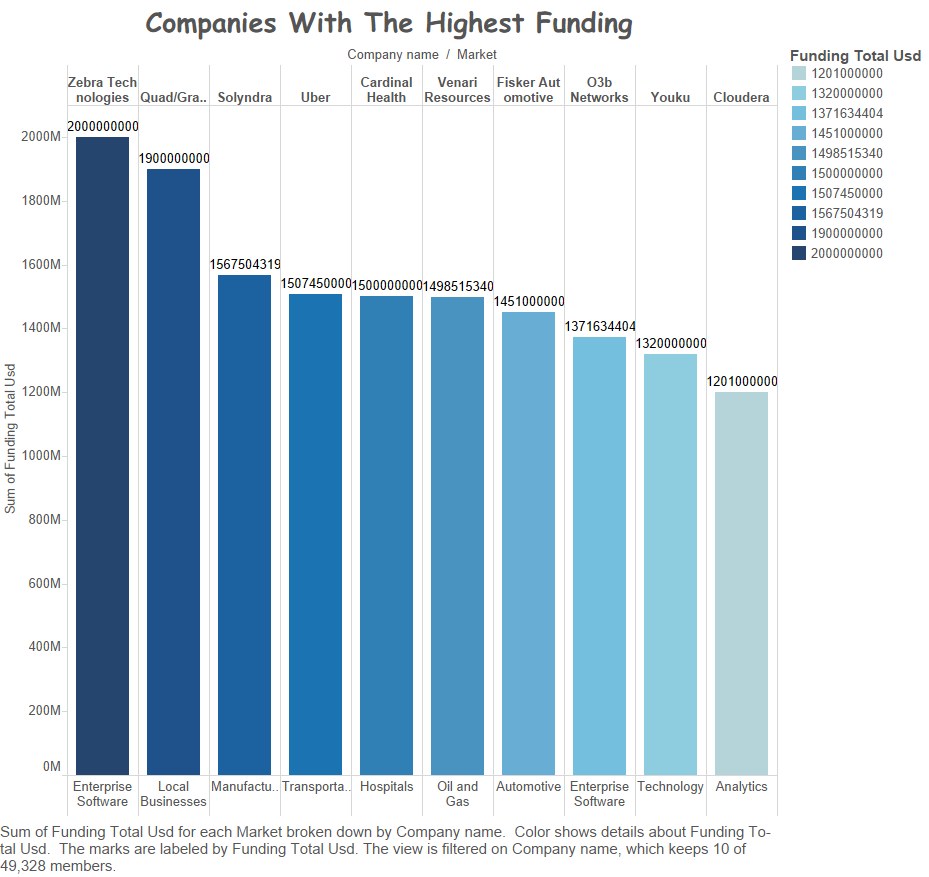
**Row:** Sum(Total counts)

**Column:** Founded Year

Add total counts to labels and select the line chart in the show me



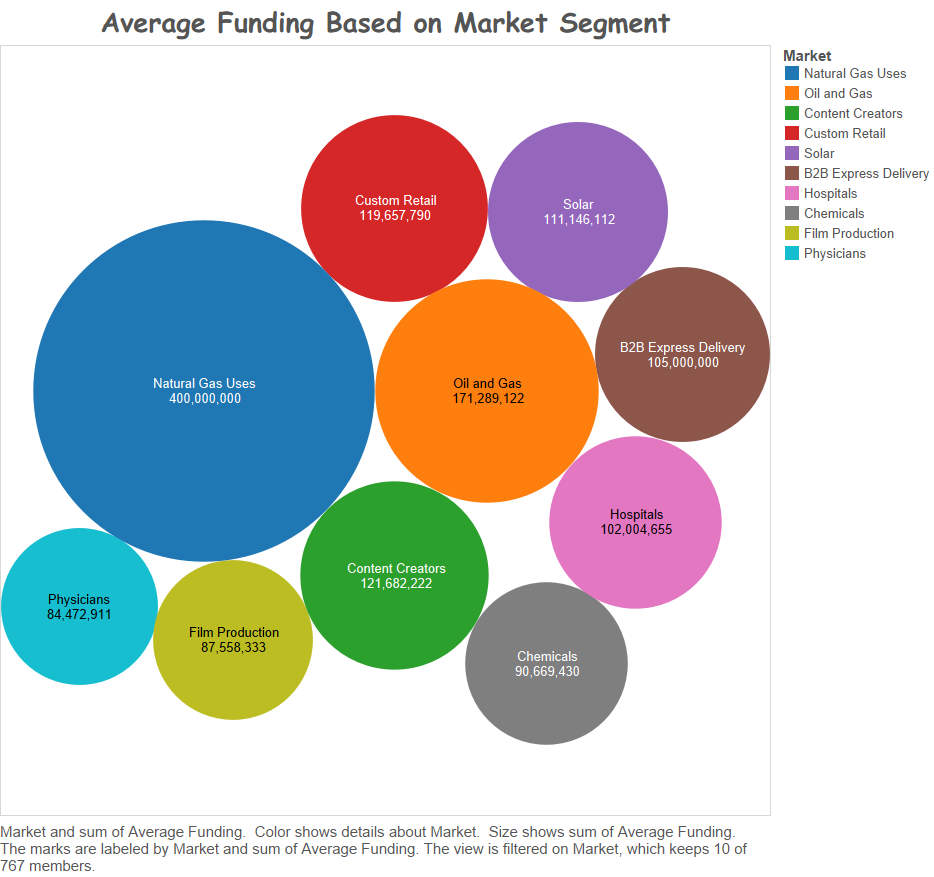
1. What are the popular companies based on highest funding?



**Column**: Company name;market

**Row:** Funding Total

Select the Bar chart

1. What is the Average funding for a company based on market segment?

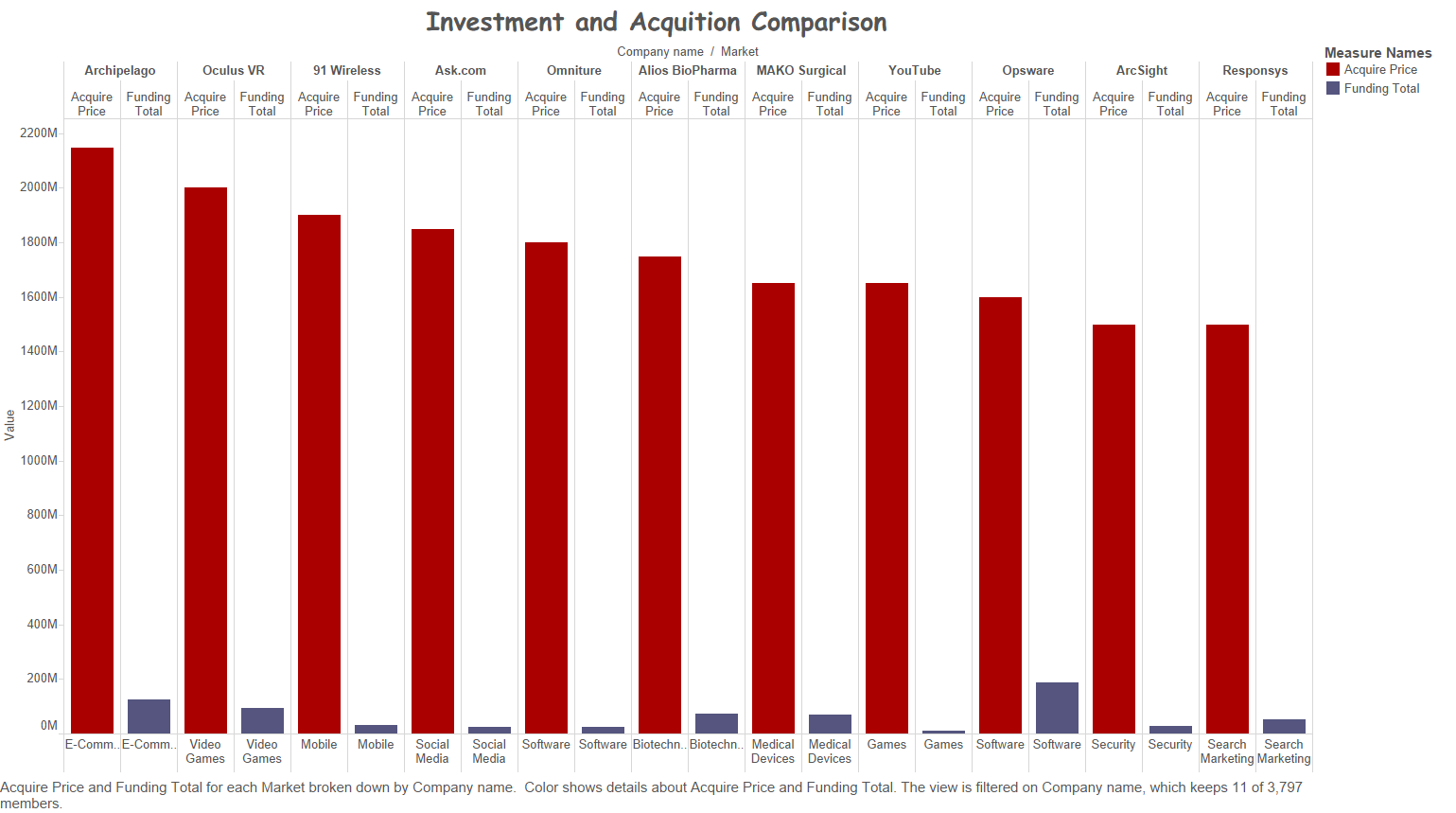
Market to colors and market to label.

Sum(Average) to Size and Sum(Average) to label

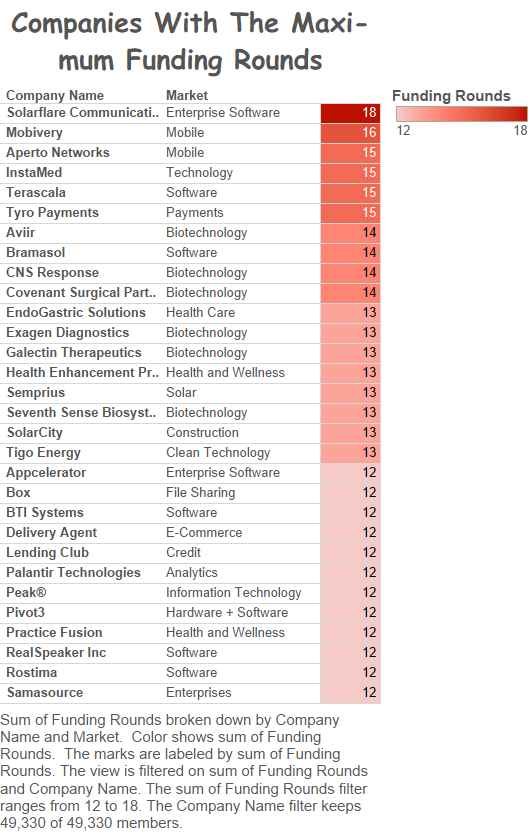
1. Which acquired company has the maximum net worth?

**Column:** Company name, Measure Name, Market

**Rows:** Measure values; Measure name to colors in marks card



1. Which are the companies with most funding rounds?



**Rows:** Company Name; Market

Sum(funding) to colors

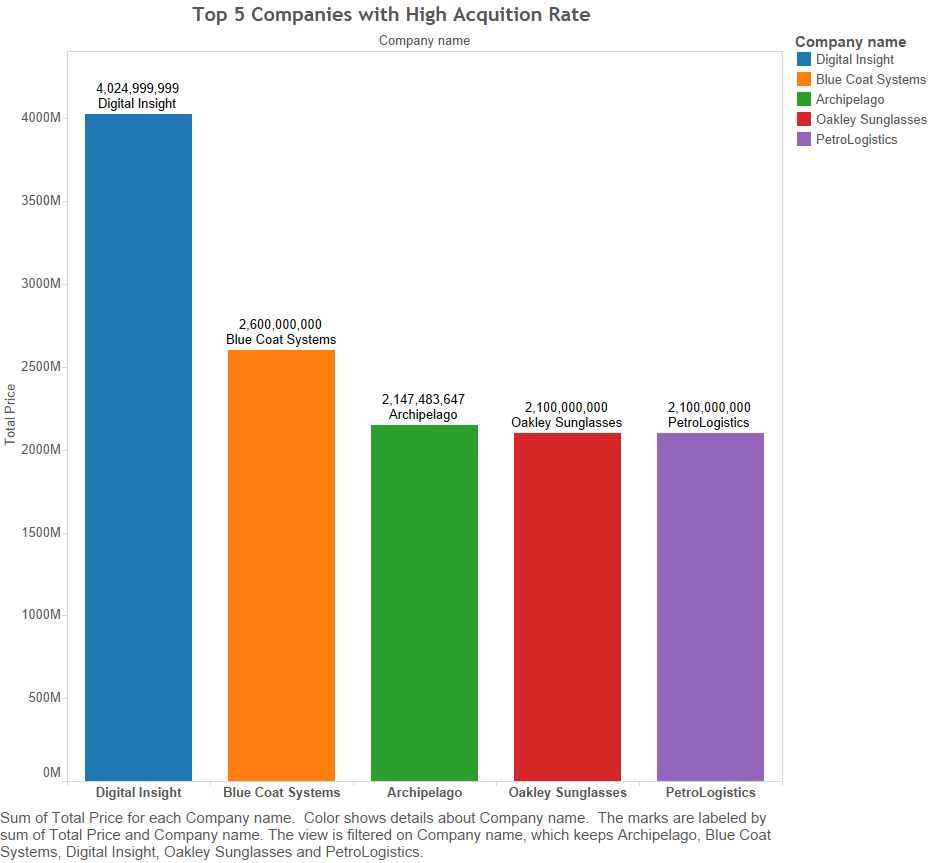
Sum(funding) to label

1. Which are the top 5 companies with highest acquisition value?

**Column:** Company Name

**Rows:** Sum(Total Price)

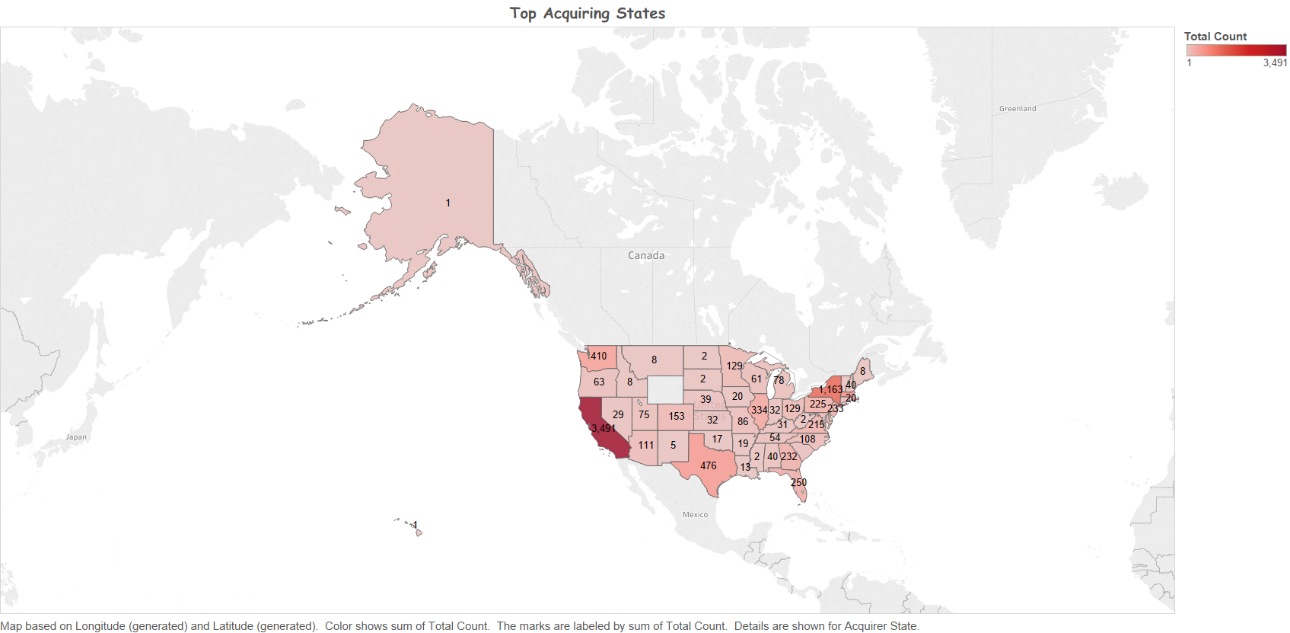
Company name to color and Label;Sum(Total Price) to label



1. Which are the Cities with Highest acquisitions were made?

Acquirer state to Details

Sum(Total Count) to Color and Label.



**GITHUB:**

**Source code:** <https://github.com/vigyr/Calstatela/tree/master/HQL%20Scripts>

**Github Link:** <https://github.com/vigyr/Calstatela>

**REFERENCES**

1. Create HDInsight at Azure: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-tutorial-get-started-windows

2. http://hadooptutorial.info/tableau-integration-with-hadoop/

3. http://www.cloudberrylab.com/explorer/microsoft-azure.aspx