MSIS 2621 | Project Report

GROUP 1

Abhishek Choudhury
Amarjit Dhal
Ameya Ghatpande
Chaitrali Naik
Omkar Gokhale
Sahil Chandawale
Vaibhav Deorukhkar

Table of Contents

EXECUTIVE SUMMARY

Venture capitalists are in a risky business. There are good reasons why VCs are tight with their investment dollars. In order to receive the large returns that they expect from investments, VCs generally ensure that their portfolio companies have a chance of growing sales worth hundreds of millions of dollars. VC's do not want to miss on the next Uber, WhatsApp or Oculus. The ways that VCs measure, evaluate and try to minimize risk can vary depending on the type of fund and the individuals who are making the investment decisions. At the end of the day, VCs are trying to mitigate risk while producing big returns from their investments. The competition in this field is intense, and it is very important for VC's to track their competitors.

Entrepreneurship comes with a host of challenges. Rewarding challenges, but harsh challenges nonetheless. Some of the challenges are - trying to establish a brand, adjust to match or exceed the competition and keep your business profitable. For new and young entrepreneurs, there are some unique challenges that are especially difficult to overcome. Nearly every entrepreneur, has to face the issue of financial crises at some point. While some businesses can be bootstrapped (operated without a cash injection), most businesses will need a cash injection of some sort to help it get the resources it needs. In the technology space, people often focus on investors. There is a high probability of securing funds in particular geographic locations depending upon the type of industry. It is very important for Startup companies to favor a particular geographic location to increase the chances of funding.

People are the most important resource in any organization. No doubt they are the costliest factor in a company. In today's globalized world, people are always looking for better job opportunities and are quite open in changing jobs. From an organization's perspective it is very important to retain the right resources. The HR people are always looking for innovative and exciting ways to reduce the attrition rate. The attrition rate for may change as per seniority level and is dependent on may factors like market conditions, overall economy, booming sectors etc. It is very important for HR team to understand the attrition rate/trends and plan policies accordingly.

PROJECT DESCRIPTION

CorpBase actively tracks an ever growing universe of companies and investors from very early stage and help connect a brilliant idea to an investor. Our application helps Venture capitalists (VC) to get an updated insight of the startup ecosystem. CorpBase also help VC to understand their competitive environment and use these insights to be on the top of the competition. It helps VC to identify companies/ domains where they should keep an eye on. It will help them understand which the best growing market is and where is the greatest potential of growth.

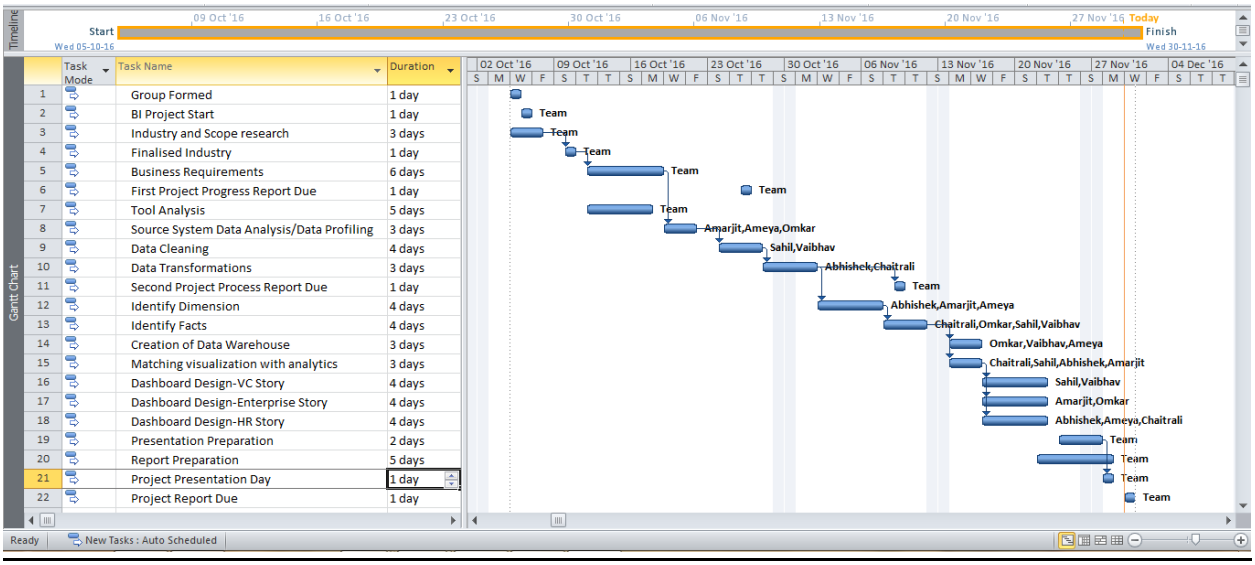
CorpBase will provide analysis about funding patterns in different domains in different geographical locations of the world. This will help Startup companies to enhance their chances of getting funded by selecting a right location. Example] a theme restaurant may want to select an exotic location rather than a heart of city. Looking at the funding patterns which other startups in the same domain have received will give a new direction for companies while approaching VC's.

CorpBase will provide an attrition trends within the company and across the industry to HR team. It will also provide experience based/ designation based attrition information which will help HR team to design and implement retention policies. Example] CorpBase helped us identify that top level executives leave the company the most within just 1-2 years of joining it. This means if the HR team can identify the root cause of this and suggest policies to cross the hurdle of 2 years, chances are these executives will stay with the company for longer time. The longer the people stay in the organization will help regain the training costs invested in them.

PROJECT TIMELINE:

We followed the startup culture and worked as a flat hierarchy. The entire team was involved in selecting the domain and shortlisting the questions which we wanted to answer. A good amount of time was invested in exploring different domains and questions. After finalizing the domain and searching the data sources, we distributed the work among the team. The following amount of time was invested

Domain identification = 10%
Requirements gathering = 10%
Data profiling = 20 %
Data cleaning = 10%
Data transformation and dimensional modeling = 30%
Creating Dashboards = 10%
Project presentation and report = 10%



PROJECT SCOPE

Our BI platform represents the whole corporate ecosystem; however, it is a daunting task to accommodate all the organizations in the project in the limited time periods. We also need more substantial data and complex algorithms to complete its integration. Therefore, we have limited our scope according to certain parameters. But our approach is consistent across the whole project and should be valid for a larger scope. Below is the scope of our project for doing analysis and generating dashboards:

For VC competitor analysis, we have integrated the investment data of major competitors (Google Ventures, Khosla Ventures, and Sequoia Capital etc.

Geography is currently confined to United States. However, our data set also comprises of companies residing outside the US.

We have limited the domains to 40. We combined the subdomains into main domain during the Data Transformation step.

We have utilized the employee data of 11 companies to interpret the attrition rate on designation and yearly basis to target the software and information technology domains.

PROJECT STAKEHOLDERS

- **USERS**

These stakeholders require increased responsiveness from the organization toward their interests or views. All demands of these stakeholders are relevant and significant and are considered during the implementation of project.

- **VENTURE CAPITALISTS**

    - Being the most targeted user group, they are the most important stakeholders.
    - They can use CorpBase to do analysis to know more about the startup ecosystem and do competitor analysis.
    - They perform analysis based on different geographical areas to see which areas are favorable for which type of industry.

- **ORGANIZATIONS (HR)**
    - CorpBase will be very useful for strategizing employee retention plans.
    - Do the trend analysis of the attrition rate on yearly and designation level basis to improve the business processes that meet the organization's and industry's requirement.

- **ENTRPRENEURS**
    - They will be using CorpBase for doing a funding analysis to find probable investors.
    - CorpBase will help them to decide a suitable location based on the analysis which will show domain specific highly productive locations to set up their businesses.

- **PROJECT TEAM**
- They will be developing CorpBase for gathering and putting in all the information together for the users to analyze and use.
- They will be the point of contact for any issue related to the application or for any amendments in CorpBase.
- They will be responsible decisively for the working and implementation of CorpBase.

BUSINESS OPPORTUNITY

In today's market, there is a rise in unpredictability due to factors like globalization, advances in technology, increased transparency, etc. New business models are being deployed considering the increase in market fragmentation. And technological advances are catalyzing disruptive innovation. All these factors, calls for the need to redefine the terms for competition which will drive the organization to mobilize global ecosystems, transform industries and markets drastically. Shaping strategies helps the organizations do exactly that.

CorpBase will help in shaping strategies. This project will help Venture Capitalists, start-up entrepreneurs and Organizations in defining and shaping their strategies.

Venture Capitalists

- VC's can have an overview of the growth pattern of the companies. CorpBase will give them a visual mapping of their investments and how the company has grown over the period.
- Being aware of what your competitors are doing is one of the key information and deciding factor required to strategize for the company to stay ahead in the competition. With CorpBase VC's will know exactly who their competitors are (domain specific) and where their competitors are investing. The VC's will get a year by year investment trend of their competitors in all the domains. This will help the VC's get the overview of the investment trend their competitors are following, which will form the base of the strategies to be planned.

- **Entrepreneurs**
  - Entrepreneurs who have just established a start-up and are looking for funding, with the help of CorpBase can easily track all the funding provided to start-ups in that particular domain over years. This will help them decide the investors they should target to gain funding for their venture.
  - Budding entrepreneurs who are planning on a start-up can be benefitted by CorpBase, as it will help them decide which location will have the highest probability of investment.
- **Organizations(HR)**
  - When good employees leave, it costs your company in many ways. From damage to morale if she was well-liked in the office, or lost skills (as well as the investments you made in helping her acquire those skills), to clients and institutional knowledge there are many risks to your company when an employee leaves. Employee Retention becomes one of the important factor for any organization and CorpBase will help plan retention strategies for organization. CorpBase will give you the trend of the employee churn, it will give you the exact data of when a particular employee based on his designation may leave the company, giving you a heads-up to plan your retention strategies way before the employee leaves.

BUSINESS PROCESS

- **Catalyzing the Due Diligence process in Venture Capital**

  Due diligence is a rigorous process that determines whether or not the venture capital fund or other investor will invest in your company. The process involves asking and answering a series of questions to evaluate the business and legal aspects of the opportunity. Once the

process is complete, the investor will use the outcomes of the process to finalize the internal approval process and complete the investment.

CorpBase will help in stage 1 of the Due Diligence process that is the Screening where predetermined criteria is evaluated to identify which opportunities to focus on as possible investments. This allows, to quickly flag the ones that fit and indicate that they will spend more time and money evaluating. The opportunity does not fit the fund's mandate or criteria (e.g., the business' stage, geographic region, size of the deal, industry sector).

- **Maximizing the change of success in a potential market**

  An organization need to analyze demographic and geographic conditions before they start their business or expand their existing business. The process involves understanding potential market and historical trend of success/failure result. CorpBase will help enterprise to select a correct market that will help them to improve their chance of success.

- **Employee Retention Strategizing Process** –

  Succeeding in your employee retention efforts requires you to think about things from employees' point of view. All employees want to see that they are appreciated by their employer and treated fairly. They want to be challenged and excited by the job they're asked to do. CorpBase will present the data of a trend of employees leaving the company which will to plan the retention strategy by giving employees incentives when you think they might leave.

DEPENDENCIES, ASSUMPTIONS, RISKS AND ISSUES

It is critical that the project team identifies the key dependencies, assumptions, risks and issues that may derail the project. At a minimum, identifying these criteria enables a project to establish an early warning mechanism to flag and address conditions as they occur. With foresight, the project team may mitigate the risks.

PROJECT DEPENDENCIES

Dependencies are the relationships among tasks which determine the order in which activities need to be performed. For most of our tasks, we have used the Finish to Start method. Predecessor task must complete before Successor task can start.

1. The preliminary point of the project is getting the correct and complete requirements.
2. Mapping it to project deliverables. Then we seek the right data source which will give us the right data.
3. Next step was to do the data integration from the identified sources.
4. Data profiling is done to ensure the right data quality.
5. Logical and business rules are validated against the data and other transformations are now done.
6. The data are structured and loaded into the target dimensional model.
7. Creation of data models with facts and dimension tables that are ready for delivery.
8. Integrating individual data marts to form data warehouse.
9. Performing business analytics on the basis of the identified insights.

PROJECT ASSUMPTIONS

- Business requirements and conditions remain stable after data collection is complete.
- Scope of the project doesn't change after the conception of data warehouse.
- Resources assigned to the project have defined tasks which need to be completed by the deadline.
- Resources will gather new skills as the project progress and will contribute to other aspects.

PROJECT RISKS

Project risks are the events or occurrences that can negatively impact the project outcomes. The more risks are identified prior to the initiation of the project, the more likely the project team and stakeholders will be able to deal with them. We identified the following project risks:

- The scope of the project can grow incrementally.
- Data cleaning is a big and significant task. Identifying what to include and clean in the conformed dataset may create further problems.
- Data modeling may be unproductive if there are no interrelated datasets.

- Usually most of the BI tools have common functionalities, but it's a risk to rely on a particular tool.
- Project schedule is very rigid, due to which sticking to it is of utmost importance and not doing so can affect the project objectives.
- Team dynamics change as the project progresses and it is very important to have good team management skills.

PROJECT ISSUES

Industry selection and Scope: Identifying the industry was the key initial task. Our team did considerable research on multiple industries and narrowed down the search to two. Another significant criterion was to check if raw data was available for the industry we select and that helped us to conclude our search on CrunchBase.

Data Gathering and Data Cleansing: Various data sources were available but majority of them were not relevant to the project idea. We approached TechCrunch to use their dataset for the project. Based upon the scope and the problems identified, we pooled data that would help us resolve the business problem. After the data was gathered, cleansing was required as there were many fields in the data that were not required or the fields did not conformed data. The size of each data source (csv files) was huge which took some additional time to clean and transform it to make it suitable based on our project requirements.

Data Modeling: Our team invested adequate time in understanding the process of linking the dimension tables and fact tables. We also analyzed to what level of granularity should we define the attributes. Understanding the concept of natural keys and primary keys in dimension table helped us in redefining and redesigning the data model.

Data Integration: After completing the process of data modeling, the other major challenge that we faced was with data integration. As the team did not have an expertise on data integration, we explored various tools such as Talend, Pentaho and Alteryx. We finalized on working with Alteryx as it matched to our project requirements and scope.

BI tool selection: Understanding and experimenting with different tools took substantial time but was an important part of the execution.
Issue Tracking: Using the Industry followed methodology - ITIL for problem and change management.

BI Project - Issue Tracker

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Resolved

Category	Issue	Owner	Table Name(Optional)	Impact	Approach to resolve	Status
Data Loading	Failed to load data due to encoding errors	Omkar Gokhale	-	High	Change Encoding Format	Resolved
Data Loading	The 'Microsoft.ACE.OLEDB.12.0' provider is not registered on the local machine. (System.Data)	Vaibhav Deorukhkar	-	High	Install Access Database Engine	Not Resolved
Data Loading	Error with Source Column [Domain]	Omkar Gokhale	-	High		Not Resolved
Data Cleaning	Removing Special Characters	Vaibhav Deorukhkar	organizations.csv	Medium	Alteryx - Formula - **REPLACE([company_name],"Ä","a")**	In Progress
Data Cleaning	Date format is not consistent	Amarjit	organizations.csv	Medium		Resolved
Data Cleaning	Unable to remove special characters using Replace command	Vaibhav Deorukhkar	organizations.csv	Medium	Use REGEX in REGEX function : [^x20-x7e]+	Resolved
Data Integration	Joining organizations.csv and Dim_Geography	Vaibhav Deorukhkar	organizations.csv and dim_geography.csv	High	Use Join Tool on Alteryx	Resolved
Data Cleaning	Currency Conversion	Vaibhav Deorukhkar	Fact_VentureCapitalist_Competitors.csv	High	Separate Data Using Alteryx and Excel	Resolved
Visualization	Need of Calculated Field	Vaibhav Deorukhkar	Fact_VentureCapitalist_Competitors.csv	High	Calculated Field formula -> **IF [Total Deal Amount] == '¥' THEN ([Total Deal Amount2] * 65) END**	In Progress / Deadlock / Not Resolved / Resolved
Data Integration	Joining Geography ID with Fact Table	Vaibhav Deorukhkar		High	Join on USA States first and later join again on Countries other than USA.	
Data Integration	Similar looking Target_industry types	Vaibhav Deorukhkar	Fact_VentureCapitalist_Competitors_3	High	Use Fuzzy Match on Alteryx	Resolved

Teamwork: Our team did not face any major issues with regard to the project work but we encountered conflicting schedules for the project meetings. However, with planning, we could overcome the issue and overall there was good communication & collaboration within the team. Everybody contributed equally in terms of time and efforts towards the successful closure of the project.

DATA ARCHITECTURE, MODELING AND DESIGN

INFORMATION ARCHITECTURE

This section explains "how" of the project and includes a high-level description of the information, data, and technology architectures.

Our **Information Architecture** defines the "what, who, where, and why" for our BI application. This is very essential part of our planning, as the information architecture defines the business context of a BI solution. It is the starting point for data modeling, design, and development to support business process needs. If there is no information architecture, the result will be silos, which prevent effective analytics and carry a high cost in time and money.

The Information Architecture will cover what information the stakeholders will get from the BI project. Our project has primarily two stakeholders: Venture capitalists and the new startup companies who seek investments.

The VCs generally want to ensure that their portfolio companies have a chance of growing becoming the next of millions of dollars' company. Although, this is a risky business, but the VC would still need some tools to mitigate this risk. They need to understand which domain is growing. Which is a conducive geography for a domain? They also need to understand how their competitors are doing. What companies they are currently investing in. All this information is important to give the VCs valuable insights into the startup ecosystem.

The new startup companies look to generate revenue for their business. But that is really tough, especially when there are thousands of people with great business ideas, but only a select few, perhaps 2-3% can take their idea and execute by turning them into something that people will pay for, over and over. Therefore, it is important for them to collect the right information from the early state of their inception. They need to understand which domain had the maximum investments in the recent times. They also need to seek the information regarding which VC has invested in their area of expertise. Besides that, some geography is favorable for some domain, and such an insight also helps in the long run. Finally, if there is information like which quarter had the most number of investments, or what the trend is for the last year, then they can plan accordingly. Besides that, our tool also helps them to know which events they should target to meet a prospective investor. Finally, the company should strive to get subsequent investments.

Summary of Information Architecture:

WHAT	What business processes or functions are going to be supported	Funding/Investments/
	What types of analytics will be needed	Historical analysis and trends.
	What types of decisions are affected	VC: If I should fund in this company or not.

		Startup: Right market, right source to get fund. HR: Retention policies
WHO	Who will have access	VC firm, Angel investor, Personal investor, Founder of the company, HR
WHERE	Where is the data now	Different data sources, public records, financial records, Investment files, employee information
	Where will it be integrated	All the data will be integrated as data marts which will be further integrated to data warehouse
	Where will it be consumed in analytical application	Users need to register on our website, and have to be authenticated before using the analytics that will be hosted on the website. These analytics will be fetched using web services from our warehouse.
WHY	Why will the BI solution(s) be built	To help companies find investments and help in brilliant ideas getting executed. Help the HR team to understand Attrition trends and design retention plans

DATA ARCHITECTURE

The **Data Architecture** defines the data along with the schemas, integration, transformations, storage, and workflow required to enable the analytical requirements of the information architecture. The scope of the data architecture starts where data is created in the source systems and ends where the business user performs data analysis. We can infer that Data Architecture is a component of Information Architecture, and is concerned with the actual design and use of information or requirements provided by customers.

For our project, we have used the Kimball's bottom up approach to design our data warehouse. We first identified the similar data sets and create dedicated data mart relevant to that group of data. We then integrated these data marts to create the data warehouse having conformed dimensions.

DATA IDENTIFICATION

The data sources for our project, we have used datasets from three different sources. Our primary data source is **CrunchBase**, which gave us the data about startup companies, investors and people. We have used **PrivCo** data source for detail investments that each company received in the recent years. We have then used **Owler** and **Statista** data to get more insight about the companies and venture capitalist firm.

Data Sources:

- https://www.crunchbase.com/
- http://www.privco.com/dashboard
- https://www.owler.com/

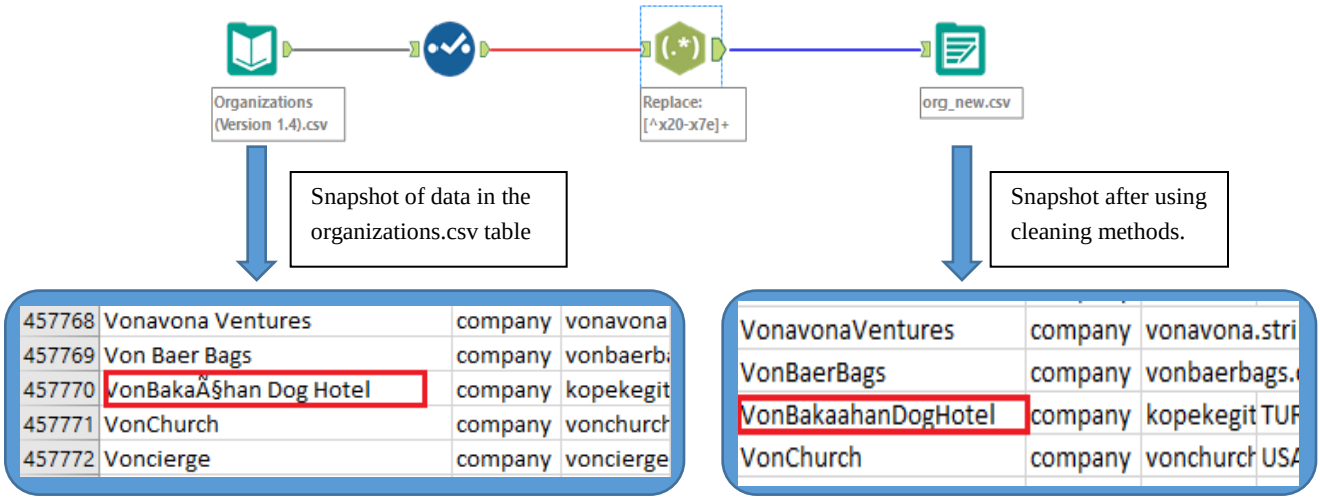
DATA CLEANING AND TRANSFORMATIONS:

The problem with using data from multiple sources is that it is almost always in a format which does complement in solving the business problem. As a team dedicated towards answering questions through data, we took ownership of carrying out data cleaning tasks. Few of the many examples given below should give an overview of the process we followed to clean this data –

- **Preliminary Data Cleaning Using Regular Expressions on Alteryx**

We had a lot of junk values and symbols in the data which could not be easily replaced as they were NON ASCII values. To remove theses NON ASCII characters we used the following regular expression in Alteryx RegEx tool to find and replace the NON ASCII values.

This RegEx means that all the ASCII characters not (^) in the range of \x20 - \x7E (Hex 0x20 to 0x7E). RegEx used - **"[^x20-x7e]"**



- **Data Cleaning using MS-Excel**

This is best represented with a screenshot.

H	I	J
Deal Type	**Status**	**Total Deal Amount**
Funding	Completed	₹200,000,000.00 [INR]
Funding	Completed	$5,000,000
Funding	Completed	$3,300,000
Funding	Completed	₹250,300,000.00 [INR]
Funding	Completed	$15,000,000

Hinder aggregations on the data.

Since we were aggregating the Total Deal Amount column to answer VC based questions, data with currency symbols such as - ₹ OR $ and additional currency information like [INR] was not helping our case.

We used MS – Excel's Text-to-Column feature to segregate the data.

Deal Type	Status	Currency_ Symbol	Total_Deal_Amount
Funding	Complete	$	36,175,009
Funding	Complete	$	4,250,000
Funding	Complete	₹	200,000,000.00
Funding	Complete	$	5,000,000
Funding	Complete	$	3,300,000
Funding	Complete	₹	250,300,000.00

This helped us in dynamic calculation of the foreign currency conversions. Screenshot of dynamic currency conversion on the visualization tool is as follows -

rrency Adjustment (in USD)

```
IF [Currency Symbol] == "¥" THEN ([Total Deal Amount] * 0.0088)
ELSEIF [Currency Symbol] == "£" THEN ([Total Deal Amount] * 1.25)
ELSEIF  [Currency Symbol] == "€" THEN ([Total Deal Amount] * 1.06)
ELSEIF  [Currency Symbol] == "₹" THEN ([Total Deal Amount] * 0.015)
ELSEIF  [Currency Symbol] == "C" THEN ([Total Deal Amount] * 0.74)
ELSEIF  [Currency Symbol] == "$" THEN ([Total Deal Amount] * 1)
END
```

In the real world, it is possible that a Business Analyst is not provided with access to data in order to add custom conversion tables to business tables, in such cases, using such calculated fields on the visualization tool itself can help solve the problem.

- **Data Transformation using Joins** –

Need: The source data had a lot of inconsistencies. The data granularity for "USA" was defined to the state level and for other regions, the data was restricted to country level of granularity.

1. Sample row for USA region was on the format:

ID	State	Country	Industry Type
1	California	United States	Big Data Company

2. And the other set of records (in the same source table) but not belonging to USA were of the format:

ID	State	Country	Industry Type
121	*Unspecified*	Brazil	Telecommunications

3. Thus, to make the data consistent and apply the geography dimensions, we had to use joins twice, once for USA based countries and another for countries that were not USA:

ID	Geog_ID	State	Joined_Column	Country	Industry Type
1	G3	California	California	United States	Big Data Company

And:

ID	Geog_ID	State	Country	Joined_Column	Industry Type
121	G215	Unspecified	Brazil	Brazil	Telecommunications

- **Data Transformations using Alteryx – Fuzzy Match:**

In order to analyze one of the VC based question, we used data from the Venture Capitalists Table. A column in the table was as follows -

Target_Industry
Artificial Intelligence & Algorithmic Automation Software / Mobile Applications
Online Travel Arrangement Services / Travel Agencies
Other Internet Services
Database Storage & Management Software / Market Research / Marketing Services / Other Internet Services
Mobile Applications / Online Ticket Services
Food Delivery Apps
Food Delivery Apps / Marketing Software / Other Billing Services / Other Billing Services
File Sharing
E-Commerce / Luggage Retail / Manufacturing & Industrial / Suitcases & Luggage
E-Commerce / Luggage Retail / Manufacturing & Industrial / Other Agriculture Farm Machinery & Support / Suitcases & Luggage
Biotechnology
3-D Printing Services
Database Storage & Management Software / Human Resources Management Software / Mobile Technology / Software-As-A-Service (Saas)
File Sharing / Software-As-A-Service (Saas)

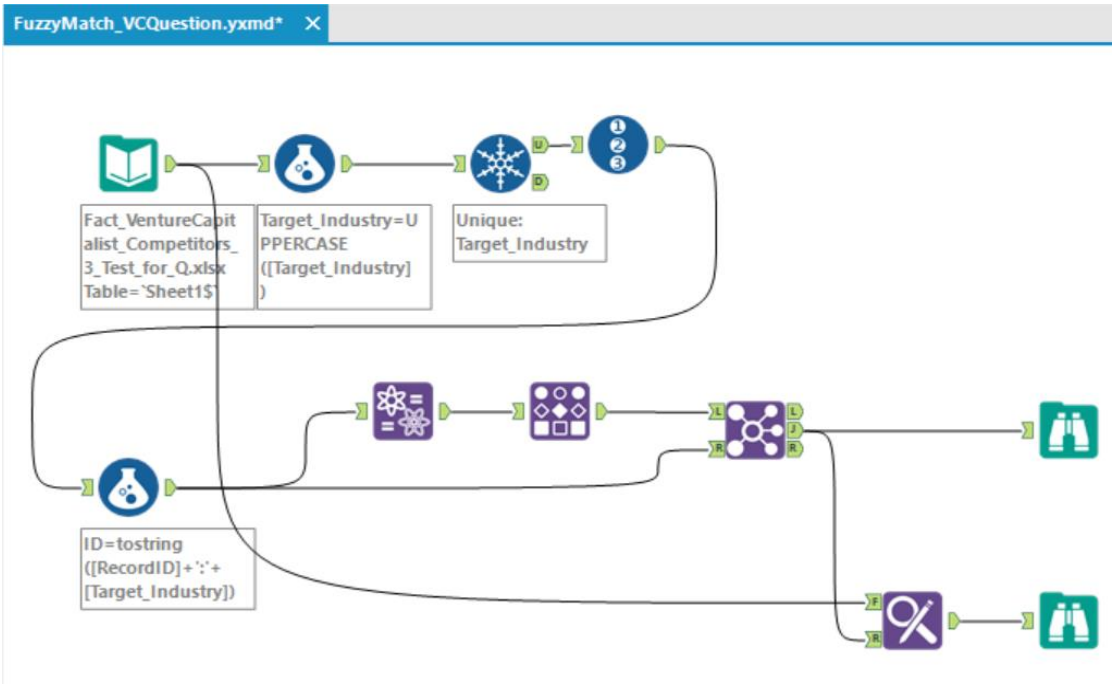
And to profile the data, it was not possible to categorize the data with multiple categories within the same row that were separated by "/". We needed the below show list to be converted to a single domain as shown –

E-Commerce / Luggage Retail / Manufacturing & Industrial / Suitcases & Luggage

E-Commerce

In order to resolve this, we used the Fuzzy Matching algorithm from Alteryx. Below is an attached screenshot of the Alteryx workflow to generate relevant Target Industry Domains. We used the following Alteryx Algorithm –

Source (with an Extra Column for Domain Name) **>** *Uppercase (for case sensitive matches)* **>** *Unique Target Industry Types* **>** *Append Record Number (facilitates Fuzzy Matching)* **>** *Fuzzy Match* **>** *Grouping* **>** *Find and Replace Target Industry and Save to Domain Name.*



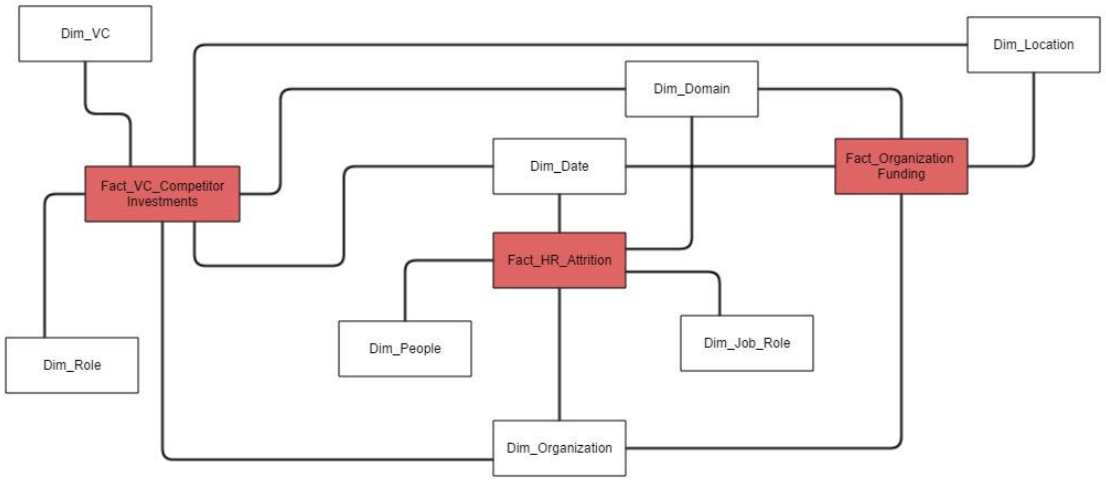
Result - We were able to profile the data according to the Industry domains within which they were operating.

F	G	H	I	
Year	**Role**	**Target**	**Target Industry**	**Domain Name**
2011	Investor	Aquion Energy, Inc.	Advanced Batteries	Hardware
2008	Investor	Aquion Energy, Inc.	Advanced Batteries	Hardware
2012	Investor	LocBox Labs Inc.	Advertising & Marketing	Advertising services
2006	Investor	AdECN, Inc.	Advertising & Marketing	Advertising services
2011	Investor	Moment Systems	Advertising & Marketing	Advertising services
2010	Investor	Madhouse Media	Advertising & Marketing	Advertising services
2011	Investor	Linkable Networks, Inc.	Advertising & Marketing	Advertising services
2006	Investor	RAPT Studio	Advertising & Marketing / Architectural Design & I	Advertising services
2013	Investor	Beijing second hand int	Advertising & Marketing / Cloud Computing / Info	Advertising services

DIMENSIONAL MODELING

Conceptual Dimensional Data Model

After data profiling, cleansing, and transformations, we designed the conceptual and logical dimensional data models for our insights. As we found multiple facts for our insights, we believe our schema is 'Fact-constellation Schema'. The process we followed for the conceptual model is to identify the major entities like facts and dimensions first and then established relationship between them.



Fact Tables –

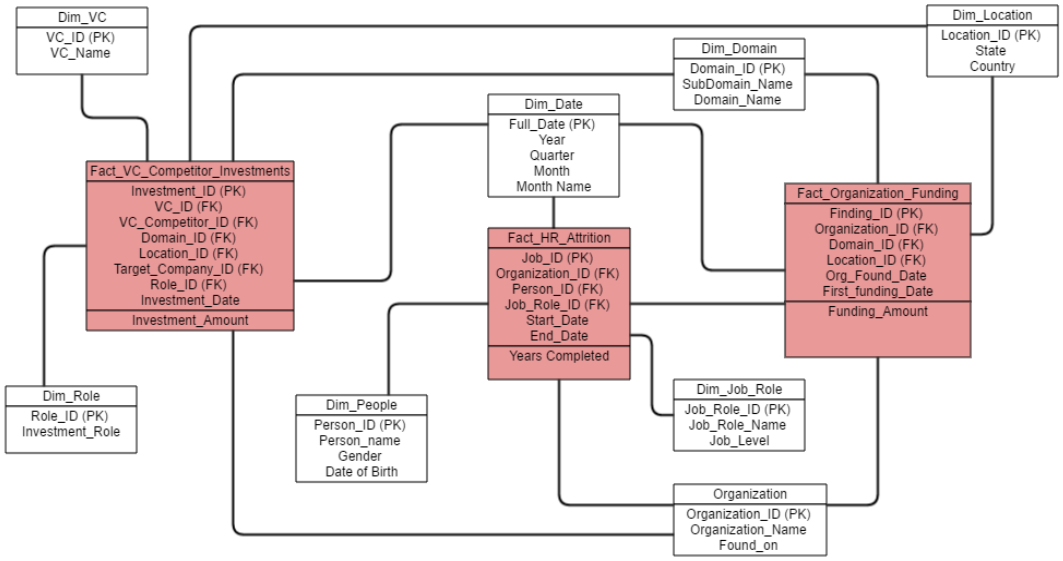
Fact_VC_Competitor_Investments	**Fact_Organization_Funding**	**Fact_HR_Attrition**
Includes the data about investments made by the competitors of some VCs.	Includes data about funding of companies with respect to industry domains and geographical locations.	Includes data about events of people leaving jobs and their companies and designation levels.

Dimension Tables –

Dim_VC	Dim_Role	Dim_Organization	Dim_Domain
Includes data about VCs (Also includes the competitor VCs)	Includes details about Role in investment – either Funding or Acquisition	Includes data about companies	Includes details about Industry domains and sub-domains
Dim_Location	**Dim_People**	**Dim_Job_Role**	**Dim_Date**
Includes details about geographical locations; states and countries.	Included data about employees	Included data about designation levels- Executive/Senior/Mid-level/Junior/Intern	Included dates and its pre-stored classification fields like year, quarter, month etc.

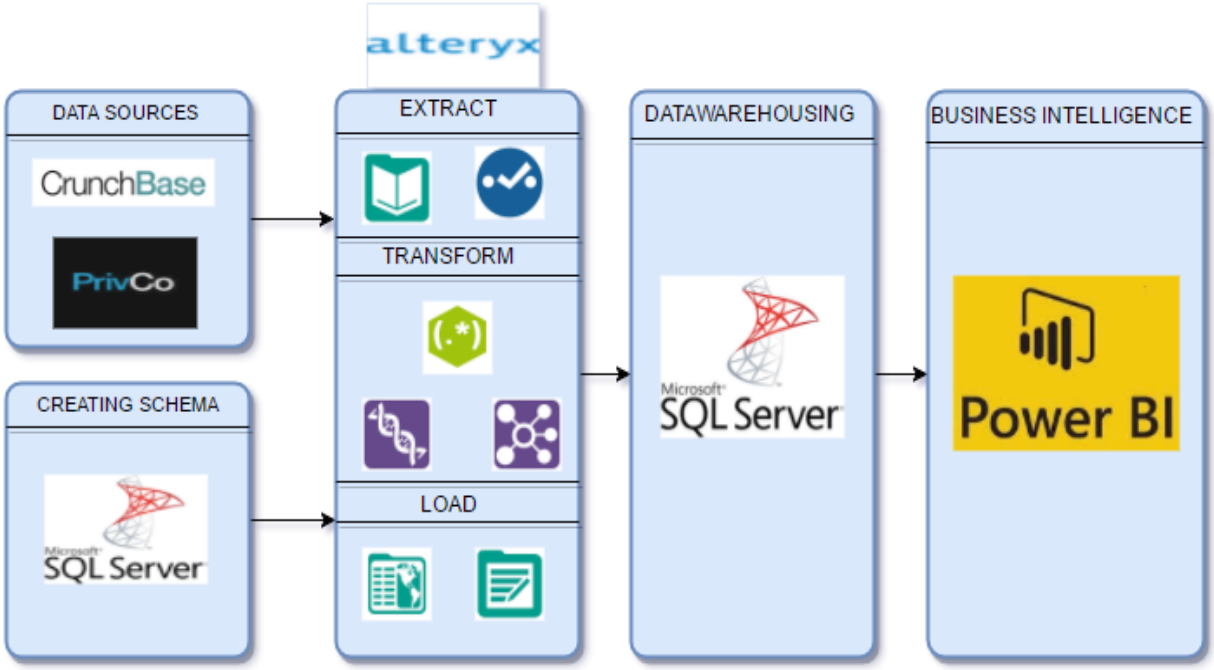
Logical Data Model –
Facts and dimensions are connected via a referential key. The facts contain only keys and measures and dimensions contain all the details about that particular entity.

Fact_VC_Competitor_Investments had investment amount as the measure. Fact_Organization_Funding has Funding_Amount as the measure. These are **additive facts** as they can be aggregated over its connected dimensions. Whereas, in the Fact_HR_Attrition, we believe it is a **pseudo fact** as we are just tracking an event of person leaving a job with respect to other dimensions.
We could see some class-concepts like **role playing dimensions** and **conformed dimensions** here in our model. For example, the dimensions Dim_Date, Dim_Location, Dim_Domain, Dim_Organization are connected to more than one fact. Hence, they are conformed dimensions. Also, the Dim_Date is a role-playing dimension as it maps Start_Date and End_Date from Fact_HR_Attrition and also, Investment_Date from Fact_VC_Competitor_Investments and Org_Found_Date from Fact_Organization_Funding.



TECHNOLOGY ARCHITECTURE

Technology architecture includes the BI software and tools used to create the project. Below are the details about the same:



DASHBOARDS

1) **VC Competitive Analysis : Click here for the** dashboard

   VC's are always inundated with different companies from different domains. This is a high risk and risk gain business. Venture Capitalists follow a due diligence process to identify a right company to invest. The complex process of due diligence is divided into three parts –

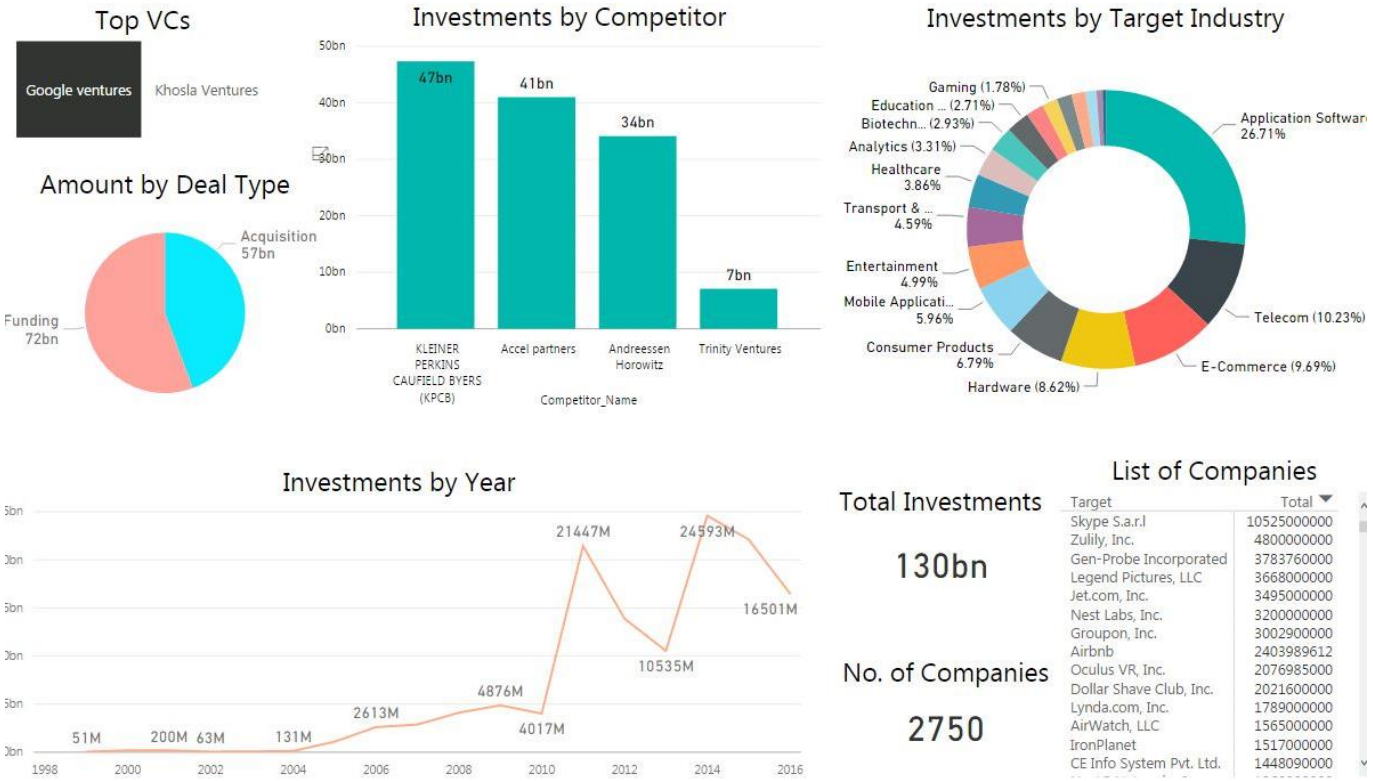
   a) Initial screening due diligence
   b) Business due diligence
   c) Legal due diligence

This dashboard will help VC's to get business insights for the initial screening due diligence, check investment trends in different domains over period of years and keep a close eye on competitor's investment patterns.

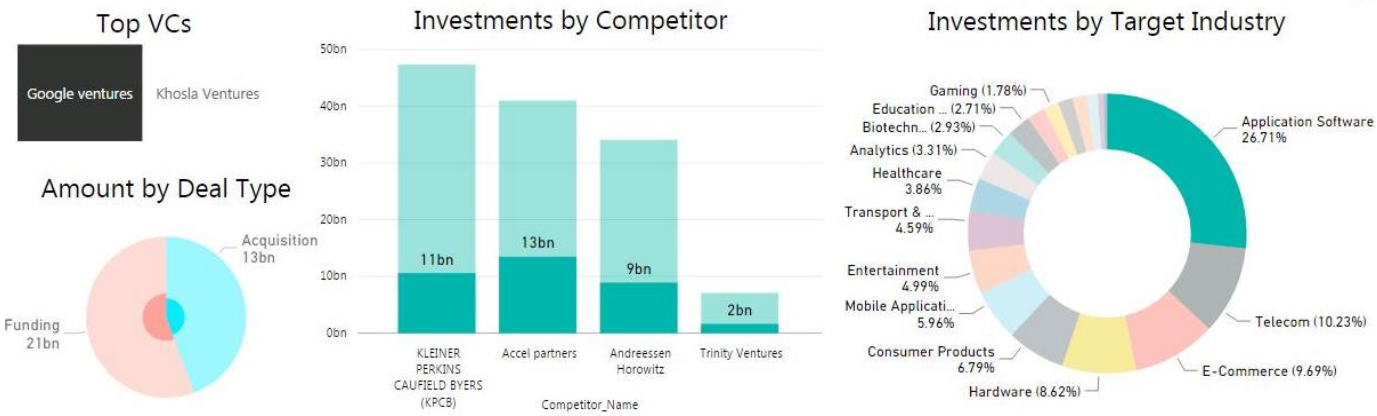
The main graphs in the dashboard show –
1] Main competitors for a selected VC
2] Domain wise investment of the competitors
3] List of the companies the competitors have invested money in

4] Year wise investment trends

<u>Directions for using the dashboard</u>:
The dashboard shows main VC's which we have targeted for demo purpose.
When you click on any VC (say Google Venture), then its main competitors based on money invested will be shown. The pie chart shows the domain wise investments the competitors have done. (We understand that pie chart is not a good for dashboard visualization; but we felt that it would be better than **tree map** as we wanted the total share to be represented and space constraints) Making a selection on pie chart will show the list of companies in that domain.



Above, we can see KPCB is the number one competitor for Google venture. If we select for a domain say Application software, we can also see which other competitors have invested in that domain and how much they have invested. Also, the investment type and amount can be seen like acquisition or funding and its respective amount.

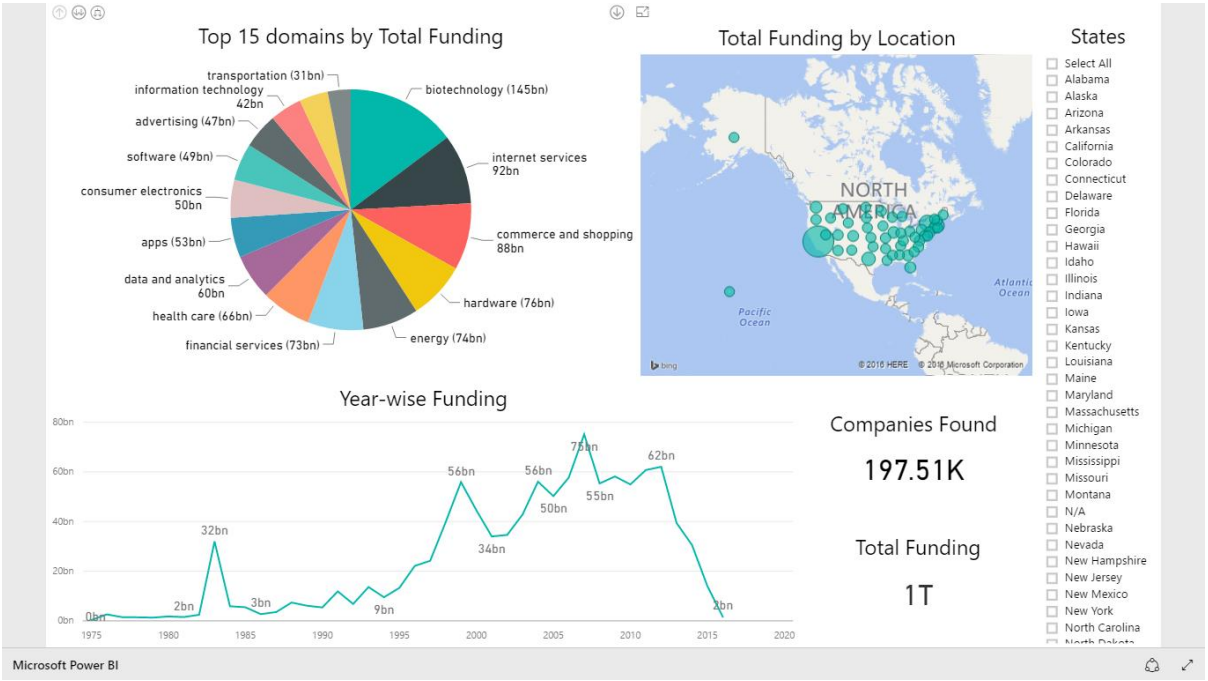


2) Decisions to be made by Location and Domain : Click here for the dashboard

The dashboard was designed keeping two different points of view in mind:

- Startup point-of-view
- Venture Capitalist point-of-view

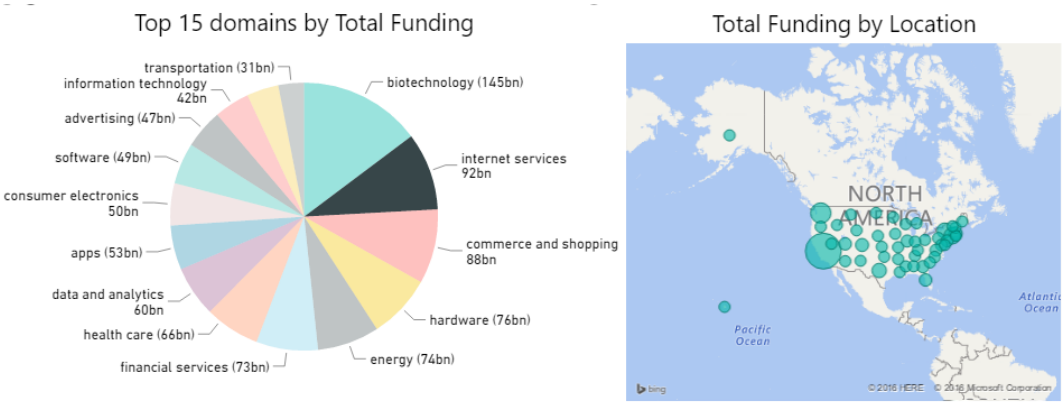
This dashboard shows data regarding all the domains taken together. We can see the market share of each doamain and the map plots the total fundings by location. The line chart plots the total funding for all the domains every year which helps us get a time series analysis of funding trend.



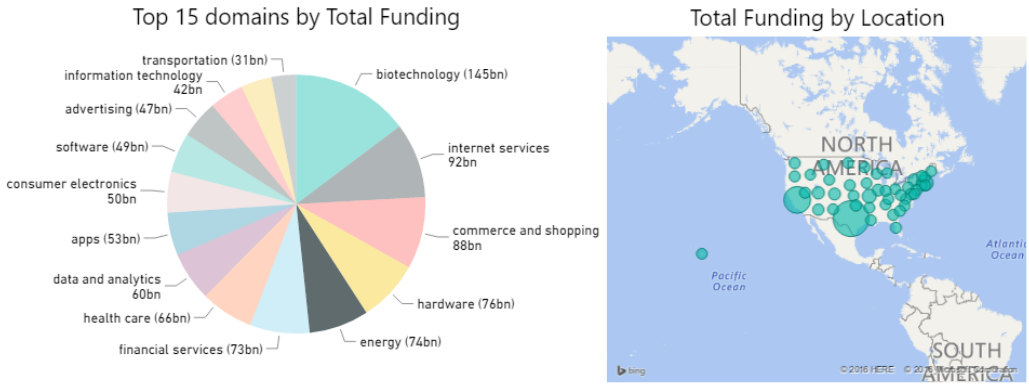
Following is a description of the ways how the dashboard can be used:

a. Startup point-of-view:

The dashboard help the startups identify the best locations for a domain. So, for example if someone has a **startup in the software domain**, from the dashboard, **California** seems to be **a promising choice**.



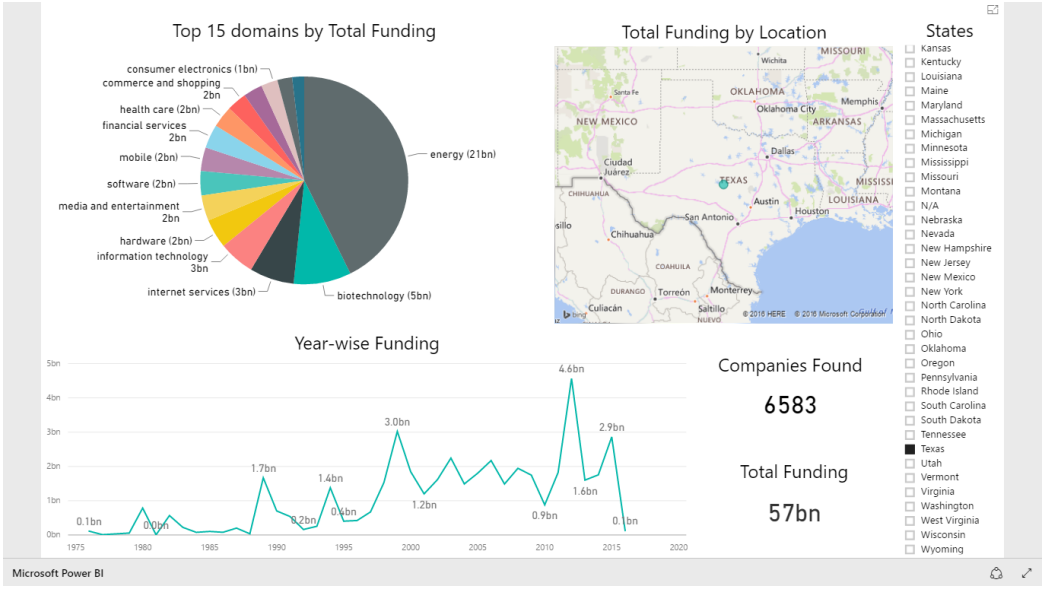
Or similarly, for **startup is in energy** sector, **Texas** seems to be a viable option.



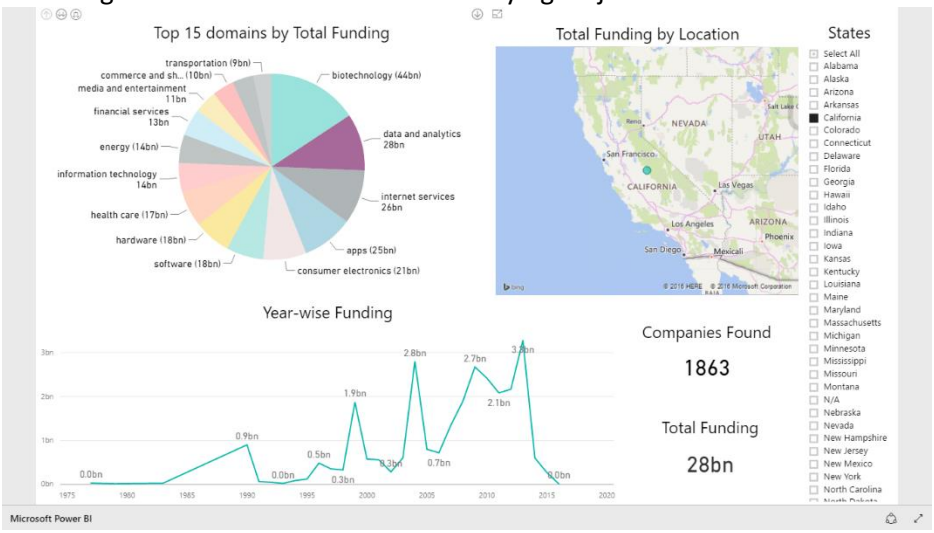
b. Venture capitalist point-of-view:

A VC can take better decisions if they know which domains are doing good in that location and what has been the trend of funding for a domain in that location. The following dashboard locks on various location and checks the funding patterns for a domain. Following are various examples a VCs are likely to come across:

i. Let's take an example of Texas. Me as a VC would like to know how various companies in the domain of Energy are doing*(Fig-3)*.



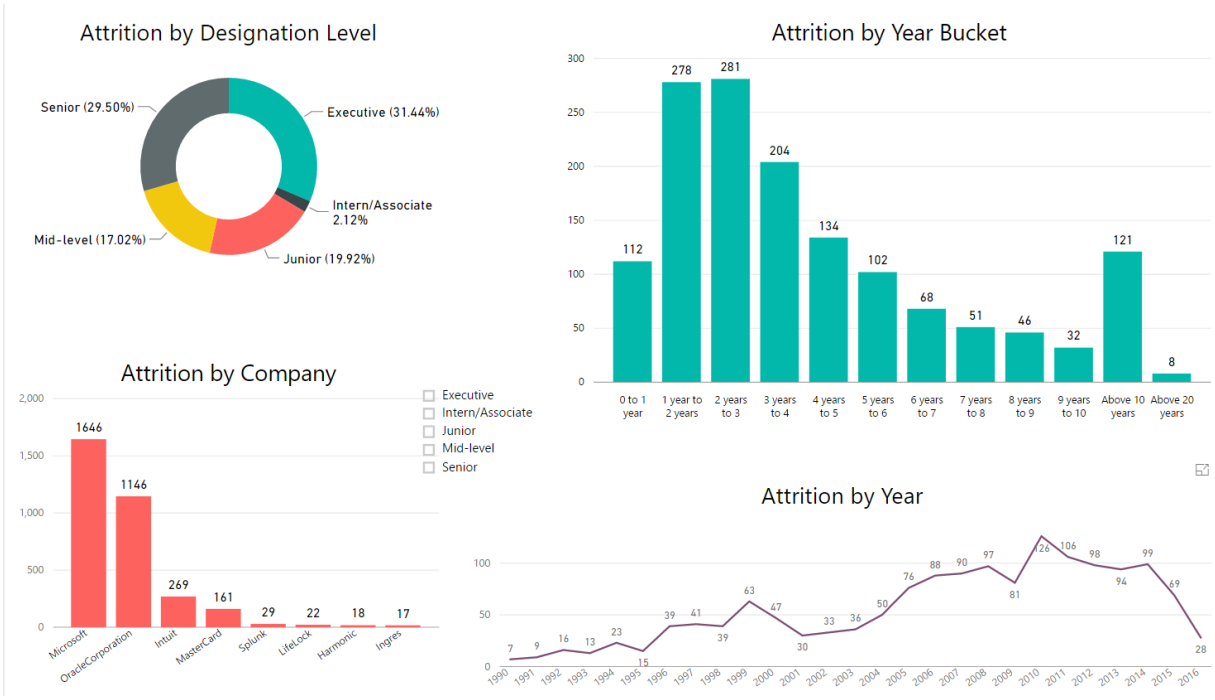
ii. But for a place like California, we see it isn't just the software industry which is at a boom. The domain of Biotechnology is also at a boom. In scenarios like this one would like to know the most recent funding trend for domain rather than relying on just the market share of that domain



So, for a certain kind of startup, which location attracts more funding, can be found here at one shot view.

3) HR point of view - Employee Attrition : Click here for the dashboard

Human Resources Department is responsible for managing the company's most valuable resources: its employees. This dashboard was made to show the HR department how the employee attrition scenario varies from company to company based on various designation categories and how it changes on yearly basis. This will help them to plan their policies and strategies according to the current trend and retain their valuable resources.

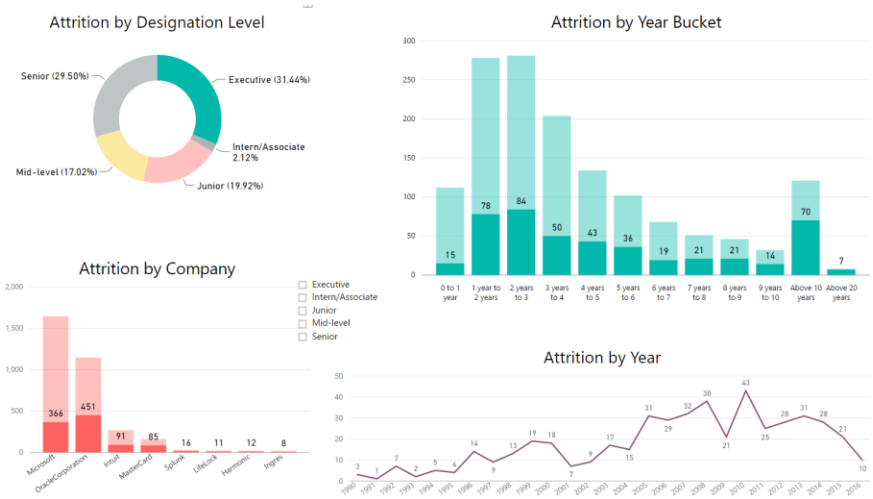


The dashboard has 4 charts that signify:

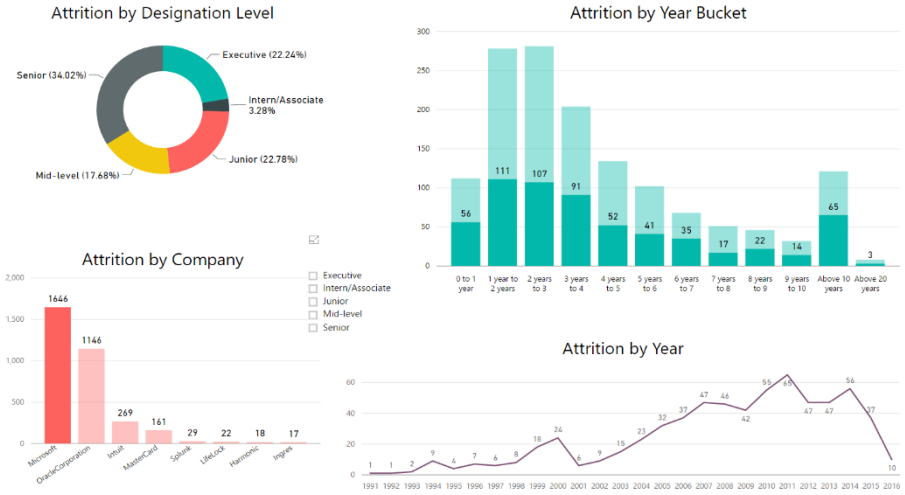
1) Pie chart with attrition rate by the designation levels. (Again, not a good practice; but we wanted to show the % share from total)
2) Bar chart with attrition rate on company basis.
3) Bar chart with the attrition rate based on the no. of years of experience of employee in that job.
4) Line chart with the attrition rate on yearly basis.

<u>Directions for using the dashboard to gain meaningful insights:</u>

1) Designation Level: By selecting a designation level, we can see the attrition rate of all the companies, on year to year basis and years bucket based on that designation level.
So, by selecting Executive level we can observe that majority of the executives leave after an experience of 1-2 and 2-3 years. Oracle Corporation has the highest attrition rate and the years 2008 and 2010 saw higher no. of executive leaving and currently the rate is decreasing.

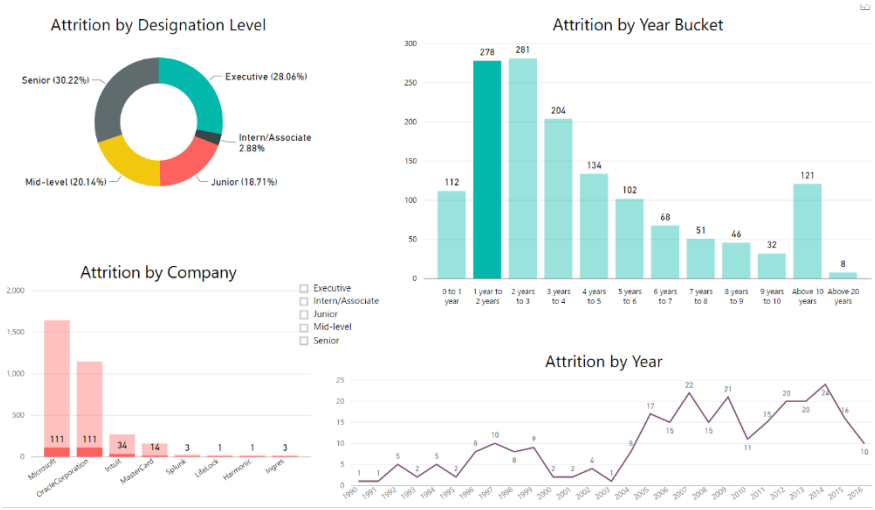


2) Company level: By selecting a company, we can see the attrition for the company at each designation level, year bucket and on yearly basis. So, by selecting Microsoft we can see that the senior level employees have the highest attrition rate and the employees tend to leave after experience of 1-4 years. The year 2011 saw the highest rate of attrition in Microsoft and currently the rate is lowering.



3) Year-bucket level: By selecting a year bucket we can see the company wise, designation level wise and year wise attrition rate for those specific no. of years of experience.

By selecting the year bucket of 1 year to 2 years we can perceive that the senior level employees have highest attrition rate closely followed by the executive level employees. Microsoft and Oracle has equal no. of employees leaving in 1-2 years and years 2014 and 2007 saw maximum no. of employees belonging to that year bucket.



These insights will help the HR department to plan and provide structure to the **company's employee performance appraisal** and position program and aid in meeting the business needs. For example, if for a job role, the attrition is happening after 3 years, a company may think of raising salary for that employee or offering promotion.

LEARNINGS AND FUTURE SCOPE

This project has been instrumental in enhancing our knowledge on how BI initiatives transpire in an organization. We undertook the planning, creation and execution of an idea and as the project progressed, we realized that there is huge potential associated with the idea.

LEARNINGS

- It is important to have a clear understanding of the requirements before you start the project plan. In case the requirements are not clear, it's better to ask questions.
- Data is available across multiple sources, but relevant data sources help in developing meaningful insights.
- Data modeling helps in creating a structure based on which the project can be developed.
- Data integration is the most important part of the BI project, and takes the maximum time and effort. We applied the concepts of pseudo facts and conformed dimensions and identified them to enhance the data model which better suited the business requirements.
- It is very easy to fall in traps while creating the data model, an insight or a dashboard. Therefore, preventive measures should be taken to avoid any probable traps.
- It is important to identify the BI personas who will use the dashboard from the application. Before creating a dashboard, thorough analysis is required. The dashboard should be ideal for the type of data and analytics we are demonstrating.

FUTURE SCOPE

- With companies collecting large amount of data and decisions becoming data driven, we can incorporate predictive analytics to help company make better decisions.
- The current scope of the companies can be extended to different geographical domains.
- Focus on companies which are going public.
- Focus on relevant events with notable keynote speakers and attendees which will help the company to identify potential investors.
- To track where the employees are going to determine the required strategies that should be implemented to retain the most valuable resource of the organization.

REFERENCES

- Business Intelligence Guidebook: From Data Integration to Analytics - Rick Sherman
- http://www.kimballgroup.com/
- http://www.mrc-productivity.com/blog/2014/05/7-reasons-why-business-intelligence-projects-fail/
- http://www.dataintegration.info/business-intelligence
- https://help.alteryx.com/10.6/index.htm
- https://powerbi.microsoft.com/en-us/documentation/powerbi-service-get-started/