REAL-TIME DATA ANALYTICS - Meetup.com
FINAL REPORT

Motivation

Data today is growing at an exponential rate and precise insights - if drawn out of it can lead to great positive decision making and competitive edge in the market. Currently, organizations are generating a huge amount of data, that is stated as the Big Data and without efficient analytics carried out - it has no value to the organization.

- **Why Real-time Analytics**
  - It is the best way to focus on the 3V's of Big Data: Volume, Velocity, and Variety
  - Since the inflow of data in continuous, Real-Time analytics is a significant way to deduce valuable statistics as soon as the data enters the systems of the organization. It allows getting the most recent trends, updates, and market characteristics
- **Why this topic**
  - Our primary motivation to take up this topic was to dig deeper into the field of real-time streaming analytics and the technologies associated - such as Kafka, Apache Spark and sophisticated Machine Learning algorithms for segmentation. In our project, we have taken up the Meetup.com website's RSVP data and have performed real-time analytics along with segmentation

Goal

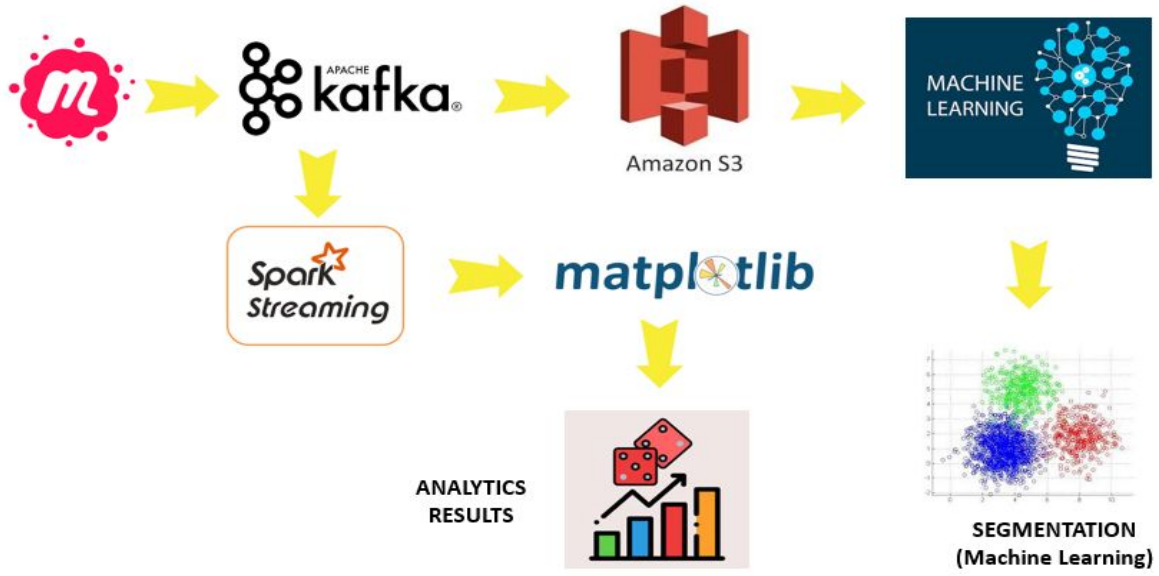
Our aim with this project is to pull live RSVP data from the API of Meetup.com website using Apache Kafka integrated with Spark Streaming and studying it to draw conclusions that will help the business grow and strategize better. With our real-time analytics and prediction, we are trying to help business in managing more effective events at the places where more meetups occur. This helps the business to increase their visibility to more customers and build more effective relationships in the networking events. With our analysis and prediction, we are planning to frame a set of KPI metrics.

Key Performance Indicator Metrics

- Which city has the highest number of meetup events?
- Which is the most famous event?
- Which type of event suits a chosen business?
- We plan to see which cities are responding most actively at a given point of time and what is the technology/event agenda which is popular throughout
- We plan to create segments of the most popular cities and technologies with the help of Clustering machine learning algorithms

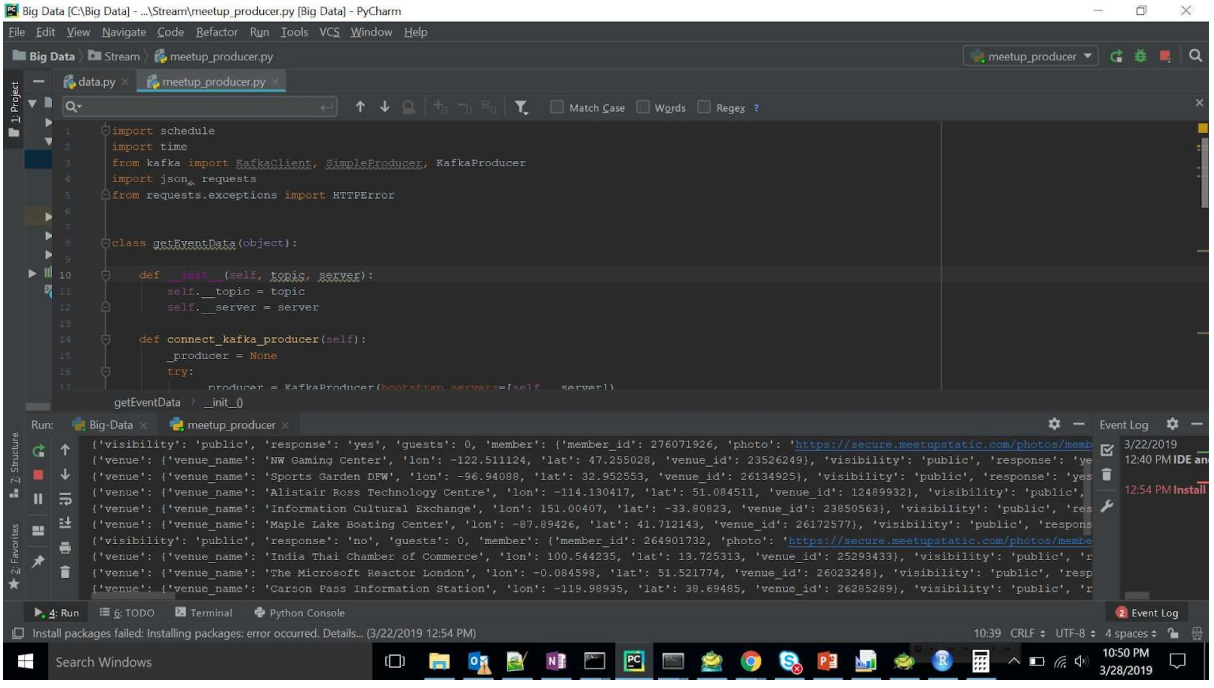
We have restricted our analysis to the cities within USA by filtering the overall data from across the world

Architecture Overview



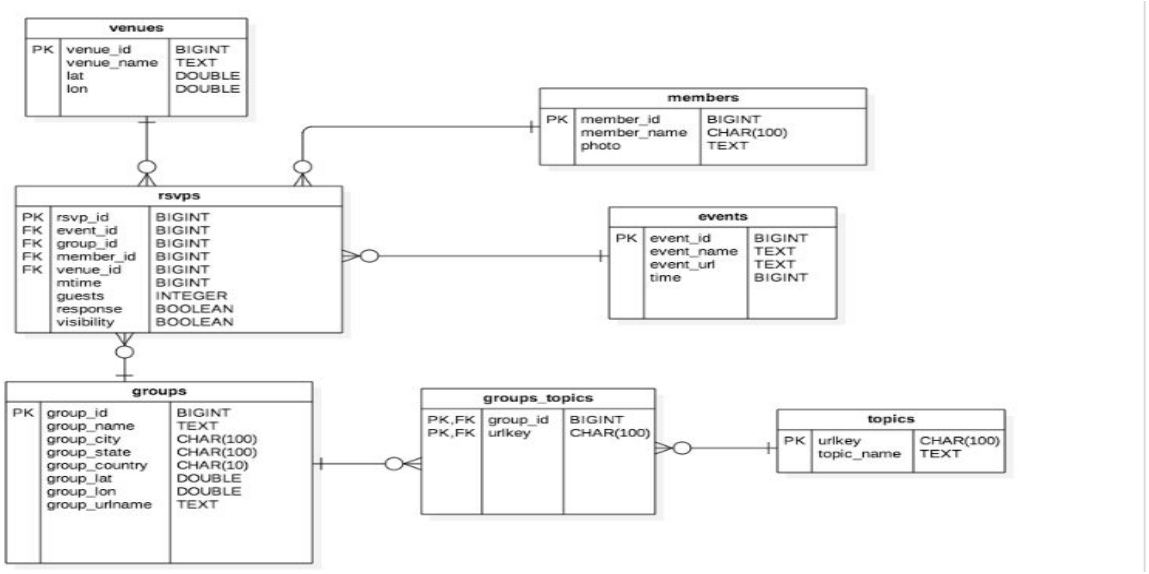
Data Source and Collection

- We have pulled the data from the **Meetup.com API** using **Apache KAFKA** and could obtain multiple JSON files which we will be working upon.



- Meetup API JSON Data - Displayed below is one amongst the entire set of JSONs we've extracted.

- Since we are getting in the data in multiple formats from the Meetup.com API - i.e. structured and unstructured - we are primarily focussing on one major segment: the JSON format data.
- To put that to our application and computation, we have converted the JSON format to an ERD schema to get a tabular form.
- From here on, we treat our unstructured data as relational (tabular) and perform the required operations.



Relevance of Data to our use case

Analytics Perspective:

- As displayed in the JSON extracted, we have variables like **city, country, state, topic_name,**

```
{
 "venue": {
  "venue_name": "Community Room in Vancity",
  "lon": -122.868164,
  "lat": 49.237938,
  "venue_id": 26175596
 },
 "visibility": "public",
 "response": "yes",
 "guests": 0,
 "member": {
  "member_id": 276905729,
  "photo":
"https:\\/\\/secure.meetupstatic.com\\/photos\\/member\\/4\\/9\\/3\\/5\\/thumb_286218741.jpeg",
  "member_name": "Gerry Chahal"
 },
 "rsvp_id": 1777786286,
 "mtime": 1553120859000,
 "event": {
  "event_name": "Pre-Dating, Know Your Worth",
  "event_id": "259763892",
  "time": 1553650200000,
  "event_url": "https:\\/\\/www.meetup.com\\/Conscious-Couplings\\/events\\/259763892\\/"
```


event_name, and **time** - to name a few

- This enables us to carry out visualization on our real-time data using matplotlib and see which city is responding most actively and to what events
- This also gives deeper insights about the popular times and specific technologies that garners the user interest

Segmentation using Clustering:

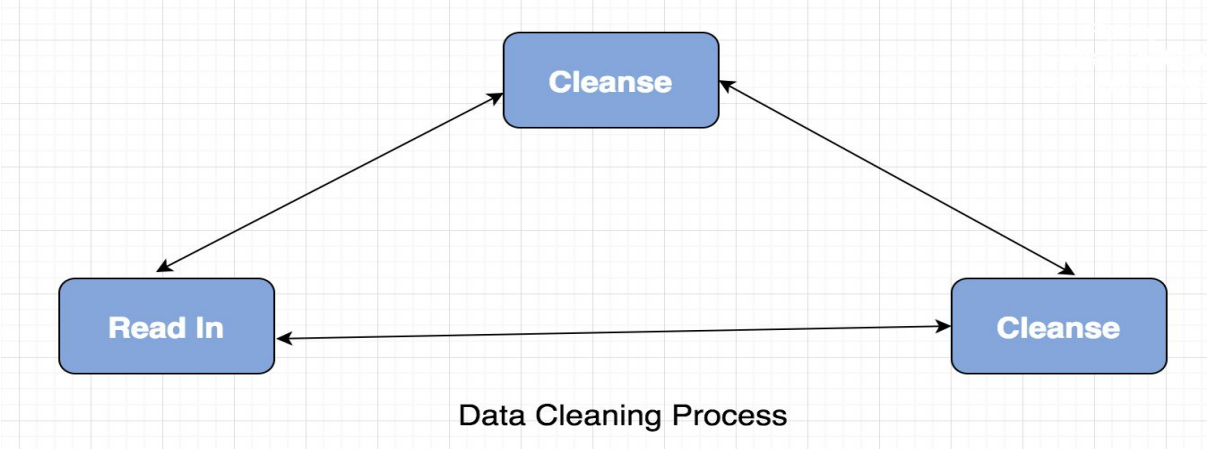
Out[7]:

	event_name	group_name	urlkeys	all_text	response
0	Nine Famous Graphs and How to Make them Interactive	SF Data Science	[data-science, big-data, data-visualization, data-mining, machine-learning, big-data-analytics, ...	Nine Famous Graphs and How to Make them InteractiveSF Data Sciencedata-science big-data data-vis...	1
1	Steilacoom Guided Walk 5 or 10km - Traditional Volksmarch	South Sound Walkers Meetup Group	[walkers, volkssport, fitness, outdoors]	Steilacoom Guided Walk 5 or 10km - Traditional VolksmarchSouth Sound Walkers Meetup Groupwalkers...	1

- We have the real-time RSVP data with information as shown above in the data snippet
- 5 important categories have been considered and the observation made by us is that each group/event relates to a particular topic category, eg tech, culture, etc
- Post studying the urlkeys closely, we can define the similarity between the groups and we see that urlkeys are mostly overlapping with the similar groups
- We plan to use this conclusion to categorize groups into Clusters

Data Cleansing Steps

Since we have streaming data, we identified four problems in our data i.e. the data was incomplete, incorrect, unstructured, and inconsistent. Incorrect or inconsistent data leads to false conclusions during analysis. Hence it is very important to clean and transform the data before doing the actual analysis. Proper handling of the text data will help us to reach more enriched conclusions from data analysis.

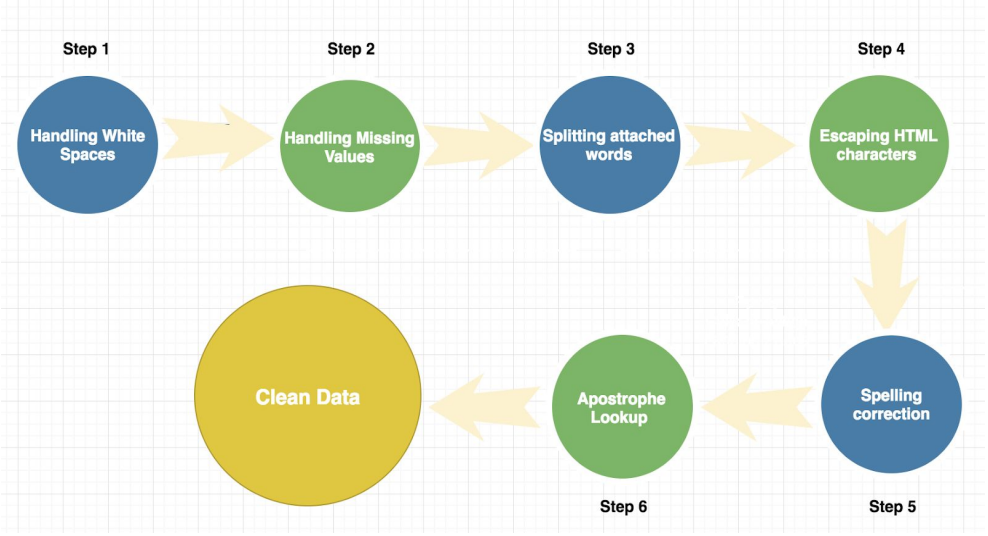


Below are the steps we followed to clean our data:

1. Handling Missing Values
It is difficult to perform analysis on data where one or more values in the data are missing. From our dataset, we choose to omit elements that contain missing values (NA).
For this purpose, na.omit and tidyr::replace_na() functions are used.

2. Removing White Spaces
The whitespace in the dataset can cause problems when attempting to sort, subset or various other common operations. We removed the white spaces in two steps:

a. To remove the leading and the trailing spaces, we used trimws() function and,
b. To remove 'extra' spaces in between, we used gsub() function.



3. Splitting attached words
Some of the streaming data contain multiple attached words. For example, WalkersMeetupGroupwalker needs to be split to Walkers Meetup Groupwalker.
This can be done using the simple strsplit function.

4. Escaping HTML character
The streaming data contains a lot of html entries like > & and these entries get embedded in the original data. To get rid of these entries, we used XML package of R, which converted these entities to standard html tags. For example & is converted to & and > is converted to >.

5. Spelling correction
The spell check process includes correcting spelling errors and finding slang terms. Social media is full of Slang words. When working with free text, these words should be converted to standard ones before doing the actual analysis.

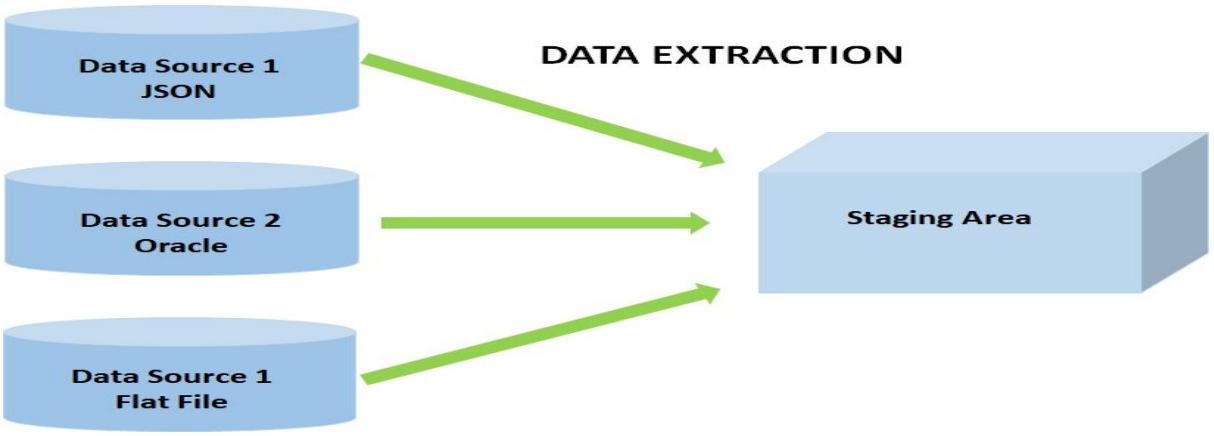
6. Apostrophe Lookup
In the case of apostrophes, the chances of disambiguation increases. For this, we can either use pre-existing dictionaries available on the internet or create our own.

ETL

Post data acquisition, the next steps are bringing in the data in a platform from where it can be accessed and utilized for analysis and visualization. To be able to put the acquired data to use, important factors are Extraction, Transformation, and Loading of the data. We will be discussing all three stages in detail and how we have incorporated them while moving ahead with the development of our tool.

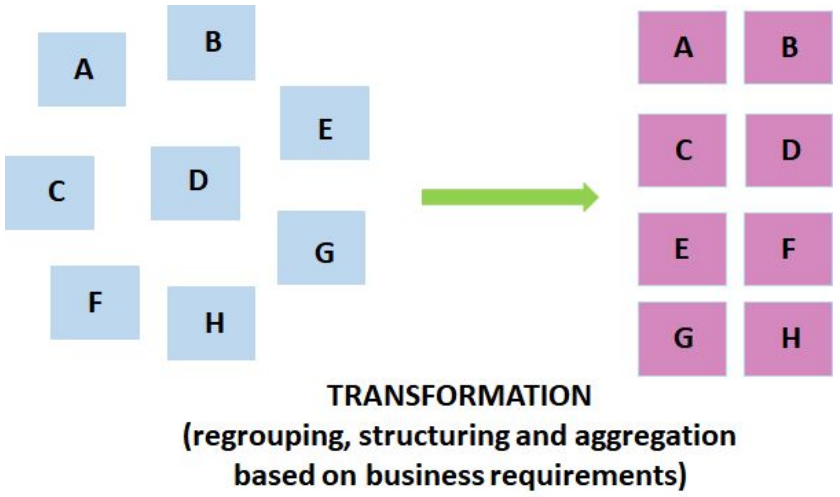
Extract

When we begin working with data, our first challenge is to get all of it on one single platform and make it accessible to manipulate and perform analysis. We have multiple data sources available out there and bringing them all on one scale is what we have to achieve as our end goal.
In our use case, we've had multiple types of data - RDBMS tables and JSON files to be specific since that's what we have brought in to our working. The former is a structured form of the data, while the latter being unstructured. We have used platforms/tools like Apache KAFKA for load distribution of our data to multiple users (depending on the requirement of every end user) and then stored the data fit to work upon at Cassandra.



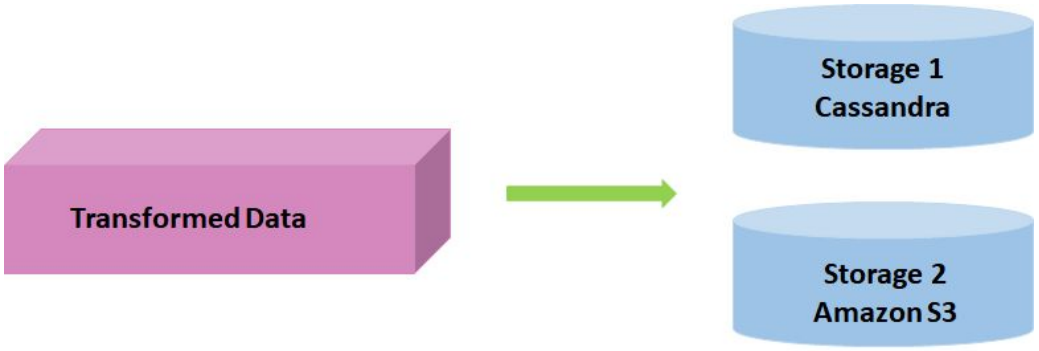
Transform

Having extracted the data from the source, we now move on to the next step which is the Transformation of the data. We now have all our data on a scalable single platform, but the challenge lying ahead is making that fit for use. Transformation of data is the process where we apply multiple techniques to restructure the information we have and remove all the unwanted elements. Transformation isn't always about the removal of incorrect/dirty data - but refinement according to the use case and relevance with the context. In our process of development - we had multiple columns of RSVP Meetup data having specifications like area, date, time, technology, people attending, etc. But, we did not need all of it for our analysis - hence we transformed our data by adding necessary filters.



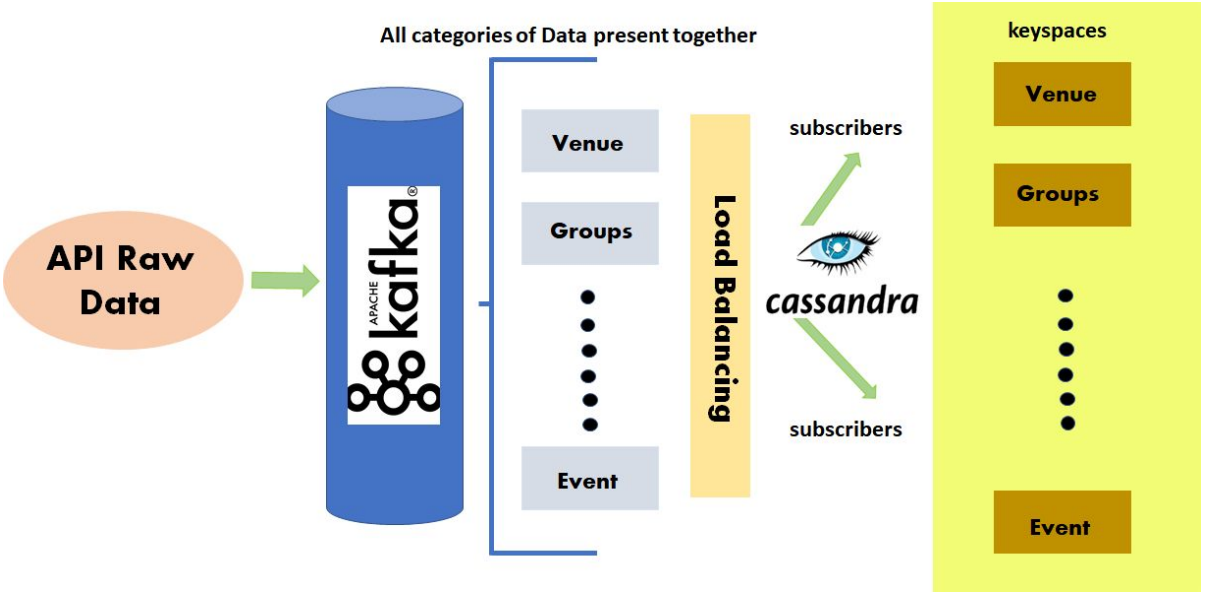
Load

This is the final leg of the ETL process and largely depends on the organization/school we are working the data for. The fully transformed data is supposed to be stored somewhere so that it can be referred to in any requirement. In our tool development - we are primarily using Pivot Tables because that is a common platform.



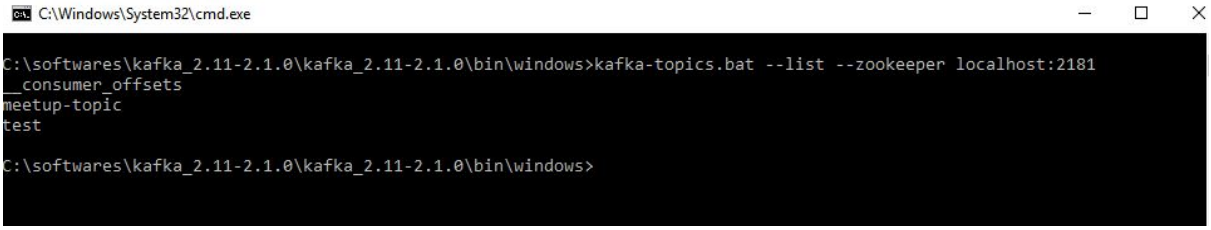
LOADING the processed data on various Data Storage Platforms

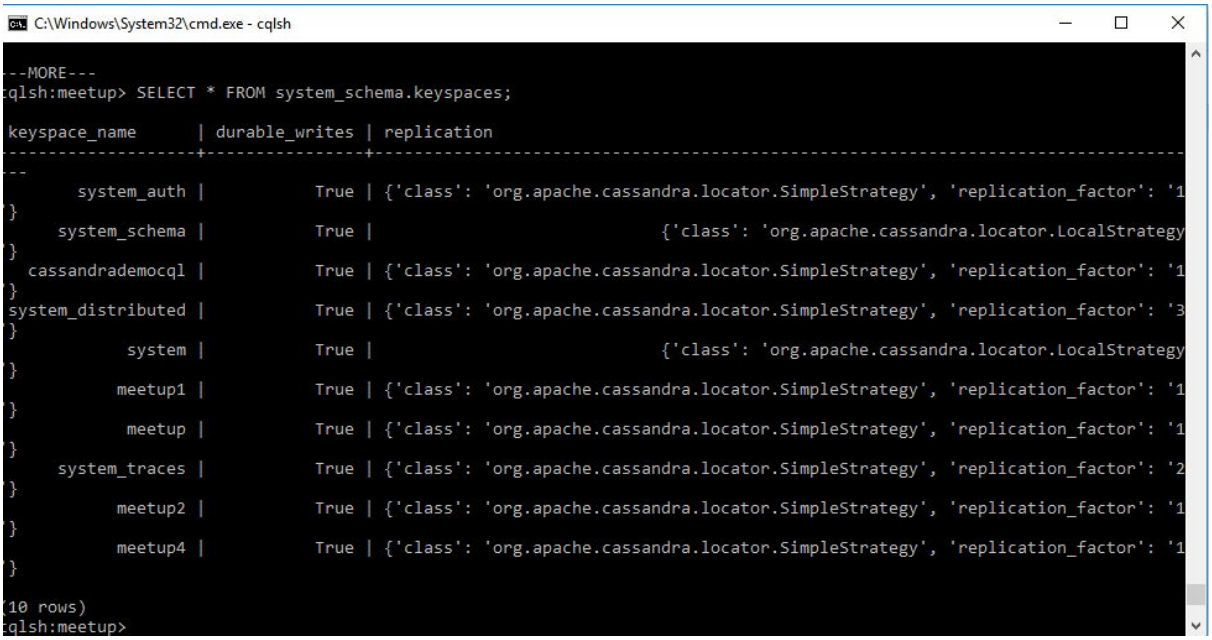
ETL and Storage : Meetup.com API Context

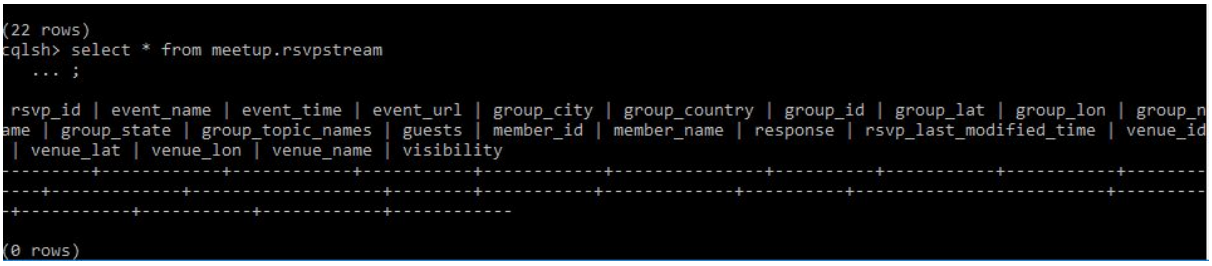


1. We have used the Meetup.com Rest API to pull the Raw API data - in tabular and JSON formats.
2. The data covers multiple categories like Venue, Groups, Topics and more. We aim at storing this diverse data according to the user requirement at separate places and here we incorporate KAFKA for load balancing.
3. The data is split into TOPICS and this is then sent over to specific SUBSCRIBERS (users), avoiding high data volume handling at a single node.
4. We have stored this in Cassandra in the form of KEYSPACES, segregated TOPIC wise.

5. At this point - our data is ready to use for Visualization and Model fitting.







```python
#!/usr/bin/env python

from cqlengine import columns
from cqlengine.models import Model


# Model Definition
class Rsvpstream(Model):
    venue_name = columns.Text()
    venue_lon = columns.Decimal(required=False)
    venue_lat = columns.Decimal(required=False)
    venue_id = columns.Integer()
    visibility = columns.Text()
    response = columns.Text()
    guests = columns.Integer(required=False)
    member_id = columns.Integer()
    member_name = columns.Text()
    rsvp_id = columns.Integer(primary_key=True)
    rsvp_last_modified_time = columns.DateTime(required=False)
    event_name = columns.Text()
    event_time = columns.DateTime(required=False)
    event_url = columns.Text()
    group_topic_names = columns.Text()
    group_country = columns.Text()
    group_state = columns.Text()
    group_city = columns.Text()
    group_name = columns.Text()
    group_lon = columns.Integer()
    group_lat = columns.Integer()
    group_id = columns.Integer()
```

```
cqlsh:meetup> select * from rsvpstream;

 rsvp_id    | event_name                                                              | event_time
        | event_url                                                                            | group_
city       | group_country | group_id | group_lat | group_lon | group_name
     | group_state | group_topic_names

                                                                                                   | guests |
member_id | member_name                          | response | rsvp_last_modified_time | venue_id | venue_lat | venue_lon | ven
ue_name                                                                                   | visibility
------------+------------------------------------------------------------------------------+------------
---------+-------------------------------------------------------------------------------------+-------
----------+--------------+---------+----------+----------+--------------------------------------------
-----+------------+------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------+--------+-
---------+------------------------------+---------+------------------------+---------+----------+-----
----------------------------------------------------------------------------+-----------
 1782869420 |                                          MEET AND MINGLE @ Retro Bar  | 2019-04-27 14:00:00.000
000+0000 |                                        https://www.meetup.com/LGBT-20s-and-30s/events/258644763/ |
    London |                | 26970769 |        51 |         0 |                                     LGBT+ 20s and
 30s |            |
                                                                                   lesbian-couples,gay-singl
es,gay-men,gay,lgbtfriends,Lesbian,gay-and-lesbian-friends,lesbian-friends,gaypros,lesbian-social-networking |      0 |
277599720 |                              Dani M |      yes |                      null | 25953180 |      null |      null |
                                                                                    Retro Bar |     public
 1782871744 |                                      DRINK & DRAW & SNACKS. SHORT POSES | 2019-04-26 18:00:00.000
000+0000 |                                     https://www.meetup.com/ART-CLUB-BARCELONA/events/260637901/ |
```

Data Visualization

Data visualization is the process of understanding patterns, trends, and insights by transforming business data into a visual context. Using data visualization, we are aiming to

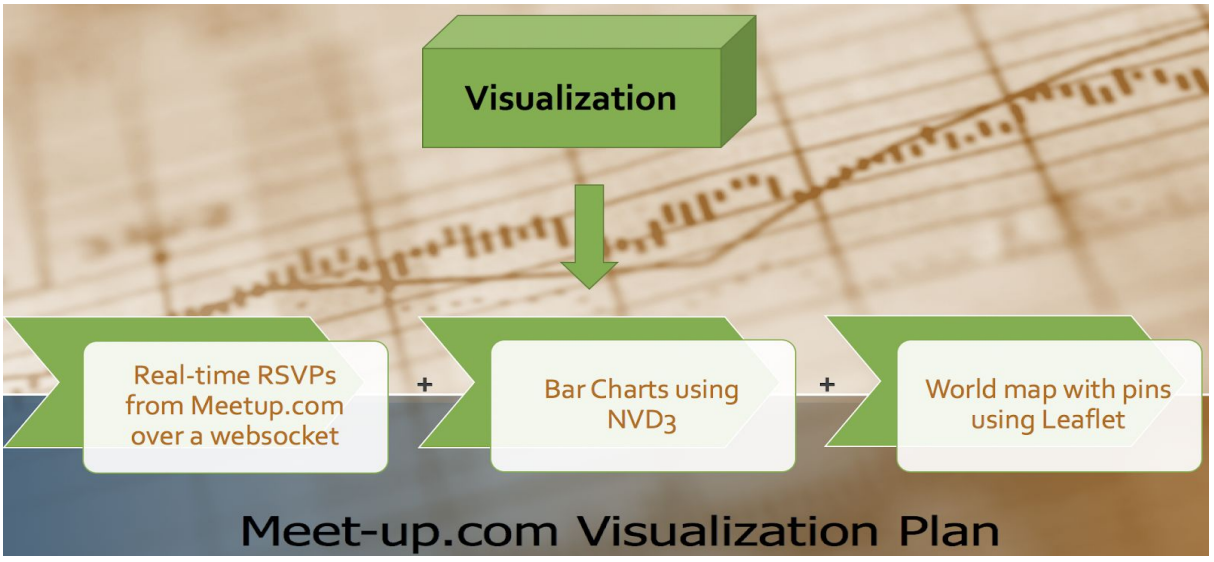
put the complex data into a graphical format enabling us to understand the state of all different types of RSVP events and identify patterns to chart out successful strategies.
For our Real-time Data Analytics, data visualization has helped in:

- ➢ Building association between seemingly unrelated RSVP data which further helps in understanding the location of events
- ➢ Predicting future trends in Meetup events (e.g. the type of people and location which participates in Meetup events)
- ➢ Guiding the end users/viewers to make more informed business decisions

Our Plan

We are doing real-time visualization of streaming data from the Meetup.com open events RSVP API using:

- React
- Flux
- D3, and
- Leaflet



Why we chose Reach and Flux:

1. Re-rendering in real-time.
2. Single source of data, multiple components.
3. Reusable components.

Real-time visualization

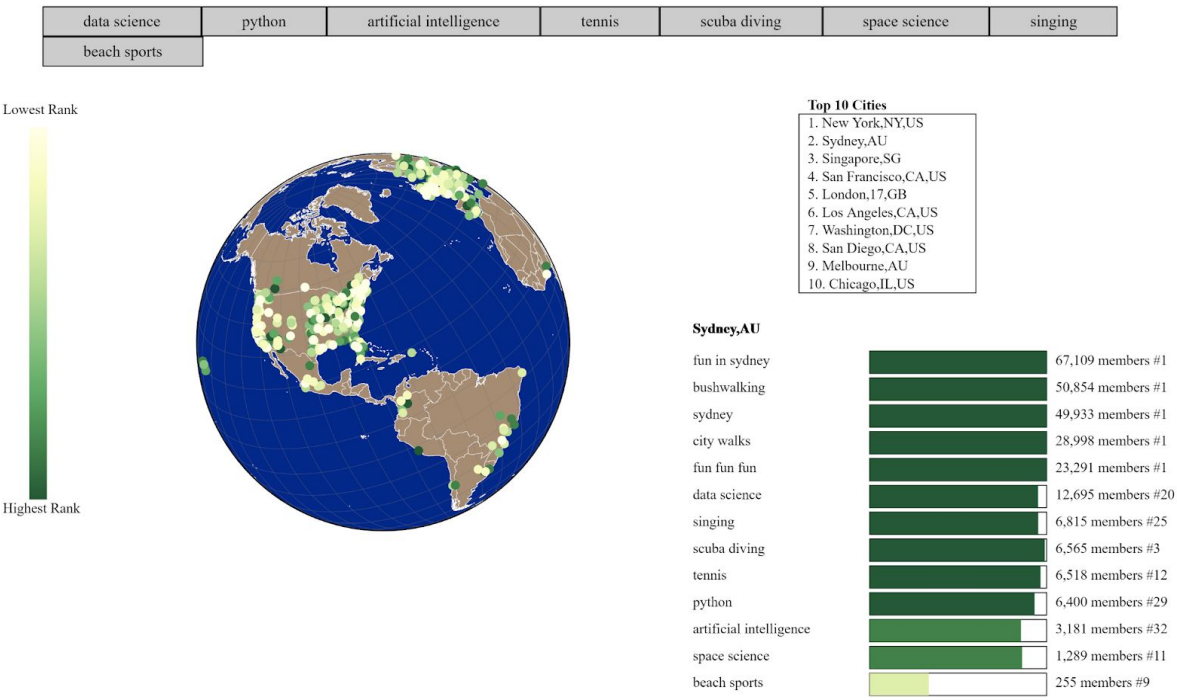
For real-time visualization, we are visualizing events as a map with markers and a bar chart showing the most frequent countries, states (USA), and names as shown below:

Since it is a real-time visualization, hence when the number of RSVPs increases, the values under Most frequent Countries, States and First names fluctuates.

Finding the best city for user specific topics

Here, the visualization is made up of 3 main components:

1. Globe with city markings indicates cities containing the user selected topics and color-coded by that cities sum total ranking in each topic. The lowest total score is the highest-ranking city for that unique list of topics.



2. Detailed data for user selected city indicating top 5 ranked topics for those cities, plus the rank for each of the user inputted topics.
3. Top 10 ranking cities for user selected list of topics.

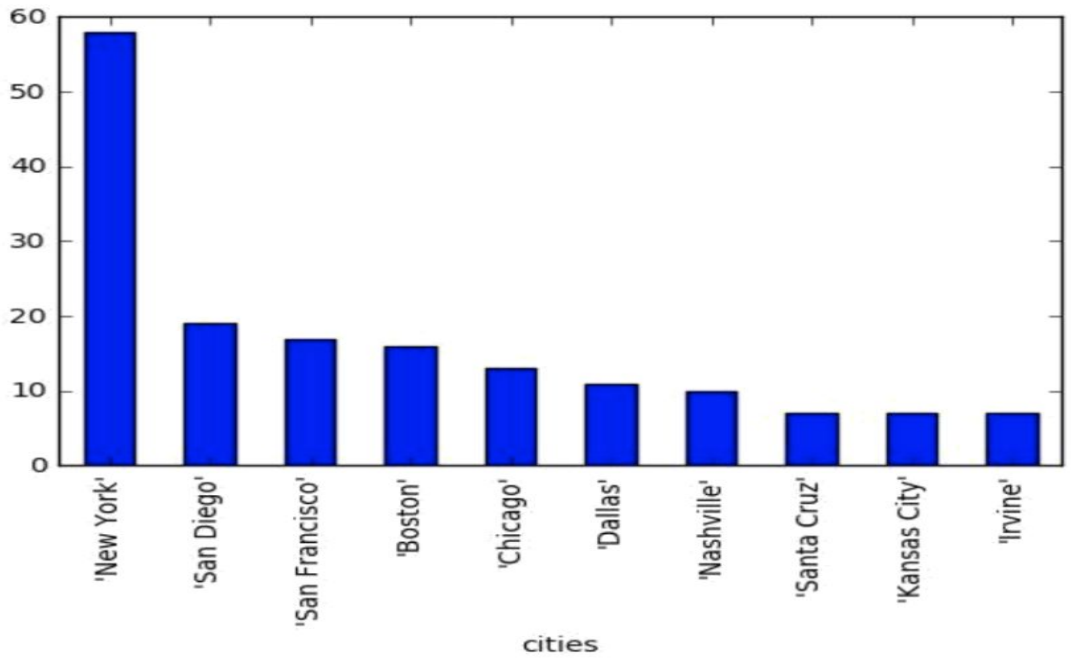
By looking at the above graph, a user can find out the city for the meetup events for his selected topics. It is clear from the above graph that the best city for all 8 topics is New York followed by Sydney and Singapore on the 2nd and 3rd place respectively.

Top 10 meetup venues in the vicinity of given co-ordinates

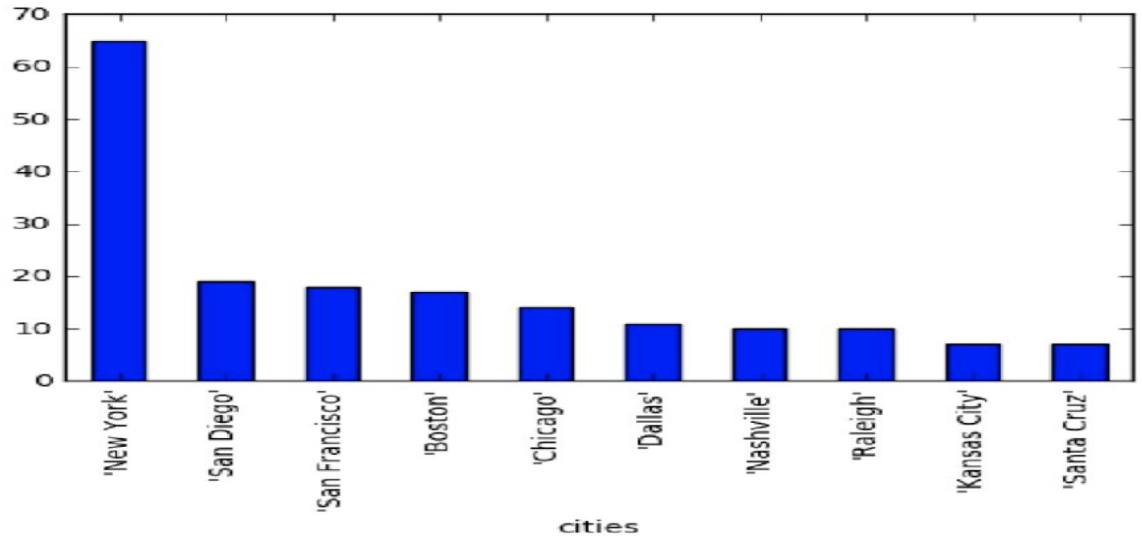
Taking co-ordinates as inputs, the graph shows ten of the highest rated meetup venues in the vicinity on an interactive map. Venue ratings were first normalized taking into account individual ratings and their count for each venue.

Top 10 cities with maximum number of RSVPs

Using the Bar Graph, we can see the top 10 cities from which RSVPs are coming.
The time difference between the plotting of each curve is 2 seconds.

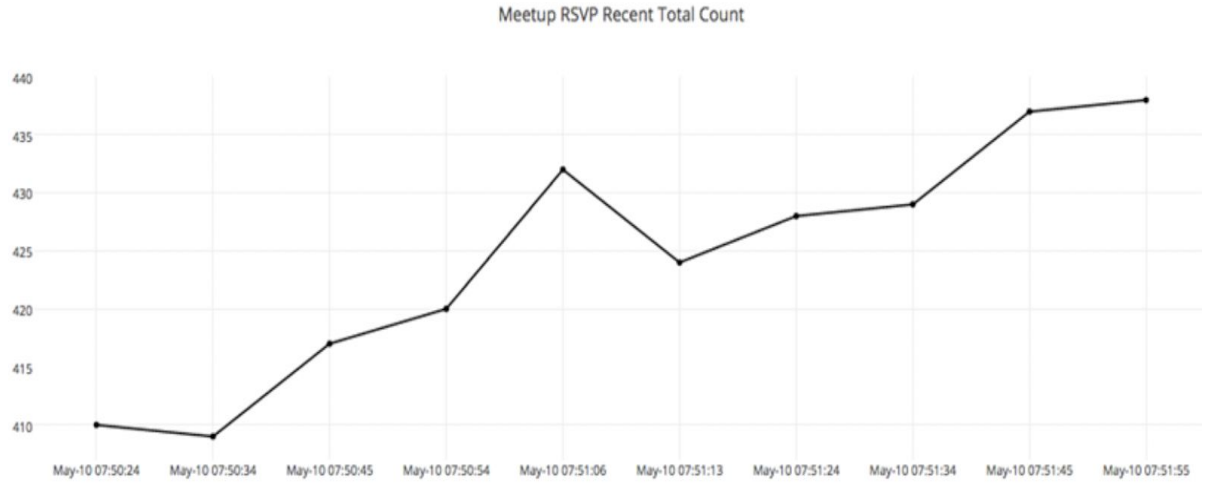


As you can see cities like ‘New York’ received more than 7 RSVPs within a time span of 2 seconds.



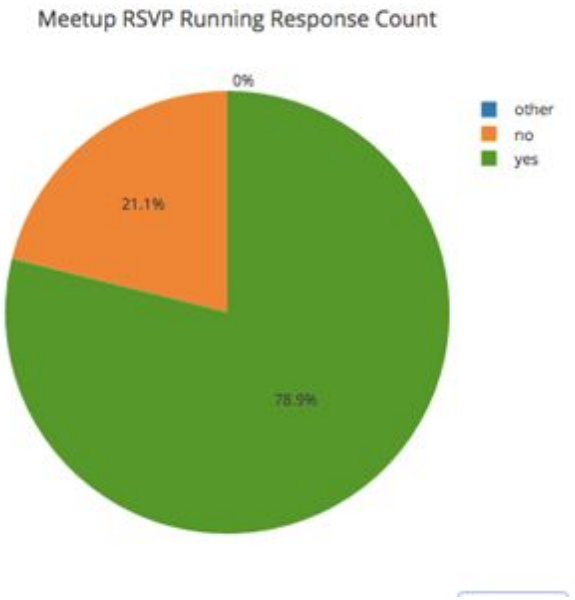
Recent Total Response Count

A line graph is used to capture the most recent (time frame taken is 5 minutes) response counts for the events hosted on Meetup.com. It is refreshed every second for the new incoming data.



Classification of RSVP Response

In every 3 seconds, RSVP responses are refreshed and classified counts of YES, NO and OTHER are generated.

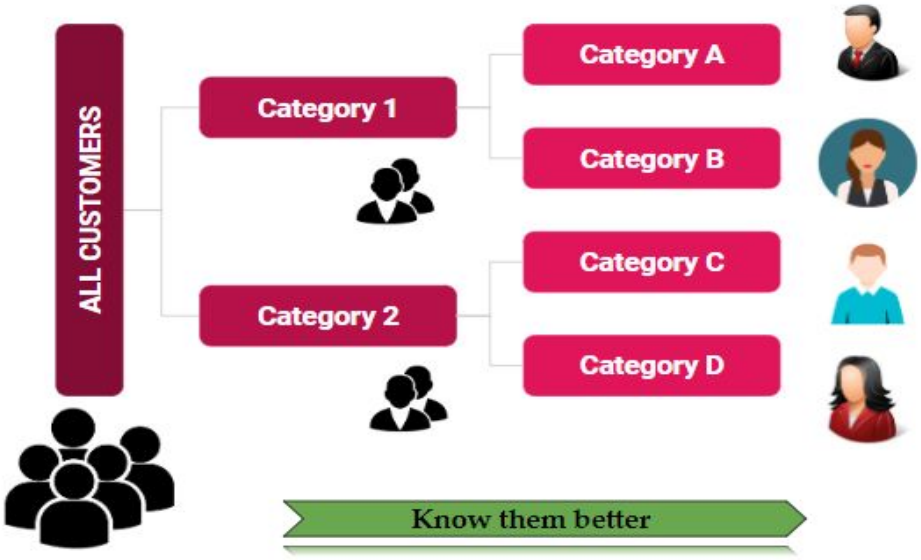


MACHINE LEARNING TO PERFORM CUSTOMER SEGMENTATION

After having successfully completed all the steps discussed till now, we could finally reach the last phase of our tool development - Machine Learning application to obtain segments of customers using clustering algorithms. These results can then be used in the creation of a recommendation engine that will help in the development of better and informed business strategies.

Customer Segmentation

Our intention to carry out clustering was to perform Customer Segmentation. When we talk about business - it's highly crucial to know who are our customers and what they want. Only when we as an organization have the granular and detailed information of our clients/customers, can make more strategic and informed decisions contributing to growth and success.



Why Customer Segmentation ?

Of the numerous advantages that this practice of knowing our customer brings us - a few of them are worth mentioning over here since they have a tremendous impact on business.

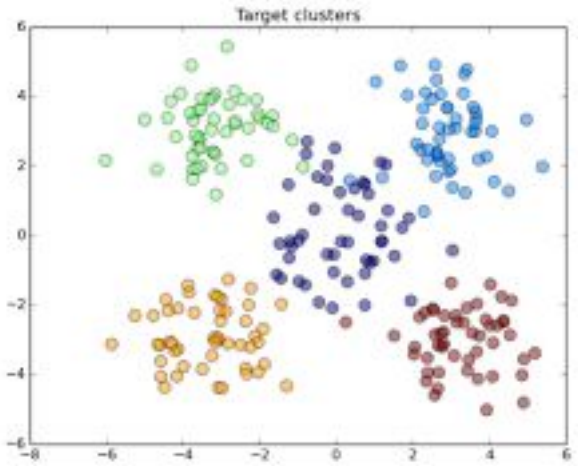
- Market Focus
- Customer Retention
- Business Expansion
- Increase in Competitiveness

In our context, when we have all the data concerning the RSVP responses for the website Meetup.com - we can leverage that to classify the users to give them personalized services, loom into the areas/cities/geographies that are doing really well and can be used for business expansion and the areas that aren't doing very well and needs to be worked upon to improve the business.

MACHINE LEARNING: k-means Clustering

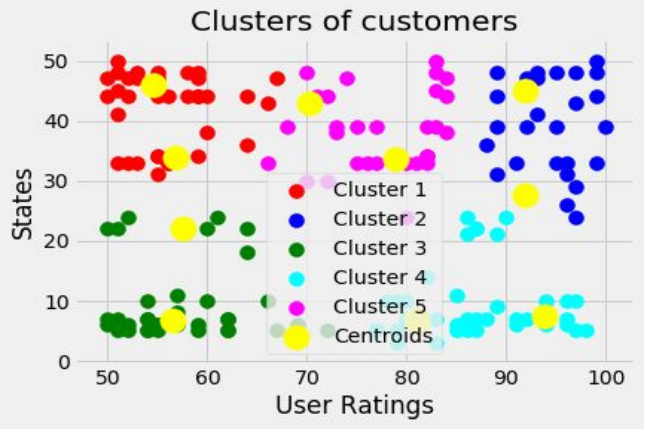
In order to classify the users on the basis of their RSVP data collected, we tried to put to use a couple of Machine Learning algorithms like Hierarchical Clustering and k-means Clustering. We could obtain desirable results using k-means clustering and went ahead with the same to get clearly demarcated user groups.

K-means Clustering: This is the most popular and widely used Unsupervised Machine Learning algorithm used for segmentation. K over here signifies the number of segments we want to break our users into, for obtaining deeper insights.

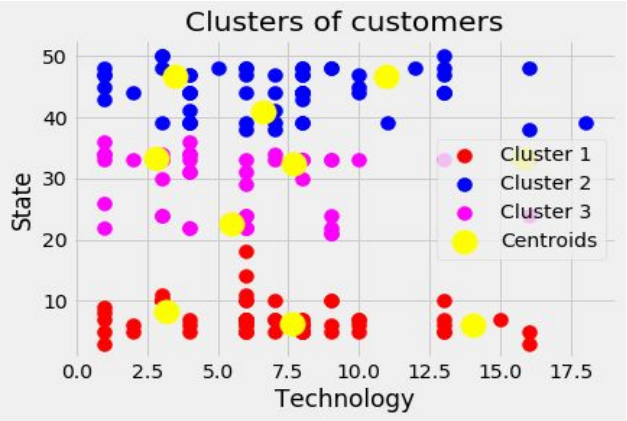


We could classify our users on the basis of multiple parameters that we got in the data acquired from the Meetup.com API. Shown below are a few of them to illustrate our work and results.

- **User satisfaction based on ratings across different states**



- **Assess the popularity of technologies across different states**



So, this way we dig deeper into the details and intricacies of the users in real time by implementing machine learning.

Conclusion:

Real-time analytics has the potential to transform businesses by offering insights that help us in making effective decisions by drawing meaningful business insights.

The businesses have become highly competitive these days and real-time analytics can greatly help serve their customers more effectively based on customer behavior and ever-changing market scenario.

Team members

- Shilpa Prakash Jain
- Pratishtha Shukla
- Shrikant Patel