GOOGLE CLOUD SERVICES PROJECT

INTRODUCTION

Doug laney was the first who coined the term Big data and with the term he associated three ways by which data is meteorically rising(Grable & Lyons, 2018). He further explained the three v's as volume (amount by which data is collected), variety (distinct type of data avaialble) and velocity (the speed at data is being collected). In the later time another term has been added with the three v's and that is veracity. Veracity simply means the noise in data. The intricate web of data makes it very challenging to evaluate and analyse. Not only analyzing this data with traditional methods become overwhelming but also storing this magannimous amount of information itself is a challenege. Nevertheless, both of these complications have been resolved with the advent of cloud based big data services. Cloud services makes our life simpler by taking care of the huge amount of data we have. In addition to storing the data, cloud services also provide their analyzing tools such as, machine leaning algorithms for predictive analytics, data preprocessing or wrangling of data and data visualizaion to name a few. The top companies that are providing cloud services are, Rackspace cloud, IBM cloud, Amazon web services, Microsoft Azure and Google cloud. The direct benefist of this cloud servics can be reaped by the small organizations that can not afford installing big data centers and analyzing tools. They can make sense of their data and can utilize it to gain insights from the data. Managing data on cloud which is termed as cloud computing removes the restriction of managing the system and scalability. Microsoft azure, IBM and Amazon Web Services are some of the cloud services which started providing cloud platforms to the users(Hashem et al., 2015). The cloud computing is based on virtualizaion technique. Virtualization is a technique in which resources are shared by keeping the hardware isolated from the user which increases the resource utilization, efficiency, and scalability.

USER REQUIREMENT

The Prime minister election is one of the most important events for a country that impacts the economy, stregthen or weakens the foreign ties and helps in making reforms suited for the people of that country. The people's view of the Prime Minister is the key element that one party keeps tabs on and it becomes more important after the results are announced. Recently, in India election for the prime Minister are ended and Prime Minister chosen is Mr. Nrendra modi, which by far is the people's choice, however, there are individuals that are unhappy with him or has legitimate complains about him being the Prime Minister. If we can analyze the people's view and check what is the pecent of people that are against him being the Prime Minister can surely help the the committe or advisors appointed by the Prime Minister to take necessary steps towards resolving their issues. The best place to get people's view on any trend is Twitter. As the elections ended on 25th May 2017, I have collected the tweets for over 5-hour period on the same day which amounts aroung 1GB or around 8 lakh rows of raw twitter data. This data was then futher cleaned and all the tweets regarding "Narendra Modi" were filtered. Then a sentiment analysis is done on the tweets to know the sentiments of people that are currently talking about the Prime Minsiter and evaluate their views.

DATA PIPELINE DESIGN

The data used in this project is twitter data which is streamed in real time using GKE (Google cloud Kubernete) on 25th May 2019 for the duration of around 5 hours. GKE engine which essentially runs pods for specific service type mentioned in '.yaml' file. The GKE will read the tweets from the twitter API and dump them into the Gcloud pub sub topic. Furthermore, Kubernetes will pull the tweets from the pub-sub topic and insert them in batches in the Big-query table. The collected data is then cleaned in big query and for some further modification sent to gcloud dataprep. After cleaning the data, all the english tweets are collected and a model is trained in AutoML Natural Language, which performs the sentiment analysis on the desired data.

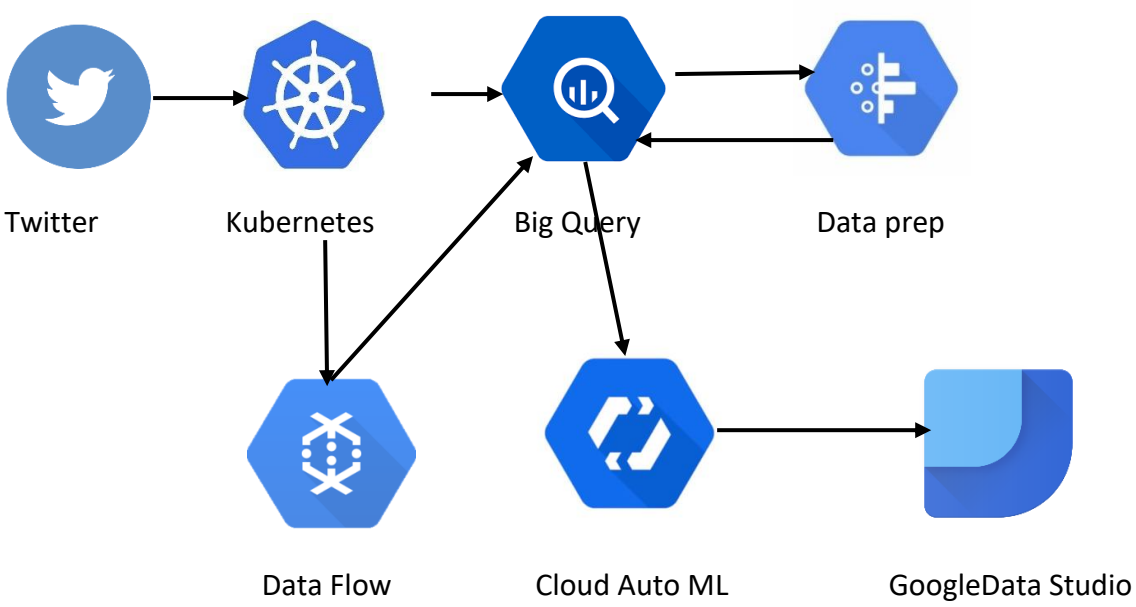


Figure 1. Data Pipeline Design

MAJOR DEPLOYMENT STEPS

STEP 1/5 STREAMING DATA FROM TWITTER AND SENDING TO PUB-SUB

To stream data from twitter, a twitter app is created in developer.twitter.com and keys and tokens are noted which will acts as a comminicator between kubernetes and twitter. In kubernetes a cluster is created (Fig. 2) and access is granted to all the Google cloud API's, so that they can intract with the kubernetes. Two pods are created, one that will stream the data from twitter and other that will write that data into a big quey table. In twitter-stream.yaml file pub-sub topic name with the current project credentials are inserted. The Fig (2) shows that the pods have streamed the data from twitter for 4 hour 21 minutes and it amounts to around 1 GB.

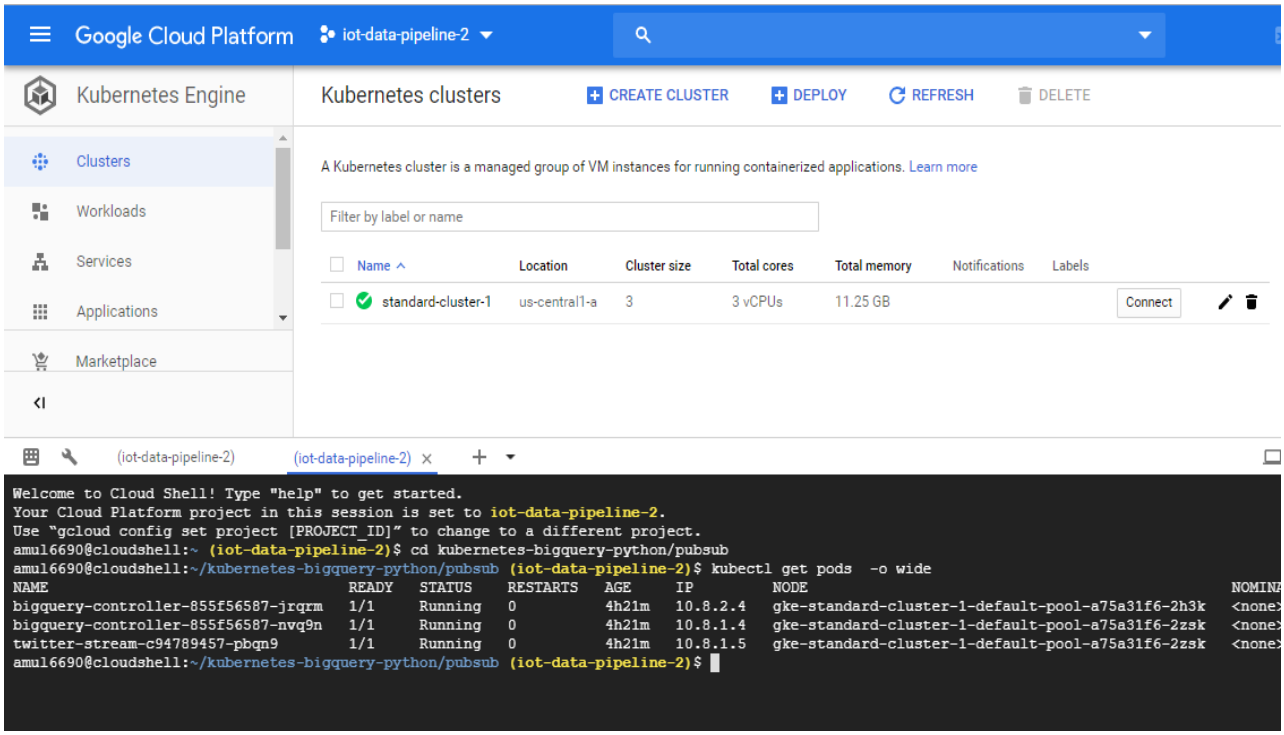


Figure 2. Kubernetes Cluster creation

STEP 2/5 PULLING DATA FROM PUB-SUB TO BIG QUERY

In big query, a dataset is created and, in that dataset, a table is created with the schema to match twitter data. In kubernetes, bigquery-controller.yaml file is created and same pub-sub topic name which was created in pu-sub and project name are enetred. In addition to that, big query data set name and table name are also entered. This will essentially pull the tweets from pub-sub and write them into a big query table. As kubernetes has no visual representation of tweets that are coming into the bigquery table. We have to write a select query to see the progres of data being collected. However, we can create a dataflow job with a template ‘pub-sub topic’ to ‘Big Query’. This will show us graphical representatikon of the data that is being written to the big query table. The fig (3) shows that currently there are 110,254 items have been added to the big qiery table which is of 504 MB.

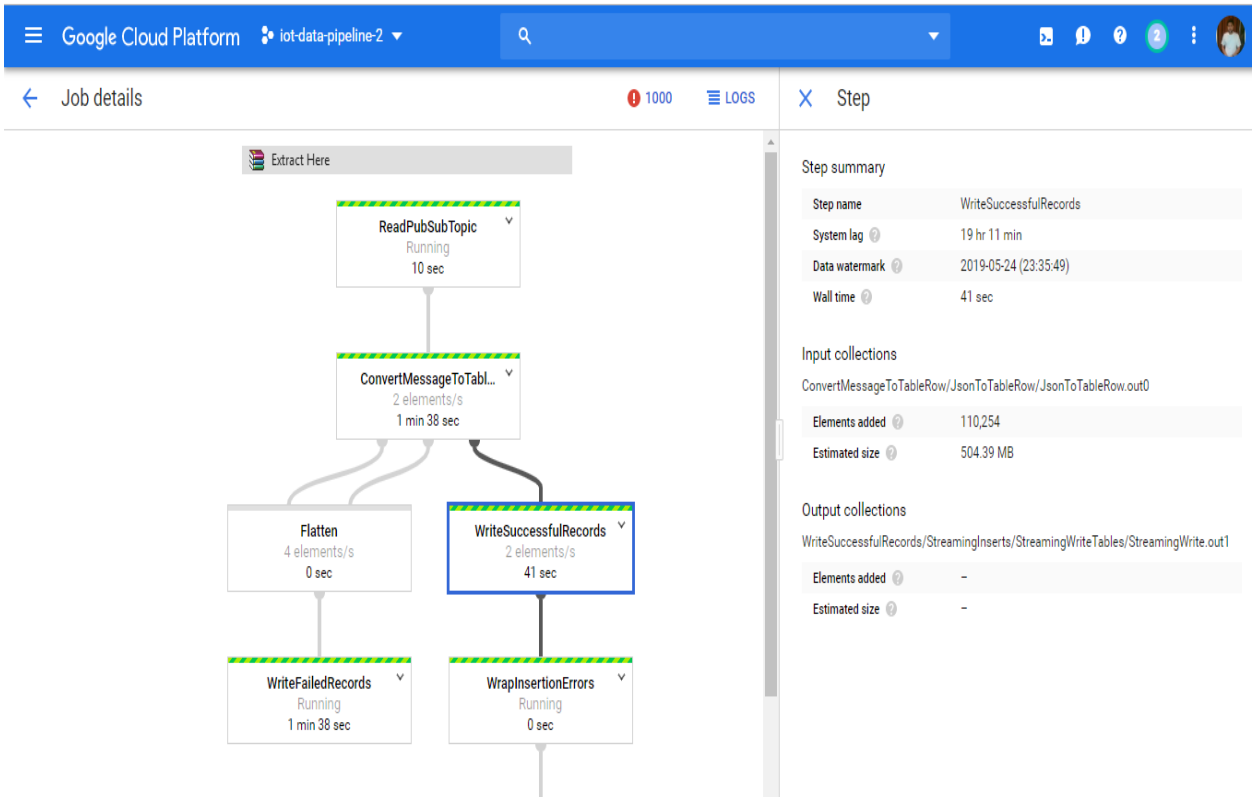


Figure 3. Pub-sub to big query

STEP 3/5 DATA PREPROCESSING IN BIG QUERY AND DATA PREP

STEP 3(A) DATA PRE-PROCESSING IN BIG QUERY

For this project we need the tweets that only contains @NarendraModi in them, hence a query is written to get the tweets that contain that same text in the ‘text ’column of gathered tweets. The (Fig 4) below shoes that the three queries that are written to clean and extract the desired data. The second query strips the words that has ‘@username’ and now we are left with the text and a URL. To get rid of the url a third query is written that will remove all the links from the tweets.

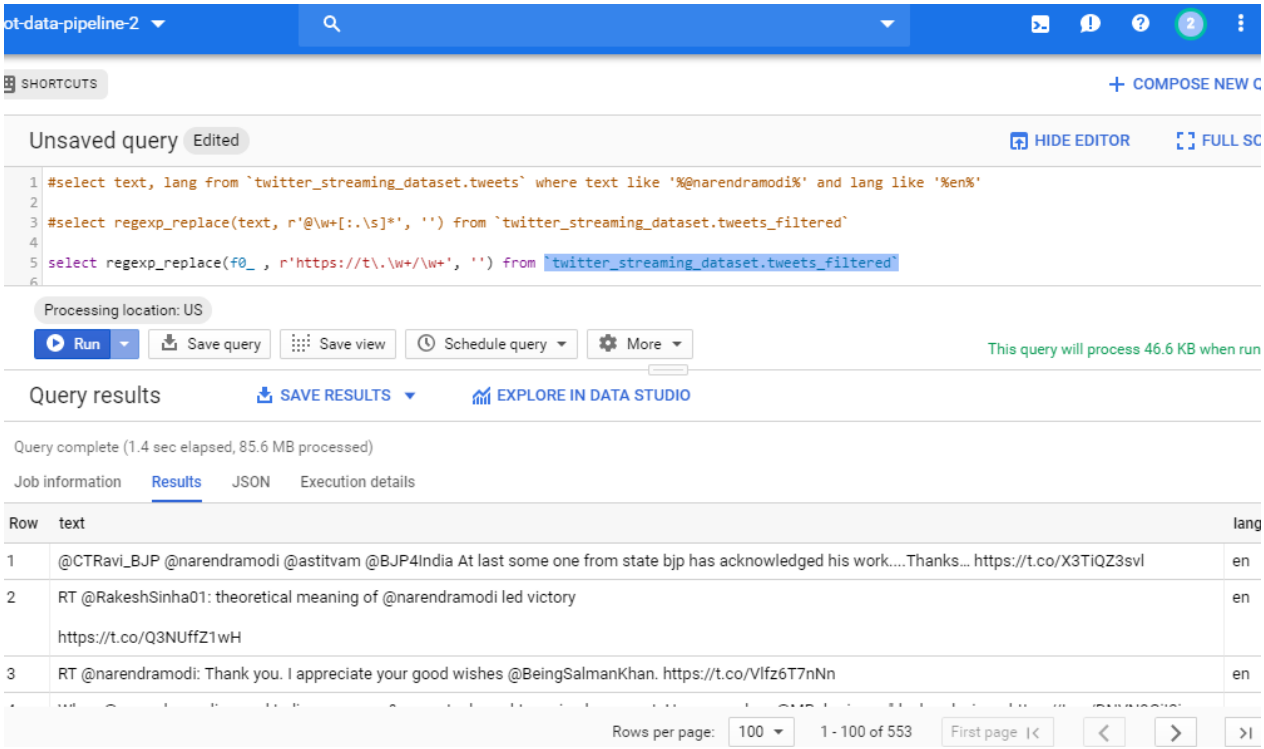


Figure 4. Data pre-processing in big query

STEP 3(B) DATA PREP CLEANING

Even though data cleaning can be totally done in big-query, Data prep is quite intutive and can be used to further enhance our data. Once the data is imported in data prep by Big-query, clicking on column name gives us the recommendations that we can use to clean the data. In data prep, duplicates, whitespace, symbols and ‘RT’ tags are removed which is shown in Fig (5).

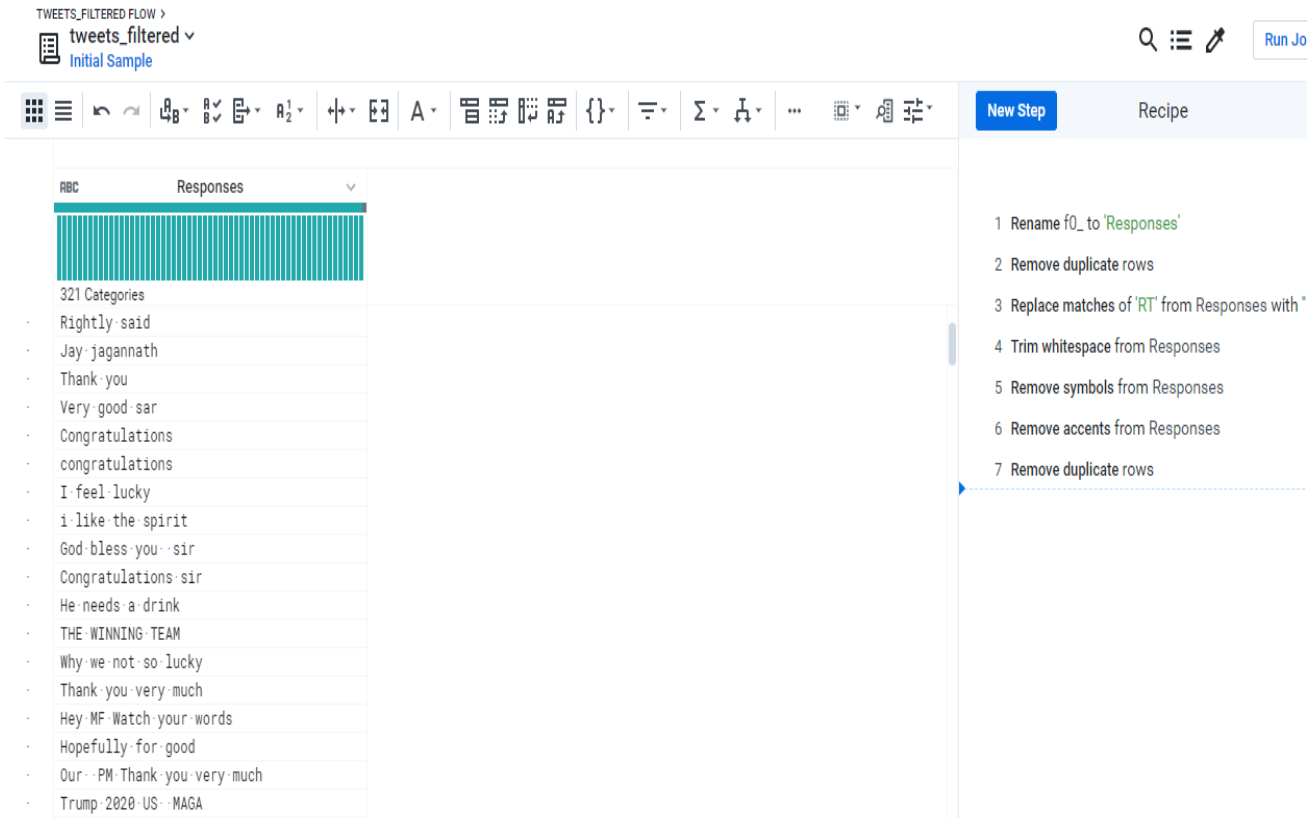


Figure 5. Data cleaning in data prep

STEP 4/5 SENTIMENT ANALYSIS IN AUTO ML

Google cloud offers the Auto ML sentiment analysis, which can be used to analyse sentiments from the text. Right now, it is in beta version and works good with english language and that is the reason we extracted the tweets that has language as ‘en’ in our first query in Fig (4). The model is trained with twitter data and it takes around 4 hours to train the model. After training is completed it comes with the text and their respective sentiment score. While initializing the sentiment analysis, we are asked to select maximum sentiment score and for this analysis, we have selected 4 as the maximum score.

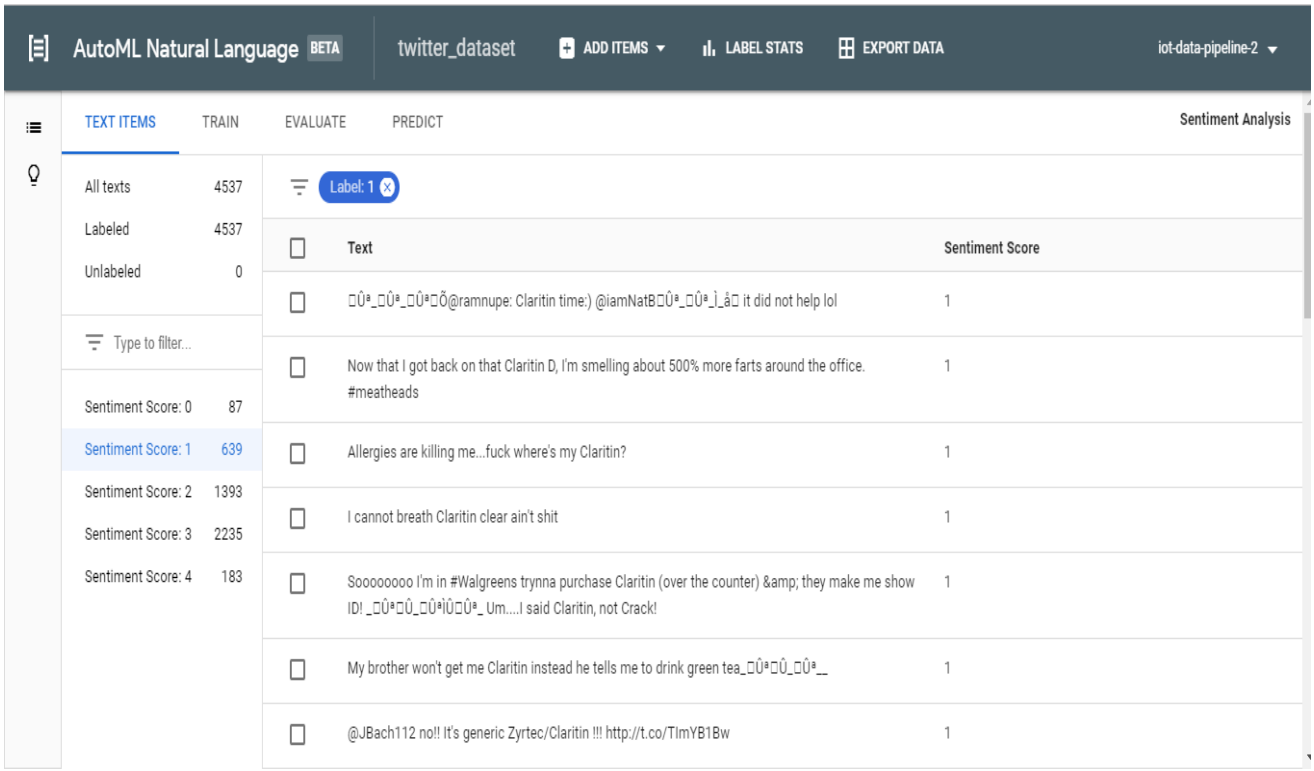


Figure 6. sentiment Analysis of twitter dat

a

RESULT OF ML AND VISULIZATION

ML RESULT

By clicking on the predict tab after training the model. A respective sentiment analysis is done on the data that we have in big query. In this the text is passed and its relative sentiment score is shown in the new column next to it. The result can be viewed in the big- query.

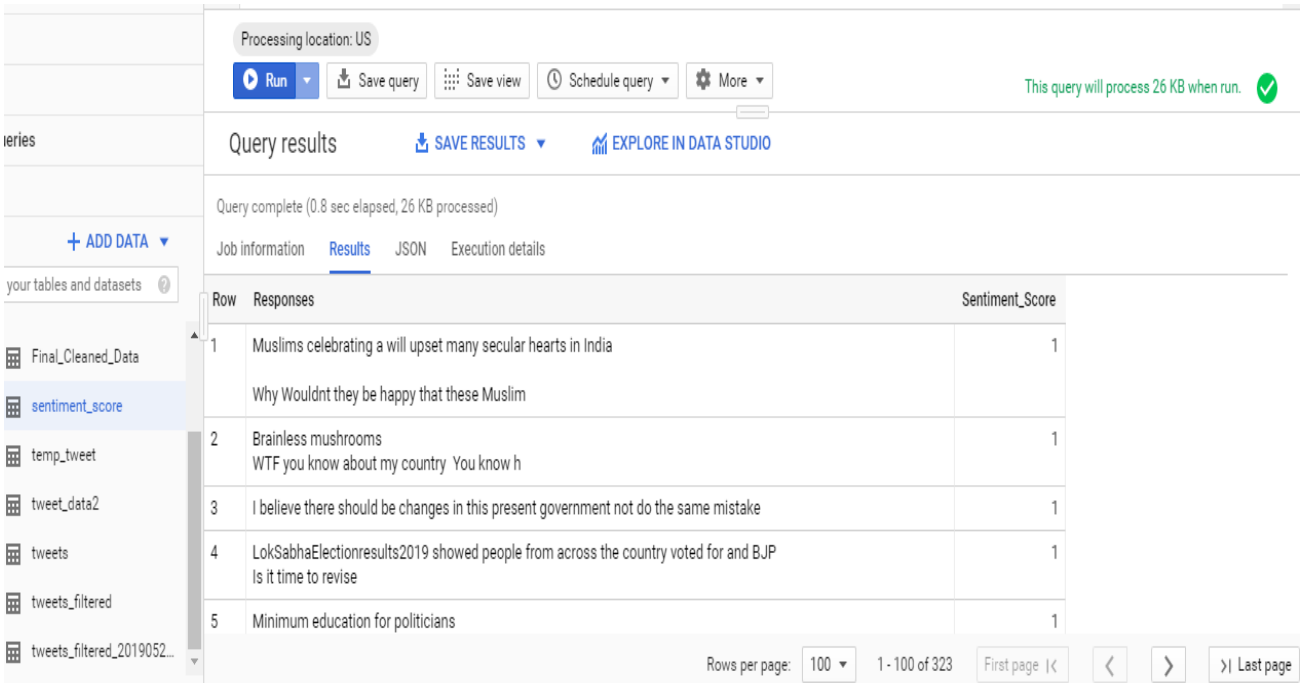


Figure 7. Sentiment Score

VIUSALIZATION

Based on the sentiment score of the tweets of the people, we can visualize the percent of people that have different sentiments regarding 'Narendra Modi' being the prime minister in four sections. 1 shows the negative emotion, 2 is being neutral, while 3 and 4 shows strong positive sentiment of people. A pie chart is drawn to get the better look of people's sentiment which is shown in below Fig (8).

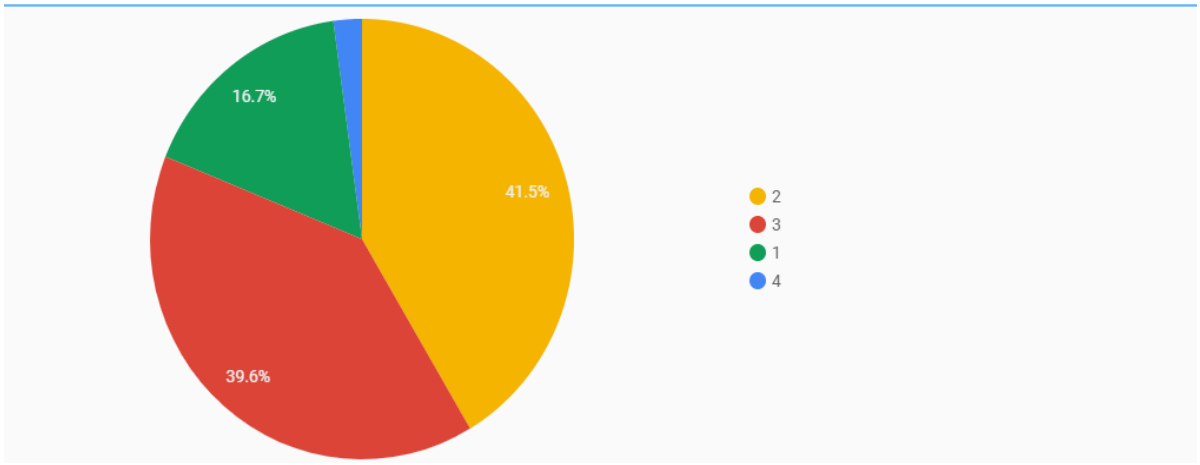
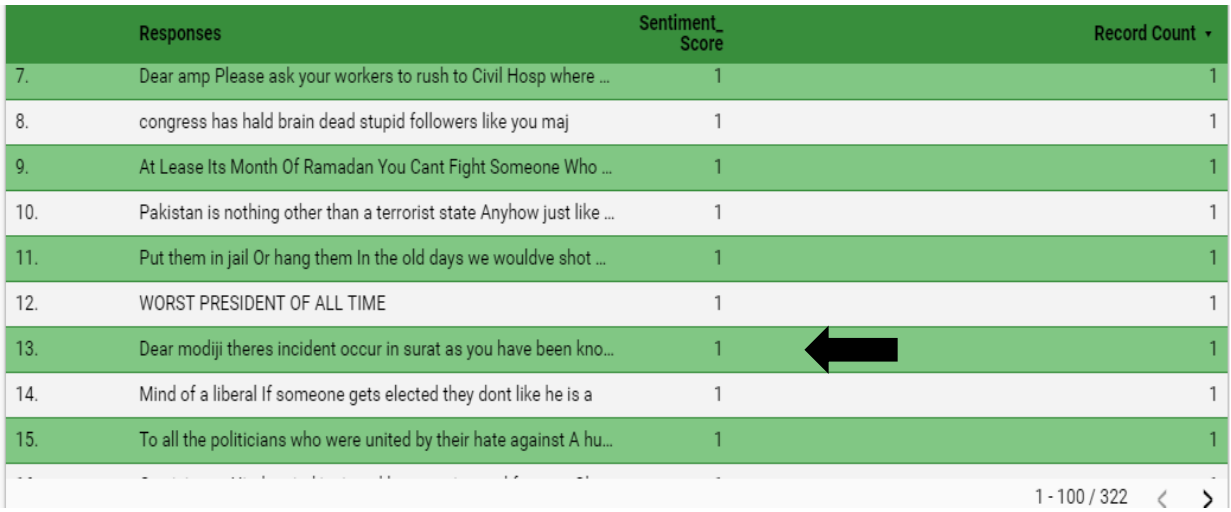


Figure 8. Sentiment percent of people

The visualization can further help us to evaluate the tweets of people that has lowest sentiment score. The visualization of the same is shown below in Fig (9). With the help of below visualization, we can see the tweet number (13), which I marked with a black arrow; this tweet specifically tells us that due to the recent fire in 'surat' just after the Prime Minsiter election, there is an unrest among the people. Hence, the committee that is backing up the Prime Minister can look for further steps that needs to be taken to appease the people and gain their trust again.



	Responses	Sentiment_ Score	Record Count
7.	Dear amp Please ask your workers to rush to Civil Hosp where ...	1	1
8.	congress has hald brain dead stupid followers like you maj	1	1
9.	At Lease Its Month Of Ramadan You Cant Fight Someone Who ...	1	1
10.	Pakistan is nothing other than a terrorist state Anyhow just like ...	1	1
11.	Put them in jail Or hang them In the old days we wouldve shot ...	1	1
12.	WORST PRESIDENT OF ALL TIME	1	1
13.	Dear modiji theres incident occur in surat as you have been kno...	1	1
14.	Mind of a liberal If someone gets elected they dont like he is a	1	1
15.	To all the politicians who were united by their hate against A hu...	1	1

1 - 100 / 322

Figure 9. Getting Tweets with low sentiment score

CONCLUSION

The big data cloud tehnologies are the best choice when it comes to put one's analytical techniques to practice with a minimum cost. The project that we have done consists of kubernetes, data flow, pub-sub, big query and ML; all of these services were present under an umbrella of google cloud. There are other providers of cloud technologies which provide all the services like google cloud, such as Amazone web services, IBM to name a few. The sentiment analysis was carried out the tweets of people containing the keyword 'Narendra modi'. The prime minister's advisor or campaing head can use this data to know the underlying cause of unrest among the people in Surat. Investigating it further shows us that there was indeed a fire incident in Surat where students were killed in horrific accident. Prime minster's committee can than organize an alleviate the situation with necessary actions. To sum up, big data cloud services not only minimizes the cost of implementing big data services but also provides greater scalability and distributed techniques.

REFERENCES

Grable, J. E., & Lyons, A. C. (2018). An Introduction to Big Data.: Discovery Service for Letterkenny Institute of Technology, 21. Retrieved from http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=3&sid=e7d66355-edfb-49d4-b7a1-bcf7bec4cf05@sessionmgr102

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015).

The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. https://doi.org/10.1016/J.IS.2014.07.006