# University of Maryland Baltimore County

*Independent Study on Big Data*

Project Title

# Meetup Streaming Data Analysis Using Apache Kafka and ELK Stack

Under the Guidance of

**Prof. Jianwu Wang**

*By*

Priyanka Patil | Srishti Patil | Sneha Rachkar

# Table of Contents

# 1. Introduction

The concept of real time analytics assumes ever increasing importance with more and more data coming from smart connected devices. The constant stream of data that is coming from these connected systems is unique in its volume, variety and velocity.

As humans we are continuously filtering and deciphering the information flowing towards us. Streaming data applications will perform impressive tasks in the same way, such as reading live location data to suggest nearby facilities, detecting machinery failures in real time, and submitting digital receipts before your customers leave the store. Recent advancements in streaming data technology and techniques allow any developer to create these apps if they have the right mindset.

There are two types of real-time analytics:

- On demand real time analytics: This is an approach that is reactive. To process a request and then deliver the analytics, it awaits a question from the end user. A web analyst, for example, tracks site traffic to prevent a possible website crash.
- Continuous real time analytics: This is an approach that is proactive. In real time, it alerts users with continuous notifications. Tracking the stock market, for example, with different representations of visualization on a website.

Real-time continuous analytics can address these problems:

- Capture the real-time and historical sensor data and analyze it further.
- Evaluate the patterns of normal and errant behavior.

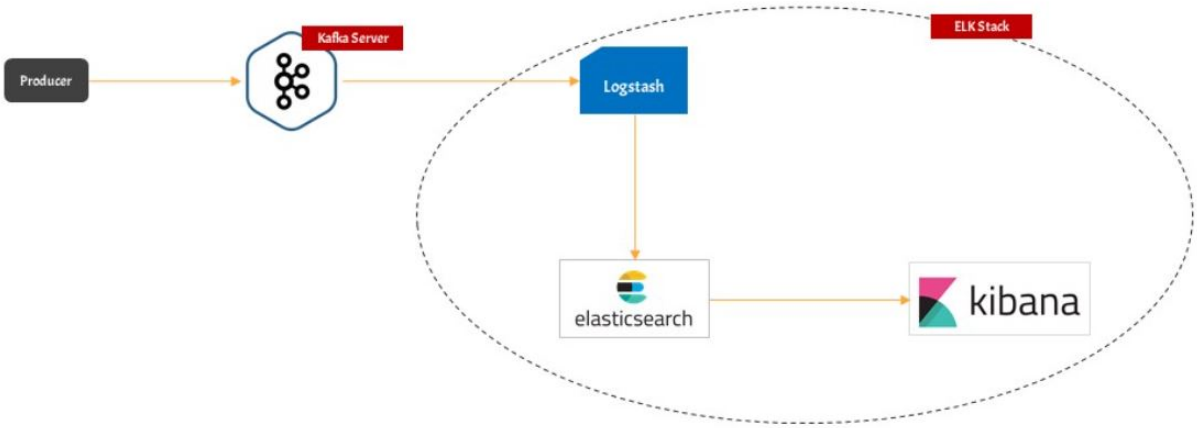
In our project we use ELK stack and Kafka where Kafka is able to handle thousands of messages per second and that too in low latency conditions with high throughput. In addition, it permits the reading and writing of messages into it at high concurrency. It is possible to integrate with the variety of consumers using Kafka

Our methodology also would allow us to produce several other business insights related to the RSVPs of Meetups. A thorough description of the real-time stream processing and the research carried out on it is provided in this report.

## 2. Problem/task

Meetup.com is the most well-known social networking website. It brings people from all the age groups together who share common passions and interests. Interested individuals from all around the world can create groups or events for a particular event or get together on Meetup.com. As of today, this website has more than 35 million meetup users, 225,000+ meetup groups in 180 countries. It is interesting to know how much real time data in the form of events is generated on a daily basis. In today's data world, it becomes important to collect, store and process such huge amounts of streaming data in order to make businesses successful. Companies need well established data pipelines or ETL flow in order to investigate this data for effective decision making. We were really interested in capturing this real time RSVP event data and to investigate the interesting facts about this website.

## 3. Background




This is the architecture we thought of implementing for the project. It consists of 4 main components.

a) Kafka server
b) Logstash
c) Elasticsearch
d) Kibana

Kafka producers written in Python publish data to the Kafka Server. Then comes the ELK Stack which consists of Elasticsearch, Logstash and Kibana. Kibana acts as a graphical user interface for Elasticsearch. Logstash plays an important role in the ELK stack. It consumes data from Kafka server and then it pushes it to Elasticsearch.

**A. Dataset:**

Apache Kafka is an open-source, event streaming platform used by thousands developed in Scala and Java. It is used for processing real-time data generated by thousands of data sources. It has three main functions which are (a) Publish and Subscribe to stream of records (b) Effectively store stream of records (c) process stream of records in real-time. It is also used in streaming data pipelines and applications that modify to the data streams.[6]

Kafka manager is basically a web-based management system for Kafka. All the topics, clusters created can be managed on one platform. It requires access to Zookeeper. So, we also need to install Zookeeper on the machine. Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. Mentioned below is a snapshot of RSVP data from the given Meetup.com link in the Kafka Producer.


{"venue":{"venue_name":" St Kilda Pier","lon":144.970978,"lat":-37.861736,"venue_id":23653470},"visibility":"public","response":"yes","guests":0,"member":
{"member_id":250039494,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/3\/9\/6\/3\/thumb_302714691.jpeg","member_name":"Saket Jain"},"rsvp_id":1858933056,"mtime":1608275157488,"event":
{"event_name":"Christmas Day on the Beach -St Kilda Beach Sunset [FREE Event]","event_id":"273474226","time":1608886800000,"event_url":"https:\/\/www.meetup.com\/exploremelb\/events\/273474226\/"},"group":
{"group_topics":[{"urlkey":"photo","topic_name":"Photography"},{"urlkey":"discover-your-neighborhood","topic_name":"Discover Your Neighborhood"},{"urlkey":"neighborhood","topic_name":"Neighborhood"},
{"urlkey":"outdoor-adventures","topic_name":"Outdoor Adventures"},{"urlkey":"adventure","topic_name":"Adventure"},{"urlkey":"outdoors","topic_name":"Outdoors"},{"urlkey":"local-activities","topic_name":"Local
Activities"},{"urlkey":"walkers","topic_name":"Walking"},{"urlkey":"newintown","topic_name":"New In Town"},{"urlkey":"travel","topic_name":"Travel"},{"urlkey":"socialnetwork","topic_name":"Social Networking"},
{"urlkey":"social","topic_name":"Social"},{"urlkey":"exploring","topic_name":"Exploring"},{"urlkey":"landscape-photography","topic_name":"Landscape Photography"},{"urlkey":"street-
photography","topic_name":"Street Photography"}],"group_city":"Melbourne","group_country":"au","group_id":20453358,"group_name":"Explore
Melbourne","group_lon":144.96,"group_urlname":"exploremelb","group_lat":-37.81}}
{"venue":{"venue_name":"Memorial Middle School","lon":-95.552528,"lat":29.777494,"venue_id":25529753},"visibility":"public","response":"yes","guests":0,"member":
{"member_id":283514572,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/5\/f\/5\/0\/thumb_289464400.jpeg","member_name":"Tommy Maoz"},"rsvp_id":1858933057,"mtime":1608275159439,"event":
{"event_name":"Memorial Middle School Soccer","event_id":"swfcnrybcqbfc","time":1608769800000,"event_url":"https:\/\/www.meetup.com\/Co-Ed-Adult-Soccer-Meetup-CypressKaty\/events\/275240633\/"},"group":
{"group_topics":[{"urlkey":"play-pickup-soccer","topic_name":"Play pickup Soccer"},{"urlkey":"soccer","topic_name":"Soccer"},{"urlkey":"pickupsoccer","topic_name":"Pick-Up Soccer"},{"urlkey":"coed-
soccer","topic_name":"Coed Soccer"},{"urlkey":"co-ed-soccer","topic_name":"Co-ed Soccer"},{"urlkey":"playing-soccer","topic_name":"Playing Soccer"},{"urlkey":"outdoor-soccer","topic_name":"Outdoor Soccer"},
{"urlkey":"futbol","topic_name":"Futbol"}],"group_city":"Cypress","group_country":"us","group_id":18443586,"group_name":"Co-Ed Adult Soccer Meetup (Cypress\/Katy)","group_lon":-95.67,"group_urlname":"Co-Ed-
Adult-Soccer-Meetup-CypressKaty","group_state":"TX","group_lat":29.99}}
{"venue":{"venue_name":"Online event","lon":179.1962,"lat":-8.521147,"venue_id":26906060},"visibility":"public","response":"yes","guests":0,"member":
{"member_id":300758611,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/a\/f\/7\/f\/thumb_296564927.jpeg","member_name":"Katrina Porter"},"rsvp_id":1858933058,"mtime":1608275160309,"event":
{"event_name":"Women Rise UP- Business Mastermind","event_id":"wbwvwrybcqbhc","time":1608321600000,"event_url":"https:\/\/www.meetup.com\/SLO-Women-Rise-Together\/events\/275107911\/"},"group":{"group_topics":
[{"urlkey":"smallbiz","topic_name":"Small Business"},{"urlkey":"women","topic_name":"Women's Social"},{"urlkey":"women-entrepreneurs","topic_name":"Women Entrepreneurs"},{"urlkey":"professional-
networking","topic_name":"Professional Networking"},{"urlkey":"business-strategy","topic_name":"Business Strategy"},{"urlkey":"entrepreneurship","topic_name":"Entrepreneurship"},{"urlkey":"startup-
businesses","topic_name":"Startup Businesses"},{"urlkey":"professional-women","topic_name":"Professional Women"}],"group_city":"San Luis Obispo","group_country":"us","group_id":31088850,"group_name":"SLO Women
Rise Together","group_lon":-120.65,"group_urlname":"SLO-Women-Rise-Together","group_state":"CA","group_lat":35.26}}
{"visibility":"public","response":"yes","guests":0,"member":{"member_id":299046152,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/7\/6\/d\/8\/thumb_302190424.jpeg","member_name":"Okba
Aljarroud"},"rsvp_id":1858933061,"mtime":1608275163318,"event":{"event_name":"Cars & Coffee","event_id":"275127246","time":1608264000000,"event_url":"https:\/\/www.meetup.com\/Cars-Coffee-Riyadh-Saudi-
Arabia\/events\/275127246\/"},"group":{"group_topics":[{"urlkey":"sports-cars","topic_name":"Sports Cars"},{"urlkey":"automotive","topic_name":"Automotive"},{"urlkey":"muscle-cars","topic_name":"muscle cars"},
{"urlkey":"classic-cars","topic_name":"classic cars"},{"urlkey":"cars","topic_name":"Cars"},{"urlkey":"exotic-cars","topic_name":"Exotic Cars"},{"urlkey":"cars-and-coffee","topic_name":"Cars and Coffee"},
{"urlkey":"cars-of-all-ages-and-types","topic_name":"Cars of All Ages and Types"},{"urlkey":"collector-cars","topic_name":"Collector Cars"},{"urlkey":"photographing-cars","topic_name":"Photographing Cars"},
{"urlkey":"antique-cars","topic_name":"Antique Cars"},{"urlkey":"cars-coffee","topic_name":"cars & coffee"},{"urlkey":"american-muscle-cars","topic_name":"american muscle cars"},{"urlkey":"cars-
motorcycles","topic_name":"cars & motorcycles"},{"urlkey":"cars-and-cappuccinos","topic_name":"Cars and Cappuccinos"}],"group_city":"Riyadh","group_country":"sa","group_id":32778500,"group_name":"Cars & Coffee
- Riyadh, Saudi Arabia","group_lon":46.77,"group_urlname":"Cars-Coffee-Riyadh-Saudi-Arabia","group_lat":24.65}}
{"venue":{"venue_name":"Memorial Middle School","lon":-95.552528,"lat":29.777494,"venue_id":25529753},"visibility":"public","response":"yes","guests":0,"member":
{"member_id":283514572,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/5\/f\/5\/0\/thumb_289464400.jpeg","member_name":"Tommy Maoz"},"rsvp_id":1858933057,"mtime":1608275159000,"event":
{"event_name":"Memorial Middle School Soccer","event_id":"swfcnrybcqbfc","time":1608769800000,"event_url":"https:\/\/www.meetup.com\/Co-Ed-Adult-Soccer-Meetup-CypressKaty\/events\/275240633\/"},"group":
{"group_topics":[{"urlkey":"play-pickup-soccer","topic_name":"Play pickup Soccer"},{"urlkey":"soccer","topic_name":"Soccer"},{"urlkey":"pickupsoccer","topic_name":"Pick-Up Soccer"},{"urlkey":"coed-
soccer","topic_name":"Coed Soccer"},{"urlkey":"co-ed-soccer","topic_name":"Co-ed Soccer"},{"urlkey":"playing-soccer","topic_name":"Playing Soccer"},{"urlkey":"outdoor-soccer","topic_name":"Outdoor Soccer"},
{"urlkey":"futbol","topic_name":"Futbol"}],"group_city":"Cypress","group_country":"us","group_id":18443586,"group_name":"Co-Ed Adult Soccer Meetup (Cypress\/Katy)","group_lon":-95.67,"group_urlname":"Co-Ed-
Adult-Soccer-Meetup-CypressKaty","group_state":"TX","group_lat":29.99}}
{"venue":{"venue_name":"Pasta e Formaggio @ Oxley Tower","lon":103.84842,"lat":1.278646,"venue_id":27023930},"visibility":"public","response":"yes","guests":0,"member":
{"member_id":314602309,"photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/8\/f\/5\/c\/thumb_300936700.jpeg","member_name":"Mark"},"rsvp_id":1858933062,"mtime":1608275167335,"event":
{"event_name":"Cashflow 101 - This Game can CHANGE your LIFE","event_id":"zlwccsybcqbzb","time":1608361200000,"event_url":"https:\/\/www.meetup.com\/CashFlow-101-the-Life-Changing-
Game\/events\/275138725\/"},"group":{"group_topics":[{"urlkey":"cash-flow-game-night","topic_name":"Cash Flow Game Night"},{"urlkey":"robert-kiyosaki","topic_name":"Robert Kiyosaki"},{"urlkey":"financial-
education","topic_name":"Financial Education"},{"urlkey":"rich-dad-cashflow-club","topic_name":"Rich Dad Cashflow Club"},{"urlkey":"cashflow-101","topic_name":"cashflow 101"},{"urlkey":"cash-flow-
game","topic_name":"Cash Flow Game"},{"urlkey":"cashflow-game-night","topic_name":"Cashflow Game Night"},{"urlkey":"cashflow-101-and-102","topic_name":"Cashflow 101 and 102"},{"urlkey":"rich-dad-poor-
dad","topic_name":"Rich Dad, Poor Dad"},{"urlkey":"fun-times","topic_name":"Fun Times"},{"urlkey":"cashflow-quadrant","topic_name":"Cashflow Quadrant"},{"urlkey":"strategy-games","topic_name":"Strategy Games"},
{"urlkey":"board-games","topic_name":"Board Games"},{"urlkey":"cashflow-game","topic_name":"Cashflow Game"},{"urlkey":"young-entrepreneur","topic_name":"Young
Entrepreneur"}],"group_city":"Singapore","group_country":"sg","group_id":20236532,"group_name":"CashFlow 101, the Life Changing Game (Powered by SBN Group)","group_lon":103.85,"group_urlname":"CashFlow-101-the-
Life-Changing-Game","group_lat":1.3}}
{"venue":{"venue_name":"Online event","lon":179.1962,"lat":-8.521147,"venue_id":26906060},"visibility":"public","response":"yes","guests":0,"member":

**B. Software techniques**

We used ELK Stack for analyzing streaming data which contains Elasticsearch as an Analytics Engine, Logstash as server-side data processing pipeline and Kibana for data visualization.

1) Elasticsearch

Elasticsearch is an open-source analytics and full-text search engine. For example, you might have a blog or webshop for which you want users to be able to search for various kinds of data. That could be blog posts, products, categories, or anything you want. You can build complex search functionality with Elasticsearch, like what you see on Google, for instance.[7]

2) Logstash

Logstash is an open-source, server-side data processing pipeline that allows you to collect data from a variety of sources, transform it on the fly, and send it to your desired destination. It has many benefits like very easy to load unstructured data, there are pre-built filters, it contains flexible plugin architecture.[10]

3) Kibana

Kibana is an open-source platform for data visualization and discovery used for analytics of log and time series, application tracking, and use cases of operational intelligence. It provides powerful features such as histograms, line graphs, pie charts, heat maps, and built-in geospatial support that are easy to use. It also offers close integration with the popular analytics and search engine Elasticsearch, which makes Kibana the default option for visualizing data stored in Elasticsearch.[11]

# 4. Solution and experiments

Explaining your code. Check in code to github and post the link in the report

Steps for installing Python:

1. Download Python version 3.7.9 from https://www.python.org/downloads/release/python-370/ on a local machine.

Steps for installing Kafka:

1. Download Kafka version kafka_2.13-2.6.0.tar from https://kafka.apache.org/downloads
2. Extract the contents of the downloaded file on a local machine.

Code Snippets:

Kafka_Producer.py

```
# coding: utf-8
from kafka import KafkaProducer
import json,requests

# Creating Kafka Producer
producer = KafkaProducer(bootstrap_servers=['localhost:9092'],api_version=(0,10,1))
# Sending a test message from Kafka Producer with topic name 'meetup'
producer.send('meetup',b'Hello, Kafka right now')

# Getting HTTP rsvp data requests from meetup.com
r = requests.get("https://stream.meetup.com/2/rsvps",stream=True)

# Looping through the stream of data and sending this data to consumer
for each_record in r.iter_lines():
# Send each record in bytes. Kafka reads in bytes
   producer.send('meetup',each_record)

# Closing Kafka Producer
producer.close()
```

By default, the Python API will decode Kafka data as UTF8 encoded strings.

Steps for installing Docker and Docker Compose:

1. Navigate to https://docs.docker.com/get-docker/ and download docker desktop and docker compose on a local machine.
2. Install docker and open it by navigating to 'Applications' → 'Docker' or execute 'open -a Docker' command in Terminal.

Steps for installing ElasticSearch and Kibana using Docker Compose:

- Create a .yml file by executing the 'touch elasticsearch-docker-compose.yml' command in the terminal.
- Open this file, copy the below contents and save the file by executing command 'vim elasticsearch-docker-compose.yml'.

```
version: "3.3"
services:
 elasticsearch:
   image: docker.elastic.co/elasticsearch/elasticsearch:7.4.0
   container_name: elasticsearch
   restart: always
   environment:
     - xpack.security.enabled=false
     - discovery.type=single-node
   ulimits:
     memlock:
       soft: -1
       hard: -1
     nofile:
       soft: 65536
       hard: 65536
   cap_add:
     - IPC_LOCK
   volumes:
     - elasticsearch-data:/usr/share/elasticsearch/data
   ports:
     - 9200:9200
 kibana:
   container_name: kibana
   image: docker.elastic.co/kibana/kibana:7.4.0
   restart: always
   environment:
     - ELASTICSEARCH_HOSTS=http://elasticsearch:9200    # address of elasticsearch
docker container which kibana will connect
   ports:
     - 5601:5601
   depends_on:
     - elasticsearch                                  # kibana will start when
elasticsearch has started
volumes:
 elasticsearch-data:
```

This file creates services for starting elasticsearch and kibana with appropriate images and containers. ElasticSearch will be running on Port: 9200 and Kibana will be running on Port: 5601.

Steps for starting zookeeper, kafka and kafka manager using Docker Compose:

1. Create a .yml file by executing the ‘touch kafka.yml’ command in the terminal.
2. Open this file, copy the below contents and save the file by executing command ‘vim kafka.yml’.

```
version: "3"
services:
# Create zookeeper service on Port: 2181
 zookeeper:
   image: zookeeper
   restart: always
   container_name: zookeeper
   hostname: zookeeper
   ports:
     - 2181:2181
   environment:
     ZOO_MY_ID: 1
 # Create Kafka service on Port: 9092
 kafka:
   image: wurstmeister/kafka
   container_name: kafka
   ports:
   - 9092:9092
   environment:
   # Local host address
     KAFKA_ADVERTISED_HOST_NAME: 192.168.0.26
     KAFKA_ZOOKEEPER_CONNECT: zookeeper:2181
   # Create kafka Manager on Port: 9000
 kafka_manager:
   image: hlebalbau/kafka-manager:stable
   container_name: kakfa-manager
   restart: always
   ports:
     - "9000:9000"
   environment:
     ZK_HOSTS: "zookeeper:2181"
     APPLICATION_SECRET: "random-secret"
   command: -Dpidfile.path=/dev/null
```

Steps for installing Logstash:

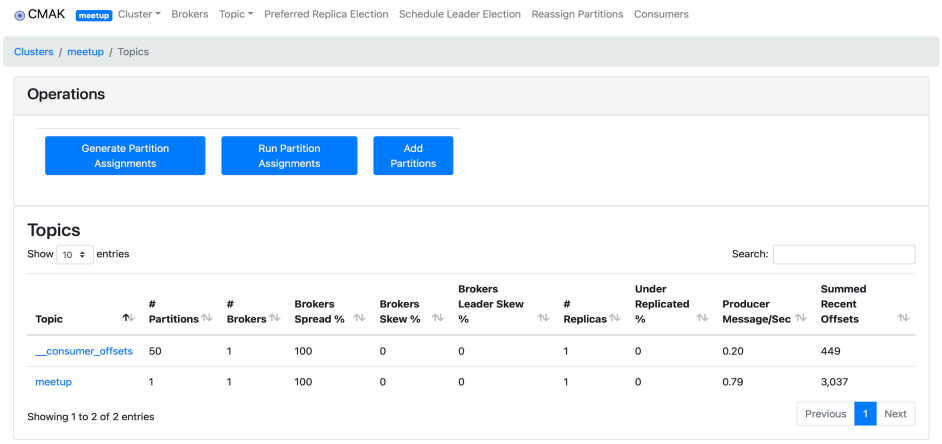
1. Download Logstash from https://www.elastic.co/guide/en/logstash/current/installing-logstash.html and install it on a local machine.
2. Create a .conf file in the logstash/config folder of your machine and copy below contents in the file.

```
        input {
  kafka {
      bootstrap_servers => "192.168.0.26:9092"
      topics => ["meetup"]
  }
}
filter {

  json {
    source => "message"
  }
}

output {
  elasticsearch {
    hosts => ["192.168.0.26:9200"]
    index => "meetup"
    workers => 1
  }
}
```

In this .conf file, logstash takes data as input from the kafka manager running on port 9092 having the topic name 'meetup'. Logstash is connected to the elasticsearch running on port 9200 which is mentioned as the output. Also, data in the message field is filtered in json format to be displayed in Elasticsearch.

Steps to consume meetup.com real time streaming data for analysis:

- Go to Terminal and run command 'docker-compose -f elasticsearch-docker-compose.yml up' for starting elasticsearch and kibana.
- Run command 'docker-compose -f kafka.yml up' for starting zookeeper, kafka and kafka manager on a new terminal.
- On the browser, navigate to http://<IP_ADDRESS>:9000
- Create a cluster with the topic name 'meetup'.




1. Run kafka_producer.py file in any python environment. We used Visual Studio Code Version: 1.51.1
2. Go to new terminal and execute command ‘logstash -f <path to .conf file>’
3. On the browser, navigate to http://<IP_ADDRESS>:5601

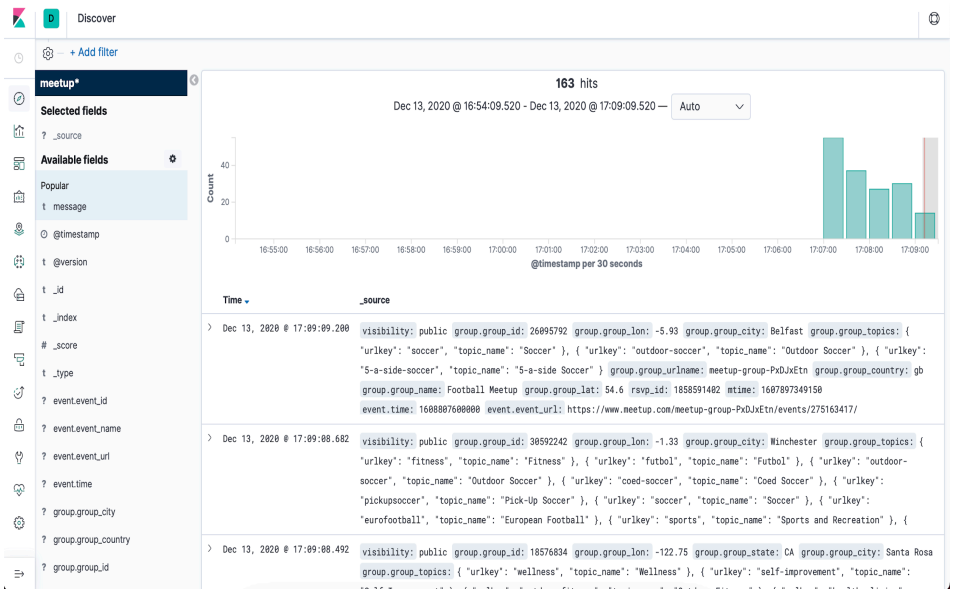



Fig1.: Elasticsearch Dashboard displaying real time Meetup.com events

Sample visualizations:

The below visualization shows the top 5 city counts from which event RSVPs are coming. This graph is set to refresh every 5 seconds.

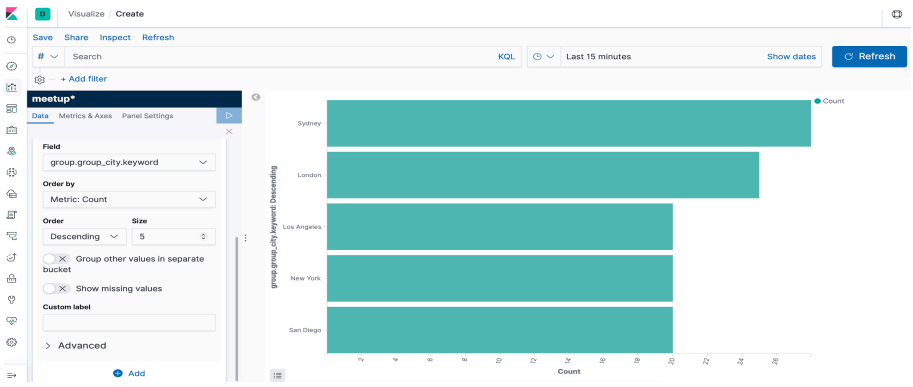



Fig2.: City counts

The below visualization shows a timeline of USA and France events.

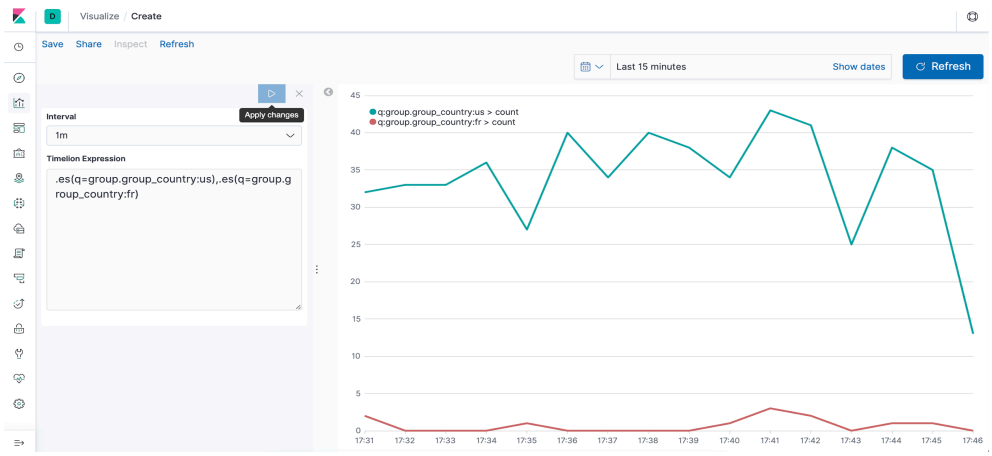



Fig3.: USA events vs France events

# 5. Journey summary:

By implementing this project, we got an idea of a real time streaming analytics platform and the challenges associated with it. For developing the data pipeline, we had to study a complete set of unfamiliar technologies like kafka and each element in ELK Stack. Online courses and YouTube videos helped us tremendously to successfully make this project work. It was challenging to match software compatibility and version for all these software's. Before processing the real-time data streams, we made sure that each and every component in the pipeline was up and running on the mentioned port numbers. Also, it was difficult to coordinate virtually with the group members for resolving obstacles that we faced.

# 6. Conclusion:

The objective of this project was to help companies capture, store and analyze real time data streams from the website. We believe that it is not easy to face the technical problems in real-time analytics, but the information generated through implementing the pipeline can prove effective for businesses. In this project, we successfully visualized the most active countries on meetup.com rsvp event streams and compared events between countries by generating a timeline chart in Kibana. Kibana can be further used to visualize a lot of business questions by using KQL, filters and charts available. In the cities where more meetings are held, we hope to help small businesses run more productive events and improve their exposure to more consumers through our analysis. Our analysis will be mainly helpful for e-commerce websites where companies can track, analyze, understand the most trending issues and create more efficient networking event relationships by processing live streaming data. For future work, such pipelines can be created on AWS using Amazon Kinesis and other cloud services for more efficient and faster stream processing.

# 7. Reference


1. *Data Flair* . (n.d.). Retrieved from Data Flair, INSTRUCTOR LED TRAINING WITH 24X7 SUPPORT: https://data-flair.training/
2. K. Aziz, D. Z. (2018). Real Time data Analysis using spark and hadoop . *4th International Conference on Optimization and Applications (ICOA)* (pp. 1-6). Mohammedia.
3. Psaltis, A. G. (May 2017 ). *Streaming Data Understanding the real-time pipeline.*
4. Mao, X. L. (2015). Real-Time data ETL framework for big real-time data analysis. *2015 IEEE International Conference on Information and Automation* (pp. 1289-1294). Lijiang.
5. *Apache Zookeper*. (2010-2020). Retrieved from https://zookeeper.apache.org/
6. *AWS*. (n.d.). Retrieved from https://aws.amazon.com/msk/what-is-kafka/
7. *Elasticsearch* . (2018). Retrieved from From Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Elasticsearch
8. *Kibana*. (2020). Retrieved from From Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Kibana
9. Logstash https://aws.amazon.com/elasticsearch-service/the-elk-stack/logstash/
10. Kibana (2020) https://aws.amazon.com/elasticsearch-service/the-elk-stack/kibana/