Cloud Computing - Major Project
Deliverable - 2

Real Time Event Monitoring System

(Using Spark)

Project ID - 14

Team Number 6

Anirudh Alampally - 201202173
Samyukta Dontireddy - 201201191
Merghoob Khan - 201405654
Nikita Gupta - 201405527

Abstract:

Event monitoring system is the process of collecting, analyzing, and signalling event occurrences to subscribers such as operating system processes, active database rules as well as human operators. These event occurrences may stem from arbitrary sources such in both software or hardware such as operating system, database management systems, application softwares and processors.

For example, every day, around 400 million tweets are sent worldwide, which has become a rich source for detecting, monitoring and analysing news stories and special (disaster) events. Since this data is very huge, analysing the entire data in a batch is almost impossible. To solve this problem, the data is to be analysed dynamically as it grows. We try to follow key words attributed to an event, monitoring temporal changes in word usage.

Proposed System gives implementation of Real time event processing on dataset from social Network and provides trending topics in real time using Machine Learning Libraries of spark a.k.a MLLIB. A further enhancement can be provided as dynamic topic visualization currently trending if time permits.

Approach:

We would like to follow the following approach to monitor the events in real time:

- Get the Live social Network stream using social Network streaming API.
- Give the live stream as an input to the spark job.
- Using spark, we analyse the keyword trends to detect events and monitor them.
- If we get a spike in any keyword count in word count vs timeline graph, we try to dig in into the text and mine the useful information in it.
- To analyse the text, we can use cosine similarity as one of the criteria to find the most relevant text regarding that particular hashtag.
- The tweet with highest pairwise similarity is the one which contains highest information regarding the event. We can present this to the user as a hot topic.
- We can limit our domain to a particular topic if we want to detect natural disasters like earthquakes, floods, hurricanes etc.
- For example, if we limit our social network stream to earthquakes and obtain tweets which contain earthquake only, a sudden raise in count of those tweets imply an event which has occurred in that time. In the same way, various events can be monitored.

Libraries:

Few libraries which might be of use are:

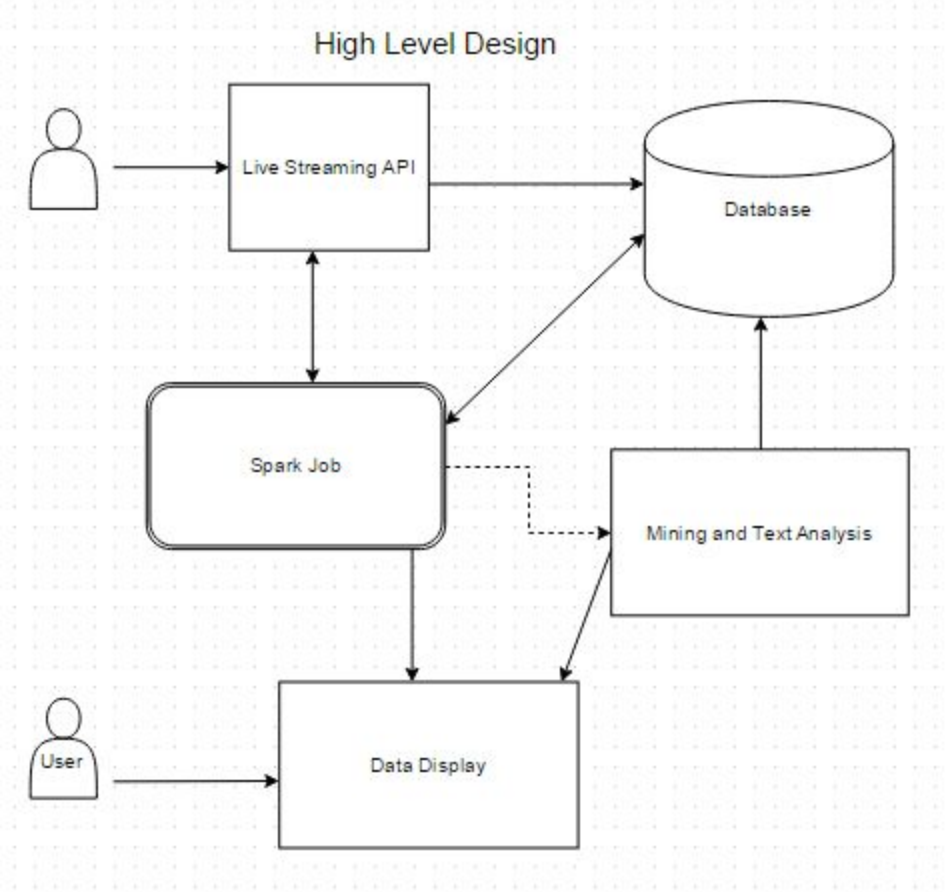
- Apache Spark.
- Python streaming API for twitter.
- MLLIB - Machine Learning library in spark.
- GRAPHX - Graph parallel computation in spark.
- word2vec - For finding similarity between two words.

Project Plan:

Some of the assumptions and decisions which we have considered as part this project are:

- Event monitoring is basically capturing the changes in behaviour of the system.
- Event monitoring can be of different types:
  - Monitoring social streams,
  - Monitoring log files to capture new changes and
  - Monitoring data obtained from various devices which capture weather, temperature etc.
- We have narrowed down the scope of the project to monitoring social streams.
- Monitoring the social streams to detect events in the real time is of high importance.

High Level Design:



Requirements:

Following are the few requirements for real time event monitoring:

- A server where a spark job can be run.
- A continuous social stream to monitor the events.
- A database to store the data to process and tweets.

Following are the few challenges which we might face:

- **Challenges:** The use of social media message streams for event detection poses a number of opportunities and challenges as these streams are: very high in volume, often contain duplicated, incomplete, imprecise and incorrect information, are written in informal style (i.e. short, unedited and conversational), generally concern the short-term zeitgeist; and finally relate

to unbounded domains.These characteristics mean that while massive and timely information sources are available, domain-relevant information may be mentioned very infrequently. Users post tweets usually from mobile devices which increases the possibilities of typing errors.

- The challenge is therefore the detection of the signal within that noise. This challenge is exacerbated by the typical requirement that documents must be processed in soft real time, such that events can be promptly acted upon.

- e.g., Consider the situation of a football match, user will post a tweet as "**GOOAALLL**" rather than a proper term "**GOAL**". Finding similarity between words becomes a challenge in these cases.

Expected outcome:

The outcome will be the information retrieved by processing twitter data in real time, on a specific domain (e.g. monitoring twitter data for a recent event).This will provide live updates for faster response times.
We are using Twitter as social data to build a working model and provide real-time responses for an event.

Uses:

We can notice application of real time event monitoring in various social networking sites like Twitter, Facebook etc. It can also in applied in monitoring various natural disasters like earthquakes, tornadoes etc. Real time event monitoring can also be used to automate the Parking system.It also helps in identifying patterns in streaming event data that are significant to business operations e.g. we can get the review of movies now in real time, that may affect their collection.