

# Real Time Event Monitoring System Using Spark

## Group:6

Anirudh Alampally

201202173

Nikita Gupta

201405527

Merghoob Khan

201405654

Samyukta Dontireddy

201201191

# OUTLINE

- **Introduction**
- **Approach**
- **Evaluation and Results**
- **Conclusion**
- **References**

# OUTLINE

- **Introduction**
- Approach
- Evaluation and Results
- Conclusion
- References

# Introduction

- **Event monitoring system is the process of collecting, analyzing, and signaling event occurrences to subscribers such as operating system processes, active database rules as well as human operators.**
- **In present scenario world is facing large huge data uploaded per minute of thousand of gigabytes data at websites, so the problem is to analyse the old data.**
- **We have to do it on the fly!**

# OUTLINE

- Introduction
- **Approach**
- Evaluation and Results
- Conclusion
- References

# Approach

- Tools Used: Apache Spark
  - To run the spark job on python.
  - The advantages of using spark is that it is lightning fast. Hadoop comes under generation 2 distributed computing where are spark comes under generation 3 distributed computing.
  - It also supports a rich set of higher-level tools including Spark Streaming for stream processing.

# APACHE SPARK

- **Apache Spark is an open source processing engine built around speed, ease of use, and analytics**
- **If you have large amounts of data that requires low latency processing that a typical MapReduce program cannot provide, Spark is the alternative**
- **Spark performs at speeds up to 100 times faster than Map Reduce for iterative algorithms or interactive data mining**

# Approach

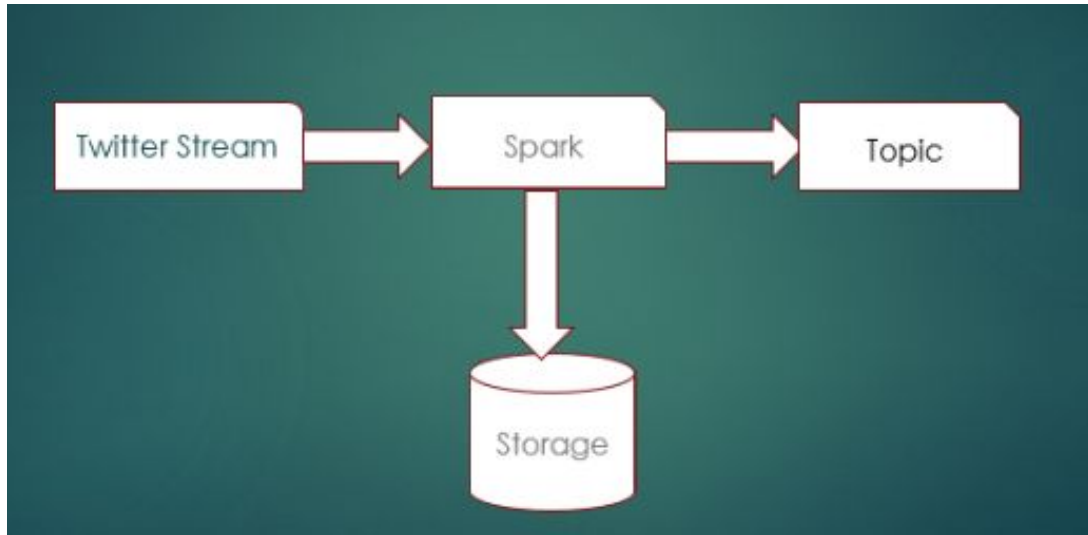
- Tools Used: **nltk**
  - It is a python package which is used to process human language data.
  - We use it find the similarity between two words.
  - The purpose of nltk is explained in the further sections.



# Algorithm

1. Twitter data is considered as streaming data. Million of tweets are everyday.
2. These tweets can be used to do real time event monitoring.
3. We obtain the tweets using twitter streaming API.
4. The input is taken in the form of Dstreams.
5. On these Dstreams, we try to apply map and reduce functions on spark by which we get the occurrences of a particular hashtag.
6. When we get a peak in the timeline vs tweet count graph, we can consider that as an event.

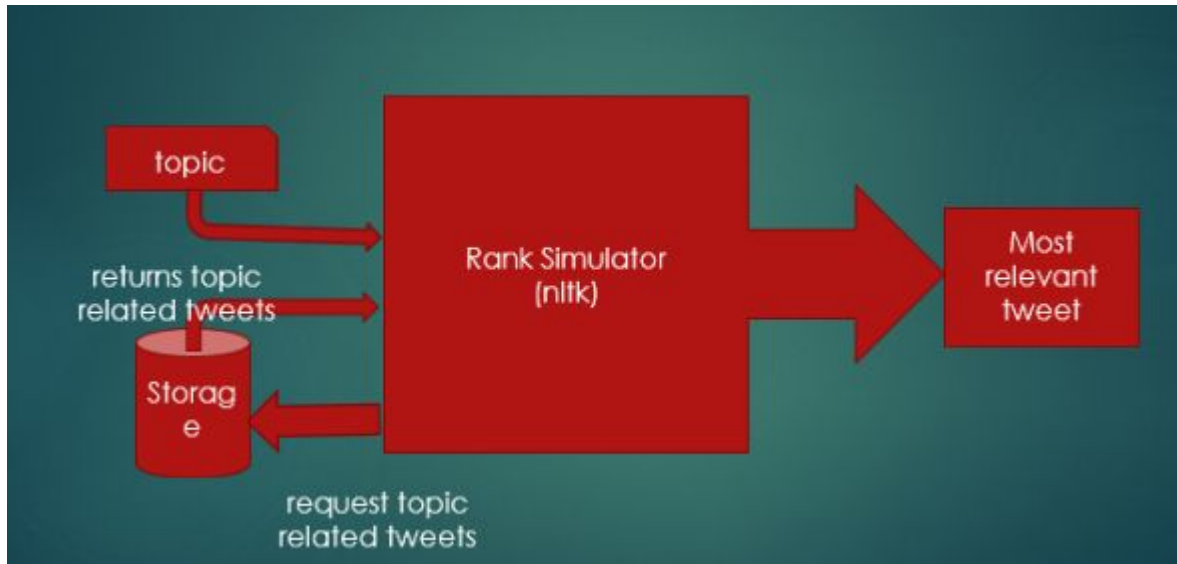
# Algorithm



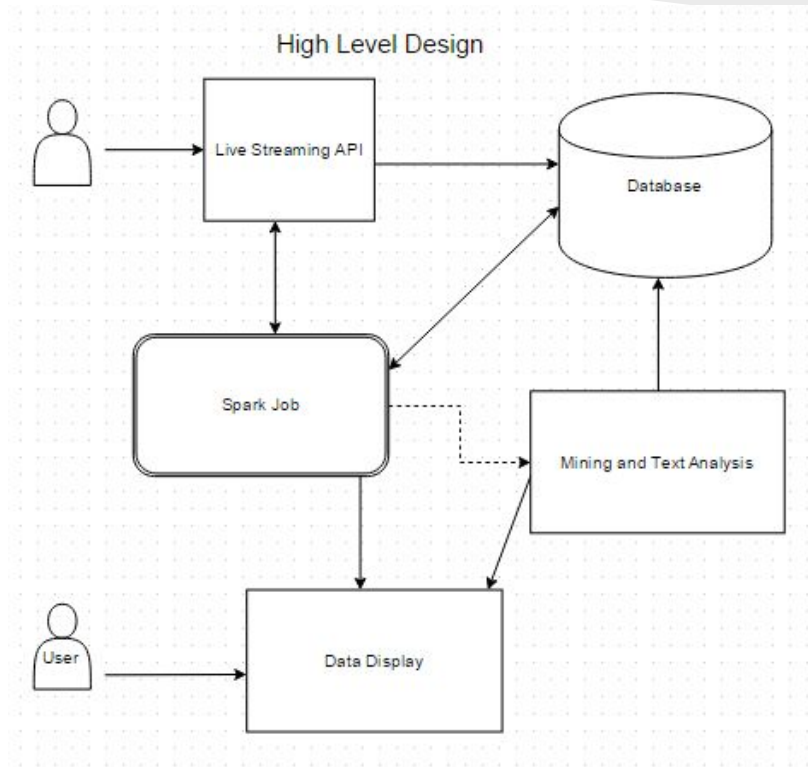
# Algorithm

1. In the next step, we do sentiment Analysis on these tweets based on hot hashtag.
2. Consider all tweets related to that hashtag and compute **tweetRank**.
3. **tweetRank**: It is defined as the the tweet with highest pairwise cosine similarity with all other tweets.
4. It is proved in many existing methods(*research paper link provided in references*)that the text with highest pairwise similarity has the highest content in it.
5. Tweet with highest rank will be displayed to the user.

# Algorithm



# High Level Design



# OUTLINE

- Introduction
- Approach
- **Evaluation and Results**
- Conclusion
- References

# Evaluation and Results

1. A live twitter stream is given as a input.
2. This code is ran at different time periods and following are the events which are detected at that time span as hot topics:
  - a. **#earthquake** A 5.9 magnitude **#earthquake** struck Afghanistan on Sunday, the US Geological Survey reported.

As we can see, various natural calamities can be monitored by using this tool. In the above example, it detected an earthquake in Afghanistan. We collected the data at 1:00AM on 22nd of November and earthquake actually occurred around 12:30AM on 22nd November.

## Results..

- #BBC #World #BBC Djokovic beats Federer to ATP title: Novak Djokovic wins a record fourth straight ATP World Tour Final

As an other example, the code also detected the trending event that Djokovic beat Federer in the ATP finals, which happened the night when we ran the code.



# OUTLINE

- Introduction
- Related Work
- Approach
- Evaluation and Results
- **Conclusion**
- References

# Conclusion

- In this, project we have devised a method monitoring events in the real time using Apache Spark.
- For evaluation, we have tested on twitter streams at different times.
- We got very good results and we were able to monitor the events.
- This tool can be applied in monitoring various natural disasters like earthquakes, tornadoes etc.

# OUTLINE

- Introduction
- Approach
- Evaluation and Results
- Conclusion
- **References**

# References

- <http://www.cs.cmu.edu/~deepay/mywww/papers/icwsm11-twiterevents.pdf> - Main reference paper which has the core idea.
- <http://adilmoujahid.com/posts/2014/07/twitter-analytics/> - Getting the twitter stream
- [spark.apache.org/docs/latest/api/python/](http://spark.apache.org/docs/latest/api/python/) - Learning how to code in spark using python
- <http://stackoverflow.com/questions/30829382/check-the-similarity-between-two-words-with-nltk-with-python> - nltk Reference for similarity between two words.

# Project links

- GitHub link

<https://github.com/anirudhalampally/Cloud-Major-Project>

- Presentation link

[https://docs.google.com/presentation/d/15IVqeBkn7UEUgHF6Zsbf\\_dmbJpqkOWtw6-oKHLOA7bw/edit?usp=sharing](https://docs.google.com/presentation/d/15IVqeBkn7UEUgHF6Zsbf_dmbJpqkOWtw6-oKHLOA7bw/edit?usp=sharing)

- Video link



THANK YOU!!