ABSTRACT

Modern technology and social media have made it possible for the voice of a single person to be heard by millions of people worldwide [1, 2]. They can share their opinions, collaborate on ideas, organize events and activities with other people, amongst others. This has affected our personality in diverse ways as people treat different users: strangers and friends in the same way [3-5]. In this study, our aim is to analyze the real-time events and conversations of people active on three social websites: Twitter, Reddit, and Meetup to find where events related to popular topics are currently happening. To achieve this, we used AWS services viz. Kinesis, DynamoDB, lambda and built a soft custom-lambda architecture. This tool can be used to identify active social gatherings and help organize targeted activities with high audience participation. Additionally, it offers insights of the data with respect to different time frames such as hour, day, or month.

1. Intro
2. Method
3. Discussion
   1. Why Reddit, Twitter, and Meetup only?
      1. Use google scholar!
   2. Design
      1. How producers put the data into the stream.
         1. Approaches
            1. Linearly

Disad: Ineffective coordination amongst producers; inefficient shards usage

Adv: Easy

* + - * 1. Randomly

Disad: Unpredictable coordination amongst producers, and shards usage

Adv: Can be efficient if very high num of shards

* + - * 1. Round-Robin

Disad: ?

Adv: All shard used approx. equally

* + 1. Table schemas
       1. Approaches (describe 52 table schemas, and current)
          1. Adv
          2. Dis-adv
       2. Live Reddit & Twitter Tables
          1. What should be the primary keys
          2. Why store raw data in live tables
          3. Discussion on attribute values (why include seconds in live, etc.)
       3. Meetups
       4. Advantages
    2. Why lambdas per minute and hour?

Effective caching: reduces latency

* + 1. Why top-K results only?

Less storage implies less cost and faster response but lower accuracy

* + 1. Why lambda functions as brokers?
       1. Scale in a constant time w.r.t. num of users.
       2. Cost: Pay only for usage
       3. Reliability: since no static broker
    2. Why own lambda arch?
       1. Own Learning: wanted to lose some sleep.
       2. Built custom features with improved dynamic reliability\* (anonymous multiple resources of AWS working to support the architecture).
    3. Why Dynamo?
       1. Auto-Scalability.
       2. System requires key-> value only.
       3. Sub-second latency
       4. Reliability
    4. Better tools, if provided enough resources.
       1. RedShift?
          1. Why not now?
          2. Why better?
          3. Will it provide additional features?
  1. Existing work
     1. Description
     2. Comparison to our model

1. Results

We used 6 shards in the Kinesis to stream real time data. The producers (Twitter, Reddit, and Meetup) put the data into these streams in a round-robin fashion. The producers write around a speed of …. per second as indicated in figure <Fig Number>. This approach allowed the all the consumers to read related data concurrently increasing efficiency of the system. Using our caching techniques in the DynamoDB, the response time for user queries (indicate the rate wrt time, geography and refer to plots).

1. Conclusion

We were successfully able to deploy Amazon Kinesis, DynamoDB, AWS lambda functions, and a custom-built lambda architecture to stream real-time data from Reddit, Twitter, and Meetup and perform analysis on it. Our analysis includes ranking data based on active communication of people on Reddit and Twitter, and finding where events related to these popular discussions are happening using Meetup. The system provides user the functionality to search in a particular time-frame, geography, or both. It can be used to organize targeted activities with high audience participation. Our data representation also allows to extend the project in multiple dimensions like finding popular venues according to the type of event, using machine learning to find patterns in the events and predict future gatherings, etc. Additionally, the architecture and design have the capacity to add more sources of data and use the same business logic to improve accuracy, and perform deeper analysis.

References

1. Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business horizons*, *52*(4), 357-365.
2. Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE.
3. Gilbert, E., & Karahalios, K. (2009, April). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 211-220). ACM.
4. Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. *Journal of computer assisted learning*, *27*(2), 119-132.
5. Sessions, L. F. (2010). How offline gatherings affect online communities: when virtual community members ‘meetup’. *Information, Communication & Society*, *13*(3), 375-395.