ETL and Analysis of Meetup Streams: What activities are communities across the United States excited about?

Karin Belsvik Brodd, Chandler McCann, Natarajan Shankar, David Watson

10/11/2016

Background

Meetup.com is a social networking website that “brings people together in thousands of cities to do more of what they want to do in life.”[1] The website hosts 27.88 million members who participate in over 258,000 Meetup groups in 178 countries.[2] The service provides a platform for people to participate in a variety of activities, such as outdoors adventures, photography, book clubs, religious studies, and entrepreneurship.[3] In an average month, about 3.96 million people sign up to participate in a local Meetup.[4] Importantly, Meetup provides access to data regarding their active Meetups via the Meetup API.[5]

Research Question

The primary question that will be answered is “What activities are communities throughout the United States excited about?”. A secondary research question we’d like to explore if time allows is, why are the active communities successful? The secondary question will potentially be answered by analyzing features that successful/unsuccessful meetups have such as meeting topic, attendee age, repeat attendees, rate of growth,

Research Purpose and Use Cases

This project will attempt to understand what activities are most popular in different areas. By categorizing and quantifying signups across various communities, it will be examined how preferences for specific activities change, or possibly remain the same, throughout the United States. Additionally, geographical areas that have similarities in attended activities, despite differences in regions and demographics, will be identified. Through this analysis, insight as to what influences an individual's choice of activities will be ascertained.

---

[1] https://www.meetup.com/about/

[2] Id.

[3] https://www.meetup.com

[4] https://www.meetup.com/about/

[5] https://www.meetup.com/meetup_api/

Gaining insight into why individuals participate in specific activities is valuable for many purposes. As the Meetup site describes it, meetups are used to convene people to:
- Do what they love
- Find others and make friends
- Get involved in their local communities
- Learn, teach, and share
- Rise up, stand up, unite, and make a difference
- Be part of something bigger - both locally and globally

Using a derivative of this work, individuals who are looking to move to a new community could find communities that share common interests and pursue activities of interest. From a business perspective, companies could determine which products and services should be marketed to specific regions based upon activity popularity, as well as identify regions of interest for employee recruitment. Local governments would find this information beneficial to understand and quantify community preferences and provide services that the community values.

What is planned to be built:

The plan of this project is to build an ETL infrastructure that can consume, extract, transform, modify and summarize meeting topics and context from data streams emanating from Meetup.com. The proposed architecture supports a process flow for data ingest, storage, schema on write, clean/transform, processing and query.

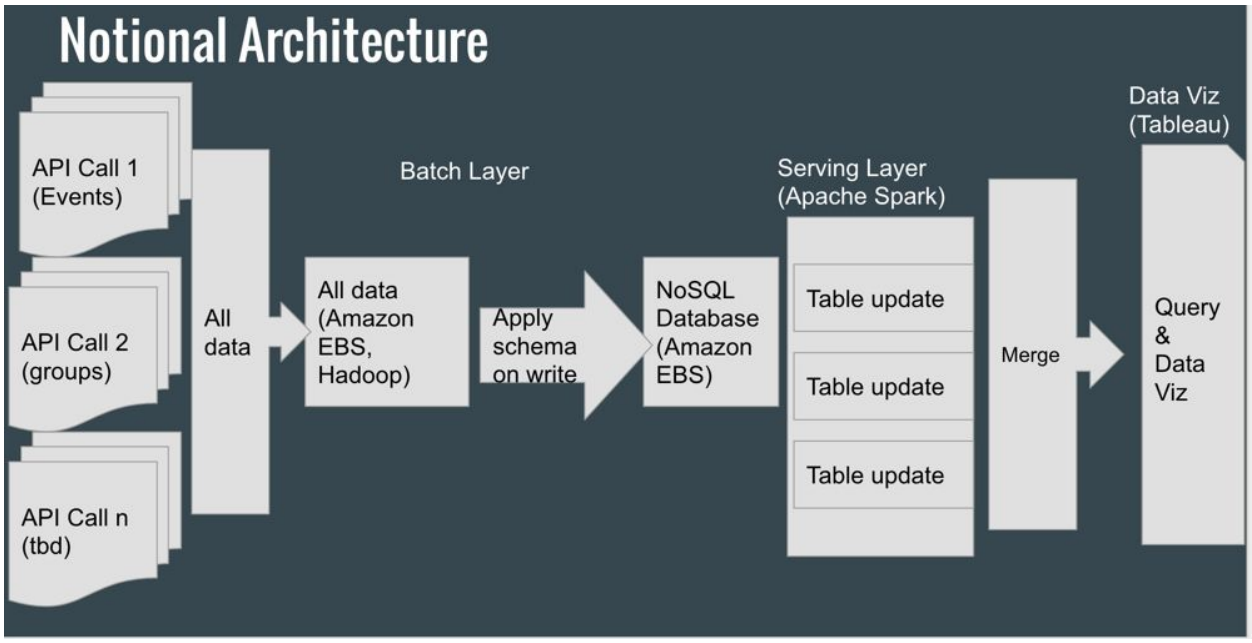

Figure 1: Notional architecture for Meet-Up data ingest, transform, process and query

To achieve the architectural goal via a functional implementation, the proposed ETL infrastructure will build upon the existing Meetup.com provided APIs. With the existing APIs, Meetups can be created, queried or updated via Meetup authorized Apps running on Mobile Web, Desktop Web or on Android/IOS appliances. Each authorized App implements the meetup API. Authorization keys are provided by registering an App with Meetup.com. The Meetup API provides a simple interface for accessing the Meetup platform from the apps and this project will build upon the interface.

- GET methods will be used to query Meetup data, POST methods can be used to create and alter resources, and streaming methods will be used to read real time and consume activity.
- Meeting CREATEs and meeting RSVPs are two continuous streams that will be consumed through one or both of REST or streaming APIs.
- A NoSQL based local storage interface a la MongoDB will be implemented for subsets of preprocessing and for postprocessing
- The ETL infrastructure is depicted in the diagram in the Notional Architecture diagram in Figure 1. The key components to be implemented are:
    - Batch Layer that is implemented using Amazon compute and storage resources with an overlay of Apache Spark and Hadoop based processing
- Each Meeting that fits a defined filter contains information that will need to be extracted and processed. This work will require textual parsing and topic aggregation. Each meeting in the Meetup stream contains key fields such as title and description. By parsing the title and the description and other critical fields, a summarized view of the topic will be created.
- A presentation interface (a la simple GUI or with Tableau) will be created if time allows. The default will be CLI based control interface.

What Technologies will be used:

- Data will be obtained from the Meetup through either REST API calls, or a streaming connection
- Hadoop File System on AWS EBS will store raw data extracts from the Meetup API calls. A Storm cluster receives data from the streaming sources and stores the complete JSON in HDFS in an EBS on AWS
- Data will be extracted in batches to a separate EBS with a MongoDB schema on write database, in order to enable table creation.
- Apache Spark will be used for Data Frame creation and analysis
- Aggregation and summarization of topics will be done through textual analysis
- With schema applied, tables will be updated in the serving layer
- Finally, related Tables will be queried using the Primary Key and Foreign Key, through a command line interface.
- Ultimately, the structured database will feed a data visualization layer with Tableau or a comparable platform. This will allow for the graphical analysis and representation.

Tasks and Assignment Plan (Names not assigned, Resource plan is preliminary)

1. Authenticated Interface with Meetup.com (Assignee)
    a. Oauth
    b. Authenticated API
2. Stream processing (Assignee, Assignee)
    a. Storm
3. Storage for Extraction phase (Assignee, Assignee)
    a. Local buffer using HDFS on AWS EBS
4. Preprocessing (Assignee)
    a. Word sorting
    b. Extract from critical fields
5. Sorting and indexing of Text (Assignee)
    a. Hadoop/Map Reduce
6. Textual Summarization (Assignee, Assignee)
    a. Call out frequent topics and sites
7. Storage for Summary and Data Visualization (Assignee)
    a. Final storage and availability and data visualization in our data visualization tool.
8. Visualization Interface (Assignee, Assignee)

Data Details

Data Volume/Velocity: This project will initially start with a static capture of data from the Meetup CREATE stream. The goal is to move the project to a streaming interface that will deliver data at a rate that can be consumed by the infrastructure. The streaming rate will be managed and throttled via data filters that define data relevance. It is estimated that the training stream will be at the rate of one or two JSON records per second. This project will test with 4K/second.

Data Format: Data will be received either through a RESTful API or through a continuous stream, each with a JSON payload . One typical record looks like the following (the projected critical fields are highlighted):

```
HTTP/1.1 200 success
{
"results": [
{
"utc_offset": -25200000,
"venue": {
"zip": "94568",
"country": "us",
"localized_country_name": "USA",
```

```
"city": "Dublin",
"address_1": "7950 Dublin Blvd",
"address_2": "Suite 103B",
"lon": -121.934731,
"phone": "925.560.0700",
"name": "The Specific Chiropractic Center",
"id": 940018,
"state": "CA",
"lat": 37.702424,
"repinned": false
},
"headcount": 0,
"distance": 3.746248245239258,
"visibility": "public",
"waitlist_count": 0,
"created": 1473798415000,
"maybe_rsvp_count": 0,
"description": "<p>Come join us for Wine, Beer and Root Beer!
(Yes we have non-Alcoholic options :) ) We will be discussing
our thoughts from "You are a Badass" by Jen Sincero</p> <p> <a
href="https://www.amazon.com/You-Are-Badass-Doubting-Greatness/
dp/0762447699/ref=zg_bs_books_39"><a
href="https://www.amazon.com/You-Are-Badass-Doubting-Greatness/
dp/0762447699/ref=zg_bs_books_39"
class="linkified">https://www.amazon.com/You-Are-Badass-Doubtin
g-Greatness/dp/0762447699/ref=zg_bs_books_39</a></a></p>",
"event_url":
"http://www.meetup.com/Books-n-Bevs/events/234100902/",
"yes_rsvp_count": 12,
"duration": 3600000,
"name": "October: Sip, Skim and Socialize!",
"id": "234100902",
"time": 1476151200000,
"updated": 1473805384000,
"group": {
"join_mode": "open",
"created": 1471853348000,
"name": "Books n' Bevs: A Club for Sipping and Skimming",
"group_lon": -121.93000030517578,
"id": 20340431,
"urlname": "Books-n-Bevs",
"group_lat": 37.709999084472656,
"who": "Book Worms"
```

```
},
"status": "upcoming"
},
```

Project Plan:

The general plan is to get a working prototype as soon as possible for the end-to-end solution. The milestones below represent go-live targets. The goal is to have a functioning beta version of each milestone scope prior to the actual milestone.

- Week 6 -8 - Acquisition and storage strategy
    - Primary and secondary research questions refined
    - Entity Relationship diagram finalized.
    - API call plan (frequency, locations, topics) supporting our Entity Relationship Diagram
    - Team AWS instance and volumes set up for our ingest and serving layer, with HDFS and MongoDB installed
- Week 9-10 - Acquisition and Storage Test
    - Pilot AWS solution and data pipeline
    - Data cleansing automation test
- Week 10-12- Data Storage and Analysis Test
    - Bulk data storage
    - Analysis algorithm test and refinement
- Week 13 on- Analysis and close-out
    - Complete analysis and results summary