ETL and Analysis applied to Meetup* Streams

What ideas are cities and communities across the US curious about?

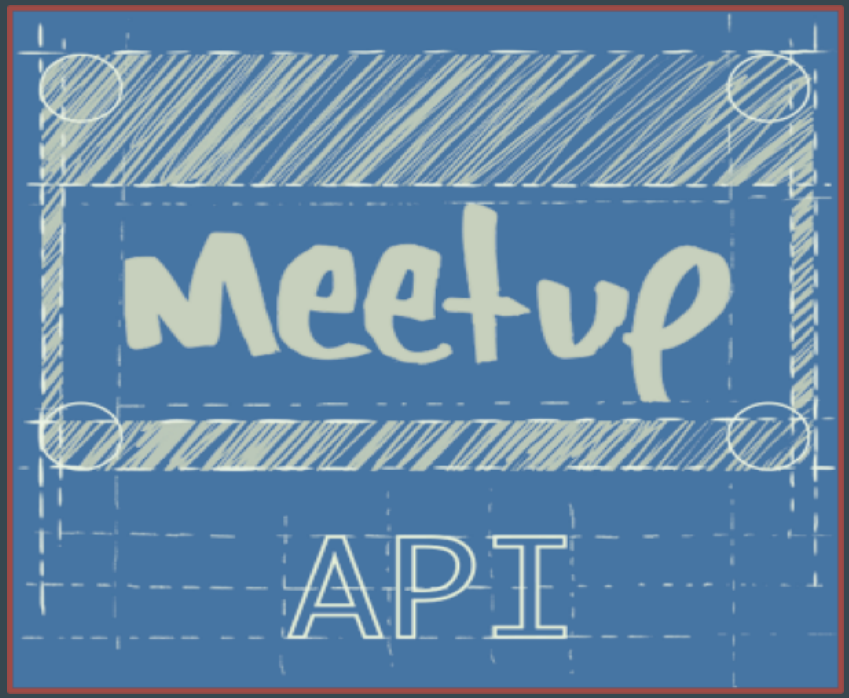
•••

Karin Brodd
Chandler McCann
Natarajan Shankar
Dan Watson

*Meetup is an online social Networking portal that facilitates group meetings. Meetup brings people together in thousands of cities to do more of what they want to do in life.

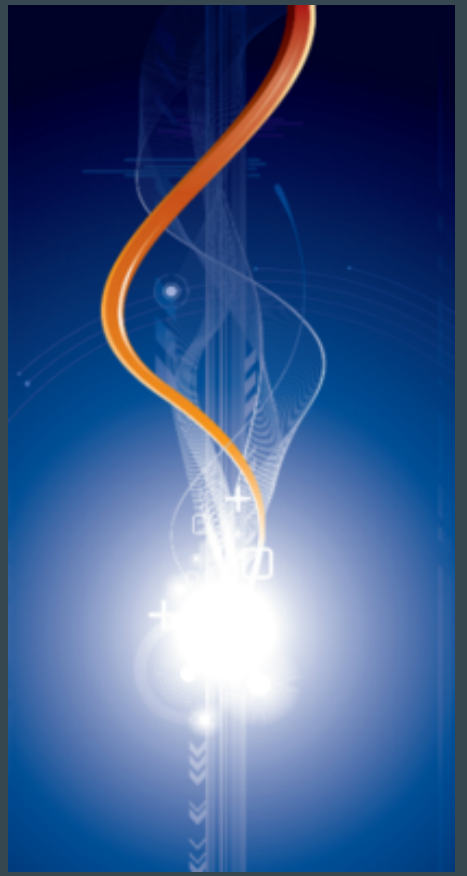
Data Acquisition

- Browser/Client authenticated with OAuth
- Streaming between server and API key protected client
- JSON formatted data predominant, other encodings available
- Event descriptors are pushed into stream in real-time
- Client can search and log Meetups by zip code, radius, groups, and number of members, allowing for Filtering
- RSVP API available for parsing
    - Rich event information
    - Date, time, location, number of members attending, number of guests attending
- Local storage in SQL-capable database, to support merging and aggregation



Anticipated Challenges

- This project is potentially programmatically intense but the core focus of this project will specifically be kept to ETL infrastructure
- Designing a system that can ingest live streaming data for storage, processing, and serving is quite involved
    - Potentially multiple streaming sources: RSVPs, Events, and comments, will need to be processed efficiently
    - May want to incorporate data from other non-streaming APIs
- Transforming JSON document stream from API into schema usable for answering research question is non-trivial
- Processing unstructured portions of data for additional features, e.g.: comments, event descriptions is critical to solution implementation
- Text processing is critical to decoding the streams and will need complex interpretative approaches



Execution Overview

- Week 6 -8 - Acquisition and storage strategy
  - Research question refinement
  - API call plan (frequency, locations, topics)
  - Storage Plan
- 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