Predicting IMDB Movie Ratings

Using social media data and open source tools

Greg, Talieh & Apekshit

Introduction

- ***Motivation***
  - Using powerful tool of Sentiment Analysis
  - Millions of users' data every day
  - Twitter feeds represent a valuable collection of human opinion
  - $$$$: Forecasting ratings which in turn -> box-office revenues in some cases

- ***Hypothesis:***
  - Movie ratings are influenced by demographics and personal factors
  - Such factors are hard to model explicitly
  - Correlations can be found between IMDb ratings and activity indicators around the same artifacts

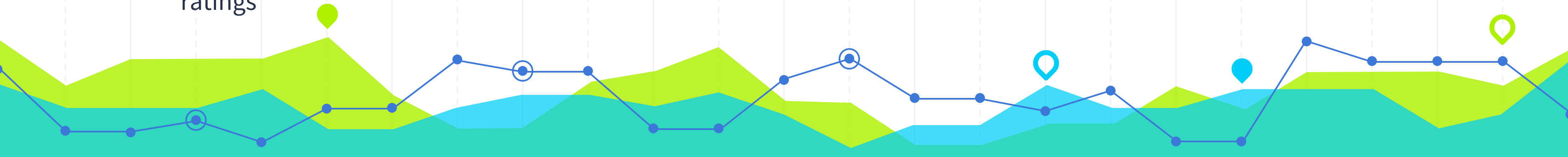
- ***Implications for a working prototype***
  - Will allow require us to deal with the 3 V's Volume, Variety and Velocity
  - Will allow for us to explore and extend prior work and analysis done in the same subject area

Problem Statement

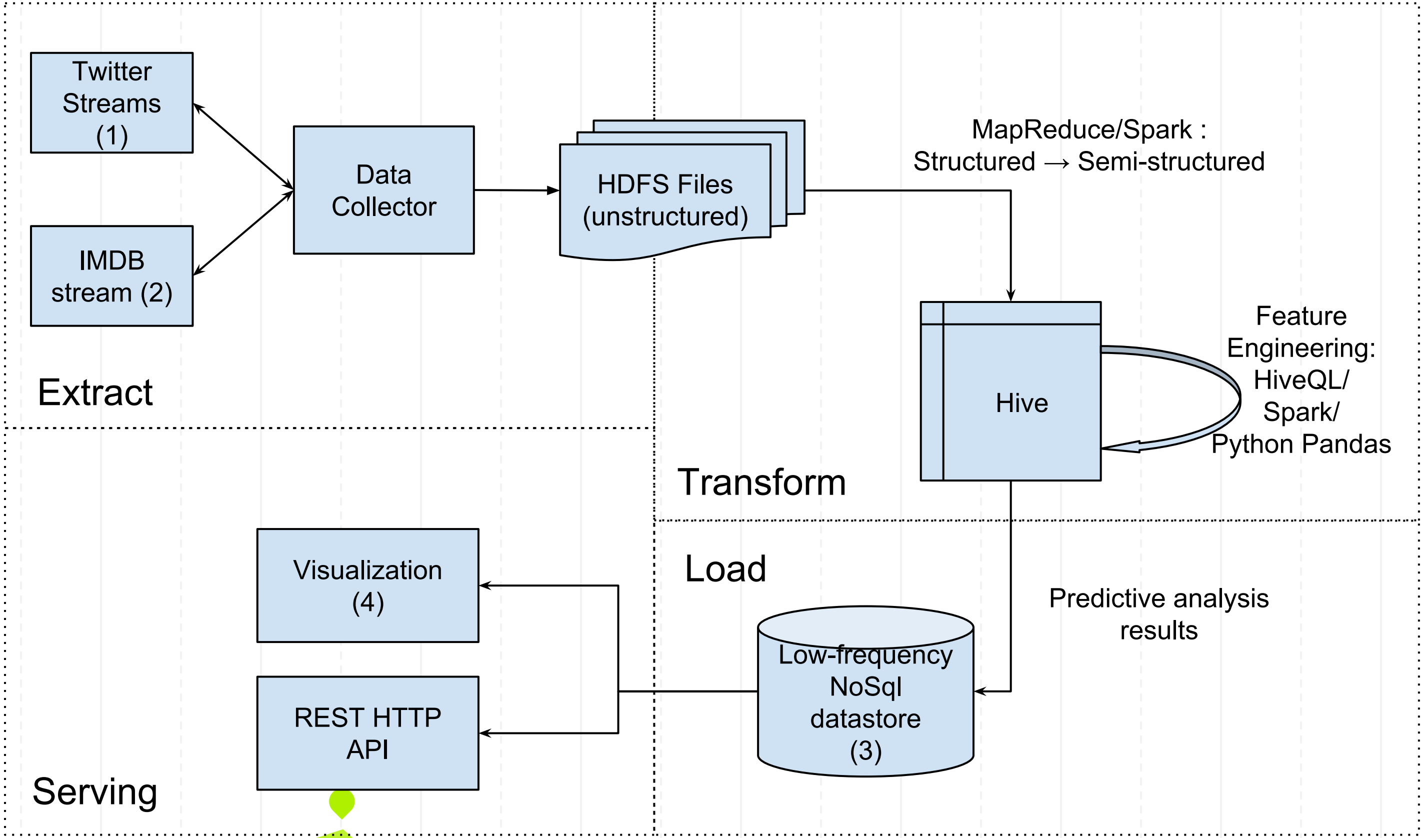
Collect, load, transform and analyze data from tweet streams* to explore and predict recently released IMDB movie ratings

Expected Sub Problems to Solve

- **Data Acquisition**
    - Implement process to ingest real time Twitter data on a particular movie
    - Implement a process to download and curate movie details from IMDB
- **Data Load**
    - Stage data in HDFS
    - Build Hive staging schema(s) that can handle near real time load + historicals
- **Data Integration**
    - Design schema model conducive to predicting movie ratings
- **Data Analysis**
    - Leverage Spark / Python ML libraries to make near real time predictions re: attributes collected in
- **Reporting**
    - Design an intuitive interface that allows for a user to visualize and explore outcomes of processing / predictions
    - Make publicly available
- **Repeatability**
    - Ensure processing pipeline and resultant analysis is repeatable and resilient



Intended Architecture & Tools


Twitter Streams (1)
IMDB stream (2)
Data Collector
HDFS Files (unstructured)
Extract
MapReduce/Spark : Structured → Semi-structured
Hive
Feature Engineering: HiveQL/ Spark/ Python Pandas
Transform
Load
Predictive analysis results
Low-frequency NoSql datastore (3)
Visualization (4)
REST HTTP API
Serving

1. Twitter Python
   http://www.tweepy.org/
2. IMDBPy
   http://imdbpy.sourceforge.net/
3. Hive/MongoDB/HBase
4. Tableau

Expected Learnings

1. ***Demonstrated ability to:***
    a. integrate a number of open source tools to solve an interesting problem
    b. experience with NLP in a non-toy setting
        i. computing both activity and meaning surrounding tweet data
        ii. sentiment analysis
    c. integrate distinctly different data sources with varying latencies

2. ***Experience working with geographically dispersed folks on a real challenge***
    a. meaningful github contributions

Sources

1. Predicting Ratings for New Movie Releases from Twitter Content, Schmit 2015

2. Predicting IMDB Movie Ratings Using Social Media, Breuss 2015.

3. MovieTweetings: a Movie Rating Dataset Collected From Twitter, Dooms 2015

4. Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site, Trusov 2009.

5. Our project documentation