Holiday Kickoff in the Twittersphere

Analyzing tweets from Black Friday

Jasen Jones, Miki Seltzer, Eric Whyne

Project Background

- Use tweets to quantify various aspects of Black Friday and lead-up to Black Friday
- Collect tweets prior to and during holiday kickoff weekend (11/27 - 11/30)
- Examine popular brands to see how they compare to each other
    - Current idea is to examine consumer electronics brands
- Examine difference between consumer electronics brands, product categories, etc.
- Further opportunity to examine sentiment of tweets to gauge customers' attitudes towards brands

Architecture

Data Gathering:

Twitter API, SupervisorD, Kafka, Crontab, AWS CLI

Data Processing:

S3, EC2, EMR, GNU Parallel, Hive

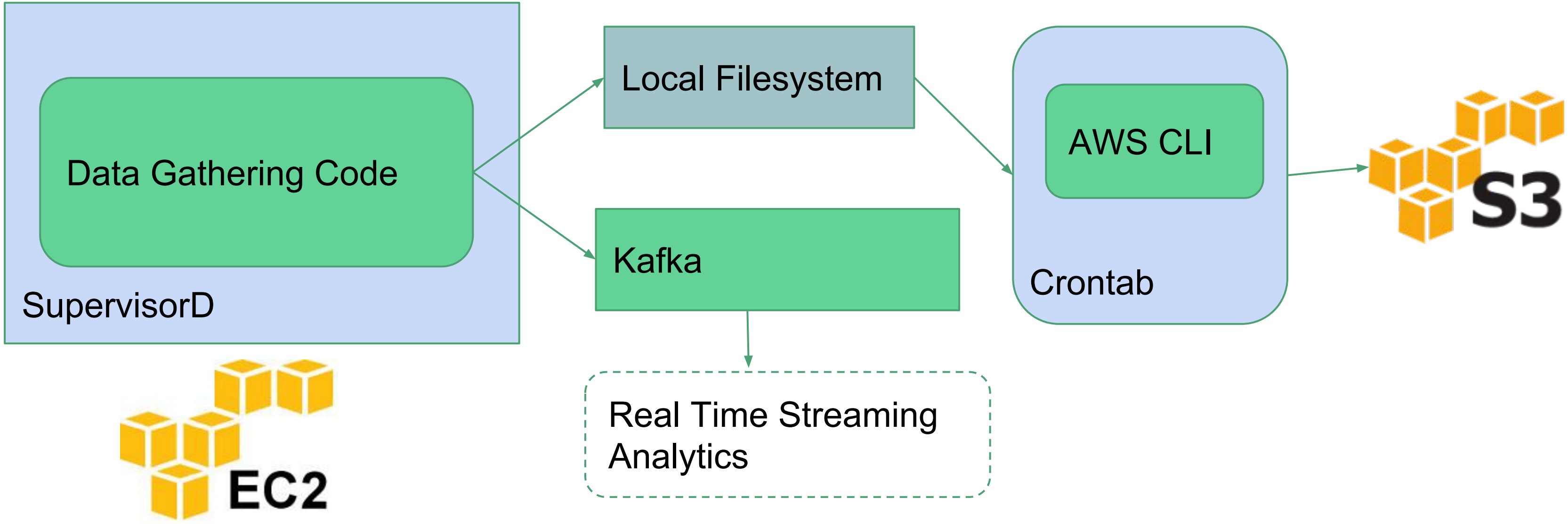
Spark, Storm

Serving:

Tableau

Data Collection Architecture

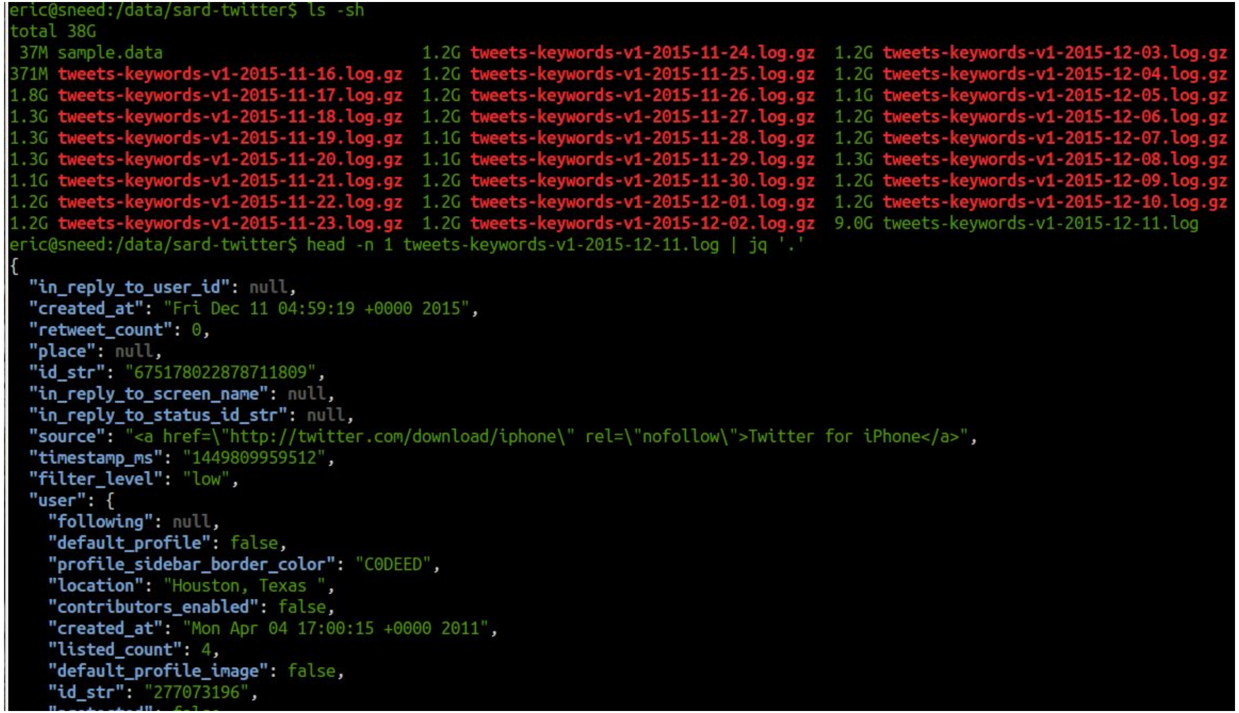
Code located here: https://github.com/mseltz/W205-Project/tree/master/data_gathering



Data Gathering Results

Keywords:

```
4k
android
apple
beats
bluetooth
camera
dslr
electronics
fitbit
gopro
headphones
ipad
iphone
jawbone
laptop
mac
macbook
microsoft
nintendo
panasonic
philips
receiver
samsung
sony
speaker
surface
tablet
theater
visio
wii
xbox
```



```
eric@sneed:/data/sard-twitter$ ls -sh
total 38G
 37M sample.data                            1.2G tweets-keywords-v1-2015-11-24.log.gz  1.2G tweets-keywords-v1-2015-12-03.log.gz
371M tweets-keywords-v1-2015-11-16.log.gz  1.2G tweets-keywords-v1-2015-11-25.log.gz  1.2G tweets-keywords-v1-2015-12-04.log.gz
1.8G tweets-keywords-v1-2015-11-17.log.gz  1.2G tweets-keywords-v1-2015-11-26.log.gz  1.1G tweets-keywords-v1-2015-12-05.log.gz
1.3G tweets-keywords-v1-2015-11-18.log.gz  1.2G tweets-keywords-v1-2015-11-27.log.gz  1.2G tweets-keywords-v1-2015-12-06.log.gz
1.3G tweets-keywords-v1-2015-11-19.log.gz  1.1G tweets-keywords-v1-2015-11-28.log.gz  1.2G tweets-keywords-v1-2015-12-07.log.gz
1.3G tweets-keywords-v1-2015-11-20.log.gz  1.1G tweets-keywords-v1-2015-11-29.log.gz  1.3G tweets-keywords-v1-2015-12-08.log.gz
1.1G tweets-keywords-v1-2015-11-21.log.gz  1.2G tweets-keywords-v1-2015-11-30.log.gz  1.2G tweets-keywords-v1-2015-12-09.log.gz
1.2G tweets-keywords-v1-2015-11-22.log.gz  1.2G tweets-keywords-v1-2015-12-01.log.gz  1.2G tweets-keywords-v1-2015-12-10.log.gz
1.2G tweets-keywords-v1-2015-11-23.log.gz  1.2G tweets-keywords-v1-2015-12-02.log.gz  9.0G tweets-keywords-v1-2015-12-11.log
eric@sneed:/data/sard-twitter$ head -n 1 tweets-keywords-v1-2015-12-11.log | jq '.'
{
  "in_reply_to_user_id": null,
  "created_at": "Fri Dec 11 04:59:19 +0000 2015",
  "retweet_count": 0,
  "place": null,
  "id_str": "675178022878711809",
  "in_reply_to_screen_name": null,
  "in_reply_to_status_id_str": null,
  "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "timestamp_ms": "1449809959512",
  "filter_level": "low",
  "user": {
    "following": null,
    "default_profile": false,
    "profile_sidebar_border_color": "C0DEED",
    "location": "Houston, Texas ",
    "contributors_enabled": false,
    "created_at": "Mon Apr 04 17:00:15 +0000 2011",
    "listed_count": 4,
    "default_profile_image": false,
    "id_str": "277073196",
    "protected": false,
```

Several overly broad keywords were removed early on, like ‘pc’.

Data Gathering Lessons

SupervisorD was absolutely critical when running for a long time. There were errors we were unable to plan for and latency/timeout issues that would cause the streaming API to disconnect. By being smart with our SupervisorD timeouts we were able to re-establish connectivity and also not burn out our Twitter API credentials by hitting the API too aggressively upon disconnect.

Avoid overly broad keywords.

The processing pipeline needs to be able to keep up with the velocity of data coming in.

Data Processing

Once the data was on S3, we had many options for processing it.

- EMR allows running Hive code directly on data stored in S3
  - http://docs.aws.amazon.com/ElasticMapReduce/latest/ReleaseGuide/emr-hive.html
- S3 allowed us to share large amounts of data within our team

To get to our end result, we applied filters and transforms using Python code and GNU Parallel to make it run faster, then loaded the filtered data into Hive for interaction and visualization. We've packaged these operations as runnable examples in the project's Github repository.

https://github.com/mseltz/W205-Project

Sentiment Analysis - Core

- Each tweet assigned a score based on message content
- Scores appended to the of tweet’s JSON entry
- Score based on number of positive and negative words
  - Dictionary of terms retrieved from: Minqing Hu and Bing Liu. "Mining and Summ
    Customer Reviews."

Toys R Us PS4 sale is awesome!
+1

Sentiment score = 1

Time to enjoy my great new 55-inch Panasonic TV!
+1 +1

Sentiment score = 2

I want to like Amazon’s Black Friday deals, but I hate how they only put cheap stuff on sale.
+1 -1 -1

Sentiment score = -1

Sentiment Analysis - Modifiers

- Added additional functionality to handle special terms that increase, decrease, or invert the meaning of the following word
- Increaser examples:
  - So, totally, completely, absolutely
- Decreaser examples:
  - Sorta, a bit, a little
- Inverter examples:
  - Not, isn’t, lack of, didn’t, don’t

Toys R Us PS4 sale is totally awesome!

2x multiplier +1

Sentiment score = 2

Wow, I don’t like the deals this year at ALL.

-1x multiplier +1

Sentiment score = -1

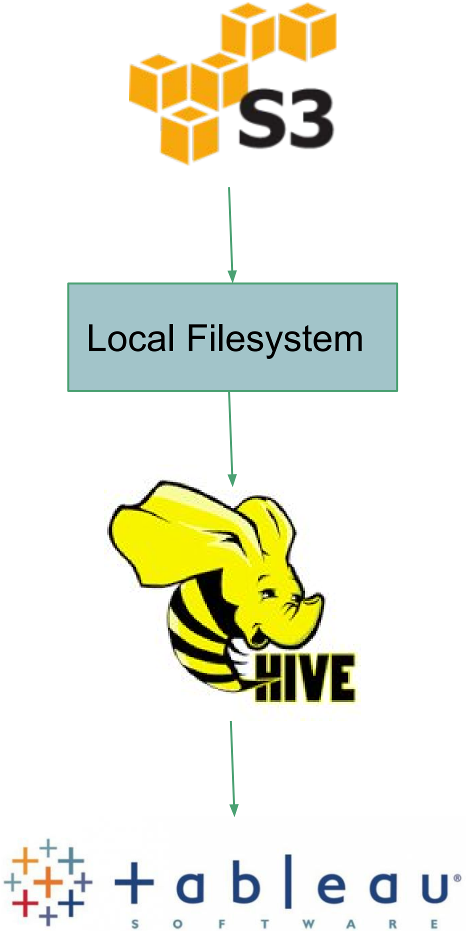
Toys R Us PS4 sale is sorta good, I guess.

0.5x multiplier +1

Sentiment score = 0.5

Serving Layer

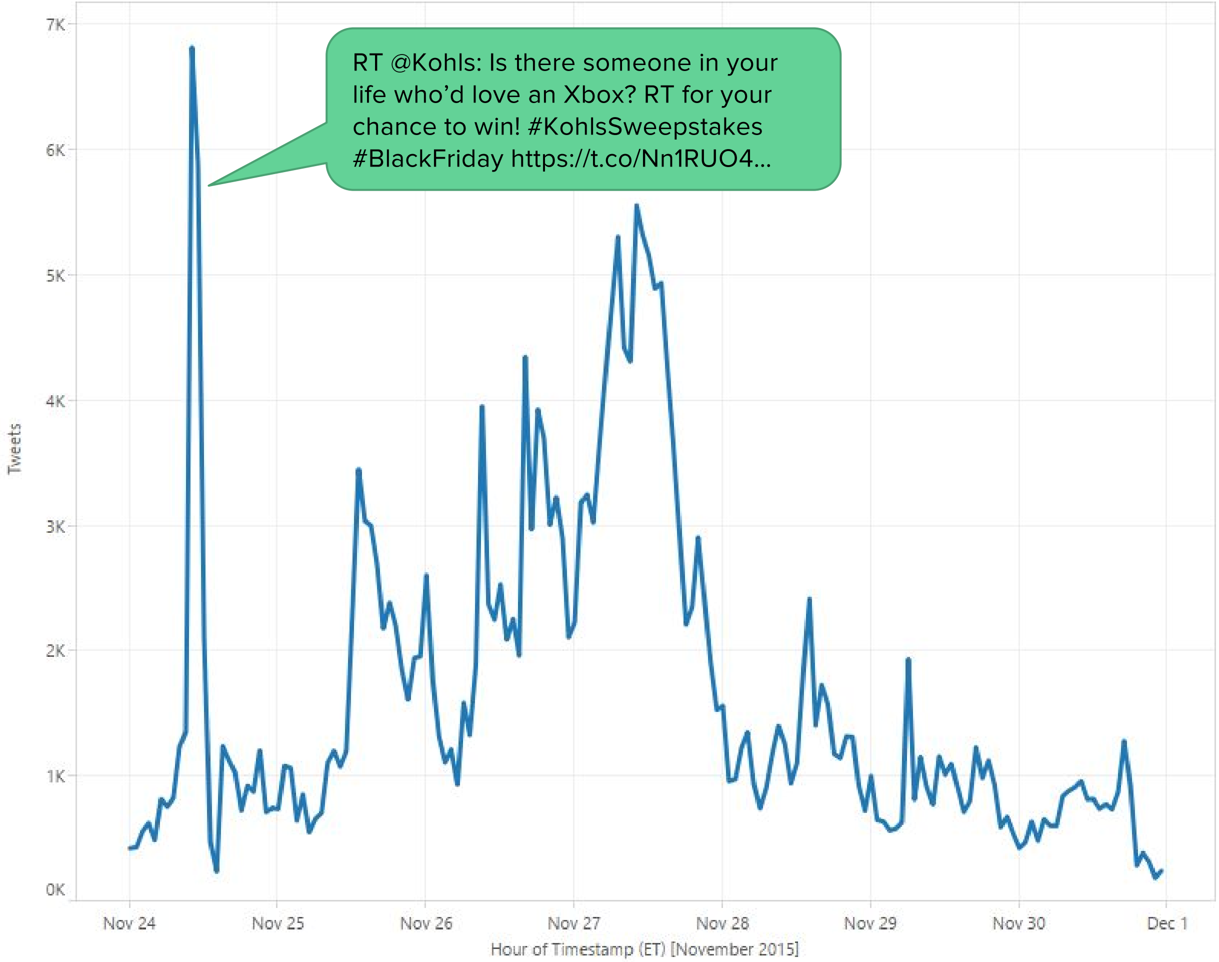
- Tweets stored in Hive table
- Options to pull insights:
  - Query directly in Hive
  - Use Hiveserver2 to connect to Tableau
- Advantages:
  - Tableau has rich visualization capabilities
  - Can store queries to Hive tables for faster rendering in Tableau
- Disadvantages:
  - Live connection between Hive and Tableau limits speed
  - Tableau cannot handle “big” data



Local Filesystem

Analysis Results

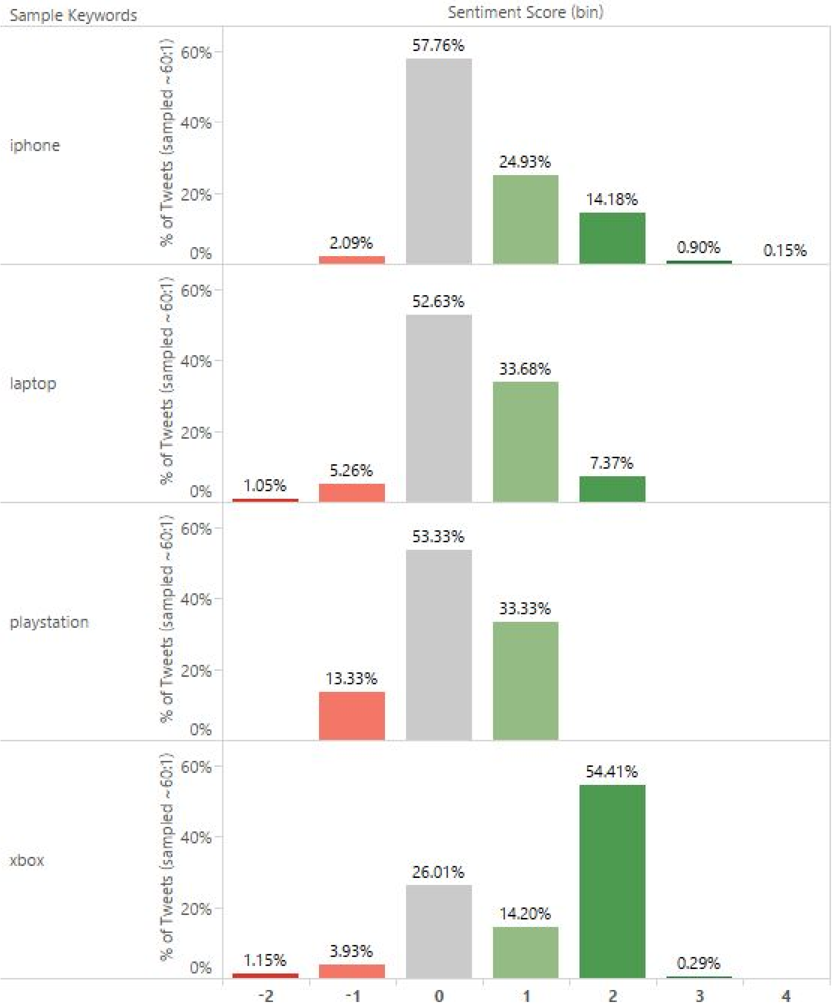
- Tweets containing consumer electronics keywords and “Black Friday” start to ramp up starting Wed Nov 25
- Large spike in tweets on Nov 24 @ 10am ET (Kohl’s Sweepstakes)
- Black Friday (Nov 27) has largest volume of tweets, then tapers down



RT @Kohls: Is there someone in your life who’d love an Xbox? RT for your chance to win! #KohlsSweepstakes #BlackFriday https://t.co/Nn1RUO4...

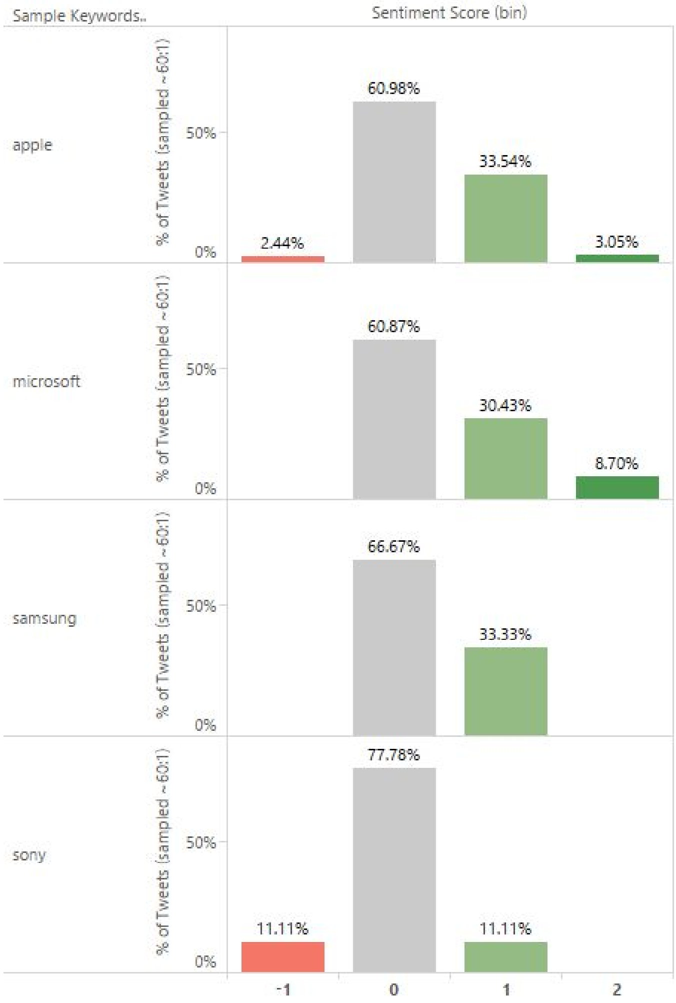
Analysis Results - Sentiment

- We can look at the breakdown of sentiment of a particular keyword
- Sample keywords for products are given in the image to the right
  - Most terms have the majority of tweets as neutral
  - Keyword “xbox” has large percentage of fairly positive sentiment



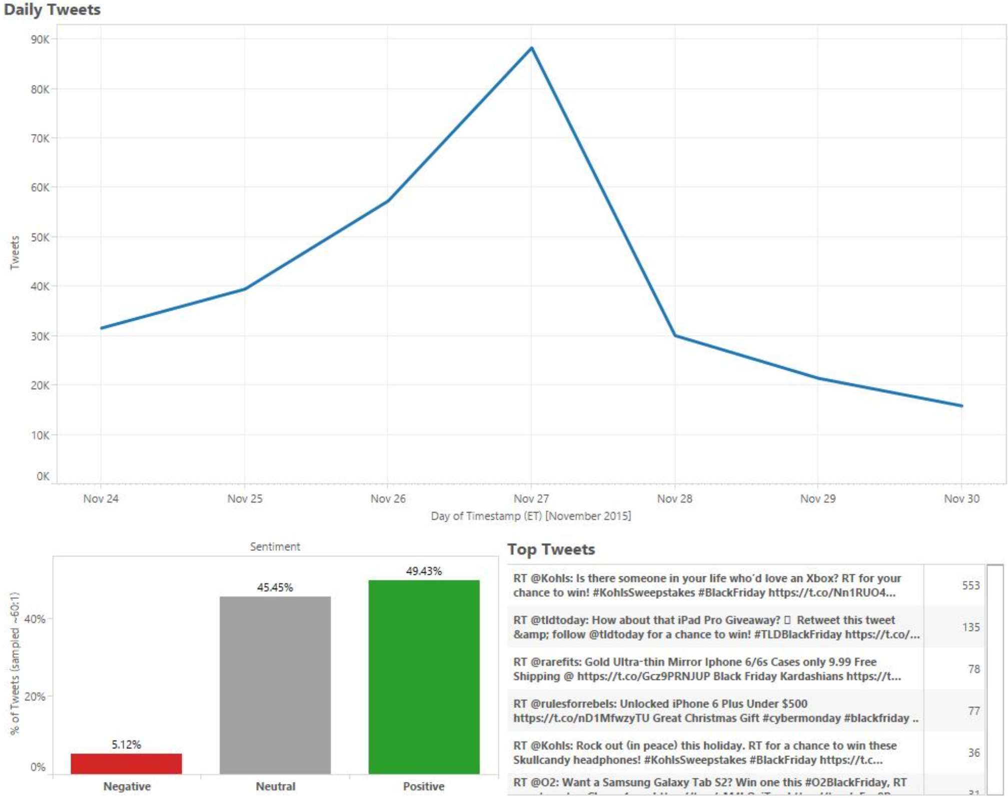
Analysis Results - Sentiment

- We can look at the breakdown of sentiment of a particular keyword
- Sample keywords for brands are given in the image to the right
    - Most terms have the majority of tweets as neutral
    - Keyword “sony” has both mildly negative sentiment and mildly positive sentiment
    - Keywords “microsoft” and “samsung” have only neutral and positive sentiments



Analysis Results - Simple Dashboard

- Volume of tweets per day
- Sentiment of tweets (sampled)
- Top tweets (sampled)
- Interactive
    - When user clicks on any element of the dashboard, the other modules are updated to reflect only the selected data



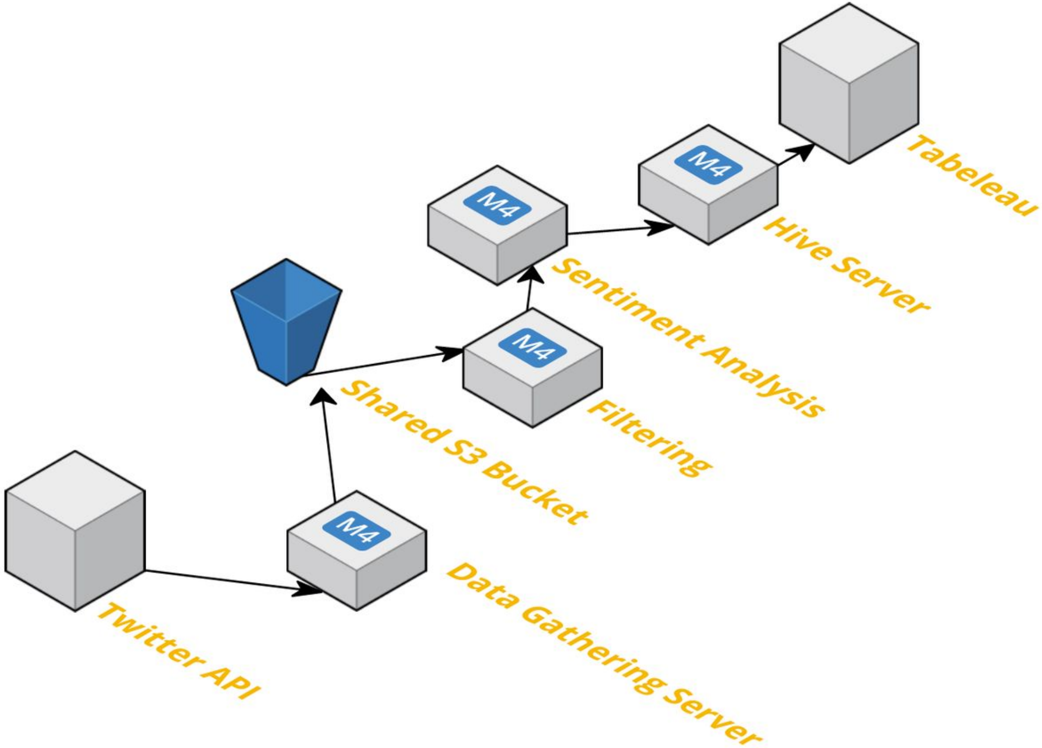
Top Tweets

RT @Kohls: Is there someone in your life who'd love an Xbox? RT for your chance to win! #KohlsSweepstakes #BlackFriday https://t.co/Nn1RUO4...	553
RT @tldtoday: How about that iPad Pro Giveaway? □ Retweet this tweet &amp; follow @tldtoday for a chance to win! #TLDBlackFriday https://t.co/...	135
RT @rarefits: Gold Ultra-thin Mirror Iphone 6/6s Cases only 9.99 Free Shipping @ https://t.co/Gcz9PRNJUP Black Friday Kardashians https://t...	78
RT @rulesforrebels: Unlocked iPhone 6 Plus Under $500 https://t.co/nD1MfwzyTU Great Christmas Gift #cybermonday #blackfriday ..	77
RT @Kohls: Rock out (in peace) this holiday. RT for a chance to win these Skullcandy headphones! #KohlsSweepstakes #BlackFriday https://t.c...	36
RT @O2: Want a Samsung Galaxy Tab S2? Win one this #O2BlackFriday, RT	31

Considerations for Scaling

- All filtering should be done at time of collection
    - We chose to collect more data than necessary, but chose to filter later -- filtering earlier would cut down on storage needs, allowing for more efficient data collection
- Sentiment analysis should be parallelized and run at time of collection
    - Currently, sentiment analysis is done as a batch process on a small subset of data
    - Current sentiment analysis algorithm is not optimized to be applied to streaming data -- scoring takes too long
        - We might be able to make this work with a larger cluster, Kafka and balanced consumers, Storm, or Spark Streaming.

Architecture Overview


Twitter API
M4
Data Gathering Server
Shared S3 Bucket
M4
Filtering
M4
Sentiment Analysis
M4
Hive Server
Tabeleau