W205 Final Project

Predicting Economic Indicators Using Open Data Sources

Evyatar, Daniel, Konniam

Problem

Economic indicators
Crucial for business decision making
(unemployment, consumer confidence)

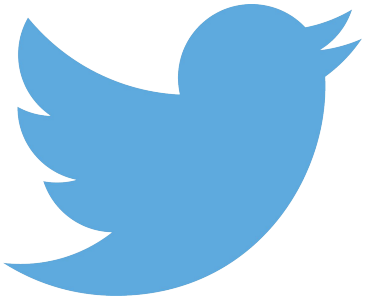
Issues
Infrequent updates
Rigid schema
Backward facing

Opportunity to create real-time and predictive indicators that provide value to stakeholders in the economy

Proposal

Data Sources

Twitter
Stock Prices
Initial Claims



Data Acquisition and Analysis

(frequency of keywords in tweets)

Outputs

Unemployment metrics
Dashboard of metrics

For this project we are implementing a subset of the overall proposal due to complexity and time constraints of the Twitter APIs

Data and Processing Dimensions

DATA

Size
6GB per day

Structure
JSON (semi-structured)

Velocity
1000 tweet/min

Source Latency
hourly updates

PROCESSING

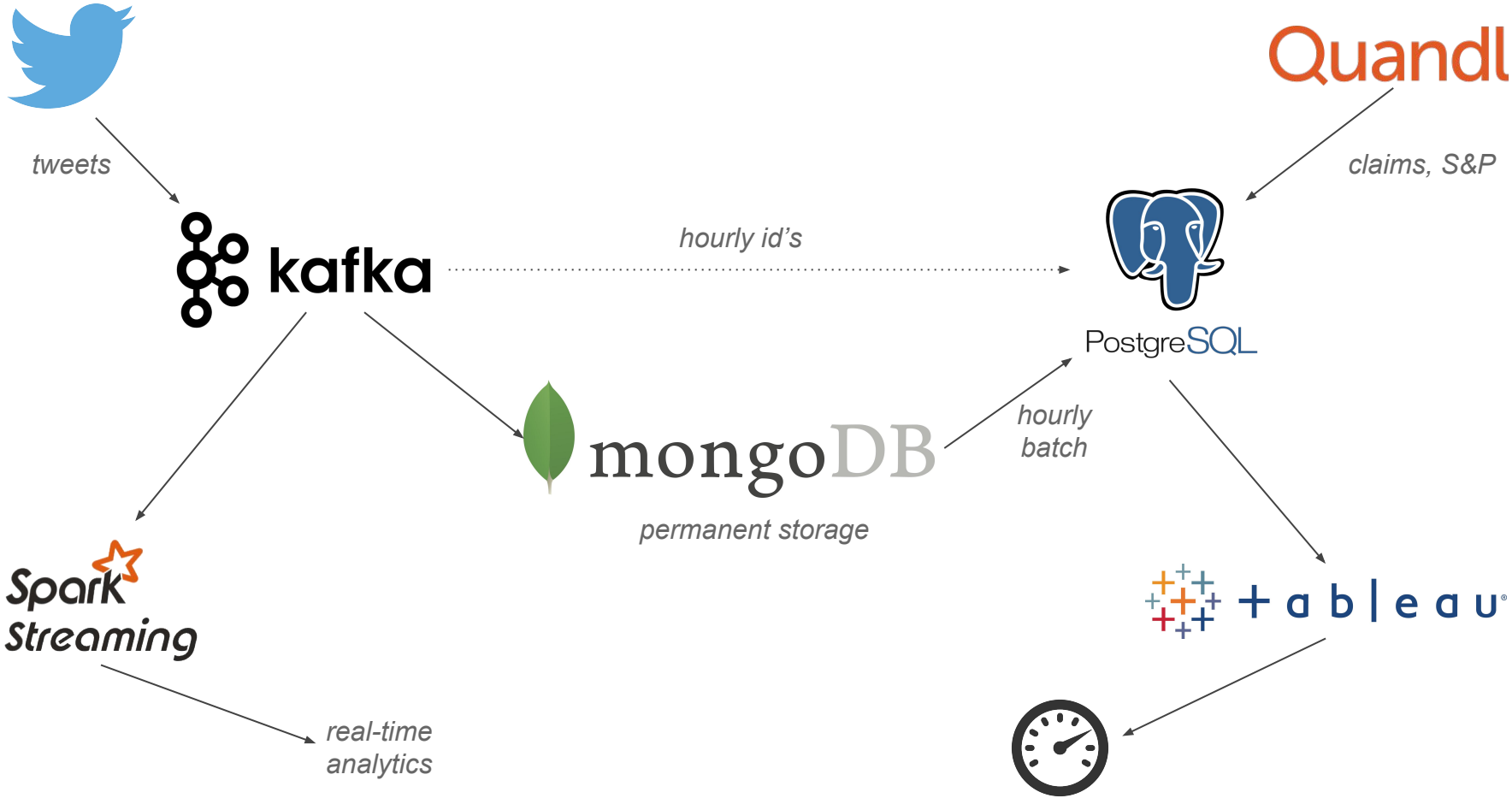
Selectivity
process every tweet

Processing Time
semi-batch

Aggregation
keyword counting

Precision
approximate is fine

Architecture



Choice of Tools

Kafka
Backbone of ingestion
Decoupled reading/writing
Distributed

MongoDB
Documents format
Horizontal sharding available

PostgreSQL
Convenient as serving layer
Great for stocks/economic data

Spark Streaming
Small batch approach

Setup

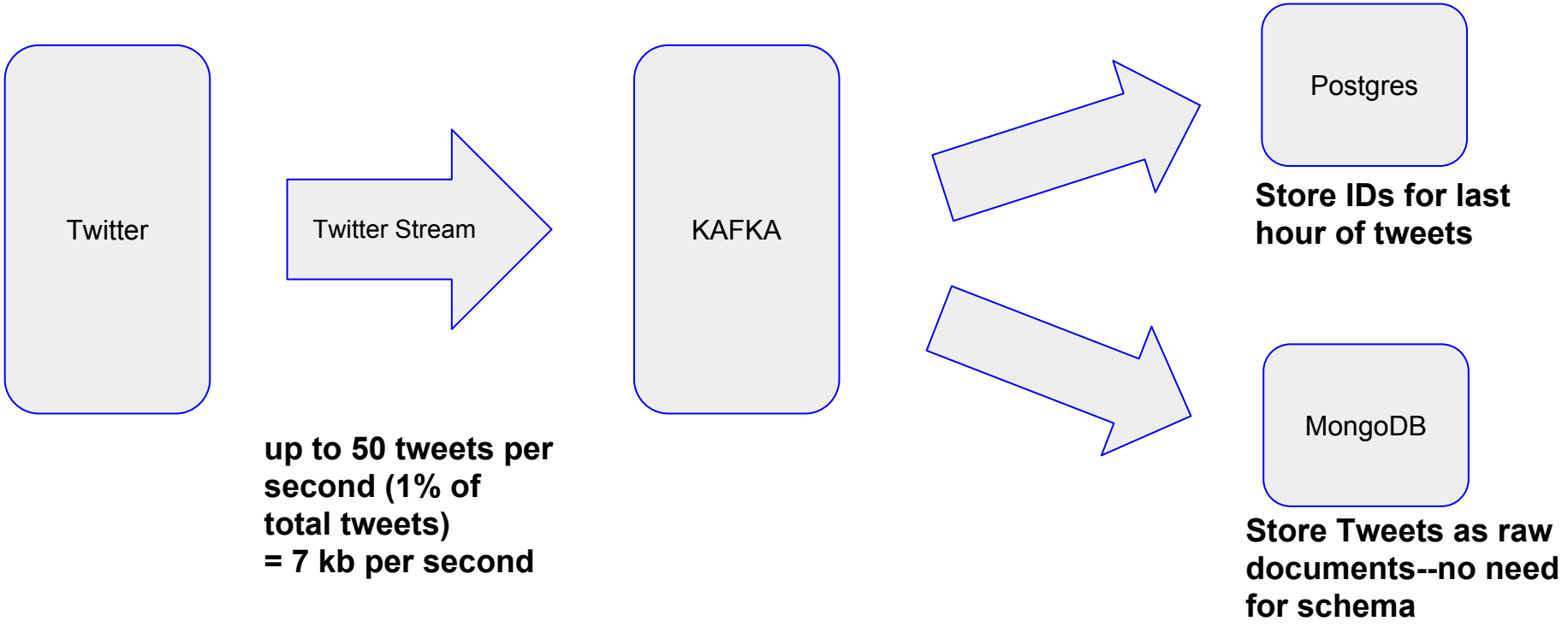
Amazon Web Services
m3.xlarge instance
15GB ram
400GB EBS (1.2K/3K IOPS)
Amazon Linux
Data Collection: 12/3-12/10

Software
Kafka
MongoDB
PostgreSQL
Spark

Python Packages
tweepy
kafka-python
pymongo
psycopg2
Quandl

Data Collected
10 Million Tweets
42GB

Data Ingestion - Twitter

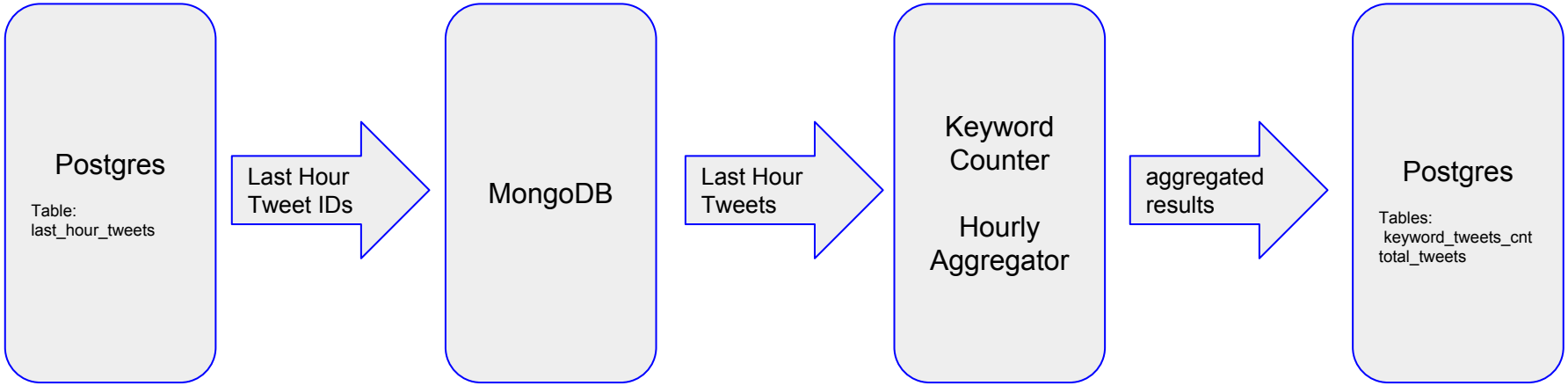


~/kafka_2.10-0.8.2.0 — -bash | java -Xmx512M -server -XX:+UsePar...ookeeper localhost:2181 --topic twitter | ...le Drive/MIDS/W205/economy-monitor/ingestion — python twitter_ingest.py

u0627\u0644\u0644\u0647\u060c\u0627\u0633\u062a\u0648\u062f\u0639\u062a\u0643 \u0627\u064a\u0651\u0627\u0647\u0627", "source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":783654877,"id_str":"783654877","name":"\u0646\u064a\u0644\u0648\u0631.","screen_name":"Nylwfr4","location":null,"url":null,"description":"http:\/\/sayat.me\/nylwfr44 \u0643\u064e\u0638\u0650\u0628\u0627\u0621\u0650 \u0645\u0643\u0629 \u0635\u064e\u064a\u062f\u064f\u0647\u064f\u0646\u0651 \u062d\u064e\u0631\u0627\u0645. http:\/\/ask.fm\/nylwfr44.","protected":false,"verified":false,"followers_count":6589,"friends_count":2329,"listed_count":6,"favourites_count":963,"statuses_count":25144,"created_at":"Mon Aug 27 02:45:35 +0000 2012","utc_offset":-28800,"time_zone":"Pacific Time (US & Canada)","geo_enabled":false,"lang":"ar","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/674408246464135169\/ByW6HJUc_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/674408246464135169\/ByW6HJUc_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/783654877\/1449626377","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":54,"favorite_count":18,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"ar"},"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[{"screen_name":"Nylwfr4","name":"\u0646\u064a\u0644\u0648\u0641\u0631.","id":783654877,"id_str":"783654877","indices":[3,11]}],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"ar","timestamp_ms":"1450167534658"}
{"created_at":"Tue Dec 15 08:18:54 +0000 2015","id":676677801739739136,"id_str":"676677801739739136","text":"Payday yay\ud83c\udf89\ud83c\udf89\ud83c\udf89","source":"\u003ca href=\"http:\/\/twitter.com\/#!\/download\/ipad\" rel=\"nofollow\"\u003eTwitter for iPad\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":423666485,"id_str":"423666485","name":"Lillie","screen_name":"Lilliefagan1","location":"England","url":null,"description":null,"protected":false,"verified":false,"followers_count":286,"friends_count":974,"listed_count":2,"favourites_count":155,"statuses_count":2665,"created_at":"Mon Nov 28 19:53:49 +0000 2011","utc_offset":0,"time_zone":"London","geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/378800000124981360\/e2865d41f6a546f91b1477e9d99f3134.png","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/378800000124981360\/e2865d41f6a546f91b1477e9d99f3134.png","profile_background_tile":true,"profile_link_color":"0084B4","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/660793876777693184\/S8UBYJit_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/660793876777693184\/S8UBYJit_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/423666485\/1434793607","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1450167534657"}
{"delete":{"status":{"id":444662280119009280,"id_str":"444662280119009280","user_id":2164310532,"user_id_str":"2164310532"},"timestamp_ms":"1450167534960"}}
{"created_at":"Tue Dec 15 08:18:54 +0000 2015","id":676677801743884288,"id_str":"676677801743884288","text":"binhowgan's video https:\/\/t.co\/VPxjAkD2Kb \u0627\u0644\u0634\u064a\u0639\u064a \u0648\u0627\u0644\u0639\u0631\u0636\u0647","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":455508394,"id_str":"455508394","name":"\u0623\u062d\u0645\u062f\u0627\u0644\u0632\u0647\u0631\u0627\u0646\u064a \u0623\u0628\u0648\u064a\u0632\u064a\u062f","screen_name":"ahmad2005223","location":"\u0645\u0643\u0629\/\u0627\u0644\u0634\u0631\u0627\u0626\u0639","url":null,"description":"\u062a\u0639\u0644\u064a\u0645 \u0645\u0643\u0629","protected":false,"verified":false,"followers_count":1963,"friends_count":2931,"listed_count":1,"favourites_count":368,"statuses_count":2339,"created_at":"Thu Jan 05 05:48:38 +0000 2012","utc_offset":10800,"time_zone":"Kuwait","geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/674853874604318721\/db-5tBLs_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/674853874604318721\/db-5tBLs_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/455508394\/1449255891","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/VPxjAkD2Kb","expanded_url":"https:\/\/instagram.com\/p\/_Q8h5uzNJC\/","display_url":"instagram.com\/p\/_Q8h5uzNJC\/","indices":[18,41]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"und","timestamp_ms":"1450167534658"}
{"created_at":"Tue Dec 15 08:18:54 +0000 2015","id":676677801756516352,"id_str":"676677801756516352","text":"Me ha gustado un v\u00edeo de @YouTube (https:\/\/t.co\/zIOZQbsLJl - \u00bfQu\u00e9 pasar\u00eda si salt\u00e1ramos todos a la vez en la Tierra?).","source":"\u003ca href=\"http:\/\/www.google.com\/\" rel=\"nofollow\"\u003eGoogle\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":239237974,"id_str":"239237974","name":"Joelfer Zav","screen_name":"JoelZB","location":"United States of America","url":"https:\/\/www.facebook.com\/JoelZuni731","description":"Me encantan @OneDirection, @LadyGaga, @ZacEfron \u2665INFINITE y @RogerGzz, y soy un gato :# \u2665take me to #Venus (8)","protected":false,"verified":false,"followers_count":142,"friends_count":181,"listed_count":3,"favourites_count":7544,"statuses_count":11582,"created_at":"Mon Jan 17 04:27:58 +0000 2011","utc_offset":-28800,"time_zone":"Tijuana","geo_enabled":true,"lang":"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"0099B9","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/548251393874608130\/ecst3rd5.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/548251393874608130\/ecst3rd5.jpeg","profile_background_tile":true,"profile_link_color":"4A913C","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/673952470813437956\/lax4TCbs_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/673952470813437956\/lax4TCbs_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/239237974\/1449517810","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/zIOZQbsLJl","expanded_url":"http:\/\/youtu.be\/yoV9eqtPP5Q?a","display_url":"youtu.be\/yoV9eqtPP5Q?a","indices":[36,59]}],"user_mentions":[{"screen_name":"YouTube","name":"YouTube","id":10228272,"id_str":"10228272","indices":[26,34]}],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"es","timestamp_ms":"1450167534661"}

Data Processing- Twitter

Hourly Batch Job



Retrieve tweet IDs from last hour

Retrieve tweets from last hour

Count key words:

```
[r'lost *(my|his|her|your|their)? *job',
r'got *fired',r'unemploy(ed)?',
r'la(y|id)?[ -]?off']
```

Store hourly tweets with key word
Store hourly total tweets

Claims and S&P500 Data

Quandl → PostgreSQL

```
w205project=> SELECT * FROM claims ORDER BY date DESC;
    date    | initial_claims
------------+----------------
 2015-12-05 |         384481
 2015-11-28 |         262628
 2015-11-21 |         305424
 2015-11-14 |         264816
 2015-11-07 |         291097
 2015-10-31 |         258436
 2015-10-24 |         245360
 2015-10-17 |         232860
 2015-10-10 |         256522
 2015-10-03 |         227176
 2015-09-26 |         215116
 2015-09-19 |         219342
 2015-09-12 |         198903
 2015-09-05 |         232507
 2015-08-29 |         230079
 2015-08-22 |         226649
 2015-08-15 |         229251
 2015-08-08 |         239326
 2015-08-01 |         224104
```

Serving

PostgreSQL

SQL users can query and get keyword trends

Join Twitter, S&P500, and claims tables

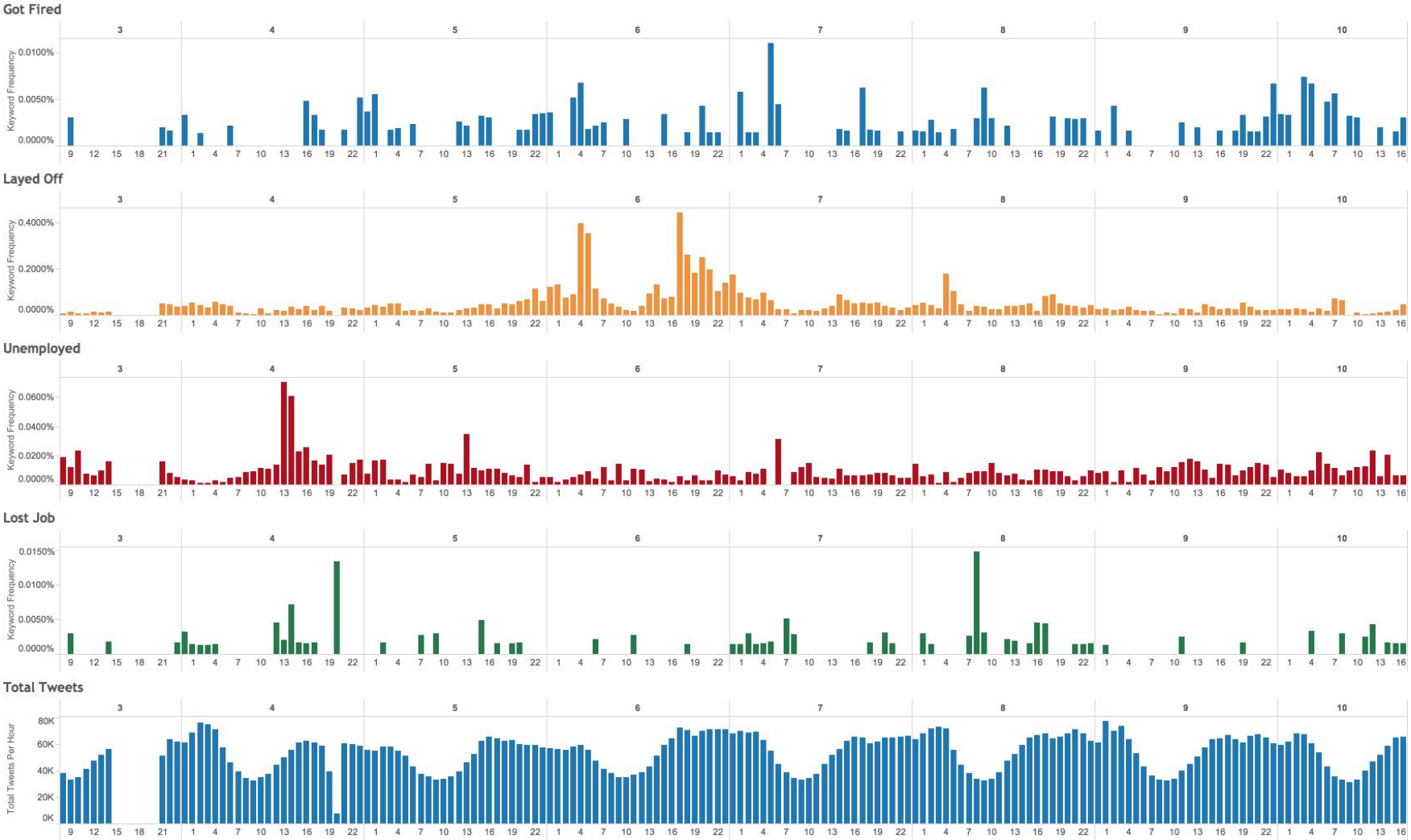
Tableau

Dashboard for quick overview

MongoDB

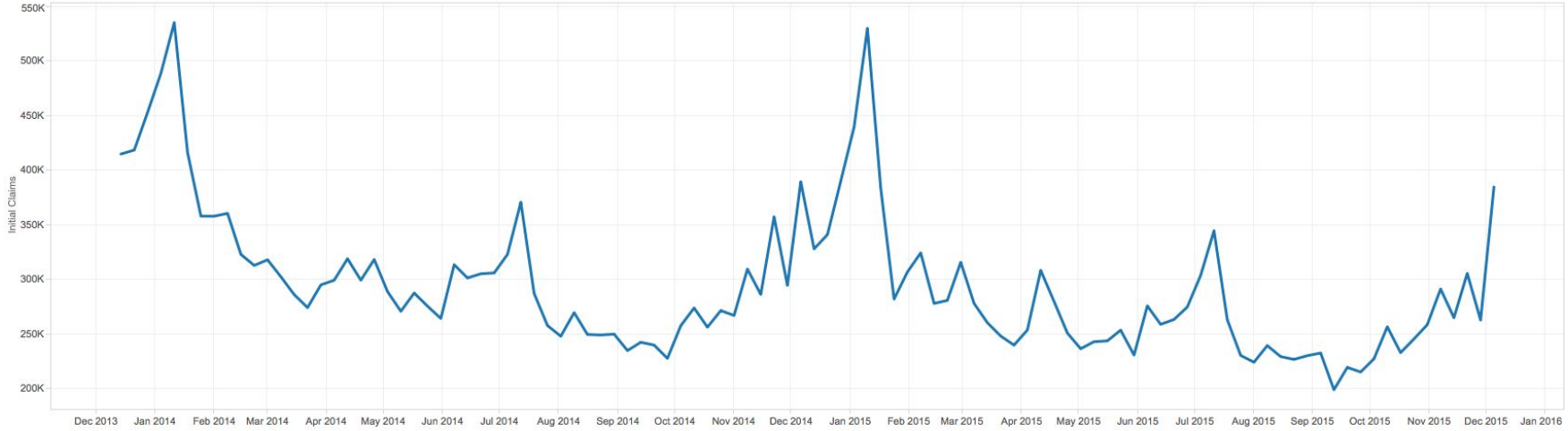
Use MongoDB for other specific queries (sort by # of hashtags)

Twitter Keyword Frequencies (Per Hour, 12/3-12/10)



Economic Indicators

Initial Jobless Claims



S&P500

Spark Streaming

search terms = [
'iphone',
'ipad',
'samsung',
'android']

```
-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
heavy duty hybrid rugged hard phone case cover for iphone 5c c+stylus+film us – bid now! o… https://t.co/ubhhcoukjk https://t.co/polulyavb7
rt https://t.co/nvyrilvbjq bell/virgin iphone 5s 32g mint https://t.co/frgx6dogoj
luxury original leather flip cover card wallet case for apple iphone 6 6s plus – bid now! … https://t.co/ubheu1oyib https://t.co/u0mnhywchf
https://t.co/tnevnp83kw #7443 apple iphone 5c 16gb "factory unlocked" 4g lte smartphone https://t.co/cyjbbbrl5e

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
4

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
0

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
samsung appealing $548m patent infringement bill in apple case https://t.co/uucxnywn27
samsung galaxy s7 to pack in pressure sensitive display, improved low-light ... – firstpost https://t.co/ic6altlliu #tech

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
2

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
i've collected 17,200 gold coins!  https://t.co/nmbdqn0chv #android, #androidgames, #gameinsight

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
1

-------------------------------------------
Time: 2015-12-15 00:31:30
-------------------------------------------
497
```

Next Steps

Machine Learning
Continue collecting tweets as training data
Predict jobless claims

Incorporate More Data
NYTimes articles, weather data