December 18, 2015

Project Shoot2Top Final Report

w205 Data Storage and Retrieval

John Bocharov, Max Shen, Alejandro Rojas

Executive Summary

Shoot2Top is a product to help video publishers optimize their social media strategy, by analyzing the social media performance of related videos in the space, and by extracting aspects of the videos and sharing strategy to maximize organic shares by the community.

Background

Back in 2011, well-known investor Mark Suster predicted that the future of the Internet will be video. His reasoning was very straight forward: in the US people spend more than 5 hours a day watching TV and less than 1 hour reading[1]. Extrapolating that proportion to online behavior provided a certain guidance that video will come to dominate the Internet.

In the last 5 years video growth has been phenomenal. A fact that has not been missed by major social networks. Facebook, Google and Twitter are all devoting significant resources to grow their video presence betting that capturing user's attention requires increasing the amount of video content at their disposal. Today, Facebook and Youtube claim to have billions of video views a day.[2]

Content creators have also been scrambling to churn out video content that strikes a note with their audiences. New media outlets like BuzzFeed and Upworthy are increasingly focusing on developing video content that becomes viral on social networks. With the rise of smartphones, consumers are also adding their set of video content to the mix. All this growth in video content publication and consumption is creating a need that Shoot2Top wants to attend.

---

[1] *See some discussion at Mark Suster's blogentry: http://www.bothsidesofthetable.com/2013/09/17/how-online-video-companies-can-increase-margin-and-build-better-businesses/*

[2] *See some discussion about video views on Facebook and Youtube: http://www.forbes.com/sites/edmundingham/2015/04/28/4-billion-vs-7-billion-can-facebook-overtake-youtube-as-no-1-for-video-views-and-advertisers/2/*

In the space there are startups like Datamnr[3] and Crowded Tangle[4] that are serving media companies with products that identify social media posts that are getting attention but none of them are exclusively focusing on video.

Questions to answer

Online media publishers need to know what type of video content is getting eyeballs across major social media networks like Facebook, Twitter and Youtube, Instagram, Vine and Snapchat.

Online media publishers need to also also be able to replay how their video content did when it was published on major social media networks against video content published by others.

Shot2Top wants to help media publishers answer what video content to publish? when to publish? and where?

Value proposition

Shoot2Top is building a platform to show live performances of videos posted on social networks, report ratings on how video content performed once it was published and predict how effective different video virality strategies may play out. Our platform essentially serves three modules that are part of a suite of products that serve different specific needs but all share similar data architecture and resources.

Product description

Shoot2Top continuously ingest videos published from major social networks like Twitter, Youtube and Facebook to analyze them real time to show media publishers how their videos are performing live. It also generates reports after-the-fact that highlights how each video and each media publisher performed. It also plans to use data collected to create analytical and machine learning tools to help predict optimal virality strategies.

Our suite of products include:
- Identifying when to post, which hashtags, and keywords to use for maximum virality
- Identifying key micro-trends in content that can enhance visibility
- Identifying opportunities to cross-promote your existing content based on current micro-trends

---

[3] *See article about Datamnr here:* *http://www.cjr.org/analysis/the_new_importance_of_social_listening_tools.php*

[4] *See article about Crowded Tangle here:* *http://www.fastcompany.com/3040951/the-secret-tool-that-upworthy-buzzfeed-and-everyone-else-is-using-to-win-facebook*

Analytics Platform

The analytics platform for Shoot2Top consist of three major components:

1. **Data Ingest Pipeline** that can consume, store, and replay social network streams
2. **Data Lake** for archiving streaming data for replay and repeat analysis
3. **Online Analytical Processing** for mining features and generating insight

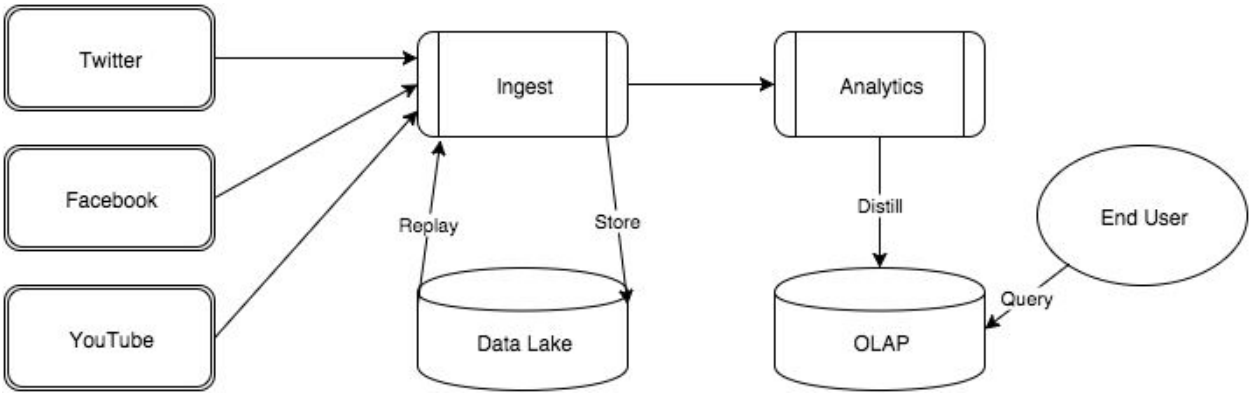
Our working hypothesis is: in order to maximize organic growth in eyeballs, shares, and followers, we had to act on micro-trends as they were happening. This meant our architecture had a **near-real-time** capability requirement, ruling out pure batch architectures.

Architecture

Motivated by the needs for near-real-time processing, we considered where the bottlenecks in the system are likely to arise. The biggest limiting factor is data acquisition: acquiring targeted streaming data from social networks as it happens is efficient are cost-effective (or free), whereas getting access to a large historical data set is cost-prohibitive.

The natural data format of our application is, therefore, streams. To develop and maintain analytics, we needed to store and be able to act on data we previously collected as it was happening. However, we wanted to avoid the cost of implementing the analytical platform twice: once in batch and once more in near-real-time. Solving for these constraints, we settled on a variant of the **Kappa Architecture**.

Production analytics happens in near-real-time, and algorithms development is enabled by the capability to **store** streaming data in the Data Lake, and the **replay** streams for development.



The technology choices that enable this were:

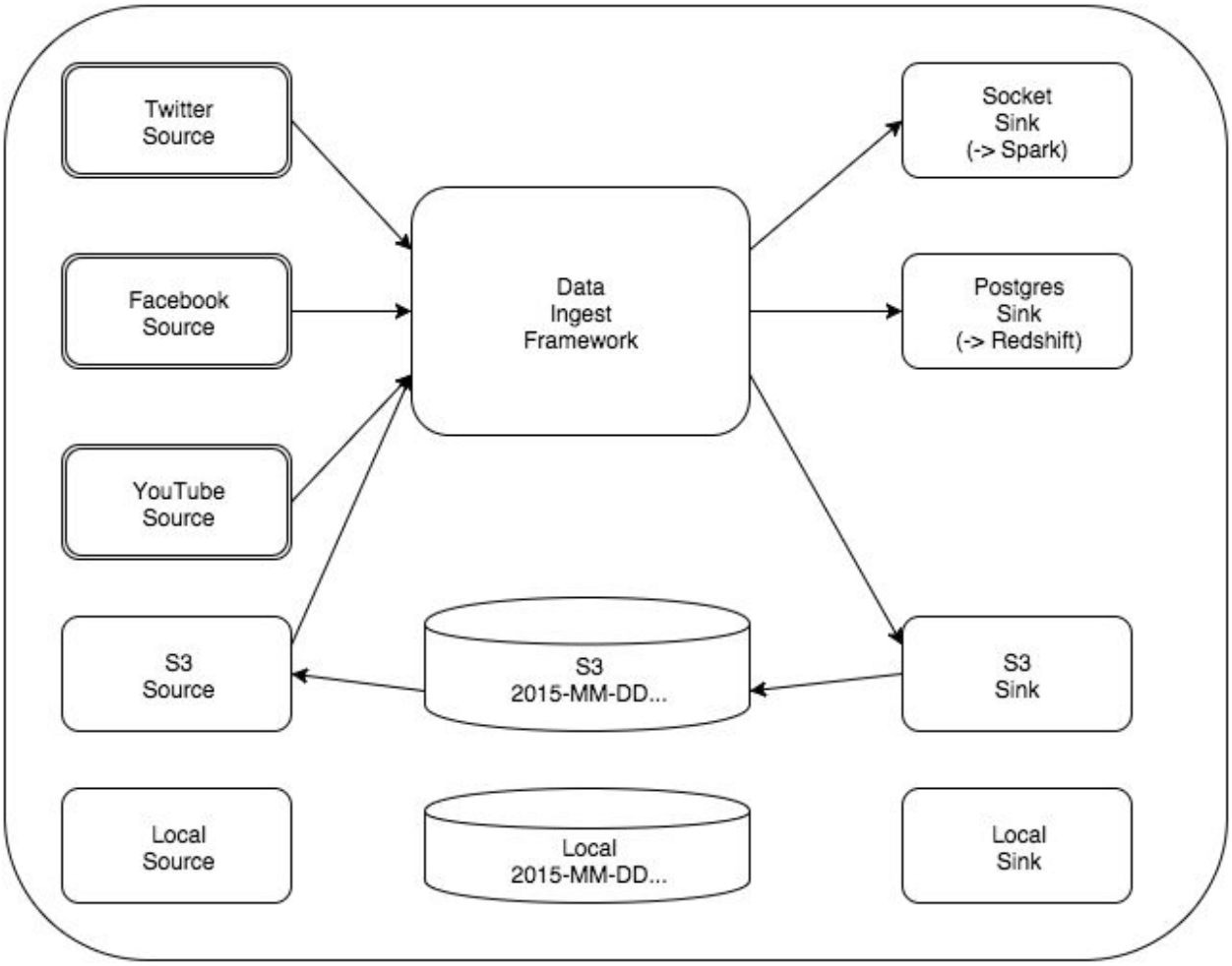
- **Ingest** - a lightweight configuration-driven Python framework (scale-out)
- **Data Lake** - S3 bucket with partitioned Hive table-like organization
- **Analytics** - lightweight Python framework (scale-out capable)
- **OLAP** - Postgres as a Data Warehouse (scaling up to Amazon Redshift DW)
- **End User** - the end user application is an IPython notebook connected to the DW

Data Ingest Pipeline

The Data Ingest Pipeline is responsible for streaming data from its primary sources (Twitter, Facebook, YouTube) or a replay source (S3, Local), and dispatching a stream to one or more streaming data sinks. The pipeline consists of several key components:

- The core **Data Ingest Framework**, which consumes a configuration file to instantiate at least one source and at least one sink, and wires them together by stream-chaining
- The **Streamable Object Primitive** - for which we chose a single social media post (tweet, post, or video, for Twitter, Facebook, and YouTube, respectively), represented as a self-describing JSON-equivalent Python object
  - For example a tweet is: { “tweet”: { “created_at” : “...”, “text” : “Hello, World”, ... } }
  - This enables the data to be context-free (since entity types are self-documenting and always contain timestamps) and the framework to be non-opinionated
- The **Stream** of posts, which is represented as a Python iterable of streamable objects
  - Enables clear semantics for consuming the stream (`for item in source:`)
- The **Sources**, which each emit a stream - if the underlying data is batched, the sources are responsible for mapping the batches to streams. The sources are:
  - **Twitter** - emits streams of tweetsbased upon topics of interest
  - **Facebook** - emits streams of posts based upon topics of interest
  - **YouTube** - emits streams of videos based upon topics of interest
  - **S3** - replays a stream captured to S3
  - **Local** - replays a stream captures to the Local filesystem
- The **Sinks**, which consume a stream - if the underlying target is batched (e.g. S3), the sinks are responsible for batching writes appropriately
  - **Postgres** - performs feature extraction and writes each post to Postgres/Redshift
  - **Socket** - send the stream to a local socket for ingest into Apache Spark
  - **S3** - captures a stream to S3 for future repeat analysis
  - **Local** - captures a stream to the Local filesystem for repeat analysis

The overall Data Ingest Pipeline component architecture is captured by the diagram below.



As requirements evolve, this framework becomes a natural candidate for migration to Streamparse, as all the sinks and sources are already de-coupled, and the coordinating component maps cleanly to a topology.

Data Lake

The data lake is stored in an S3 bucket. A structure similar to a partitioned Hive table in HDFS is used to ensure that multiple instances of the Data Ingest Pipeline can write to the Data Lake without coordination in advance and so that the data can be retrieved by partition.

Each instance of the S3 sink generates a unique folder name based on the timestamp the partition was created and a unique identifier. Inside each folder, the S3 sink generates sequential numbers for each micro-batch that it writes, similar to a MapReduce output. For example, three instances that all begin a new partition at midnight on December 18, 2015, will generate a structure like this:

```
/w205-social-media-dec <- S3 bucket
  /2015-12-18T00:00:00-afa5ac3a-e695-4f96-aca2-d24c7527aed7
    0000000
    0000001
    0000002
    ...
  /2015-12-18T00:00:00-761df5cd-ca48-412f-a860-6d01022ff9b6
    0000000
    0000001
    0000002
    ...
  /2015-12-18T00:00:00-b0126517-8996-4e89-b0af-6a5eca83c404
    0000000
    0000001
    0000002
    ...
```

This structure (especially the unique identifier) enables an arbitrary number of instances of the Data Ingest Pipeline to operate in parallel without advance coordination, and for consumers to replay a consistent partition across all source instances, for example by evaluating the glob `2015-12-18T00:00:00-*`. A side-channel partition index can be added if S3 list performance becomes a limiting consideration.

Local data for replay is stored in an equivalent structure in the file system, enabling sampling for development by downloading select files using standard tools (for example, s3cmd).

Each file in the Data Lake contains a single micro-batch of posts, organized as a JSON array for easy manipulation with standard tools. For example, a micro-batch from Twitter might looks as follows (white-space added for clarity).

```
[
  { “tweet”: { “created_at” : “2015-12-18T00:00:01”, “text” : “See you later, alligator!”, … }
},
  { “tweet”: { “created_at” : “2015-12-18T00:00:02”, “text” : “In a while, crocodile!”, … } },
  …
]
```

On-Line Analytical Processing

Our On-Line Analytical Processing layer extracts key features of each post and stores them in a table for aggregation and analysis. In the current implementation, feature extraction is integrated into the Postgres Sink.

This program contains two parts: first part of the program parses out key attributes for data analysis. For Twitter, the key attributes or “entities” are: created_time, userid, retweets, texts,

friend count, followers, and urls. The parsing program also serves as a simple ETL process that organizes JSON formatted file from Twitter into columnar format with updated data type so the data can be easily uploaded into a database for query later.

The second part of the program makes a connection to a Postgres database to: 1). create a table in a pre-defined database and insert data 2). if the table already exists, append new records to existing dataset (in other words, it's not the first time user launch the program ***PostgresDataIngestSink.py).***

Sample Analysis on Twitter data:
We let the data collection process run for few days and collected about 1.5G of data. Out the 1.5G of data, we select a sample file with 5,000 records for initial trend and statistical analysis.

Convert our data into Postgres database enables us to easily query the data. Having a database also allows easy connection and access for tools like R and Tableau.

Snapshot of Postgres table:
Below is a screenshot of sample data from the Postgres database:
*(SELECT * FROM tweets_large limit 30;).*

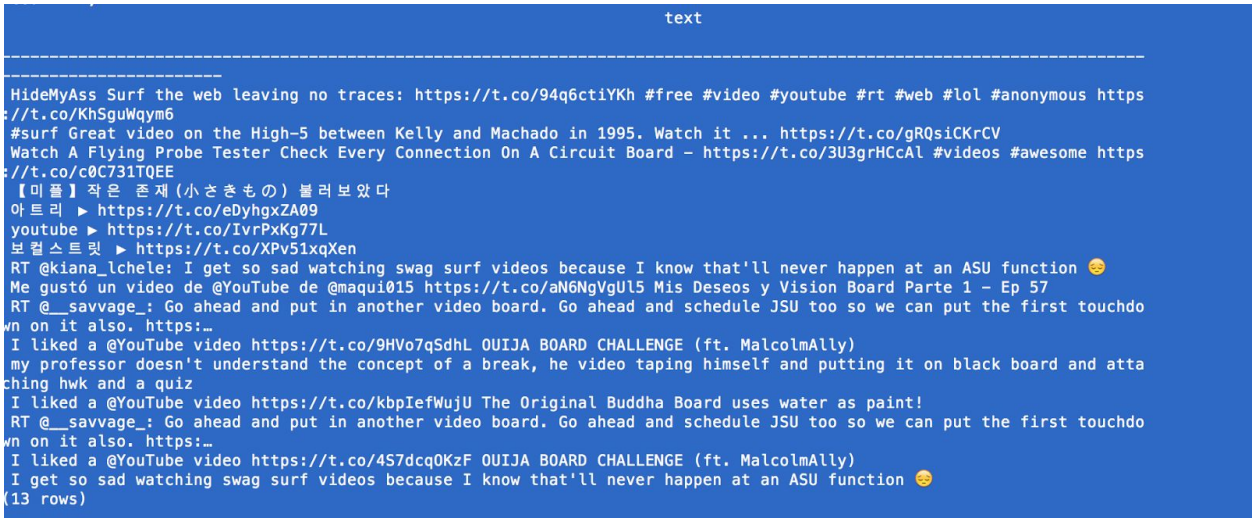
```
tcount=# SELECT * FROM tweets_large limit 30;
 index |      created_at      |        userid        | retweets |                                   text
                                                     | friendcount | followers |                          urls
-------+----------------------+----------------------+----------+-------------------------------------------------------------------------------------------
-----------------------------------------------------+-------------+-----------+---------------------------------------------------------------------------
     0 | 2015-12-18 05:43:52 | 677725950717657088 |        0 | ＣＭ動画制作サービス開始しました  https://t.co/drpXgmzMDB
                                                     |        4174 |      3691 | http://www.becreates.net/
     0 | 2015-12-18 05:43:53 | 677725952428802048 |        0 | RT @andrettixx: TRUEST SHIT IN THE WORKD https://t.co/UFOqlw9bfb
                                                     |         665 |       980 |
     0 | 2015-12-18 05:43:53 | 677725953708060672 |        0 | RT @nicaignacio33: "@ALDUBTalks: Now on board going to Davao! #ALDUBTheSearch https://t.co/
ZTJzbIdgow"                                          |          27 |        21 |
     0 | 2015-12-18 05:43:53 | 677725955125825536 |        0 | RT @KardashianReact: if DJ Khaled can update his story while lost at sea on a jet ski at ni
ght I think you can text back                        |         290 |       304 |
     0 | 2015-12-18 05:43:54 | 677725956740546560 |        0 | RT @BuckFitches420: They don't wanna see you jet ski
                                                     |         811 |      1458 |
     0 | 2015-12-18 05:43:54 | 677725957751504897 |        0 | DEMAIN JE PARS AU SKI PUTAIN !!€€(?!7€!!
                                                     |         437 |      1290 |
     0 | 2015-12-18 05:43:54 | 677725958636478464 |        0 | RT @DJKhaledSpeaks: If DJ Khaled can update his story while lost at sea on a jet ski at nig
ht I think you can text back                         |         318 |       321 |
     0 | 2015-12-18 05:43:55 | 677725963455733760 |        0 | The LITuation! https://t.co/Kz085ejKvv
                                                     |         653 |      1145 |
     0 | 2015-12-18 05:43:56 | 677725965263486976 |        0 | RT @washingtonpost: Washington Post Editorial Board: "For Republicans, bigotry is the new n
ormal" https://t.co/ZJ3wnCVh8a                       |         264 |       419 | http://effervescentlights.net
                                           https://t.co/TzhxcFI6iQ . 😍 جنتل حتى في رسايله | 0        | 677725966169427969 | 05:43:56 2015-12-18 | 0
                                                     |         110 |       401 |
     0 | 2015-12-18 05:43:56 | 677725967301890048 |        0 | RT @barishtian: Düşünsene california da doğmuşsun elinde surf tahtası.türkiye dediklerinde
"bilmiyorum ortadoğuda galiba" diyorsun.ne güzel… |          20 |       308 |
     0 | 2015-12-18 05:43:56 | 677725967767465984 |        0 | RT @addictedtoCFC: The board got it Wrong the players are to blame. #CFC https://t.co/PIaSR
NY7RG                                                |         290 |       529 | http://pkcoopie.tumblr.com
     0 | 2015-12-18 05:43:58 | 677725976017674241 |        0 | RT @DatGirlMoss: This is so beautiful 😭😭 RT @deray: Swag. Surf. 🔥 http://t.co/MaMZCOoi0b
                                               |         653 |      1145 |
     0 | 2015-12-18 05:43:59 | 677725978676830208 |        0 | Mick Fanning praised for bravery in World Surf League finale at Pipeline Masters https://t.
co/5psbEIHLfk                                        |        2187 |      1495 | http://news.anotao.com
     0 | 2015-12-18 05:44:01 | 677725985530232832 |        0 | RT @VivaArtists: OTWOL taping Baby girl @hellobangsie enjoys the hover board like hubby ❤️#
```

Here are some sample statistics we observe from this dataset:

friendcount: max 255,595, min 0, average friend count is 1,478.
followers: max 18,327,567, min 0, average followers count is 9,765.

When we conduct a search on tweets contains the keyword "video" or "youtube", we found 13 records in our sample set of 5,000 records. We will expect to include other keywords that will allows us to identify those tweets that contain video links or video post attached to them.
SQL Syntax:

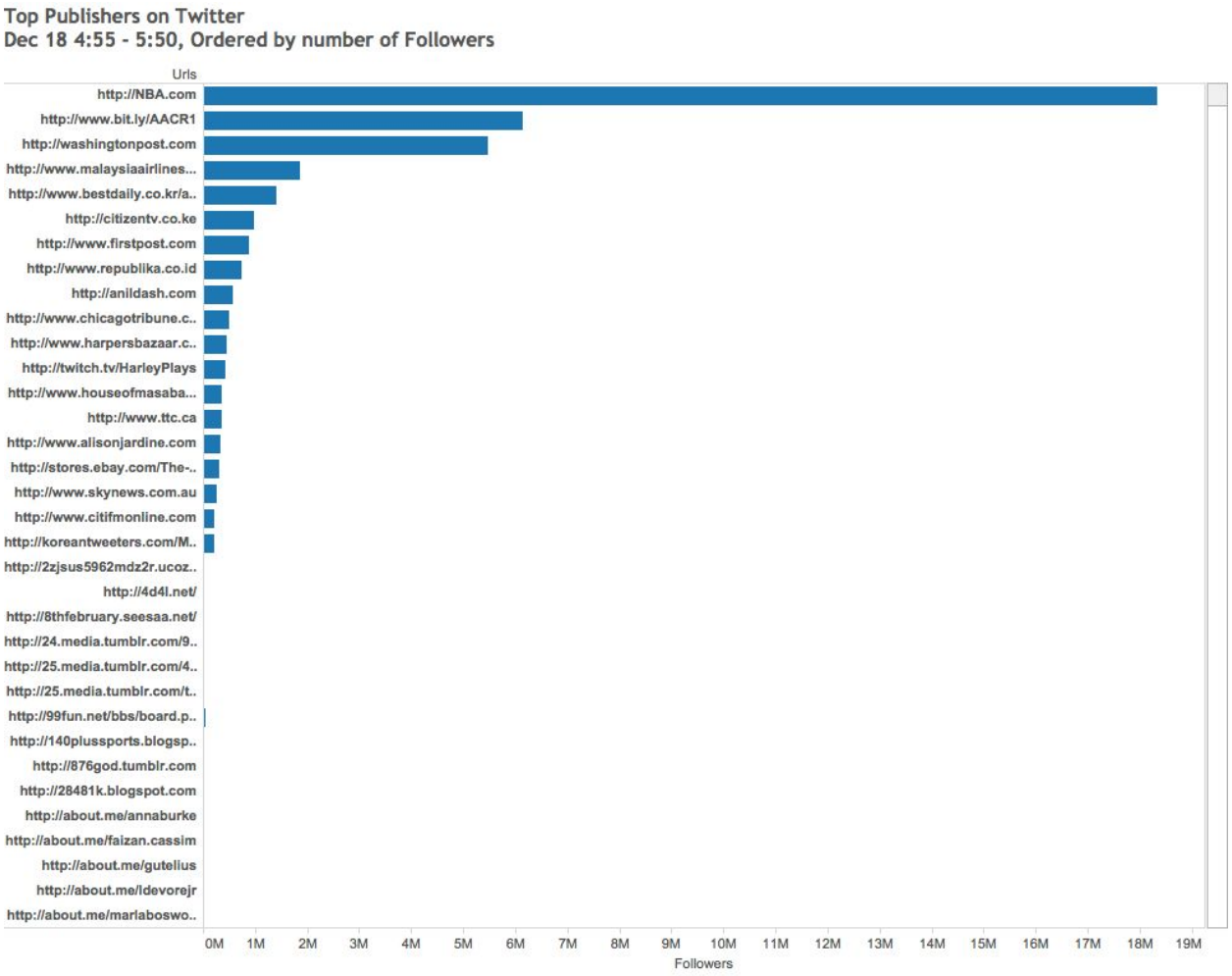
SELECT text FROM tweets_large WHERE (text like '%video%' or text like '%youtube%');



```
                                                                text

------------------------------------------------------------------------------------------------------------------------
----------------------
 HideMyAss Surf the web leaving no traces: https://t.co/94q6ctiYKh #free #video #youtube #rt #web #lol #anonymous https
://t.co/KhSguWqym6
 #surf Great video on the High-5 between Kelly and Machado in 1995. Watch it ... https://t.co/gRQsiCKrCV
 Watch A Flying Probe Tester Check Every Connection On A Circuit Board - https://t.co/3U3grHCcAl #videos #awesome https
://t.co/c0C731TQEE
 【미플】작은 존재(小さきもの) 불러보았다
 아트리 ▶ https://t.co/eDyhgxZA09
 youtube ▶ https://t.co/IvrPxKg77L
 보컬스트릿 ▶ https://t.co/XPv51xqXen
 RT @kiana_lchele: I get so sad watching swag surf videos because I know that'll never happen at an ASU function 😔
 Me gustó un video de @YouTube de @maqui015 https://t.co/aN6NgVgUl5 Mis Deseos y Vision Board Parte 1 - Ep 57
 RT @__savvage_: Go ahead and put in another video board. Go ahead and schedule JSU too so we can put the first touchdo
wn on it also. https:…
 I liked a @YouTube video https://t.co/9HVo7qSdhL OUIJA BOARD CHALLENGE (ft. MalcolmAlly)
 my professor doesn't understand the concept of a break, he video taping himself and putting it on black board and atta
ching hwk and a quiz
 I liked a @YouTube video https://t.co/kbpIefWujU The Original Buddha Board uses water as paint!
 RT @__savvage_: Go ahead and put in another video board. Go ahead and schedule JSU too so we can put the first touchdo
wn on it also. https:…
 I liked a @YouTube video https://t.co/4S7dcqOKzF OUIJA BOARD CHALLENGE (ft. MalcolmAlly)
 I get so sad watching swag surf videos because I know that'll never happen at an ASU function 😔
(13 rows)
```

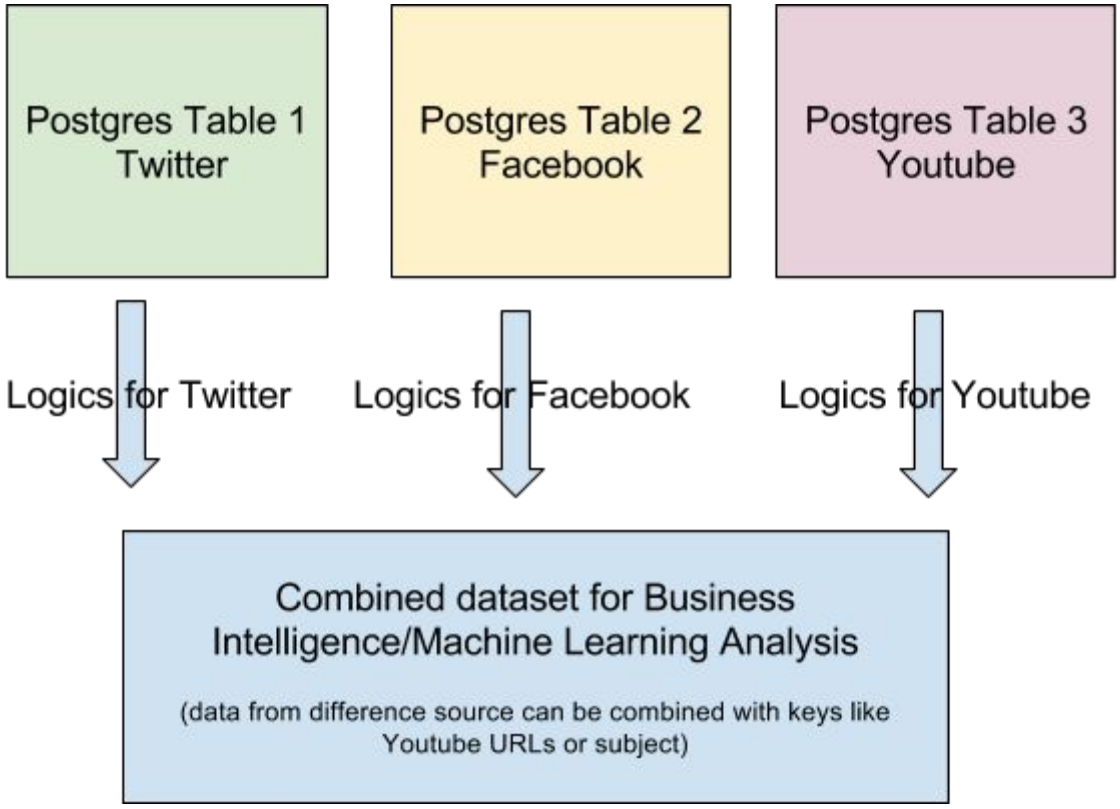
Sample Report: Assessing top publishers on Twitter at a specific time window.



Analysis integration and future improvements

Currently, the Twitter piece of the project is fully functional. The ultimate goal is to collect information from Facebook and Youtube, repeat the process above to store data in separate Postgres database. We can the query, combine, manipulate, and the data for in depth analysis.

Enhancement for analytics platform



The combined can be fed into a statistical tools like R or Spark; utilizing advanced techniques like regression or machine learning to provide more insights for the end users. Given the time constraint, the project will not reach this stage. However, we intend to enhance our application so it can reach this stage eventually.

Resources

The code for Shoot2Top can be found at:
https://github.com/W205-Social-Media/w205-data-ingest

The output (end user iPython notebook) can be found at:
https://github.com/W205-Social-Media/w205-data-ingest/blob/develop/Shoot2Top.ipynb