Twitter Spam & Fraudulent Websites

Kyle Hamilton, Carlos Rodriguez, and Sharmila Velamur


300M ACTIVE USERS
5% ARE SPAM BOTS
9,100 TWEETS PER SECOND


SNAKE
OIL

Concept stage

Browser Plug-in

Dashboard

Distributed Workflow & Scheduling Coordinator

Data Storing & Retrieval:

Hive,
NoSql (or)
RDBMS

Data Analysis & Precomputation:

Text Analysis
- Scheduled Job

Data Ingestion:

Custom Implementation -
Scraping
- Scheduled Job

HDFS - Batch Layer

52.7.131.84:10000/dashboard/

yourtrustedhacks.com

Many people are tired of filling out the surveys on those fake hack sites. We completely understand that so we created this hack site. We wanted to establish a hacker community with **no BS approach!**
Even my dad is a fan of this hack site, so you KNOW it's cool! :)

Here's the list of the recently updated hacks as of today:
[UPDATED!] Clash of Clans Unlimited PRO (v15.9.1): Your Trusted Clash of Clans Hack
[UPDATED!] PayPal Unlimited PRO (v12.4.2): Your Trusted PayPal Hack
[UPDATED!] Facebook Account Hacker PRO (v16.7.2): Your Trusted Facebook Hack
[UPDATED!] Twitter Unlimited PRO (v14.5.3): Your Trusted Twitter Hack
[UPDATED!] InstaGet PRO (v17.5.2): Your Trusted Instagram Hack
[UPDATED!] Hay Day Unlimited PRO (v17.8.3): Your Trusted Hay Day Hack
[UPDATED!] Dragon City Unlimited PRO (v12.5.4): Your Trusted Dragon City Hack
[UPDATED!] Cracked Steam PRO (v13.7.1): Your Trusted FREE Steam Games

Here's the list of the recently updated codes as of today:
[UPDATED!] Your Trusted FREE iTunes Gift Card Codes
[UPDATED!] Your Trusted FREE Steam Wallet Codes

Enjoy hacking! :)

Review Overview

Integrity	★★★★★
Credibility	★★★★★
Reliability	★★★★★
Trustworthiness	★★★★★

Search

YTH Achievement Awards

Follow us!

Click here to go to the download links section instantly!

Highlight suspicious content

Project home

Honeypot

Following/Following

Recent Activity

Following/Following

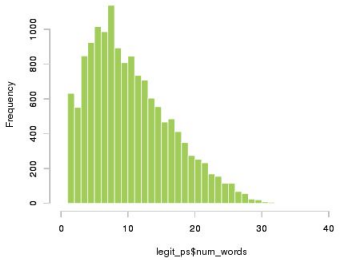
Word counts

Tweet counts

Word count

N = 14457

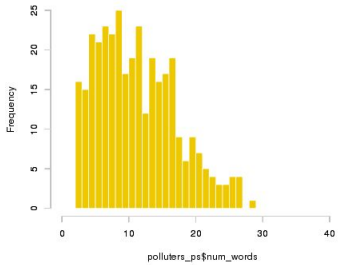
Legitimate Users



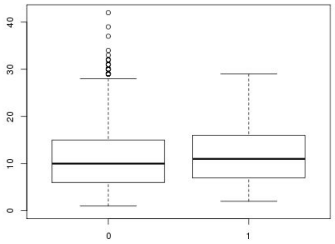

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00   10.00   10.89   15.00   42.00
```

N = 341

Content Polluters

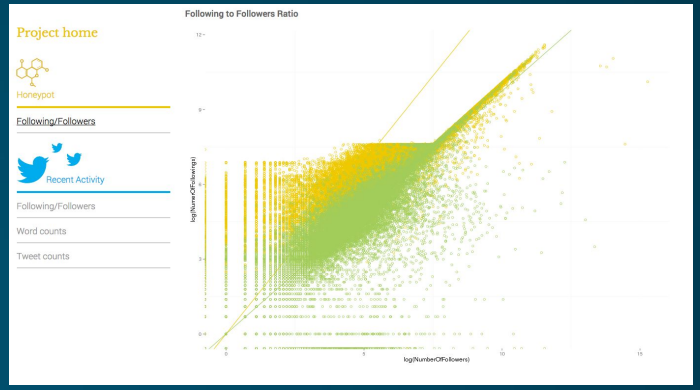


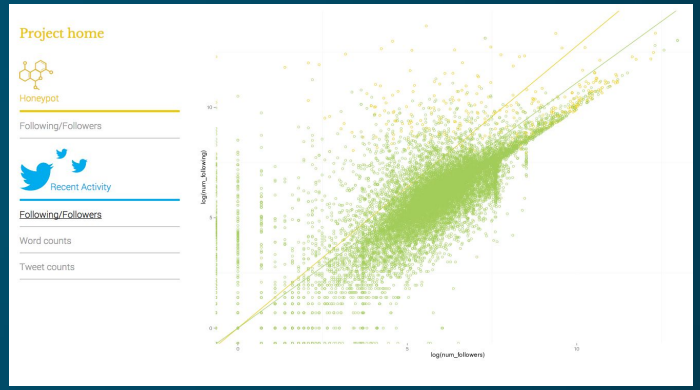

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    7.00   11.00   11.74   16.00   29.00
```



Legitimate Users Vs Polluters

	Dependent variable:
	is_polluter
num_words	0.022**
	(0.009)
Constant	-3.993***
	(0.114)
Observations	14,798
Log Likelihood	-1,619.610
Akaike Inf. Crit.	3,243.221
Note:	*p<0.1; **p<0.05; ***p<0.01



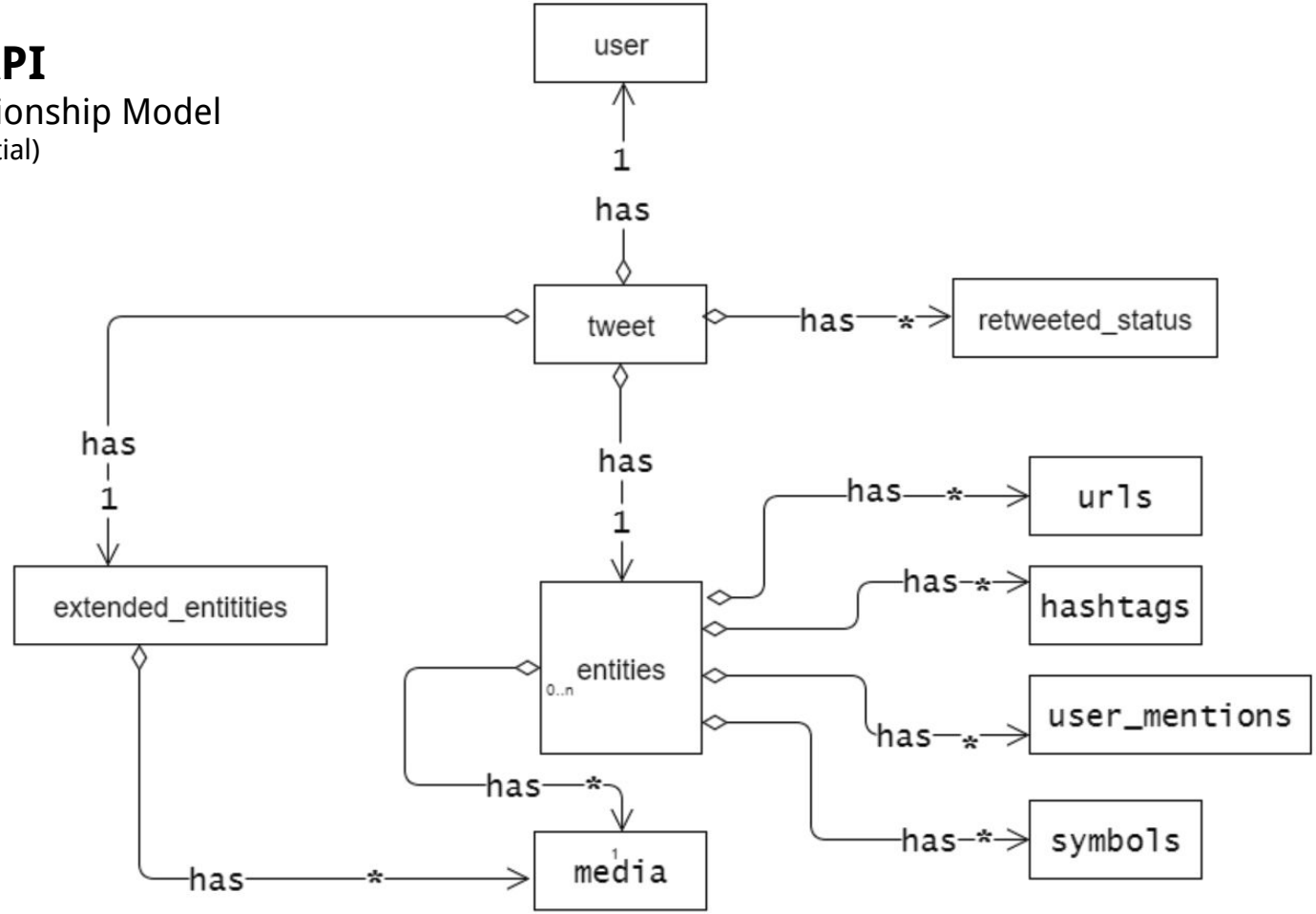


ARCHITECTURE

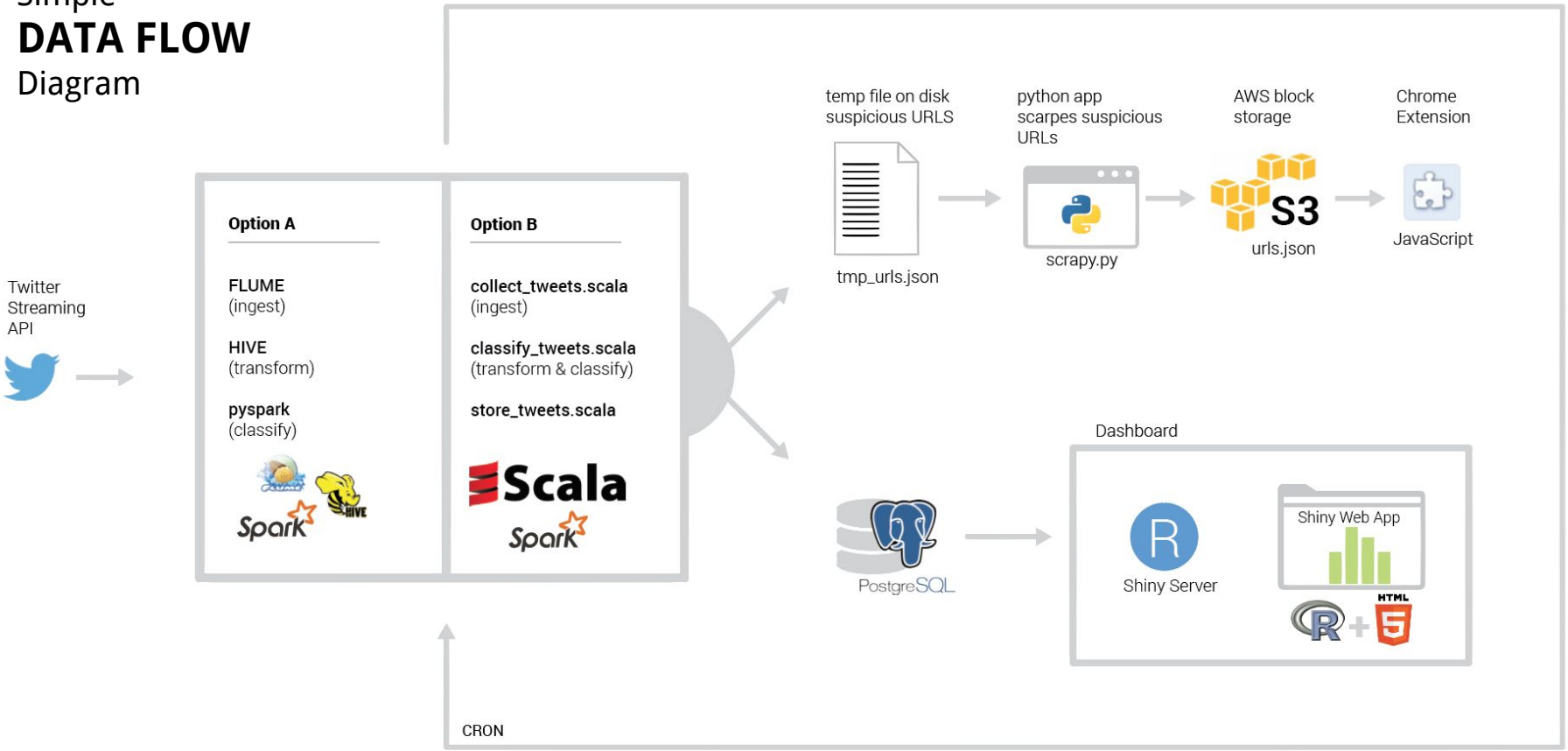
Twitter API

Object-Relationship Model

(Selective & Partial)

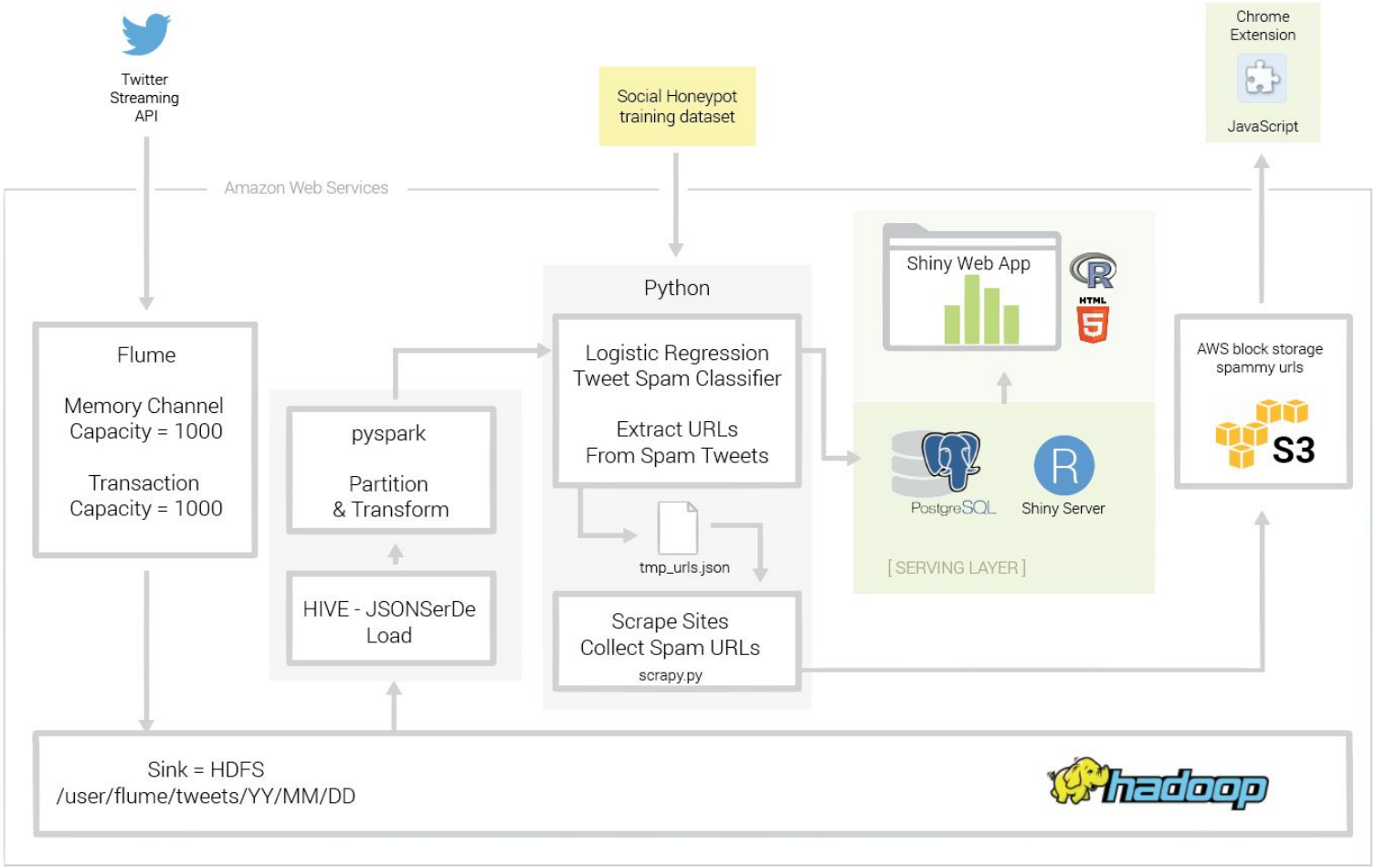


Simple
DATA FLOW
Diagram



OPTION A
Architecture

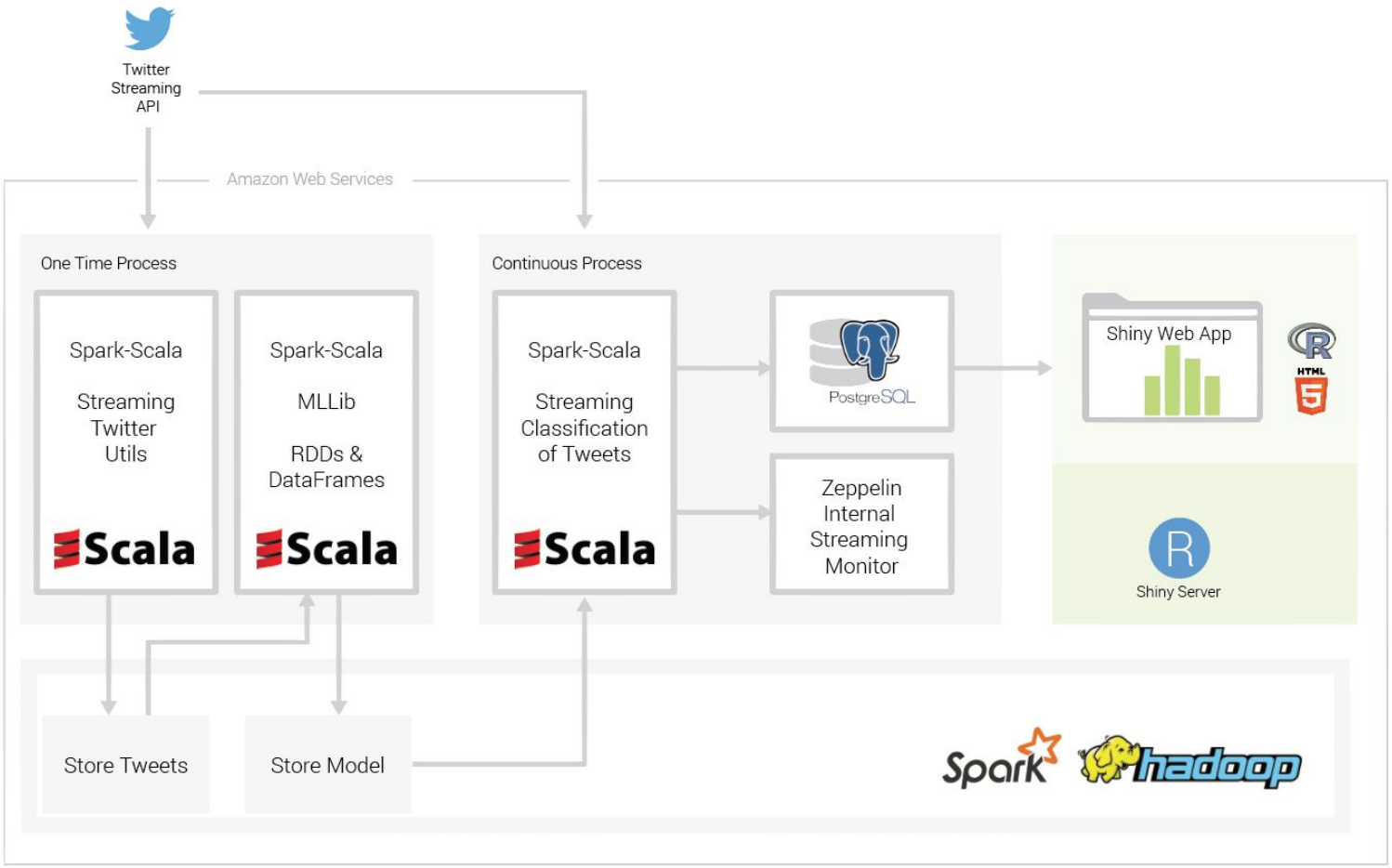
Hadoop-HDFS
Flume
Hive
Postgres
Spark-SQL
Spark-Pyspark
S3
Cron
Shiny App
(R, HTML5)
Javascript



OPTION B
Architecture

Hadoop-HDFS
Spark-Scala
Streaming Lib
Twitter Utils
MLLib/ML
Spark-JDBC
Shiny App
(R, HTML5)
Zeppelin

Optimization:
Spark RDDs
Scala (JVM)
Streaming



GO-DAWG-GO!

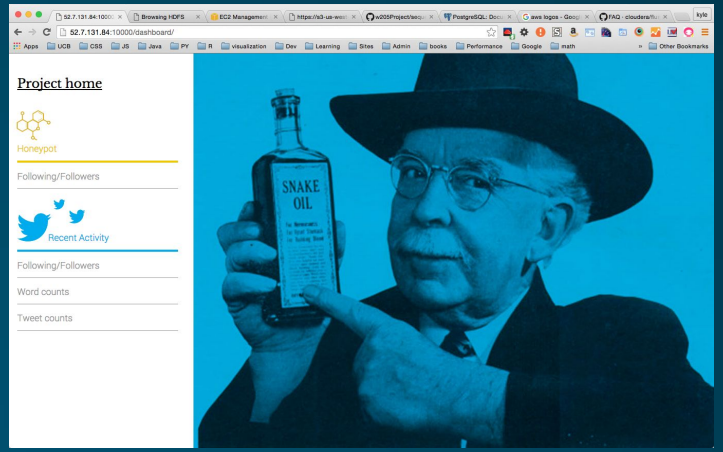
w205Project_V2.0 AMI

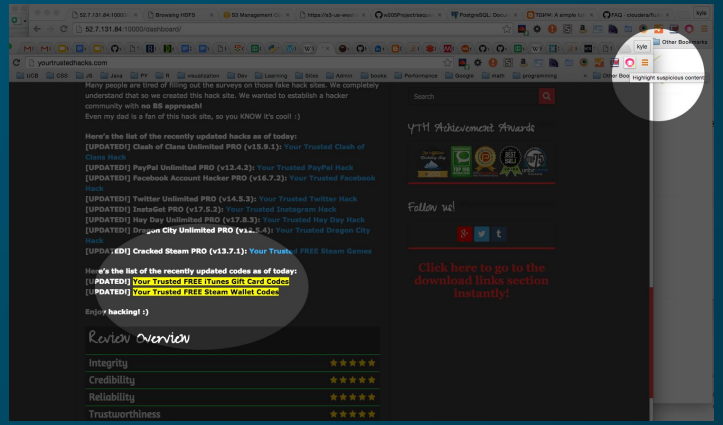
```
wget https://s3-us-west-2.amazonaws.
com/w205twitterproject/provision.sh

. provision.sh

. bootstrap.sh

crontab simple_sched_cron.txt
```





FUTURE WORK

- Clustering and parallelization; using EMR.
- Recovering from errors automagically.
- Tweaking config such as heap space and other JVM parameters.
- RESTful API for dashboard as well as browser plug-in.
- Better classification algorithms.
- Magnify “spam” URL extension
- Utilizing the same data pipeline for multiple areas:
  - Online marketing - semantic analysis.
  - Political campaigns - popular topics, alternative polling.
  - Social experiments on user behavior.
  - Geolocation based analysis.

ANY QUESTIONS?

THANK YOU