November 22, 2015

Project Progress Report

w205 Storing and Retrieving Data

Kyle Hamilton, Carlos Rodriguez and Sharmila Velamur

Hypothesis:

By scanning twitter public stream data we can identify suspicious websites proliferated by alleged spammers in social media.

Project Goals:

- Develop a big data solution to identify suspicious websites by scanning for spam in twitter public stream data.
- Develop a dashboard to view and explore pertinent aspects of the data.
- Develop a browser plugin that can highlight suspected user tweets and links to fraudulent websites.

The Environment:

- We built a new AMI upon the UCB W205 AMI3 and added the following software:
  - Python 2.7 and virtualenv
  - Scrapy
  - Flume

Architecture Updates:

We refined the architecture and aligned the components based on our project goals. See appendix-1 for the updated architecture diagram.

Data Ingestion:

We chose Flume to read the public tweets from the twitter stream. Flume allows for the collection, aggregation, and movement of large volumes of data in an efficient and reliable way. Also, the concept of a source, channel and sink within flume's architecture is intuitive and straightforward to configure.

Data Processing Pipeline:

In this part, we will apply necessary transformations to identified fields in the stored dataset. Then we will use an algorithm based on other ML papers, to identify tweets that have a high likelihood of being spam[1]. We will then consider the URLs from these tweets for scraping and subsequently, crawling. If a particular URL is classified as Spam by our algorithm, then we will consider the discovered, associated URLs as suspicious. Other strategies such as reverse DNS lookups and searching through known lists of spam sites are also being considered for classification of URLs[2].

The committed scope of this project is to use fewer (user as well as tweet related) attributes in the initial filtering stages (processing data on a daily cadence to identify suspicious tweets) and simple classification heuristics to identify suspicious tweets and the associated URLs.

Workflow Manager:

The workflow manager has two tasks in this system:

1. Manage data ingestion from twitter throughout the day to maintain a predetermined data growth rate.
2. Kick-off data processing pipeline at regular intervals. These intervals will be determined based on need as well as the time taken to complete the data processing pipeline.

We are using the Cron daemon for scheduling. We don't see a need for a DAG scheduler at this point.

External Systems:

External systems interfacing with the architecture are: twitter streaming API and browsers.

Project Timeline:

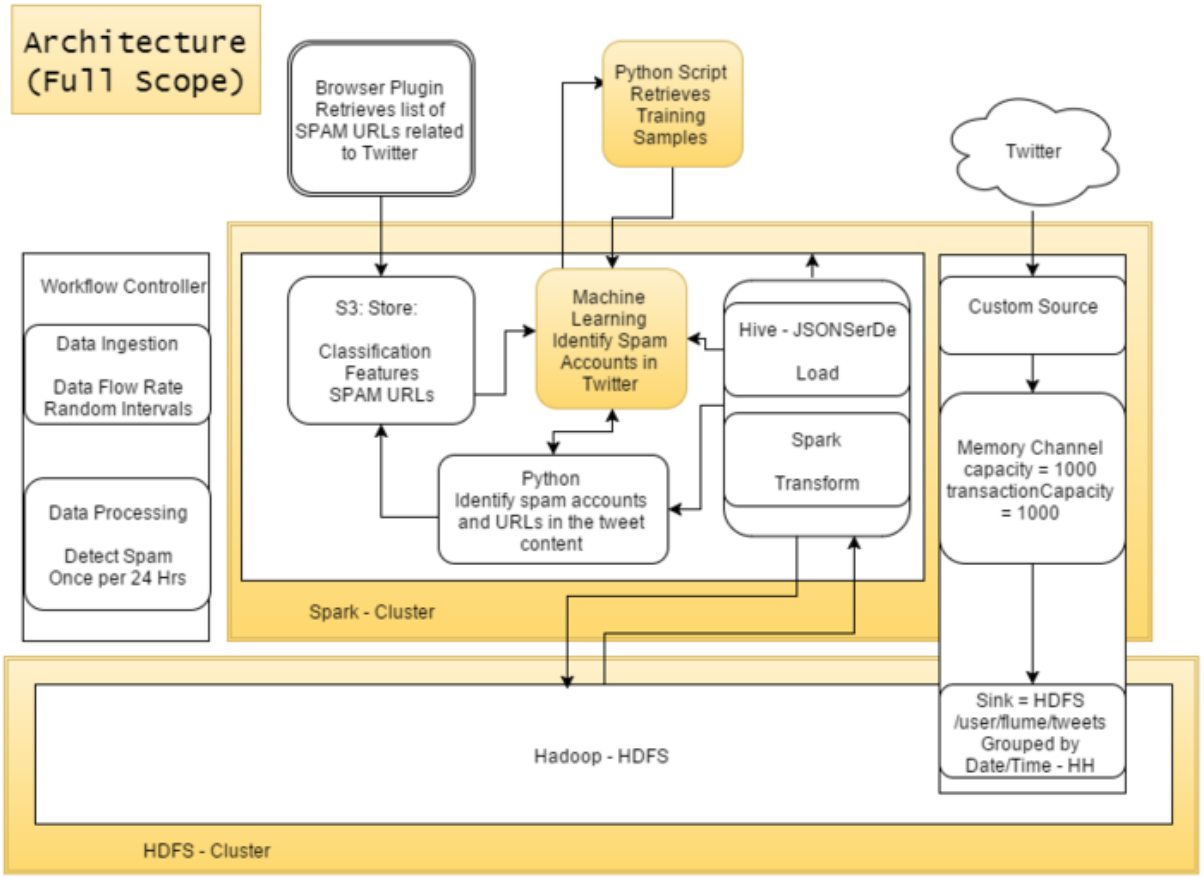
Week	Tasks	Status
10/19 - 10/31	● Configure Dev environment (non-cluster) ● Get twitter data and store in HDFS	Completed
11/01 - 11/07	● Twitter API analysis; identify key fields ● Load into Hive ● Python scraper set up ● Progress Report	Completed
11/08 - 11/14	● Cron setup - workflow manager ● Non ML based minimal scope classification of URLs as spam or not. Store Spam URLs in S3	Work in Progress Work in Progress
11/15 - 11/28 (2 weeks)	● Complete (minimal scope) python coding ● Browser plugin and dashboard ● End to end integration ● Identify optimization needs and opportunities	Work in Progress Complete/Not Started Not Started Not Started
11/29 - 12/05	● Optimize, review code and config, test ● Project presentation	Not Started

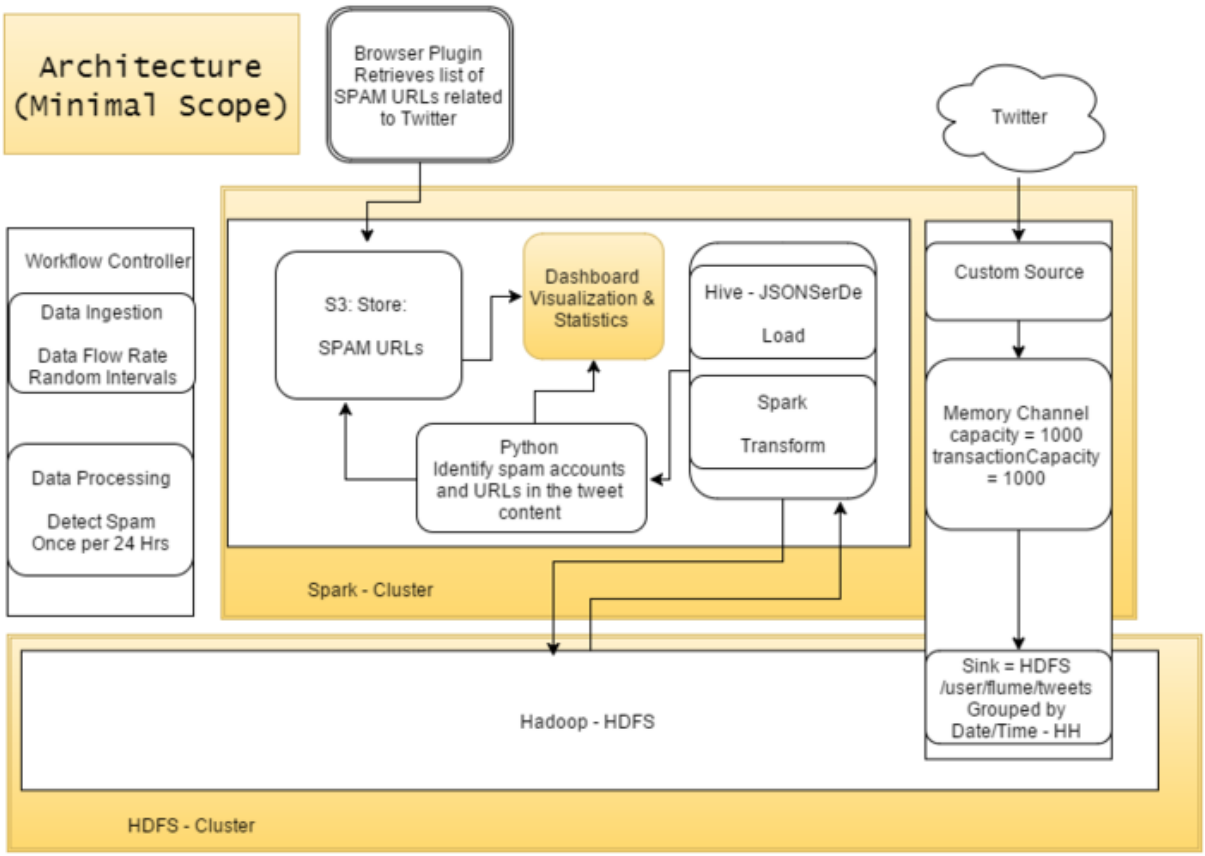
---

[1] http://www.wseas.us/e-library/conferences/2014/Florence/CSCCA/CSCCA-23.pdf

[2] http://www.icir.org/vern/papers/twitter-susp-accounts.imc2011.pdf

Appendix - 1


Architecture (Full Scope)
Browser Plugin Retrieves list of SPAM URLs related to Twitter
Python Script Retrieves Training Samples
Twitter
Workflow Controller
Data Ingestion
Data Flow Rate Random Intervals
Data Processing
Detect Spam Once per 24 Hrs
S3: Store:
Classification Features SPAM URLs
Machine Learning Identify Spam Accounts in Twitter
Hive - JSONSerDe
Load
Spark
Transform
Python
Identify spam accounts and URLs in the tweet content
Custom Source
Memory Channel capacity = 1000 transactionCapacity = 1000
Spark - Cluster
Sink = HDFS /user/flume/tweets Grouped by Date/Time - HH
Hadoop - HDFS
HDFS - Cluster


Architecture (Minimal Scope)
Browser Plugin Retrieves list of SPAM URLs related to Twitter
Twitter
Workflow Controller
Data Ingestion
Data Flow Rate Random Intervals
Data Processing
Detect Spam Once per 24 Hrs
S3: Store:
SPAM URLs
Dashboard Visualization & Statistics
Hive - JSONSerDe
Load
Spark
Transform
Python
Identify spam accounts and URLs in the tweet content
Custom Source
Memory Channel capacity = 1000 transactionCapacity = 1000
Spark - Cluster
Sink = HDFS /user/flume/tweets Grouped by Date/Time - HH
Hadoop - HDFS
HDFS - Cluster