

November 11, 2015

Project Progress Report

w205 Storing and Retrieving Data

Kyle Hamilton, Carlos Rodriguez and Sharmila Velamur

Hypothesis:

By scanning twitter public stream data one 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:

- A new AMI was built upon the UCB W205 AMI3. The following software were added:
 - Python 2.7 and virtualenv
 - Scrapy
 - Flume

Architecture Updates:

The architecture has been refined, and components aligned based on the project goals. See appendix-1 for the updated architecture diagram.

Data Ingestion:

In this project, flume will be used to read the public tweets from the twitter stream. Flume was chosen because it allows the collection, aggregation, and movement of large volumes of data in an efficient and reliable way. The concept of a source, channel and sink within flume's architecture is intuitive and straightforward to configure.

Data Processing Pipeline:

Data processing will apply necessary transformations to identified fields in the stored dataset and apply rules to identify tweets that have a high likelihood of being spam¹. The URLs from the tweet content will be scraped and subsequently crawled for discovering associated suspicious URLs. Other strategies

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

such as reverse DNS lookups and searching through known lists of spam sites are also being considered for classification of URLs².

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.

The Cron daemon will be used for scheduling.

External Systems:

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

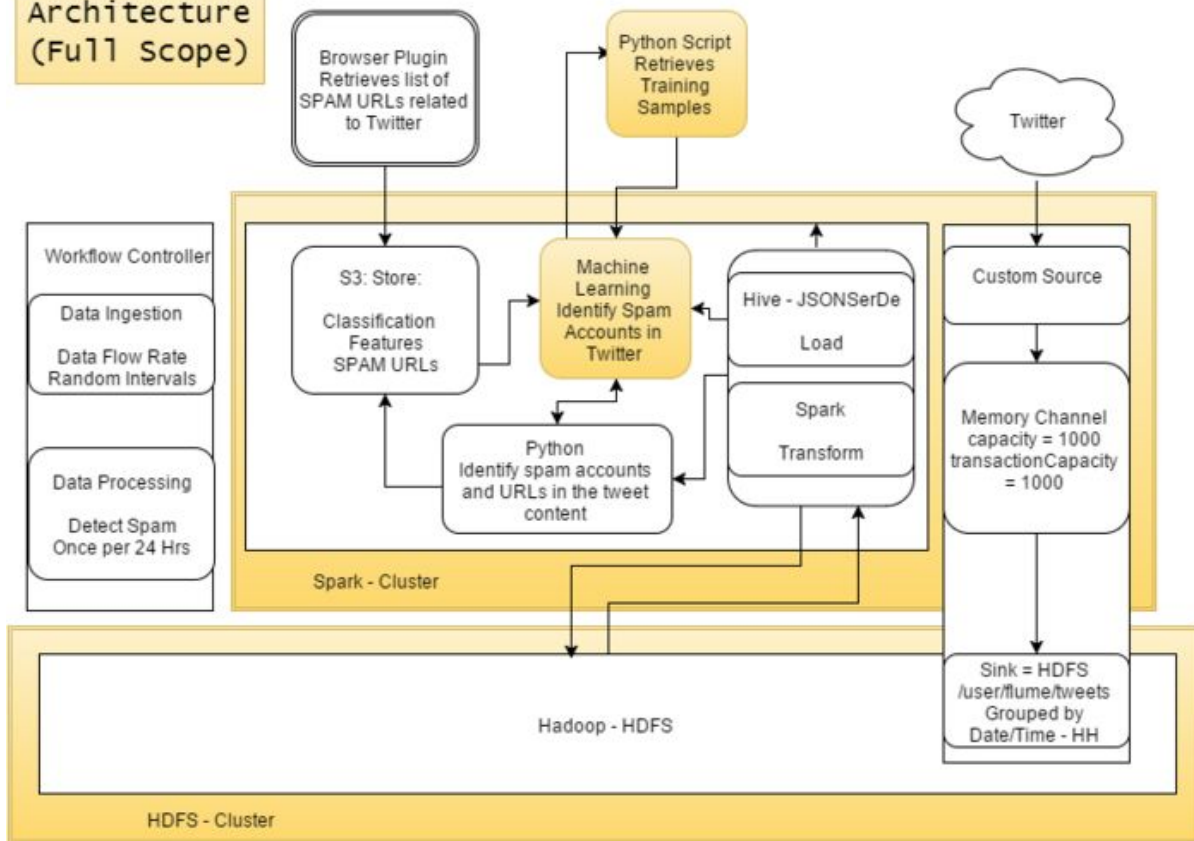
Project Timeline:

Week	Tasks	Status
10/19 - 10/31	<ul style="list-style-type: none">● Configure Dev environment (non-cluster)● Get twitter data and store in HDFS	Completed
11/01 - 11/07	<ul style="list-style-type: none">● Twitter API analysis; identify key fields● Load into Hive● Python scraper set up● Progress Report	Completed
11/08 - 11/14	<ul style="list-style-type: none">● Cron setup - workflow manager● Non ML based minimal scope classification of URLs as spam or not. Store Spam URLs in S3	Work in Progress
11/15 - 11/28 (2 weeks)	<ul style="list-style-type: none">● Complete (minimal scope) python coding● Browser plugin and dashboard● End to end integration● Identify optimization needs and opportunities	Not Started
11/29 - 12/05	<ul style="list-style-type: none">● Optimize, review code and config, test● Project presentation	Not Started

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

Appendix - 1

Architecture (Full Scope)



Architecture (Minimal Scope)

