Project Proposal

MIDS W205 Storage and Retrieval

Kyle Hamilton, Carlos Rodriguez and Sharmila Velamur

The Data

For this project, we will use web scraping technologies to collect data from sources such as social media (i.e. tweets) outlets, online news media comments, and blogs.

The Question

We are interested in using textual analysis on data obtained from various online sources to identify suspicious websites that lure people into scams.

The exploration for suspicious behavior would seek to analyze the relationship between keywords associated with free offers and prizes, the geographic location of the content as well as the metadata that is obtainable such as user accounts and websites associated with those publishing suspected content.

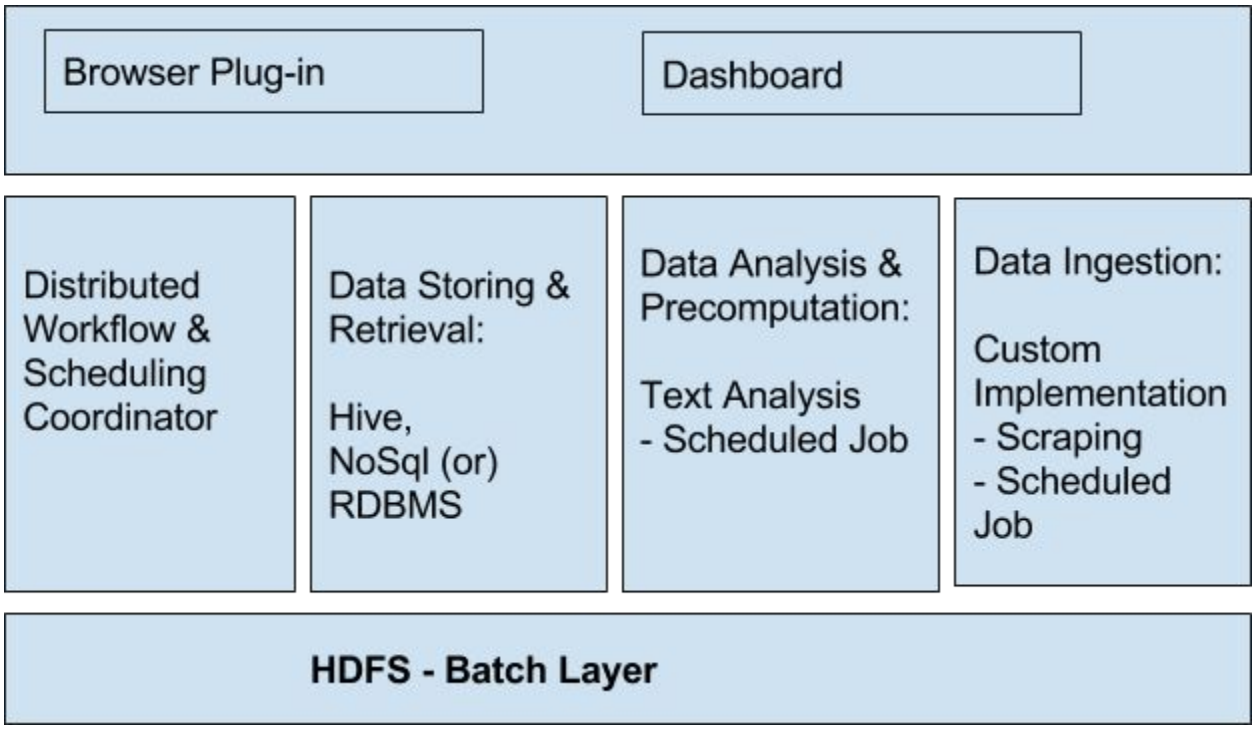
- How can we effectively use web scraping and text analysis for fraud detection, especially scam bait?
- How can we disseminate this information to the public on time?

The Challenges

At this time, the main risks and challenges that we see with the project involve:

- **Scope definition and scope creep** - need to ensure that the project is properly scoped such that it is focused and realistically achievable by week 15 of the fall semester.
- **Data collection** - because we are collecting our own data, to have a meaningful amount of data to manipulate and gain insights from, we will need to implement data collection early on in the project.
- **AWS data ingestion costs** - since this project will require constant and unpredictable inflows of data to an AWS server, we are not sure that this will fit comfortably within the $50 AWS credits allocated per person for the class.

The Proposed Architecture



The Deliverables

1. Github repo with all the code and documentation.
2. Final Project Presentation.
3. A dashboard to review the analysis (visualizations and querying).
4. A browser plugin to highlight possible scams.
5. Optional: Allow users to choose additional sources for analysis apart from the baseline that the application provides.