MICHIGAN STATE UNIVERSITY

Project Plan

Email Classification using Machine Learning

The Capstone Experience

Team Accenture

Sofia Colella
Varsha Odapally
Griffin Carr
Kevin Wilson
Yuyu Su

Department of Computer Science and Engineering
Michigan State University

Fall 2019

From Students…
…to Professionals

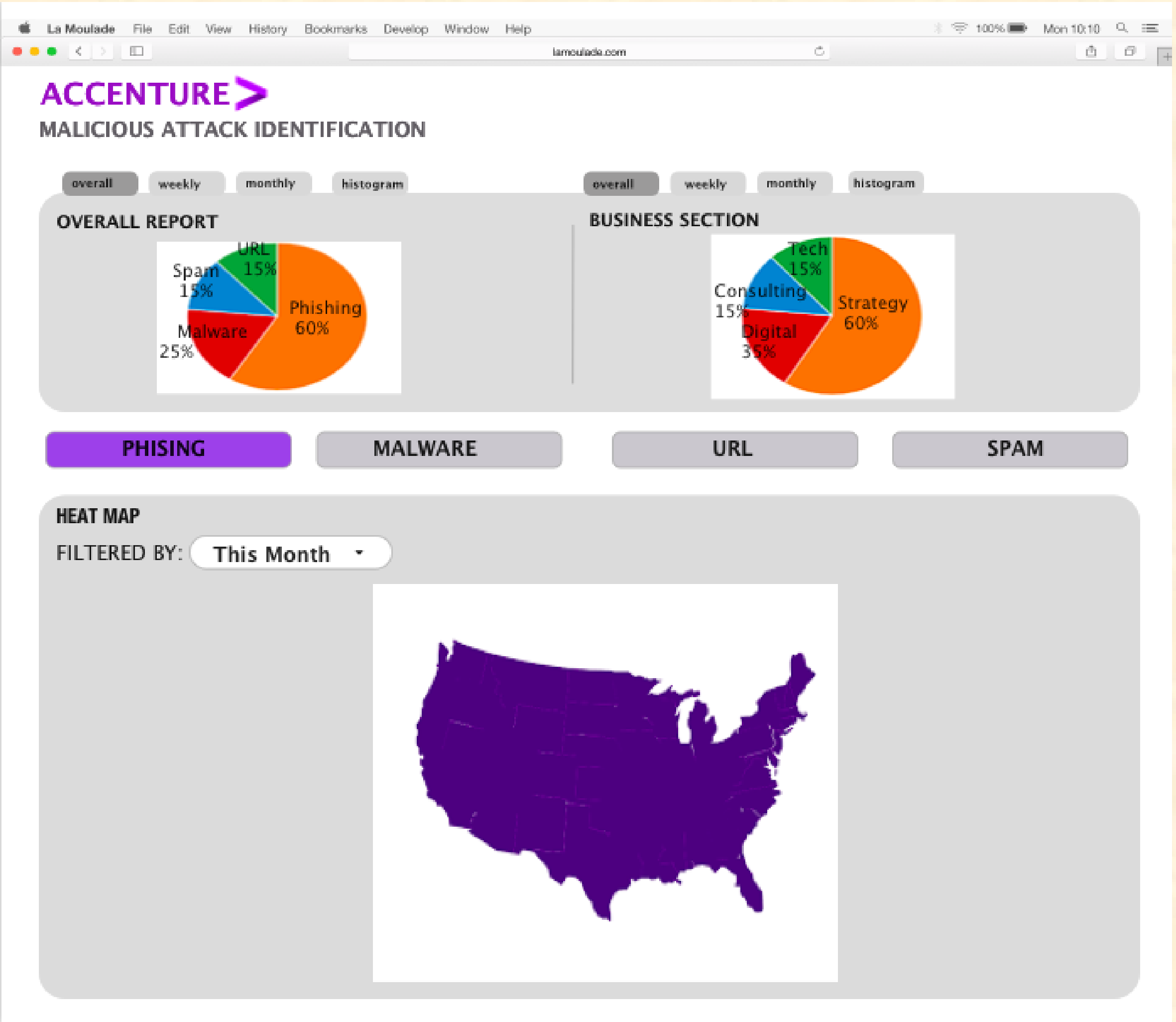
Functional Specifications

- Problem: There are about 15 billion spam and phishing emails per day.
- Solution: Create a web application dashboard for analysts
- Solution: Create and enhance classification and clustering models to help triage incoming emails, such as malicious attachment, emails with URL that have dangerous payloads, emails that lead to credential phishing and spam

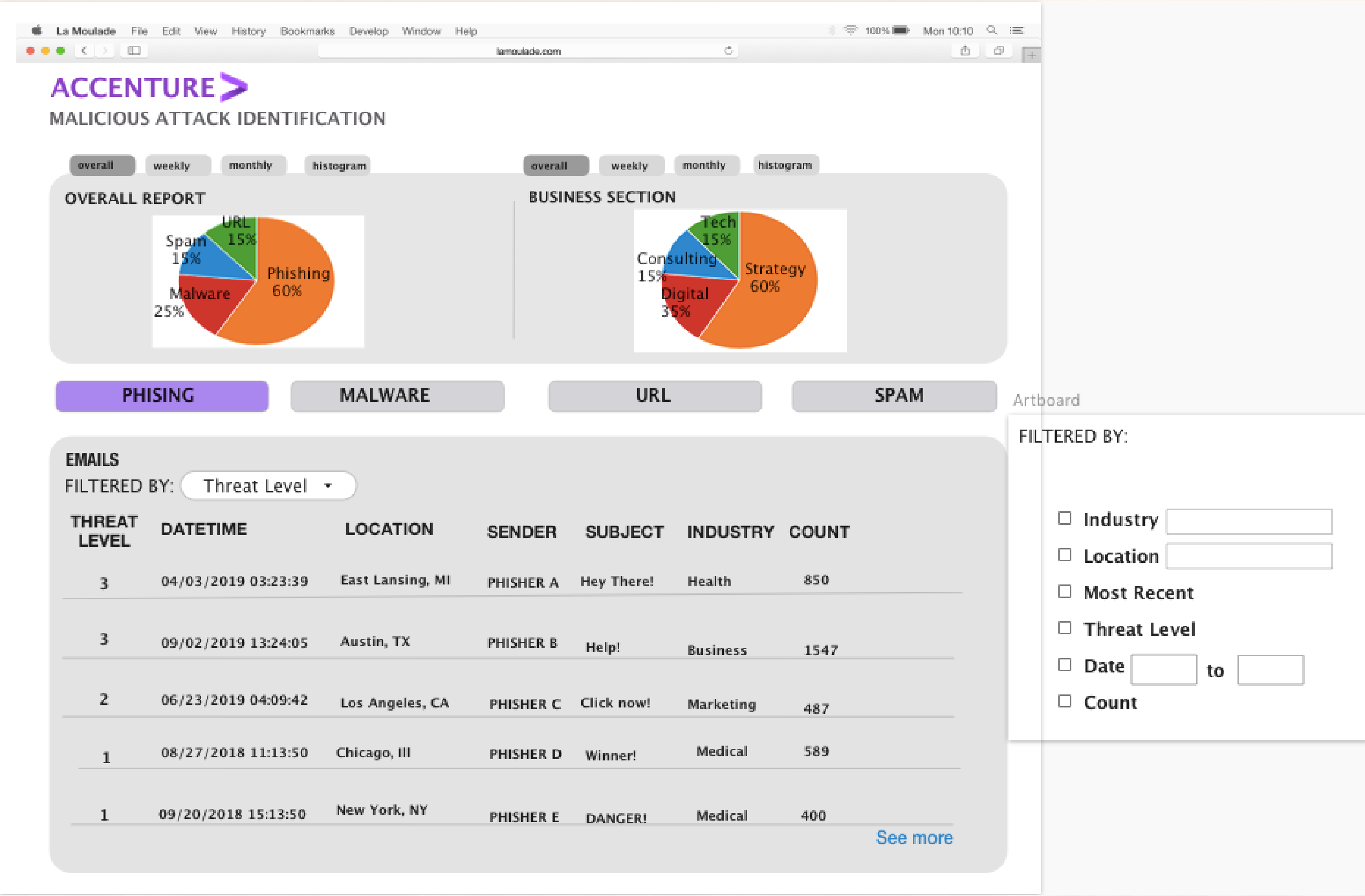
Design Specifications

- Results of ML algorithm will be displayed on the user’s dashboard
  - Overall statistics and examples of malicious emails
- Three types of users: strategic, operational and tactical
- Strategic users see the overall results in the form of pie charts and a heat map
- Operational users are given more information regarding the malicious emails in addition to pie charts
- Tactical users are given the most information. They also have the option to take a course of action for each individual malicious email

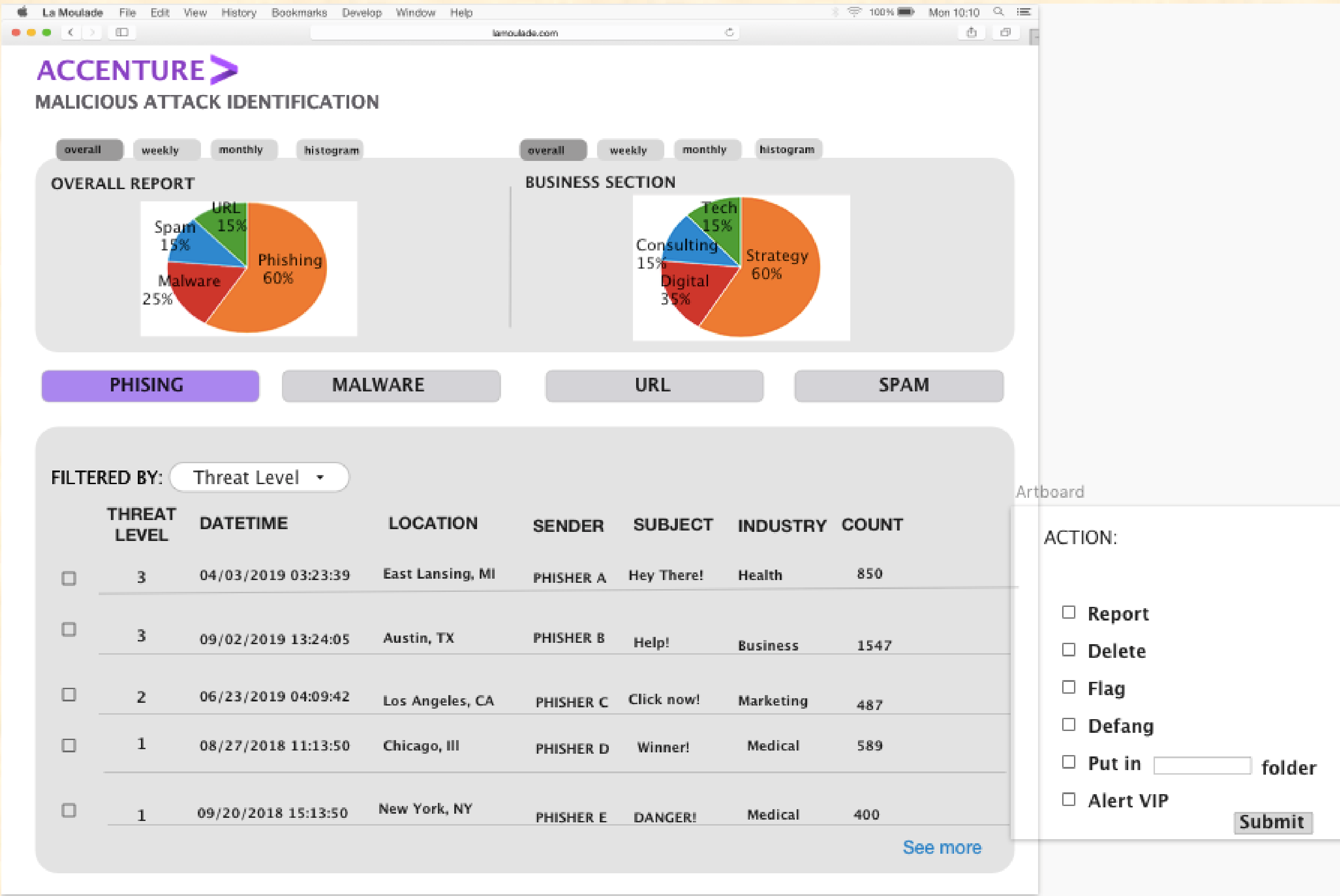
Screen Mockup: Strategic User Dashboard



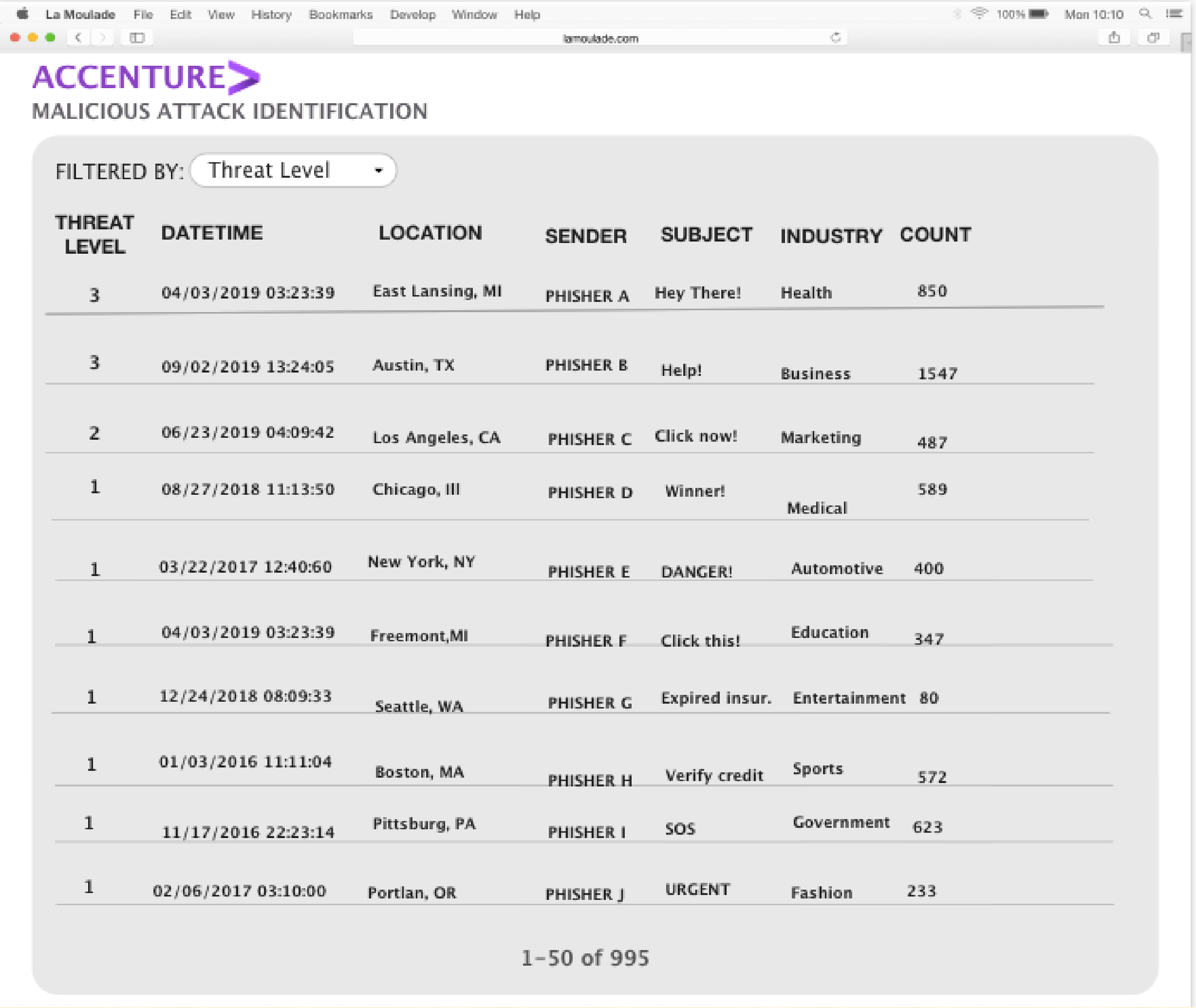
Screen Mockup: Operational User Dashboard



Screen Mockup: Tactical User Dashboard



Screen Mockup: See More Page

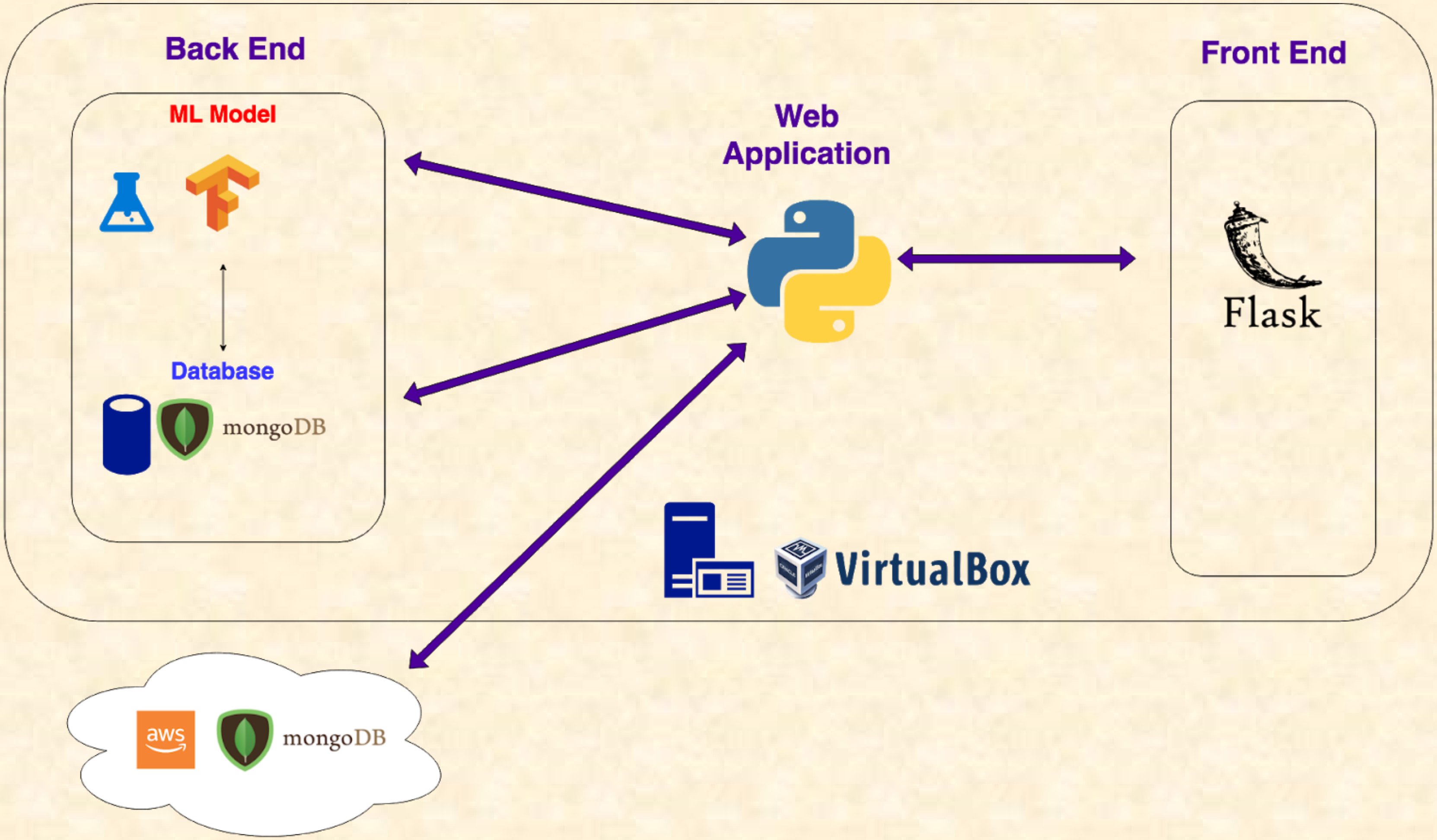


THREAT LEVEL	DATETIME	LOCATION	SENDER	SUBJECT	INDUSTRY	COUNT
3	04/03/2019 03:23:39	East Lansing, MI	PHISHER A	Hey There!	Health	850
3	09/02/2019 13:24:05	Austin, TX	PHISHER B	Help!	Business	1547
2	06/23/2019 04:09:42	Los Angeles, CA	PHISHER C	Click now!	Marketing	487
1	08/27/2018 11:13:50	Chicago, Ill	PHISHER D	Winner!	Medical	589
1	03/22/2017 12:40:60	New York, NY	PHISHER E	DANGER!	Automotive	400
1	04/03/2019 03:23:39	Freemont,MI	PHISHER F	Click this!	Education	347
1	12/24/2018 08:09:33	Seattle, WA	PHISHER G	Expired insur.	Entertainment	80
1	01/03/2016 11:11:04	Boston, MA	PHISHER H	Verify credit	Sports	572
1	11/17/2016 22:23:14	Pittsburg, PA	PHISHER I	SOS	Government	623
1	02/06/2017 03:10:00	Portlan, OR	PHISHER J	URGENT	Fashion	233

Technical Specifications

- Web app, database, machine learning model hosted on CentOS VM
- Machine learning back-end
- Python API call back-end
- MongoDB on VM with parsed email data, MongoDB on AWS hosted cluster for login info
- Flask front-end to display dashboards

System Architecture



System Components

- Software Platforms / Technologies
  - Python 3.7
  - Tensorflow 1.14.0
  - AWS
  - MongoDB
  - VirtualBox VM - CentOS
  - Flask
  - PyCharm

Risks

- Ranking Harmful Emails (High)
  - Determining a ranking system to list emails in terms of threat level
  - Threats will be evaluated using the “Threat Triangle.” Also known in cyber threat intelligences as capability, intent, and opportunity.
- Output Interpretation(High)
  - Model outputs run the risk of being misinterpreted, based on misunderstanding/incorrect assumption of how the desired model was built.
  - Clear idea of how model must be built, what assumptions must be made, and what the output should tell the audience.
- Under/Overfitting Model(Medium)
  - Having “High Bias”(underfitting) or “High Variance”(overfitting) which can lead to poor predictions for future used data sets.
  - Train-Test Split of our data (Train model on 70% of data, measure error rate on remaining 30% of data)
- Data(low)
  - Poor data quality, lack of data, lack of variability of data
  - Client provided us mock up data

Questions?

?

?

?

?

?

?

?

?

?