CS 22-317
Machine Learning Tool for Identifying Anomalies in Application Logs

VCU Capstone Project 2021-2022

Team Introductions

Adrienne Hembrick
Dylan Pierce
Sean Stitzer
Sam Sunvold

Dr. Changqing Luo(advisor)
Ben Polk (Sponsor)
Gurpreet Singh Sandhu (Sponsor)
Bryce Freshcorn (Sponsor)

Project Introduction

Problem Description

- Software applications, network devices, etc. create a deluge of log data that humans cannot (realistically) fully inspect.
- Downstream, log data is consumed by additional applications and (most critically) security tools and reporting capabilities.
- Sometimes, the data provider changes the format or content of the logs in an unexpected way, causing unexpected and (potentially) undetected impacts on the downstream consumers.
- Over the long-term, this creates a high risk of failure, which is particularly dangerous with security systems.

Solution

- Develop a solution that can consume a stream of log records without prior knowledge of the log source or record format and adapt over time to recognize the log entries and then detect potential material changes in their contents or format.

- Work at high volume without disrupting dependent data pipelines with significant delay.

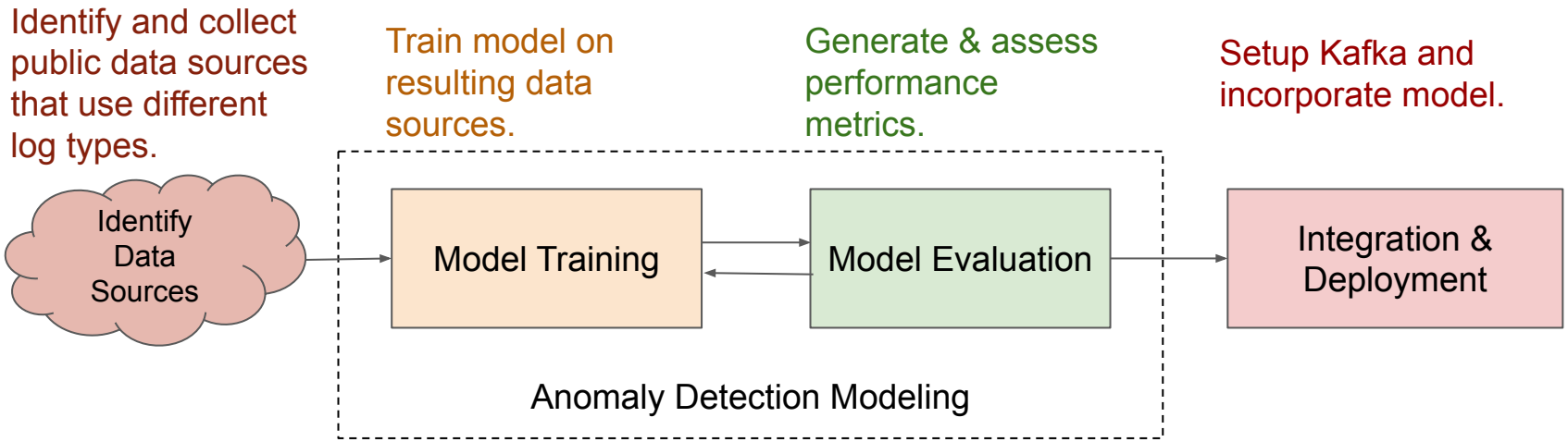
Solution - How do we do this?

- Anything (so long as it is not “hard-coded” to a particular data source)!
- Machine learning could be helpful…
- Potential relevant machine learning papers:
  - https://www.sciencedirect.com/science/article/pii/S2405959520300643
  - https://arxiv.org/pdf/2108.01955.pdf
  - https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1640&context=etd_projects
  - https://www.sciencedirect.com/science/article/pii/S0167404818306333
- Labelled data will not be available… but that does not preclude a “bootstrapping” approach...

- Feel free to experiment with different ideas and bounce them off Capital One sponsors!

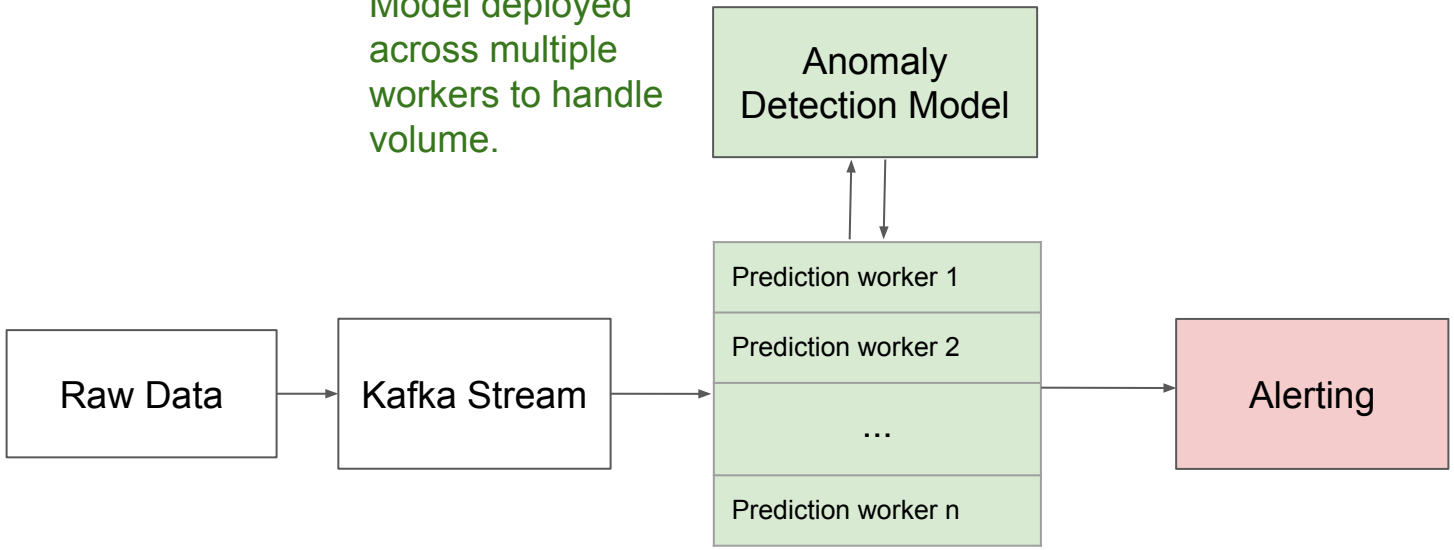
Architecting Our Solution

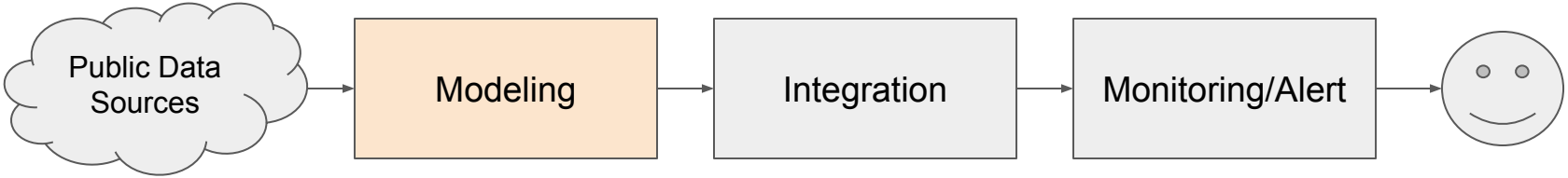
Building an Alert and Monitoring Pipeline



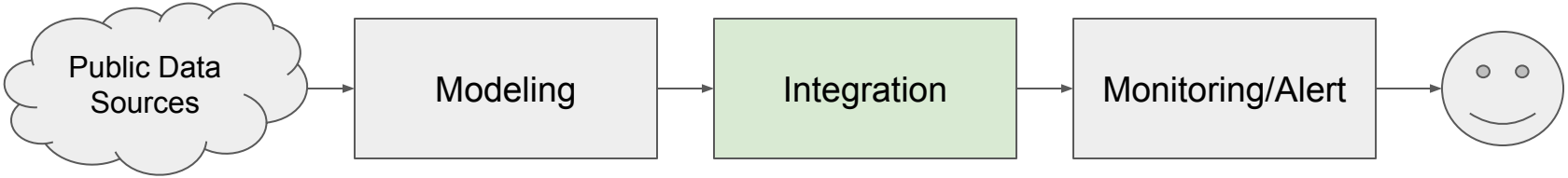
Model Integration w/ Kafka

Model deployed across multiple workers to handle volume.

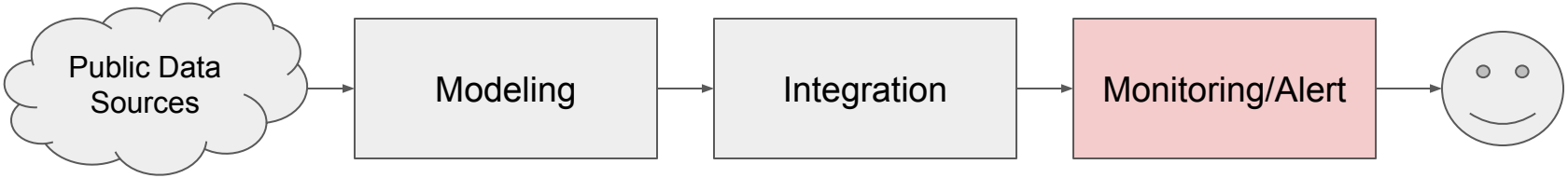




- Helpful resources:
  - Modeling Best Practices
  - Intro to NLP
  - Outlier Detection
- Questions to ponder:
  - Is there a way to do this robustly without machine learning?
  - If we use machine learning, which model(s) should we use?
  - Is an end-to-end model the best solution? Do we need multiple models?
- How are similar problems being solved? Research existing literature.



- How would raw data be ingested into Kafka stream? Is a simulation needed?
- What language would Kafka clients be written in?
- How could we approach this “cloud first” to make it scalable and easily deployable in AWS?
    - In house Kafka server or AWS managed MSK?
- How many concurrent workers?



- What existing services are available for monitoring and alerting?
- How to window the alerts? Hourly or Daily?
- Dashboard visualization?

Developing in an Agile Environment

- Weekly team meetings.
  - 15 minutes: “stand up” status discussion
    - What did I accomplish last week?
    - What do I plan to work on this week?
    - What is standing in the way of me making progress?
  - 15 minutes: demo completed work
  - 15 minutes: groom upcoming tasks so they are ready to start.
- Trello board for tracking tasks
  https://trello.com/b/zHUeGJp5/vcu-log-anomaly-detection
  - Please keep the board updated, document questions or progress in comments.
  - Feel free to tag others in comments.
  - Update your Trello profile with your picture
- Team GitHub for all complete or in-progress development
  - Please have one peer review for all merged pull requests.

Developing in an Agile Environment

- How can we deliver value to our customer incrementally, and learn as we go to improve our final results?
- What are some landmarks for success we can aim for along the development timeline?
- How can we build with a “stop starting, start finishing” mentality?

Developing in an Agile Environment

Technical Lead
Ben Polk

Help the team negotiate technical challenges and implementation decisions. Focused on creating delivered value.

Product Owner
Gurpreet Singh

Help the team prioritize work and ensure it meets the needs of customers and downstream consumers.

Scrummaster

Help the team focus on how they deliver using best practices, collaboration, and communications to do their best work.

Let’s Get Started!

Week One Objectives:

1. Discussing the technical approach and how you will plan for incremental delivery.
2. Start building out the task backlog in Trello based on the plan the team develops.
3. Decide on a team name!