NHL Win Prediction

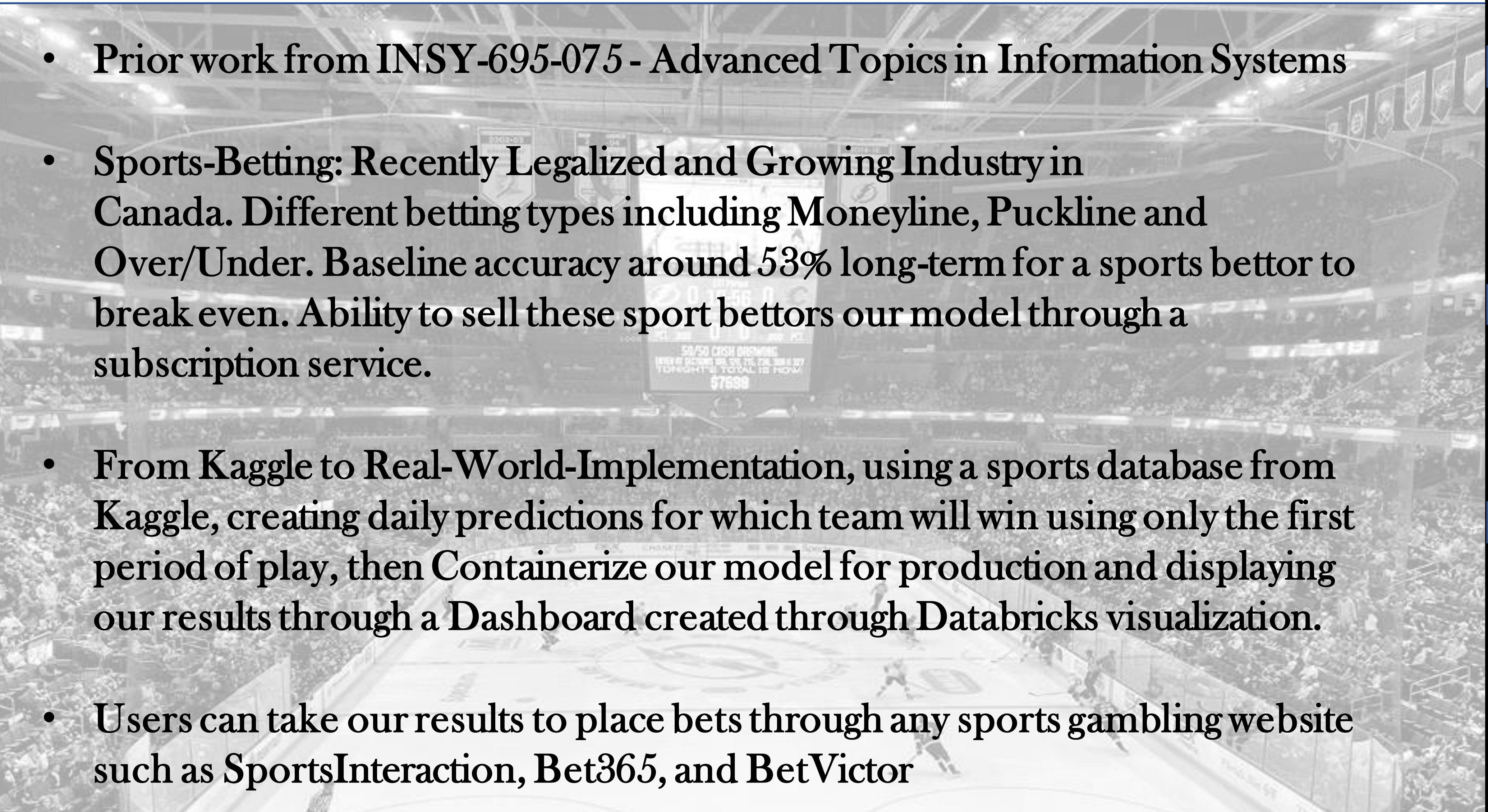
Team No.	Group #2
Team Name	NHL Pros
GitHub Repository	NHL-Game-II

Product Manager	Anqi Chen	angelach99
Business Analyst	Matthew Buttler Ives	matt-buttlerives
Data Engineer	Vaibhav Visual	vvaibhav11
	Hadyan Fahreza	hifahreza
	Mesaye Bahiru	mesaye3
	Rebecca Mukena Yumba	beccarem
Data Scientist	Louis D'Hulst	Louis-dhulst

Agenda

- #1 Business Case Overview
- #2 Streaming Real Time Data
- #3 Machine Learning Model
- #4 Deployment
- #5 Dashboards

Business Case Overview



- Prior work from INSY-695-075 - Advanced Topics in Information Systems
- Sports-Betting: Recently Legalized and Growing Industry in Canada. Different betting types including Moneyline, Puckline and Over/Under. Baseline accuracy around 53% long-term for a sports bettor to break even. Ability to sell these sport bettors our model through a subscription service.
- From Kaggle to Real-World-Implementation, using a sports database from Kaggle, creating daily predictions for which team will win using only the first period of play, then Containerize our model for production and displaying our results through a Dashboard created through Databricks visualization.
- Users can take our results to place bets through any sports gambling website such as SportsInteraction, Bet365, and BetVictor

Sports Model Time Period

Forecheck: 62% Accuracy
Public for only 2017-2018 season
Before being shelved due to costs
Built in Python

Period 1

Our Model: 68% Accuracy
Uses historical dataset provided by Kaggle and Streams data through SportsRadar.
Built the prediction model in Python

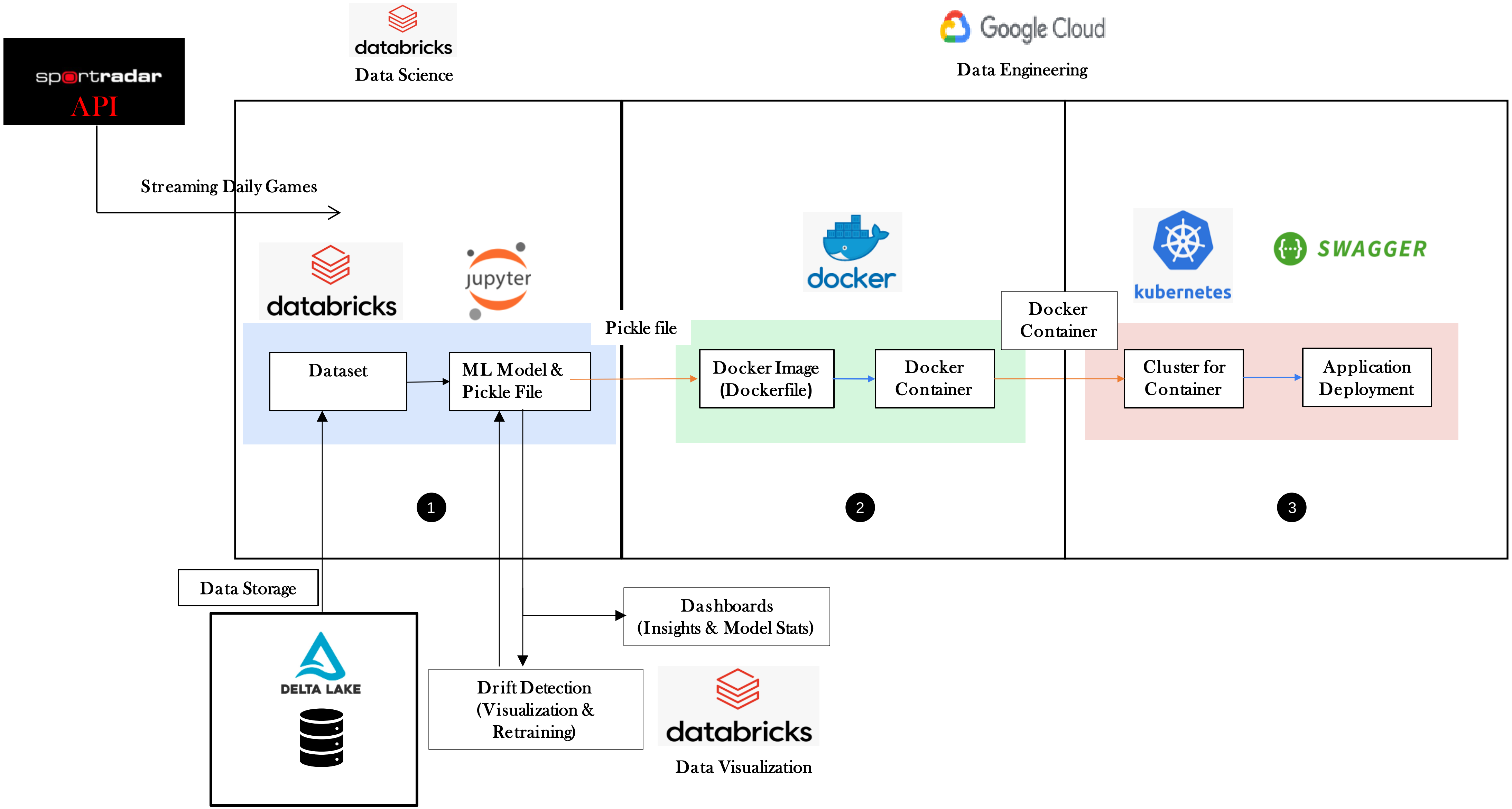
Period 2

Experimentation: 76% Accuracy
Brief investigation for using second period of play.
Better accuracy but less Opportunities.

Period 3

End of game.

Architecture (Overview)



API

Data Science

Data Engineering

Streaming Daily Games

Dataset

ML Model & Pickle File

Pickle file

Docker Image (Dockerfile)

Docker Container

Docker Container

Cluster for Container

Application Deployment

1

2

3

Data Storage

Dashboards (Insights & Model Stats)

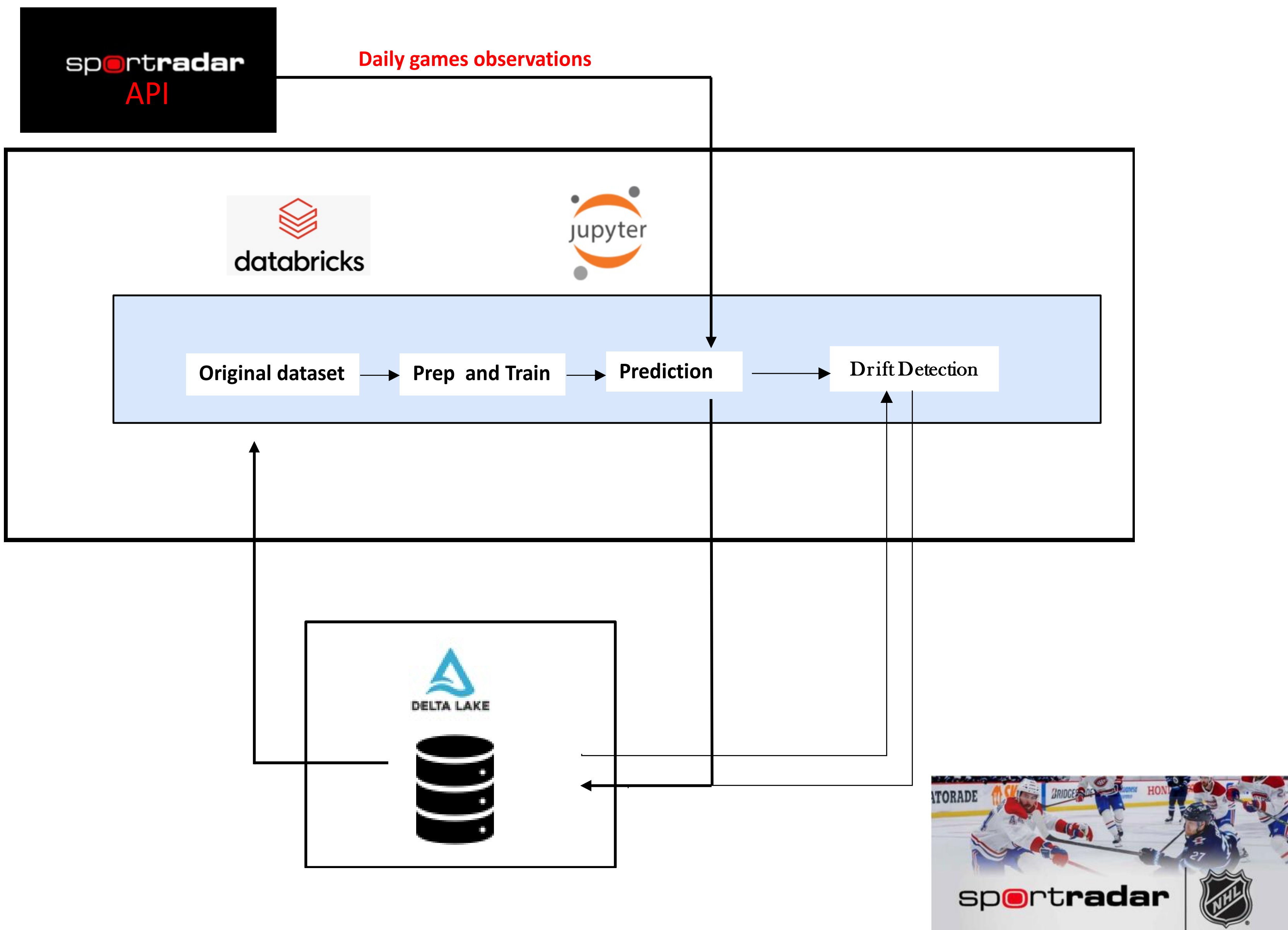
Drift Detection (Visualization & Retraining)

Data Visualization

SportRadar

- Official Partner of NHL, NBA, MLB
- Global Sports Coverage
  - 750,0000 events annually across 83 sports
- Costs Related to API Services:
  - Premium: $1,500 - $2,000
  - Real Time: $3,000 - $5,000
- SportRadar API provides:
  - Play-by-play stats for home and away teams
  - Season game schedule
  - Prediction takes place in the first period of the game

SportRadar & Streaming Data Architecture



Machine Learning Model

Predictors:

Shots	Shots_against	Goals	Goals_against	Takeaways
Takeaways_against	Hits	Hits_against	Blocked shots	Blocked shots against
Giveaways	Giveaway_against	Missed shots	Missed shots_against	Penalties
Penalties_against	#Won Faceoffs	#Lost Faceoffs	HoA_away	HoA_home

Target Variables: Won

Final Model: LightGBM LGBMClassifier model

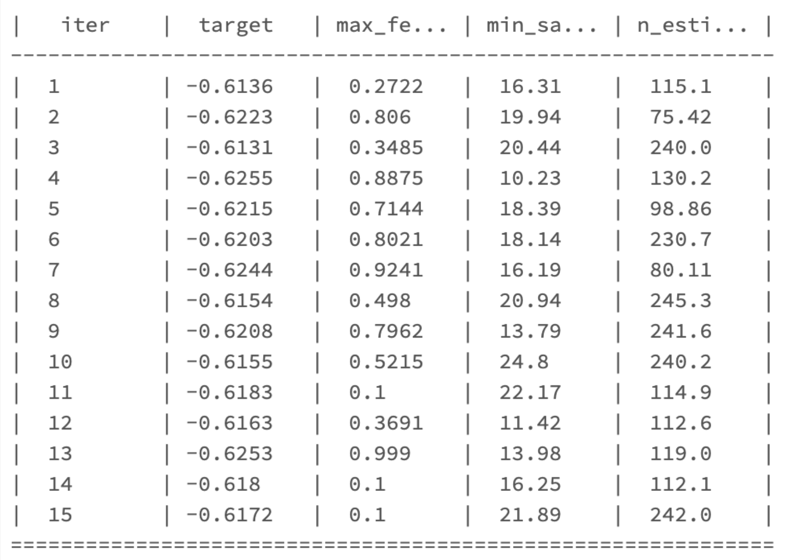
Results: See dashboard section

Hyperparameter Tuning

GridSearchCV

Learning_rate: 0.06
Max_depth: 10
Num_leaves: 31

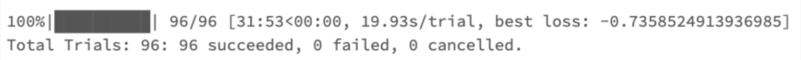
Bayesian Optimization



```
|   iter    |  target   | max_fe... | min_sa... | n_esti... |
-------------------------------------------------------------
| 1         | -0.6136   | 0.2722    | 16.31     | 115.1     |
| 2         | -0.6223   | 0.806     | 19.94     | 75.42     |
| 3         | -0.6131   | 0.3485    | 20.44     | 240.0     |
| 4         | -0.6255   | 0.8875    | 10.23     | 130.2     |
| 5         | -0.6215   | 0.7144    | 18.39     | 98.86     |
| 6         | -0.6203   | 0.8021    | 18.14     | 230.7     |
| 7         | -0.6244   | 0.9241    | 16.19     | 80.11     |
| 8         | -0.6154   | 0.498     | 20.94     | 245.3     |
| 9         | -0.6208   | 0.7962    | 13.79     | 241.6     |
| 10        | -0.6155   | 0.5215    | 24.8      | 240.2     |
| 11        | -0.6183   | 0.1       | 22.17     | 114.9     |
| 12        | -0.6163   | 0.3691    | 11.42     | 112.6     |
| 13        | -0.6253   | 0.999     | 13.98     | 119.0     |
| 14        | -0.618    | 0.1       | 16.25     | 112.1     |
| 15        | -0.6172   | 0.1       | 21.89     | 242.0     |
=============================================================
```

Max_features: 0.35
Min_samples_split: 20.44
N_estimators: 240

HyperOpt with MLFlow

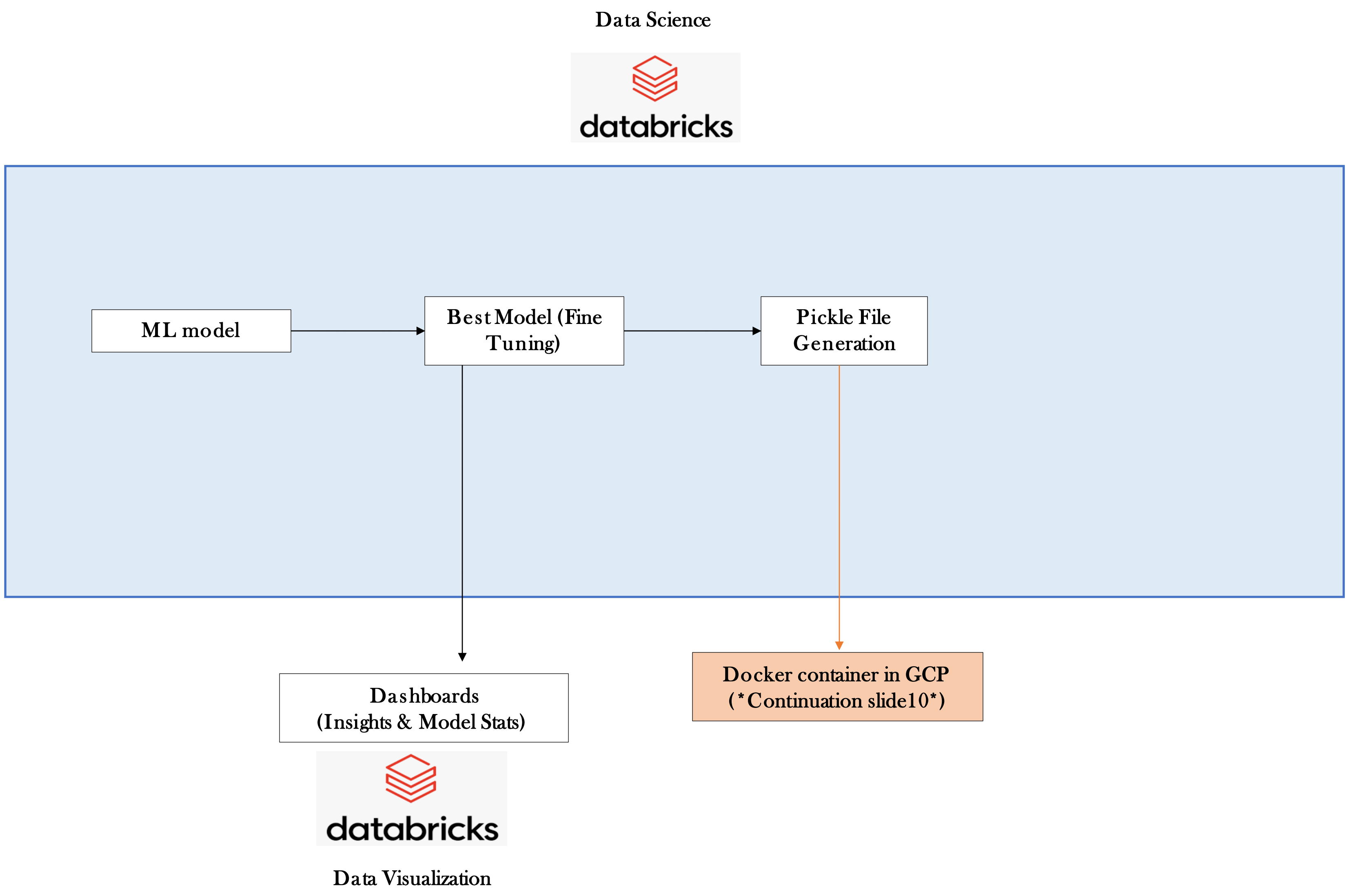


```
100%|██████████| 96/96 [31:53<00:00, 19.93s/trial, best loss: -0.7358524913936985]
Total Trials: 96: 96 succeeded, 0 failed, 0 cancelled.
```

Best Loss: -0.736

Note: There was some problem applying different methods to the LightGBM model, thus other models will be used as a demonstration.

Machine Learning Flow



Data Science

ML model

Best Model (Fine Tuning)

Pickle File Generation

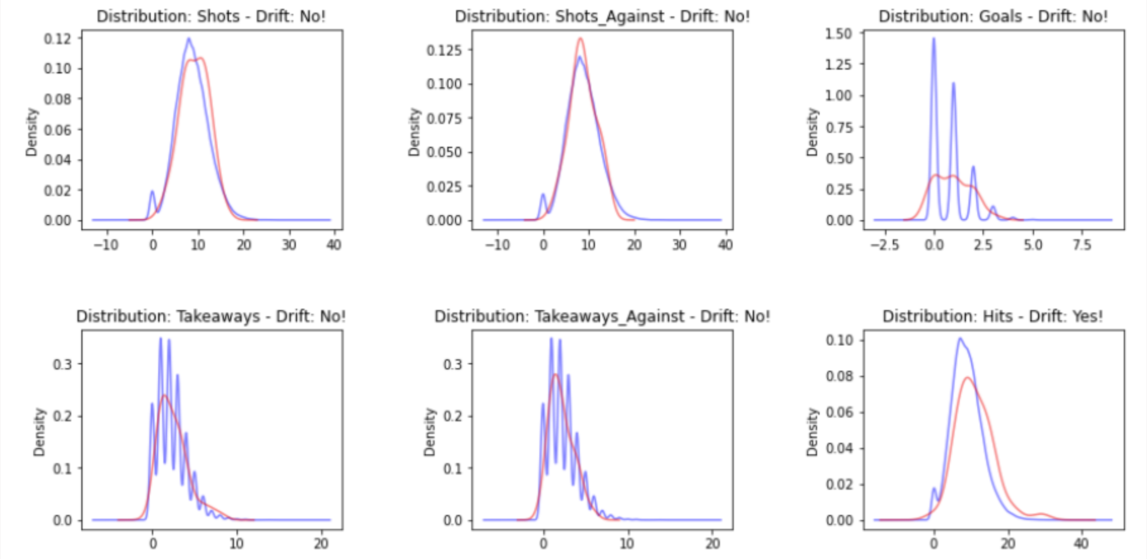
Dashboards (Insights & Model Stats)

Docker container in GCP (*Continuation slide10*)

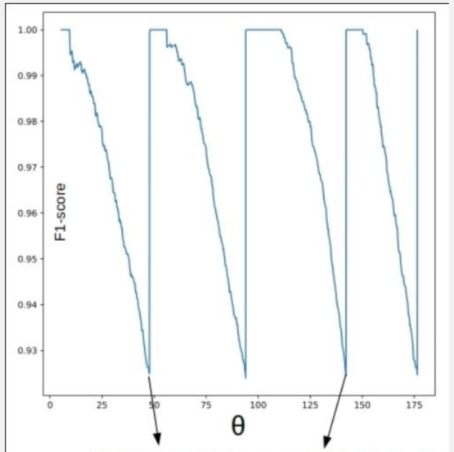
Data Visualization

Detecting Drift

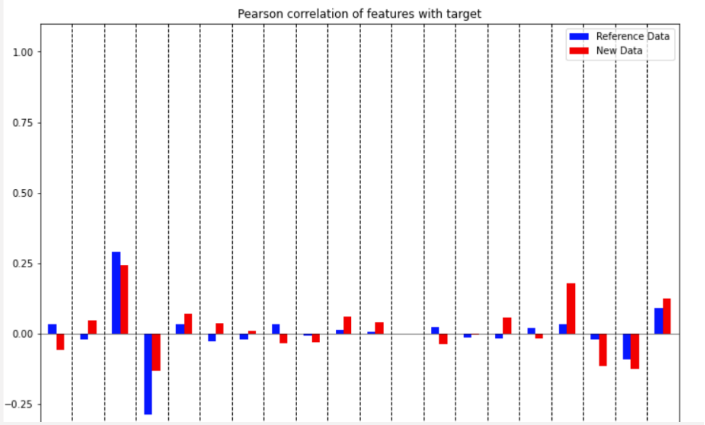
- **Detect data drift using Kolmogorov-Smirnov tests and learned Random Forest classifier**
    - Monitor changes in data over time
    - Make sure model is up to date with current data trends
    - Use last week's games



- **Monitor Concept/Prediction Drift**
    - Pearson correlation between target and features
    - Accuracy over last week
    - Compare to fixed decision rule baseline
    - Trending decrease in accuracy triggers model retraining



Model rebuilt at this time on all re-labeled data as the performance threshold has been breached



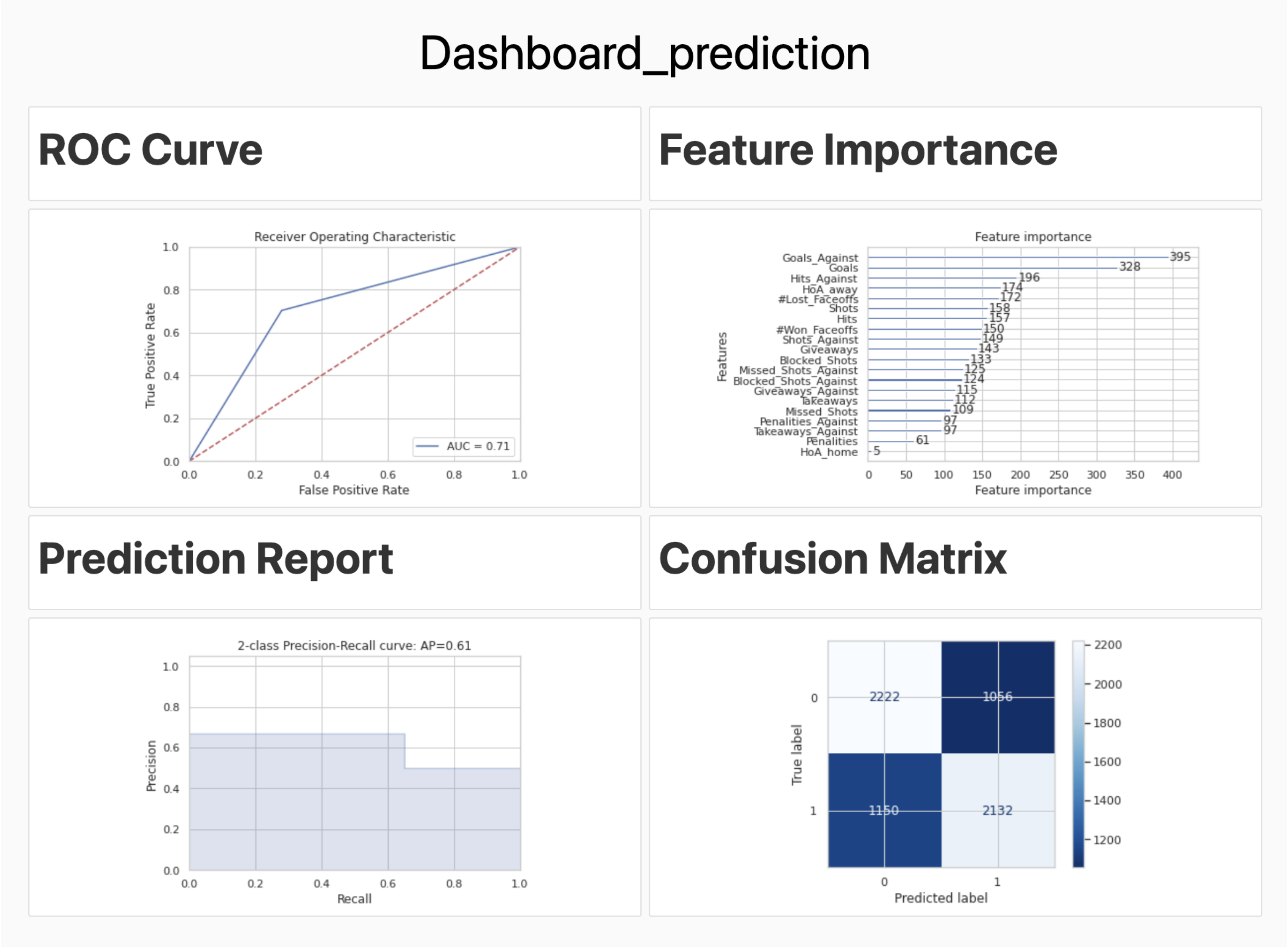
Unit Testing

- **Prediction Tests:**
  - Ensure predictions are logically consistent

- **Data Tests:**
  - Ensure data types saved are the same
  - Ensure new values fit in logical range

- **API Tests:**
  - Ensure API calls return consistent values

Dashboard of Predicted Results



Dashboard_prediction

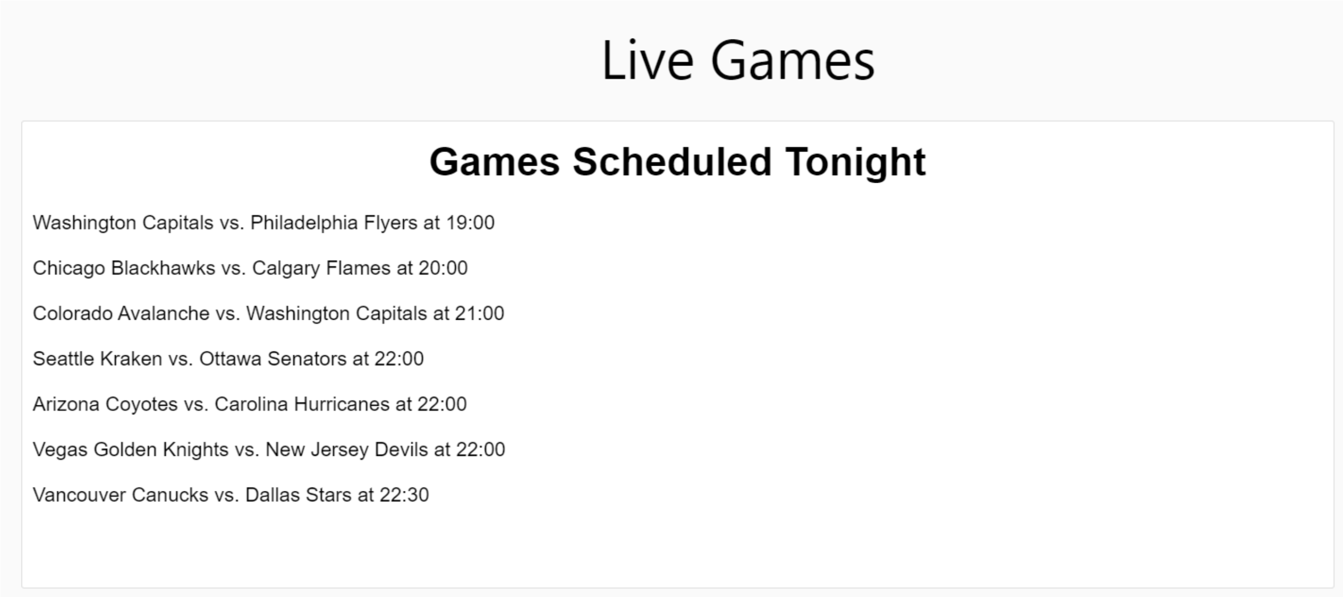
ROC Curve

Feature Importance

Prediction Report

Confusion Matrix

Dashboard of Live Data



Live Games

Games Scheduled Tonight

Washington Capitals vs. Philadelphia Flyers at 19:00

Chicago Blackhawks vs. Calgary Flames at 20:00

Colorado Avalanche vs. Washington Capitals at 21:00

Seattle Kraken vs. Ottawa Senators at 22:00

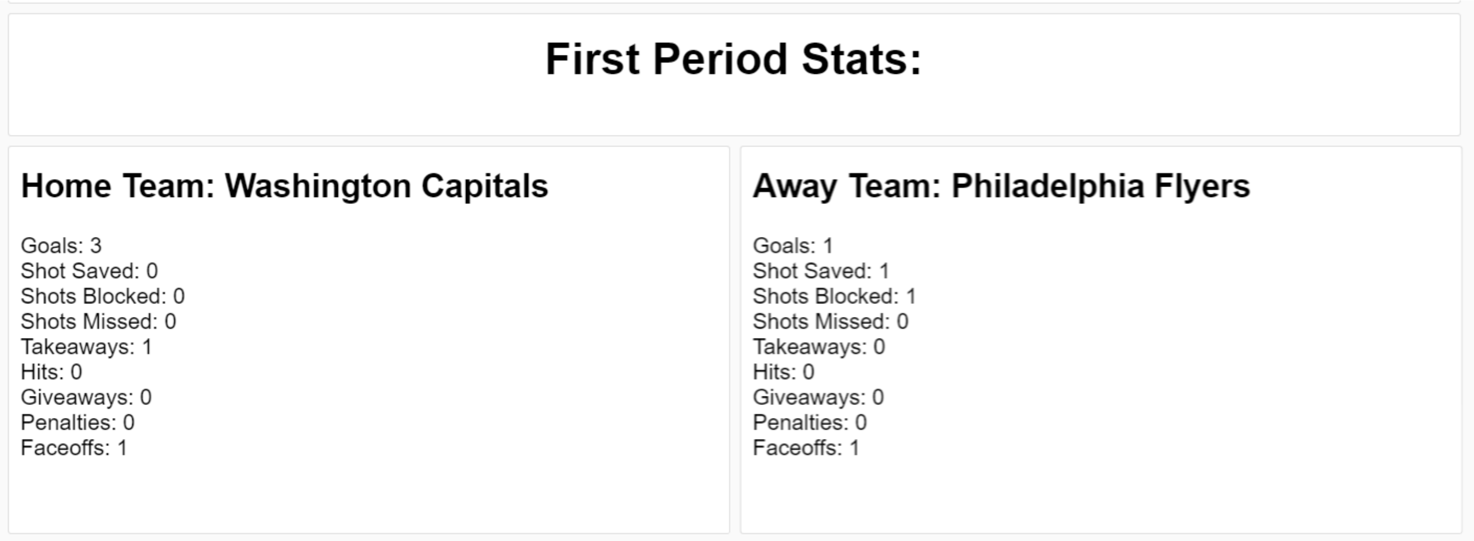
Arizona Coyotes vs. Carolina Hurricanes at 22:00

Vegas Golden Knights vs. New Jersey Devils at 22:00

Vancouver Canucks vs. Dallas Stars at 22:30

Daily Schedule

First Period Stats for Games in Progress



First Period Stats:

Home Team: Washington Capitals

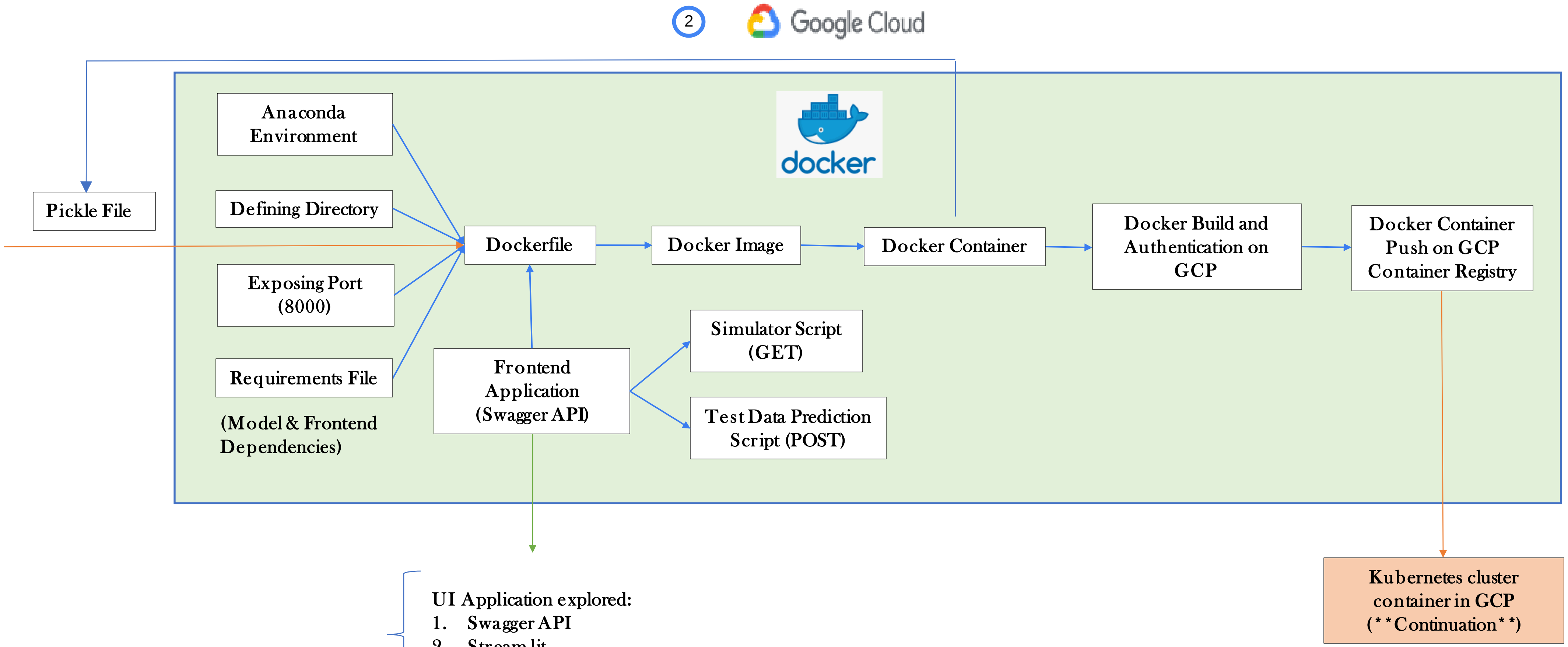
Goals: 3
Shot Saved: 0
Shots Blocked: 0
Shots Missed: 0
Takeaways: 1
Hits: 0
Giveaways: 0
Penalties: 0
Faceoffs: 1

Away Team: Philadelphia Flyers

Goals: 1
Shot Saved: 1
Shots Blocked: 1
Shots Missed: 0
Takeaways: 0
Hits: 0
Giveaways: 0
Penalties: 0
Faceoffs: 1

Docker

Data Engineering


2
Google Cloud
docker
Pickle File
Anaconda Environment
Defining Directory
Exposing Port (8000)
Requirements File
(Model & Frontend Dependencies)
Dockerfile
Docker Image
Docker Container
Docker Build and Authentication on GCP
Docker Container Push on GCP Container Registry
Frontend Application (Swagger API)
Simulator Script (GET)
Test Data Prediction Script (POST)
Kubernetes cluster container in GCP (**Continuation**)

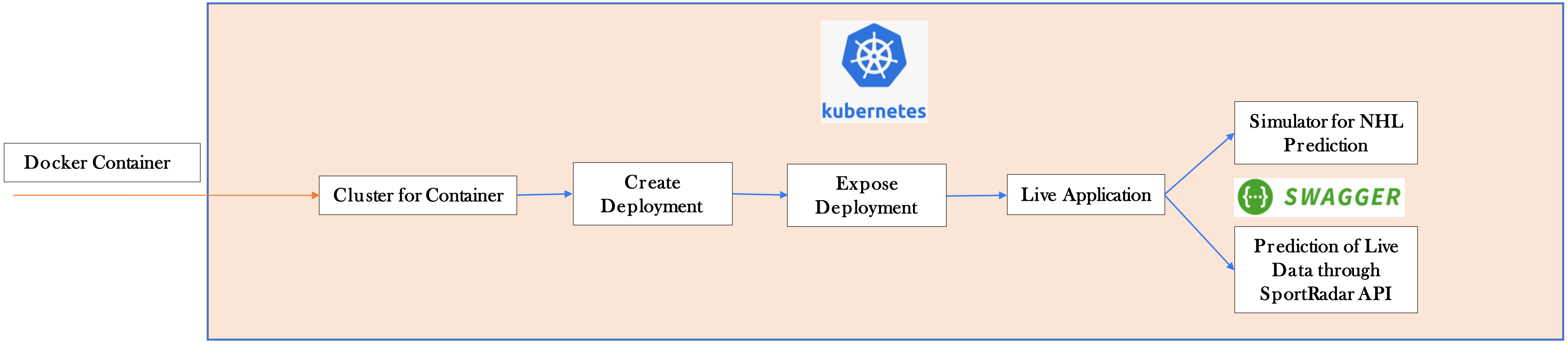
UI Application explored:

1. Swagger API
2. Stream lit
3. Flask API and Postman

Kubernetes

Data Engineering

3 Google Cloud



kubernetes

Docker Container

Cluster for Container

Create Deployment

Expose Deployment

Live Application

Simulator for NHL Prediction

SWAGGER

Prediction of Live Data through SportRadar API

Future Scope

- Connecting live database storing data from sports radar into Delta lake of Databricks with GitHub repository which will be updating at regular interval (every day before game starts), currently facing restrictions with Data Bricks community edition

- Upgrading Sports radar subscription to premium to get detailed dataset required for prediction

- Work on User Interface to make it more intuitive and user friendly

- Implementing cron jobs / other pipeline processes on the container and GCP to regularly update the pickle file with live data pulled from SportRadar API

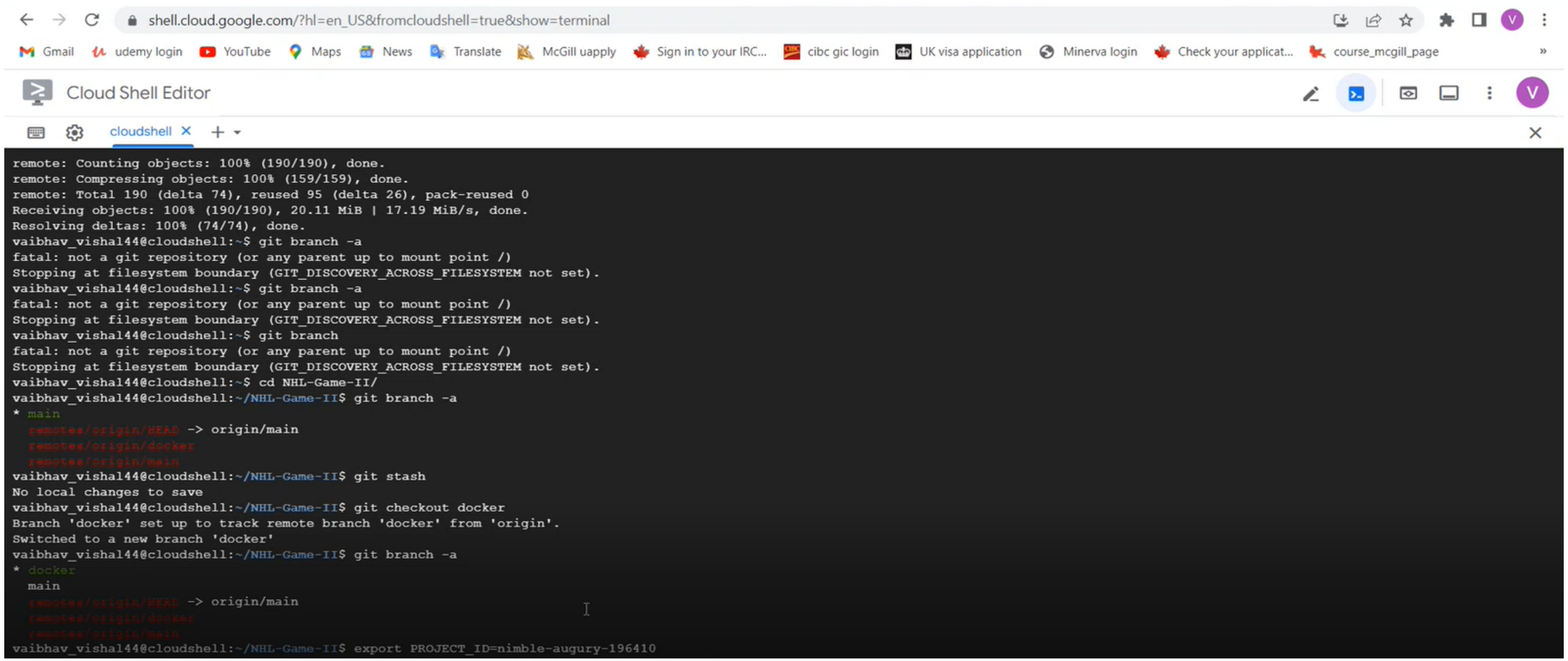
- Further, fully automation/scheduling of the application pipeline which loads data, run model and use it for containerization and building dashboards/applications. At present, there are restrictions due to community or free trial editions and would need additional funds for it to be successful

Thank You!

Appendix 1

NHL Prediction Platform

GCP screenshots



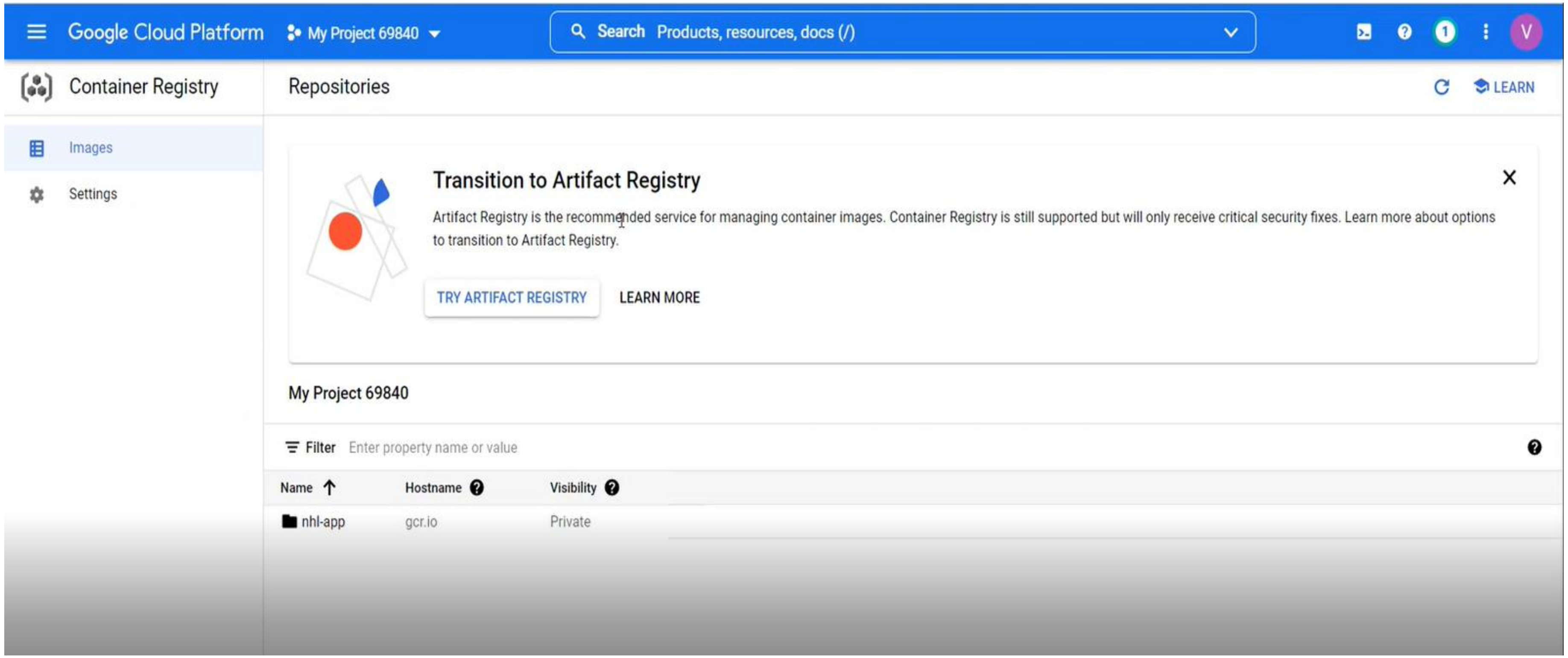
```
remote: Counting objects: 100% (190/190), done.
remote: Compressing objects: 100% (159/159), done.
remote: Total 190 (delta 74), reused 95 (delta 26), pack-reused 0
Receiving objects: 100% (190/190), 20.11 MiB | 17.19 MiB/s, done.
Resolving deltas: 100% (74/74), done.
vaibhav_vishal44@cloudshell:~$ git branch -a
fatal: not a git repository (or any parent up to mount point /)
Stopping at filesystem boundary (GIT_DISCOVERY_ACROSS_FILESYSTEM not set).
vaibhav_vishal44@cloudshell:~$ git branch -a
fatal: not a git repository (or any parent up to mount point /)
Stopping at filesystem boundary (GIT_DISCOVERY_ACROSS_FILESYSTEM not set).
vaibhav_vishal44@cloudshell:~$ git branch
fatal: not a git repository (or any parent up to mount point /)
Stopping at filesystem boundary (GIT_DISCOVERY_ACROSS_FILESYSTEM not set).
vaibhav_vishal44@cloudshell:~$ cd NHL-Game-II/
vaibhav_vishal44@cloudshell:~/NHL-Game-II$ git branch -a
* main
  remotes/origin/HEAD -> origin/main
  remotes/origin/docker
  remotes/origin/main
vaibhav_vishal44@cloudshell:~/NHL-Game-II$ git stash
No local changes to save
vaibhav_vishal44@cloudshell:~/NHL-Game-II$ git checkout docker
Branch 'docker' set up to track remote branch 'docker' from 'origin'.
Switched to a new branch 'docker'
vaibhav_vishal44@cloudshell:~/NHL-Game-II$ git branch -a
* docker
  main
  remotes/origin/HEAD -> origin/main
  remotes/origin/docker
  remotes/origin/main
vaibhav_vishal44@cloudshell:~/NHL-Game-II$ export PROJECT_ID=nimble-augury-196410
```

GCP cloud shell for setting up docker container, Kubernetes containers, and deploying application

Appendix 2

NHL Prediction Platform

GCP screenshots



Repositories

Transition to Artifact Registry

Artifact Registry is the recommended service for managing container images. Container Registry is still supported but will only receive critical security fixes. Learn more about options to transition to Artifact Registry.

My Project 69840

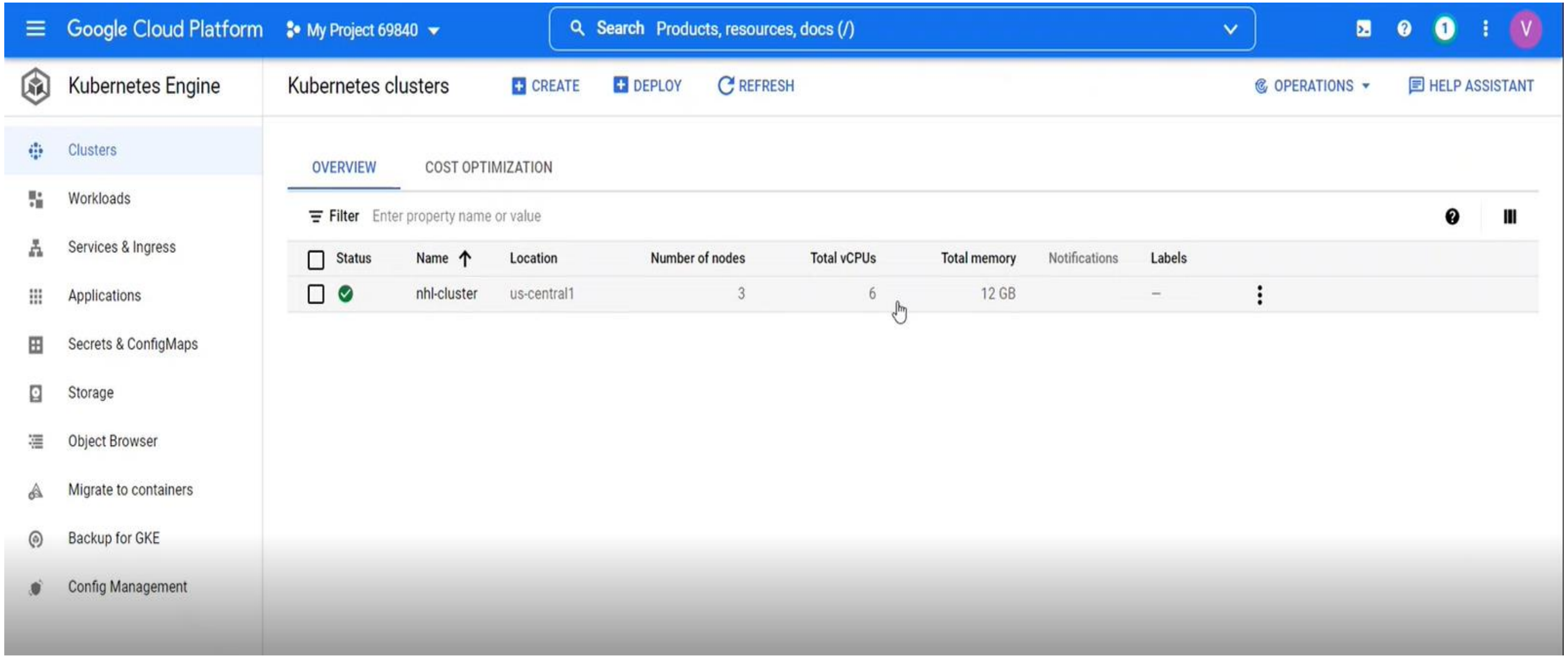
Name ↑	Hostname	Visibility
nhl-app	gcr.io	Private

GCP cloud registry storing docker container built

Appendix 3

NHL Prediction Platform

GCP screenshots



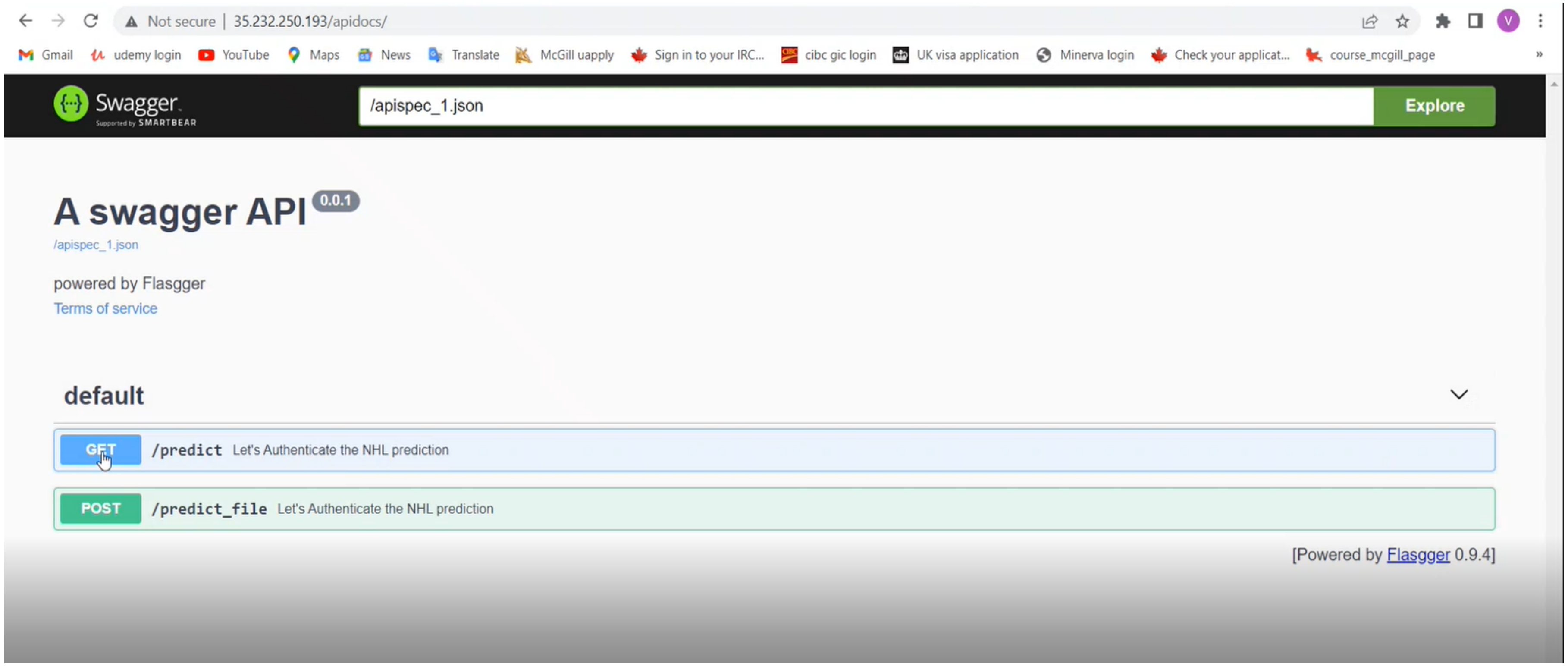
□	Status	Name ↑	Location	Number of nodes	Total vCPUs	Total memory	Notifications	Labels
□	✓	nhl-cluster	us-central1	3	6	12 GB		–

GCP Kubernetes cluster deploying application

Appendix 4

NHL Prediction Platform

GCP screenshots



A swagger API 0.0.1

/apispec_1.json

powered by Flasgger

Terms of service

default

GET /predict Let's Authenticate the NHL prediction

POST /predict_file Let's Authenticate the NHL prediction

[Powered by Flasgger 0.9.4]

Application hosted through Kubernetes cluster