Team Deliverable 3

Team 4: Ahmed Farid Khan, Bennett Blanco, Dhruv Shah, Yu-Chin (Alyssa) Chen

Recap of Phase 1 and Phase 2:

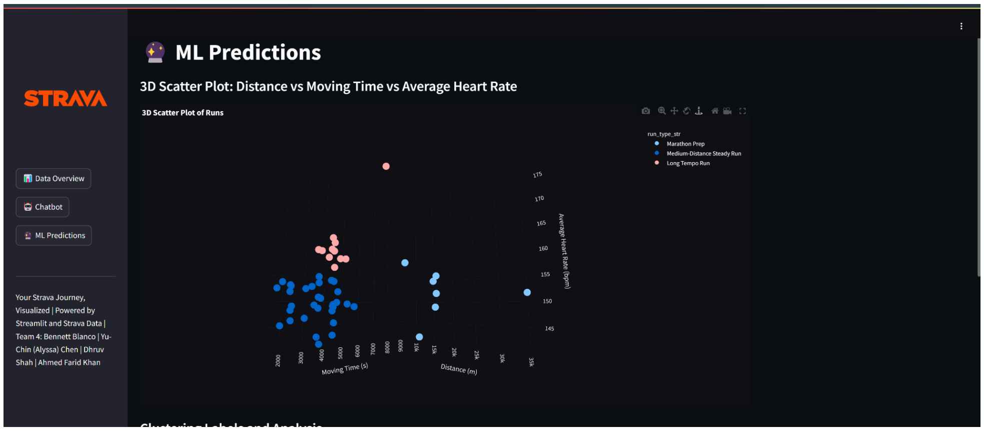
- Developed a robust ETL pipeline to pull data from the API and store it in BigQuery
- Set up a webhook function to activate the pipeline every time the athlete completes an activity
- Utilized Prefect to orchestrate the pipeline
- Deployed Apache Superset on Google Compute Engine to visualize BigQuery data
- Trained a k-means clustering model to predict run type based on three metrics (distance ran, average heart rate, moving time)
- Include this prediction as a user feature by posting it to the description of that run inside the Strava App.

Phase 3:

Phase 3 aims to offer additional features to the Strava athlete.

- **Streamlit Dashboard:**
    - In this phase of the project, we focused on developing an interactive and visually engaging Streamlit dashboard to display and analyze Strava activity data. The dashboard was designed to provide a user-friendly interface with distinct tabs for Data Overview, Chatbot, and ML Predictions, each serving a specific purpose. Using Google BigQuery, we fetched relevant data about Strava activities and integrated it into the dashboard. Additionally, we incorporated a chatbot, powered by Vertex AI, to answer user queries about the data, making the application both informative and interactive.
    - A key decision we made was to limit the data processed and displayed to the most recent 50 activities. This choice ensured the dashboard remained responsive and efficient, especially given the potential for larger datasets to slow down real-time performance. This approach allowed us to focus on creating a proof of concept that prioritized visualization and interactivity without overwhelming the application or the user. Limiting the dataset to 50 rows also aligned with the immediate needs of this phase, while leaving room for scalability in future iterations.
    - From a design perspective, we chose a clean, card-style layout for presenting Key Performance Indicators (KPIs) like Total Distance, Calories Burned, and Average Speed. These cards included hover effects and animations to make the metrics more engaging. For data visualization, we implemented a 3D scatter plot to analyze relationships between distance, moving time, and heart rate, ensuring its full visibility by carefully adjusting its dimensions. We opted for a consistent color scheme using black and Strava orange to maintain a modern aesthetic while aligning with the brand. These decisions helped us create a dashboard that is both functional and visually appealing, laying the groundwork

for future enhancements.



STRAVA

Data Overview

Chatbot

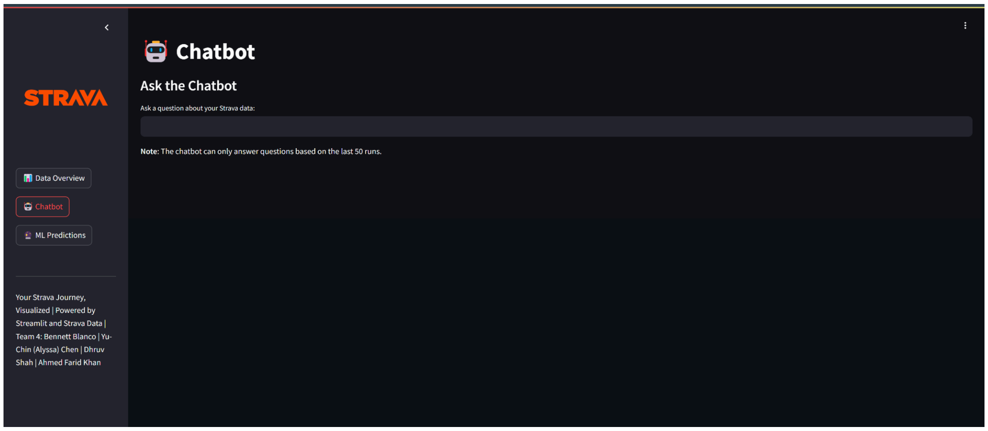
ML Predictions

Your Strava Journey, Visualized | Powered by Streamlit and Strava Data | Team 4: Bennett Blanco | Yu-Chin (Alyssa) Chen | Dhruv Shah | Ahmed Farid Khan

ML Predictions

3D Scatter Plot: Distance vs Moving Time vs Average Heart Rate

Clustering Labels and Analysis



STRAVA

Data Overview

Chatbot

ML Predictions

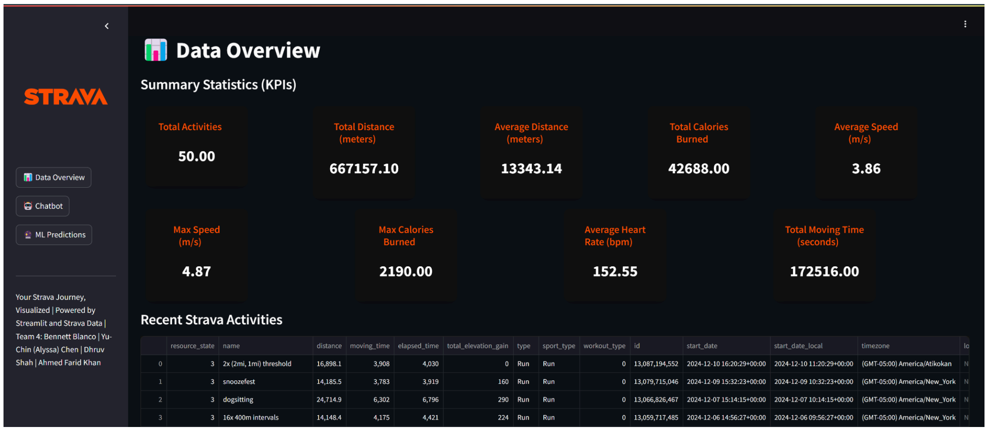
Your Strava Journey, Visualized | Powered by Streamlit and Strava Data | Team 4: Bennett Blanco | Yu-Chin (Alyssa) Chen | Dhruv Shah | Ahmed Farid Khan

Chatbot

Ask the Chatbot

Ask a question about your Strava data:

Note: The chatbot can only answer questions based on the last 50 runs.



STRAVA

Data Overview

Chatbot

ML Predictions

Your Strava Journey, Visualized | Powered by Streamlit and Strava Data | Team 4: Bennett Blanco | Yu-Chin (Alyssa) Chen | Dhruv Shah | Ahmed Farid Khan

Data Overview

Summary Statistics (KPIs)

Total Activities	Total Distance (meters)	Average Distance (meters)	Total Calories Burned	Average Speed (m/s)
50.00	667157.10	13343.14	42688.00	3.86

Max Speed (m/s)	Max Calories Burned	Average Heart Rate (bpm)	Total Moving Time (seconds)
4.87	2190.00	152.55	172516.00

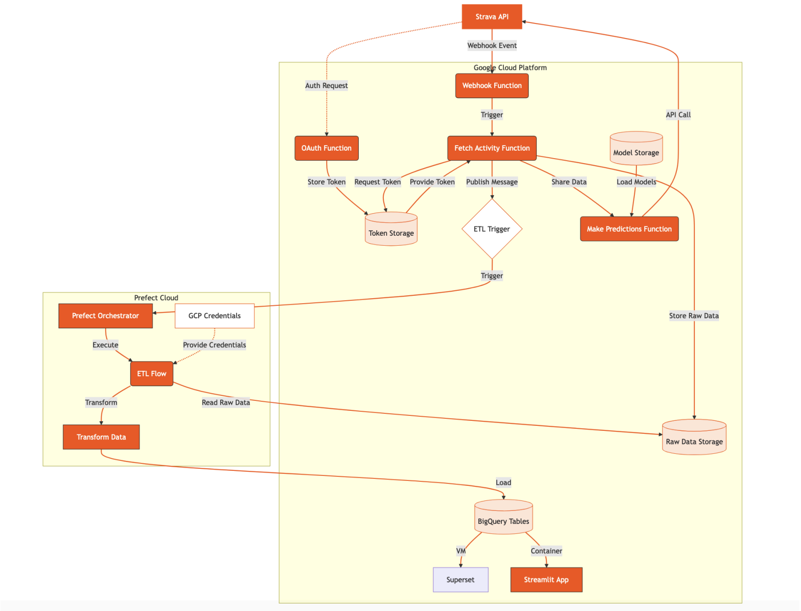
Recent Strava Activities

	resource_state	name	distance	moving_time	elapsed_time	total_elevation_gain	type	sport_type	workout_type	id	start_date	start_date_local	timezone	lo
0	3	2x (2mi, 1mi) threshold	16,898.1	3,908	4,030	0	Run	Run	0	13,087,194,552	2024-12-10 16:20:29+00:00	2024-12-10 11:20:29+00:00	(GMT-05:00) America/Atikokan	N
1	3	snoozefest	14,185.5	3,783	3,919	160	Run	Run	0	13,079,715,046	2024-12-09 15:32:23+00:00	2024-12-09 10:32:23+00:00	(GMT-05:00) America/New_York	N
2	3	dogsitting	24,714.9	6,302	6,796	290	Run	Run	0	13,066,826,467	2024-12-07 15:14:15+00:00	2024-12-07 10:14:15+00:00	(GMT-05:00) America/New_York	N
3	3	16x 400m intervals	14,148.4	4,175	4,421	224	Run	Run	0	13,059,717,485	2024-12-06 14:56:27+00:00	2024-12-06 09:56:27+00:00	(GMT-05:00) America/New_York	N

- **Backfilling**:
  - To compare performance over a larger range of time and to train a more robust model, we want data over a greater range of dates. In phase 2, we attempted to backfill data till 2022. We faced a two-fold memory challenge - within Prefect's free tier orchestrator and Strava's daily limits. Although we had no control over Strava data limits, we altered the way we backfilled data by running backfilling outside the Prefect orchestrator and doing this locally instead. This was done over a period of five days to circumvent Strava's daily limits. Once all the data was loaded locally, we added it to BigQuery to make it available for analysis and model training.

- **Additions and Adjustments to Machine Learning Model:**
    - To pull data into the dashboard, a new table is created in BigQuery called ‘clustering_labels’. This table has six columns: activity_id, start_time, distance, moving time, suffer score, and run type. Firstly, this table was backfilled with the last 50 runs with an empty run_type column. Then, the Google Cloud Function ‘populate-existing-runs’ filled this column by pulling in the k-means model and the scaler from Google Cloud Storage. The model returned the appropriate run type for each run and this data is now available to view in the Streamlit dashboard as a 3d scatterplot.
    - Additionally, in phase 2, there was an error in the labeling of the run types. In this case, ‘Low-Intensity Runs’ were being classified as ‘Marathon Preps’ and vice versa. This error has now been rectified in phase 3.

Updated Pipeline:



Limitations:

Currently the entire system relies on the fact that we only have one athlete; all max concurrencies are set to one, so cloud functions cannot run simultaneously. This is a crucial missing piece to scaling, but since this is abstracted in GCP, it’s an easy fix. Additionally, our local backfill ran successfully, but the logic did not fully match that of the ETL, resulting in some missing data points. We were on a bit of a time crunch so we didn’t give much thought to the backfill and this resulted in that issue. Carefully reimplementing the ETL logic outside of prefect will be the solution to this error.

Lastly, another limitation is the access to data that streamlit has. We limited it to the most recent 50 rows of data to keep streamlit from running too slow. A better solution would likely be to implement a custom API with more efficient endpoints regarding time and space complexity.

Future Work:

For future work, we aim to scale the system to support multiple users seamlessly. This involves enhancing the authentication process to manage individual user data securely and ensuring the infrastructure can handle increased activity data efficiently. Additionally, we plan to personalize the dashboard experience for each user, offering tailored insights and visualizations that cater to their specific training goals.

We also plan to incorporate weather data into our machine learning models to improve prediction accuracy. By analyzing factors such as temperature, humidity, and wind conditions, we aim to provide more context-aware insights and better understand their impact on a runner's performance.

We also aim to develop a mobile-friendly version of the dashboard or a companion app to increase accessibility and engagement.