Redback Operations
Data Science and Analytics Team
M.TELLEY

DOCUMENTATION
Data Procurement, Cleansing and Organisation

Purpose

Redback Operation core product will collect performance (output) data from a number of sensors while the user is using the bike and wahoo kickr trainer. Moreover, establishing a working environment that will align with the core product is critical in establishing continuity in data-related tasks and advising other core teams about Data Science and Analytics (DSA) Team's requirements. The initial effort is to set up a core data location for the team that house all key datasets.

By the end of this task, further work on data analysis, visualisation and using a number of machine learning algorithms will be possible.

Scope

Initialise a BigQuery (Google) environment that houses all of the DSA's data; three key datasets have been identified:

1. Sales device data
2. Oxygen (O2) Uptake related data
3. Wahoo/Garmin workout data.

Sales and O2 datasets were located within the DSA team's GitHub repository, however, the Garmin and Wahoo data was procured, cleansed and organised independently. This document outlines the specific details relating to independent data procurement.

Wahoo and Garmin fitness activity data procurement

What is / Who are Garmin and Wahoo?

Wahoo
Founded in 2009 by Chip Hawkins in Atlanta, GA, Wahoo creates innovative solutions to make hard-fought goals attainable and lives better. Wahoo was built on the foundation of simplicity and the mindset that "there's got to be a better way. Moreover, Wahoo produces a suite of devices that record/collect performance-related data from Athletes.

Key devices involved in the data procurement process:

1. Wahoo Bolt
2. Wahoo Element

Garmin
Gramin makes products that are engineered on the inside for life on the outside. We do this so our customers can make the most of the time they spend pursuing their passions. Garmin brings GPS navigation and wearable technology to the automotive, aviation, marine, outdoor and fitness markets.

Key devices involved in the data procurement process:

1. Edge 1030
2. Edge 1040

File output - FIT files

The Flexible and Interoperable Data Transfer (FIT) protocol is designed specifically for the storing and sharing of data that originates from sport, fitness and health devices. The FIT protocol defines a set of data storage templates (FIT messages) that can be used to store

information such as activity data, courses, and workouts. It is designed to be compact, interoperable and extensible. Extensive documentation is available here regarding FIT files.

The FIT file protocol was designed to provide:

- Interoperability of device data across various platforms
- Scalability from small embedded devices to cloud platforms
- Forward compatibility, allowing the protocol to grow and retain existing functionality
- Automated compatibility across platforms of different native endianness

Referenced from *Garmin*

FIT files are not immediately in a format that a human can read such that a conversion to a Comma-separated values file (CSV) is required. This document outlines the steps required to conduct a CSV conversion.

Where was the data procured from?

I provided over 400 FIT files to the project and also procured data from our Cyclists in Melbourne by way of convenience sampling. Therefore, the activity files can't be used for population sampling activity; demographics are extremely narrow i.e., Male, 30-60 years etc.

Accessing the data

Computer used to pull FIT file: Macbook Pro (iOS)

Garmin Devices: A USB-c cable was used to access the files on the device. The device is immediately accessible in the device listing; files are then transferred (copied) off the device.

Wahoo: To access a Wahoo device, Android File Transfer is required. The device also needs to be turned ON. Thereafter, the device is immediately accessible in the device listing; files are then transferred (copied) off the device. More information is available here about connecting a Wahoo device to a computer.

Other information collected

In addition to acquiring FIT files from a participant, their age, biological gender (male/female), weight and functional threshold power (FTP) were also recorded.

Privacy

FIT files contain GPS data and device data such as serial numbers; these data points have been considered sensitive and have been removed from the datasets entirely. Moreover, participants signed a statement of release which outlined what the data was going to be used for and how the data was going to be anonymised.

FIT File Conversion, Data cleansing and Organisation

FIT Conversion

Initially, Garmin Development documentation was assessed to understand the basics. Garmin's FIT software development kit (SDK) was installed and the FitCSVTool was used to convert FIT files to CSV format. For a demonstration, please watch this video. Note: the FIT SDK was downloaded and used on a Computer running a Windows operating system.

Testing was conducted initially using a Wahoo-derived FIT file. On review, the output wasn't organised in a way that instantly supported data analysis-related tasks. Further research was conducted to refine the conversion process.

A key resource used established a pathway to securing improved conversion techniques using a python script to convert FIT files in a highly organised manner. Moreover, the python script for wahoo-derived FIT files also included two additional functions that renamed the files to secure standardised naming conventions, removed sensitive data fields and also combined all of the CSV files into one master CSV file i.e, one master file per participant.
Key resources:

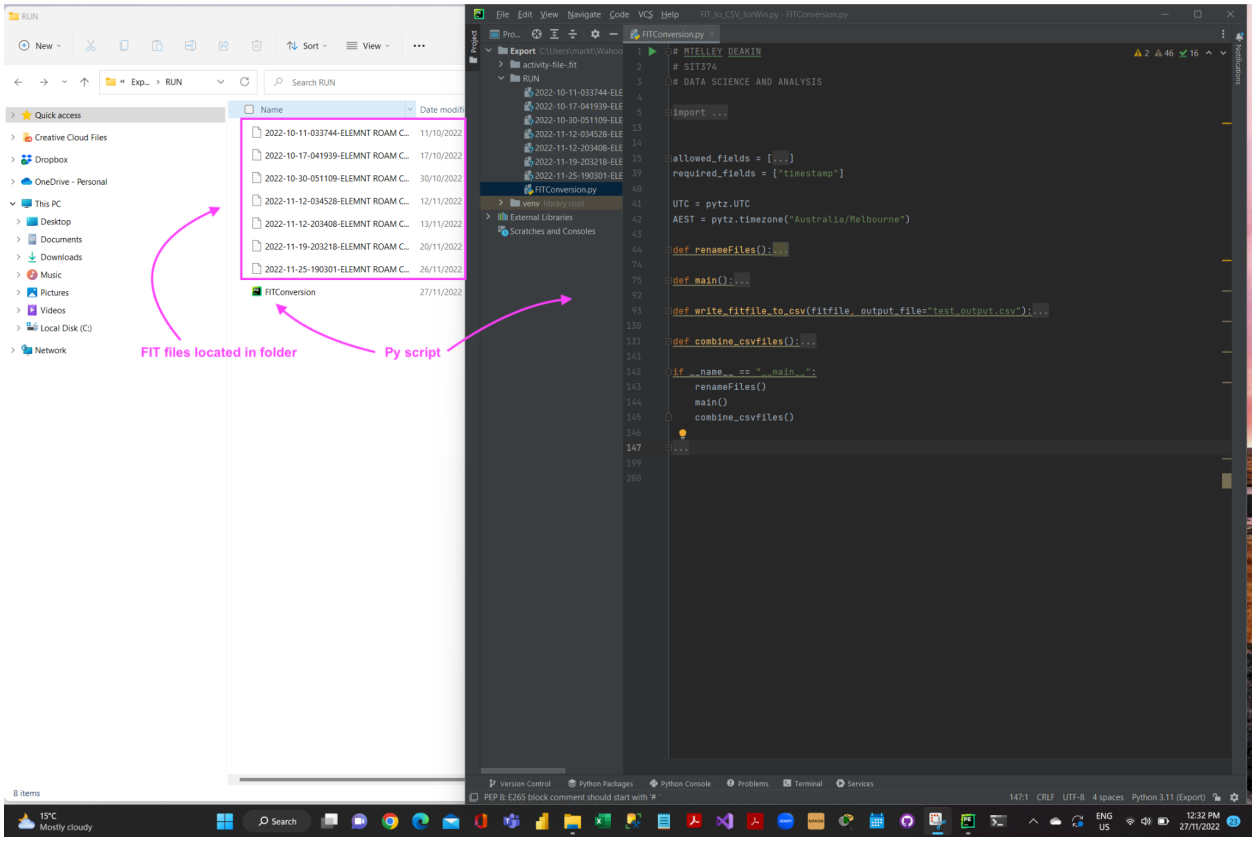
- https://github.com/rdchip/FIT-to-CSV-converter-for-windows
- https://www.python.org/downloads/release/python-3110/
- https://www.jetbrains.com/edu-products/download/#section=pycharm-edu
- https://pypi.org/project/fitparse/#description
- https://github.com/dtcooper/python-fitparse
- https://github.com/ekapope/Combine-CSV-files-in-the-folder/blob/master/Combine_CSVs.py

Data Cleansing and Organisation

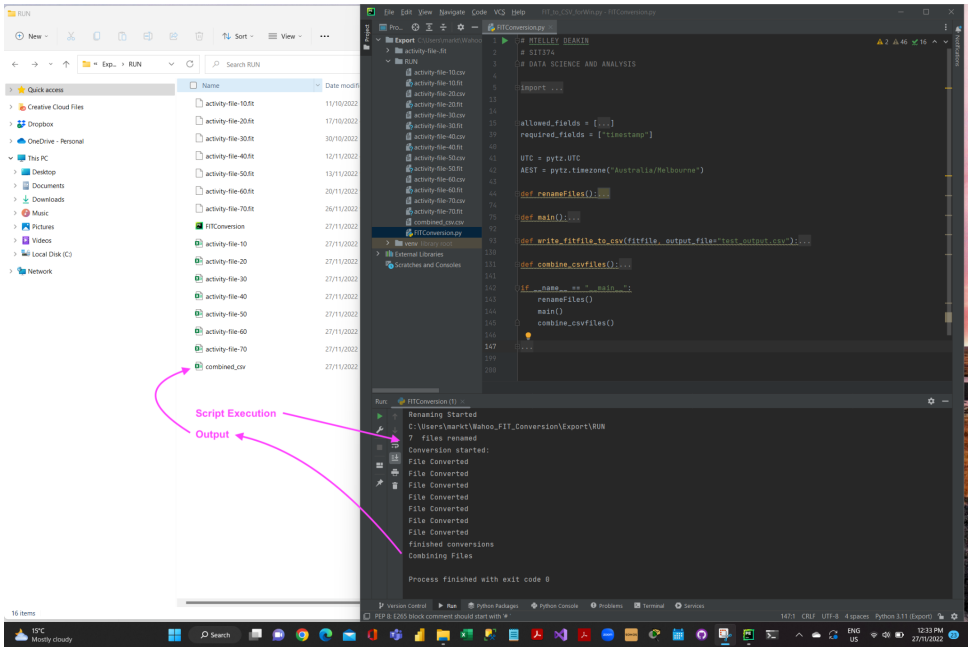
Given the scale of data being procured, BigQuery (Google) was used to house and organise the data. Due to the size of the user CSV files, they were stored in a Google Cloud Storage bucket in preparation for housing the data in a data table. On review, the schema of the CSV file relating to data types also presented challenges; all data types were set as STRING values initially to avoid data table creation errors. The schema file used is located within the Github repo.

Key Operations (Wahoo)

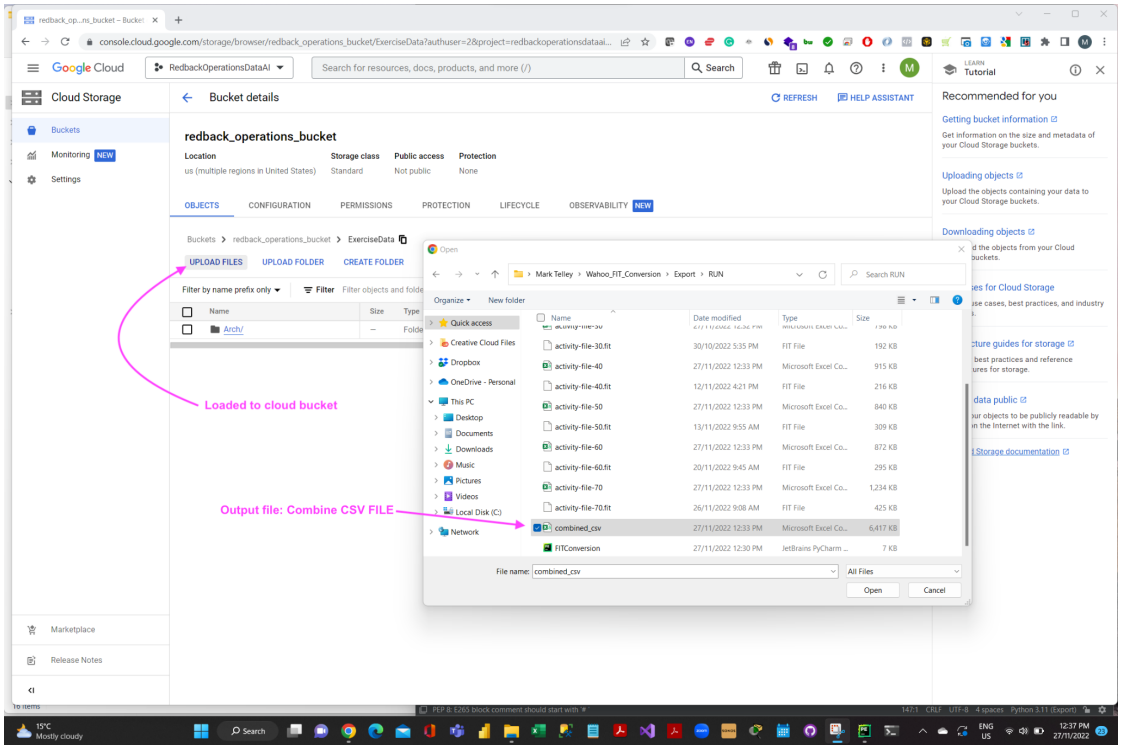
Open the script file, drop FIT files into the correct folder and run.



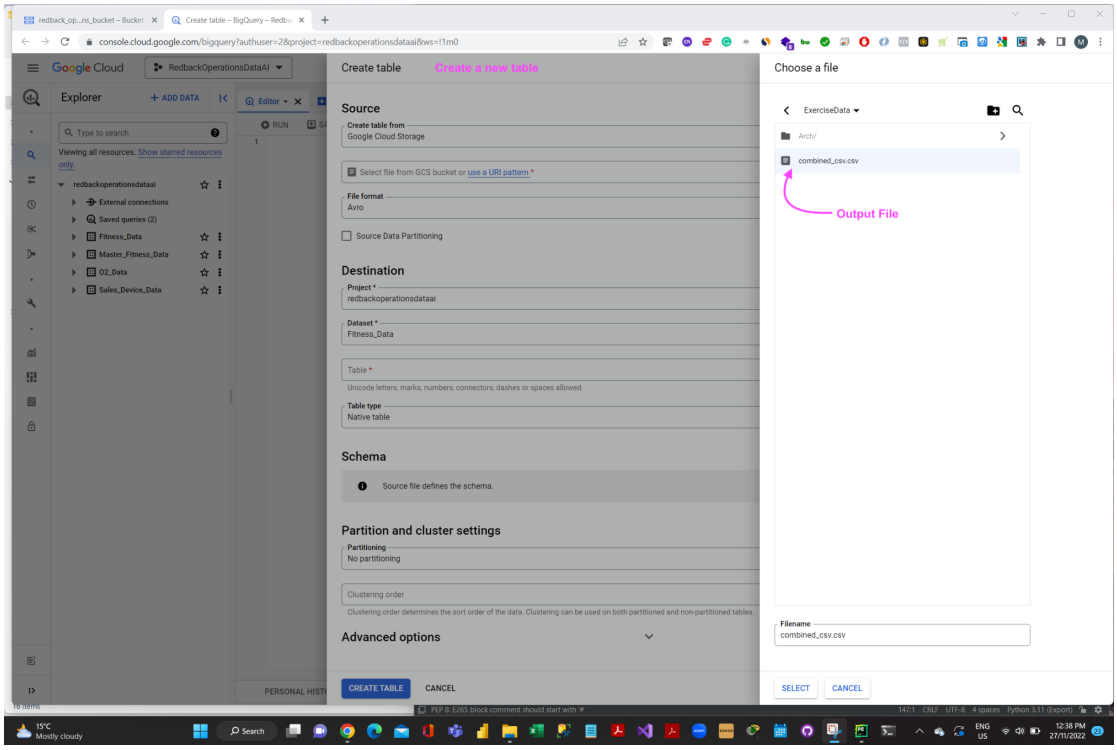
After the script has finished, a combined CSV file will be available in the same location + all single FIT files have been renamed and converted. Upload the combined CSV file to a Google Cloud storage bucket:



Upload the CSV file to a Google Cloud Storage bucket:

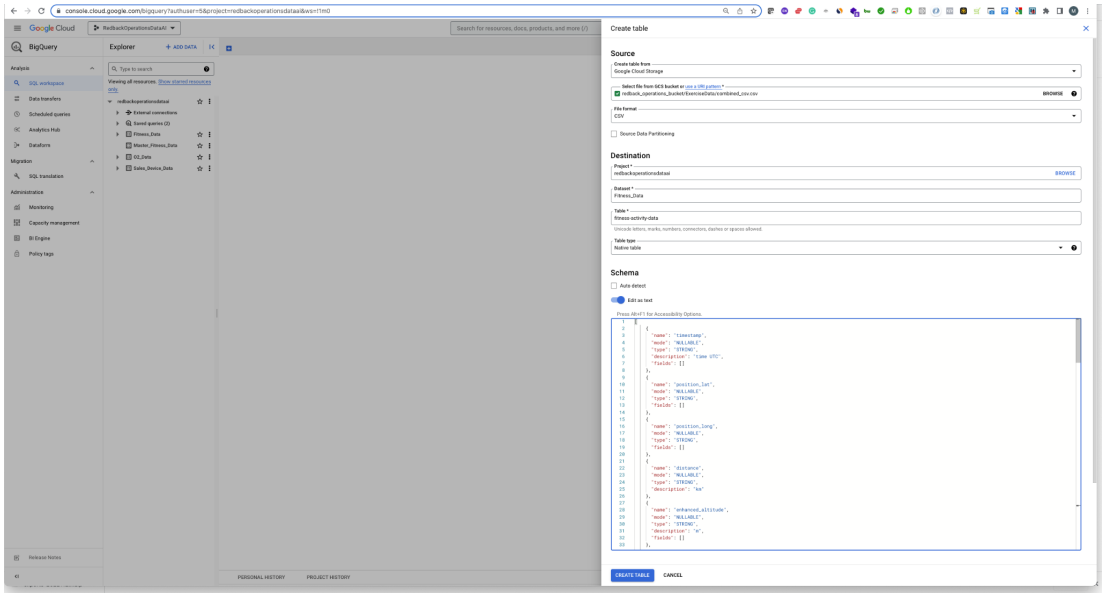


Create a data table: Load from the cloud storage bucket:

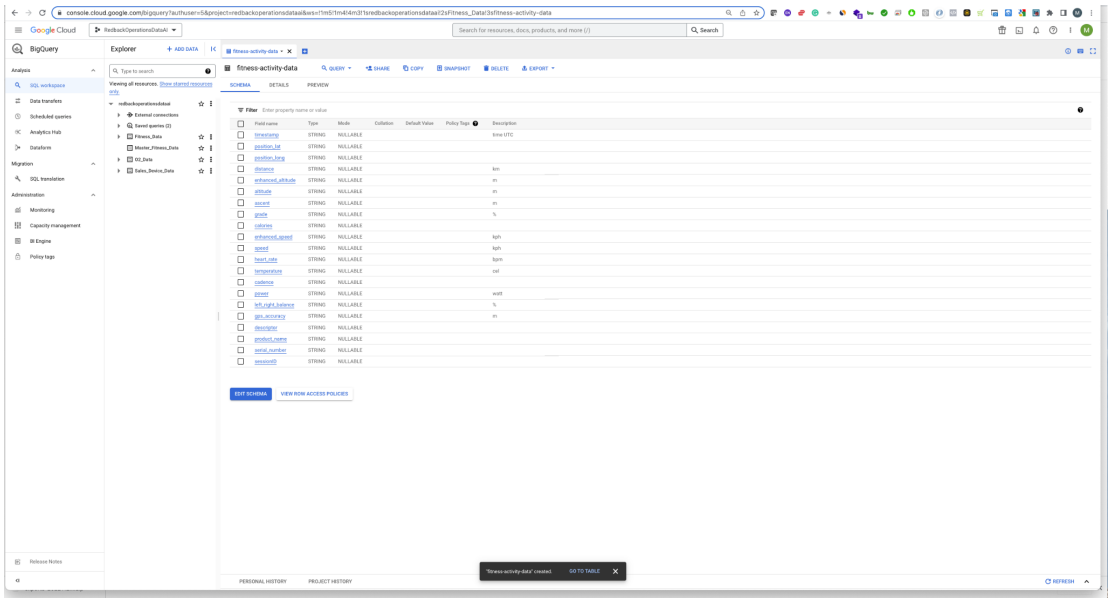


When creating a data table, local files up to 100MB can only be loaded, hence the need to store them in a Cloud bucket.

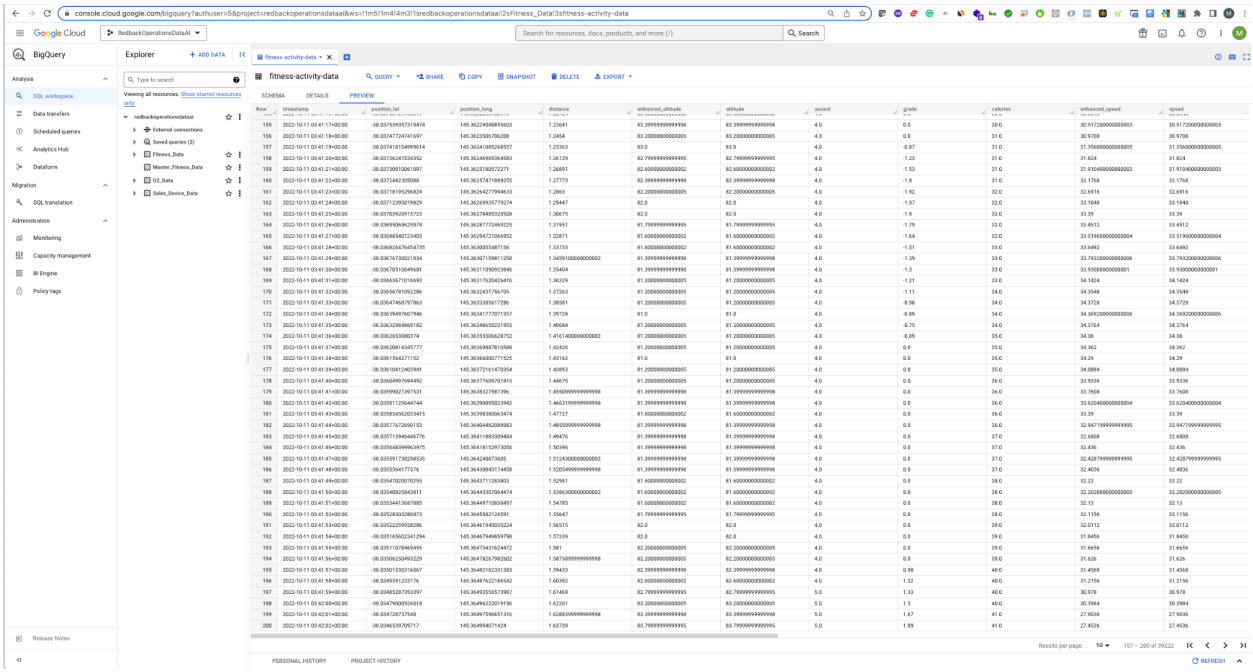
Copy in the table schema JSON code:



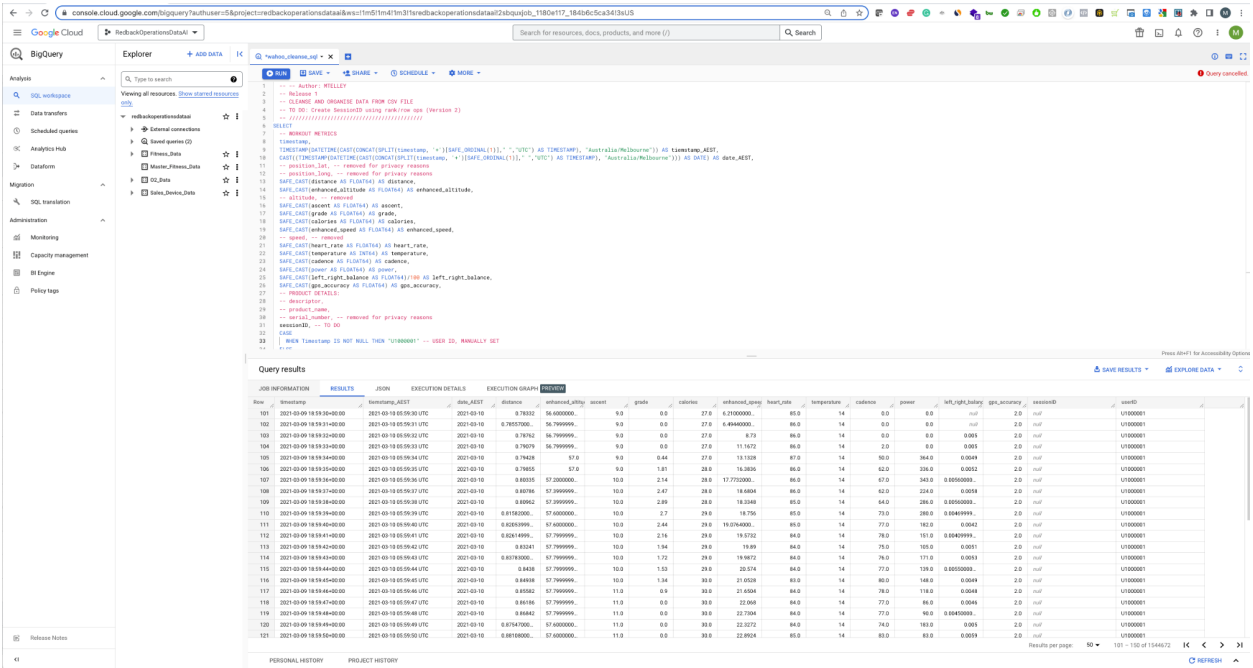
The table has been created, with singular data types:



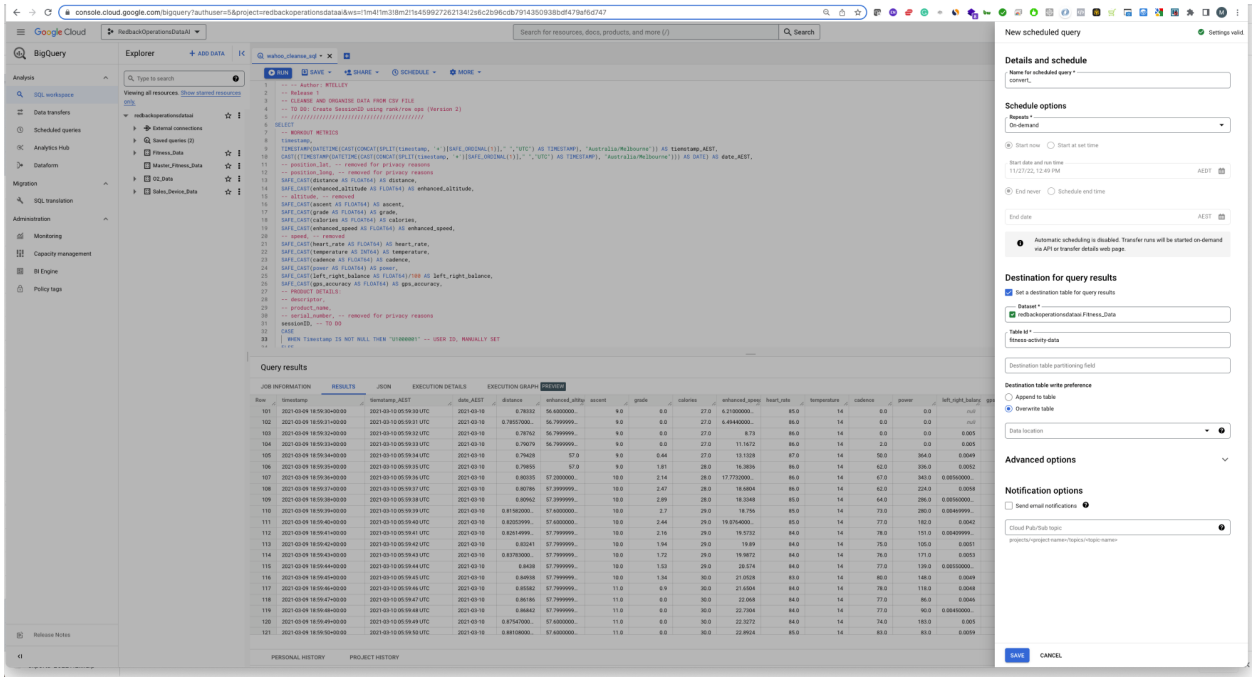
Preview of data:



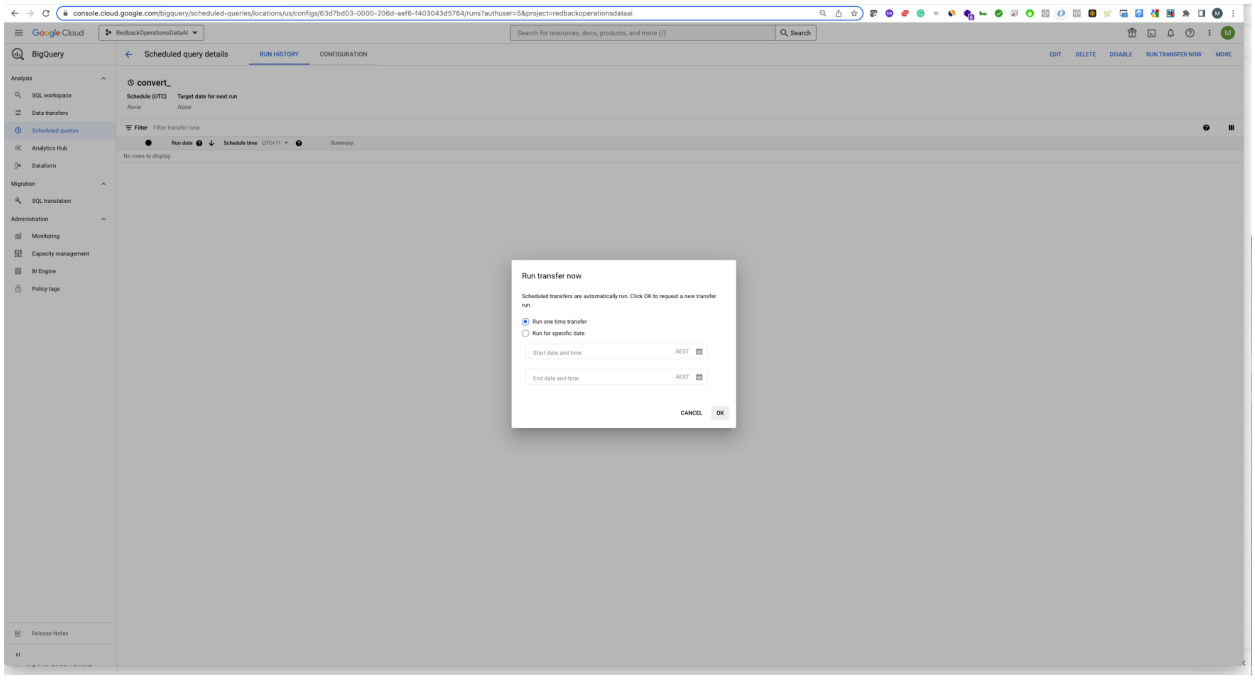
Run script to clean up the data table; The script will correct data types, key user data manually added, sensitive data removed and some filtering logic applied:



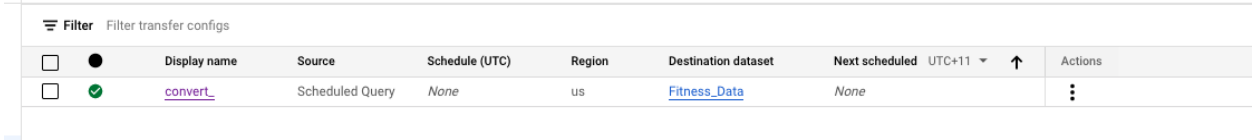
Schedule the query to create (overwrite) a new table



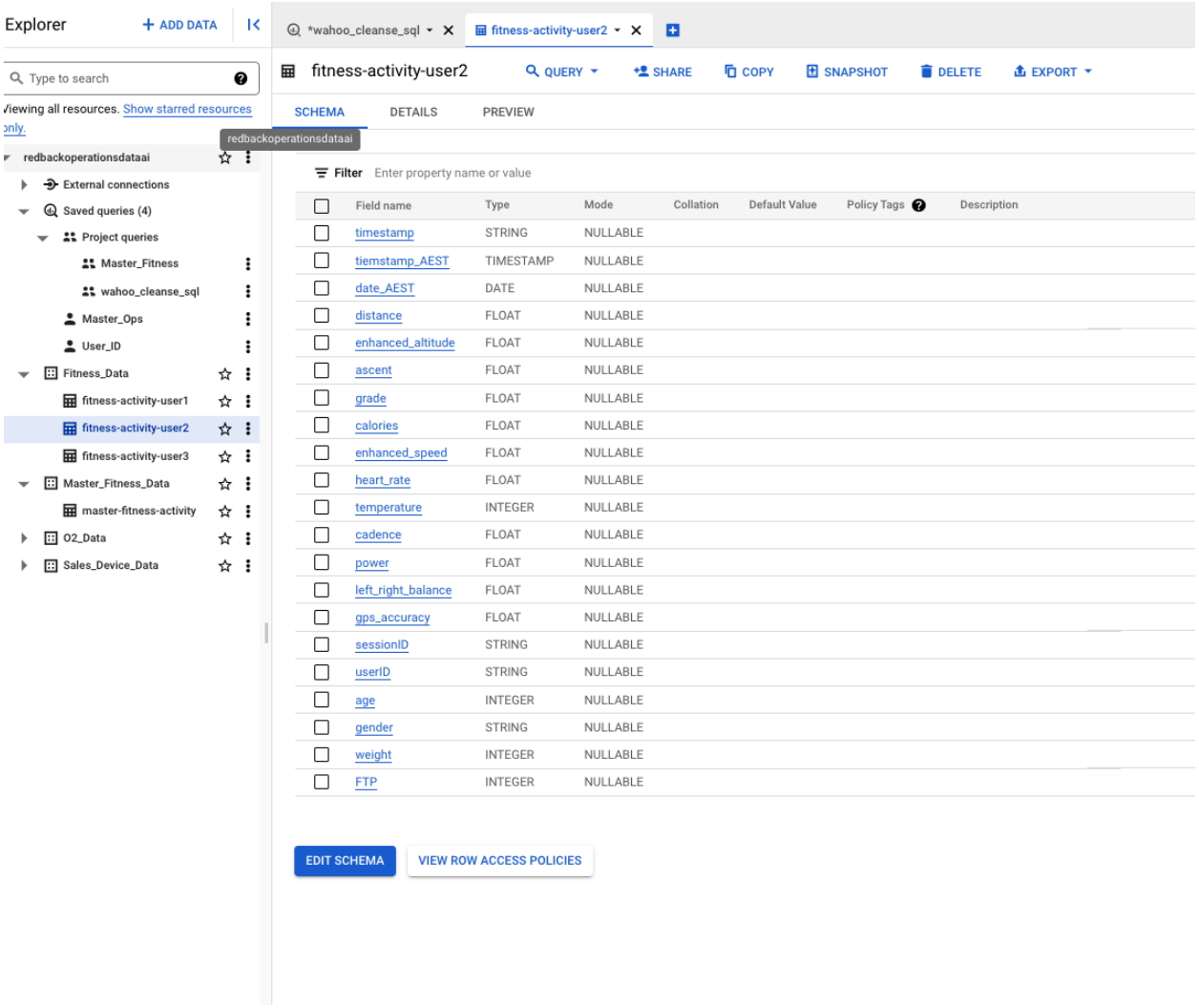
Then run the transfer:



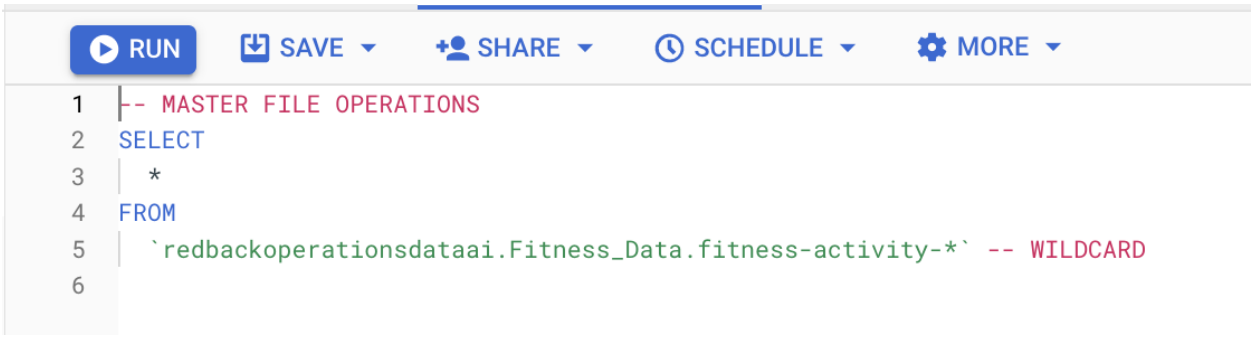
When complete, the green tick will indicate the transfer is complete:



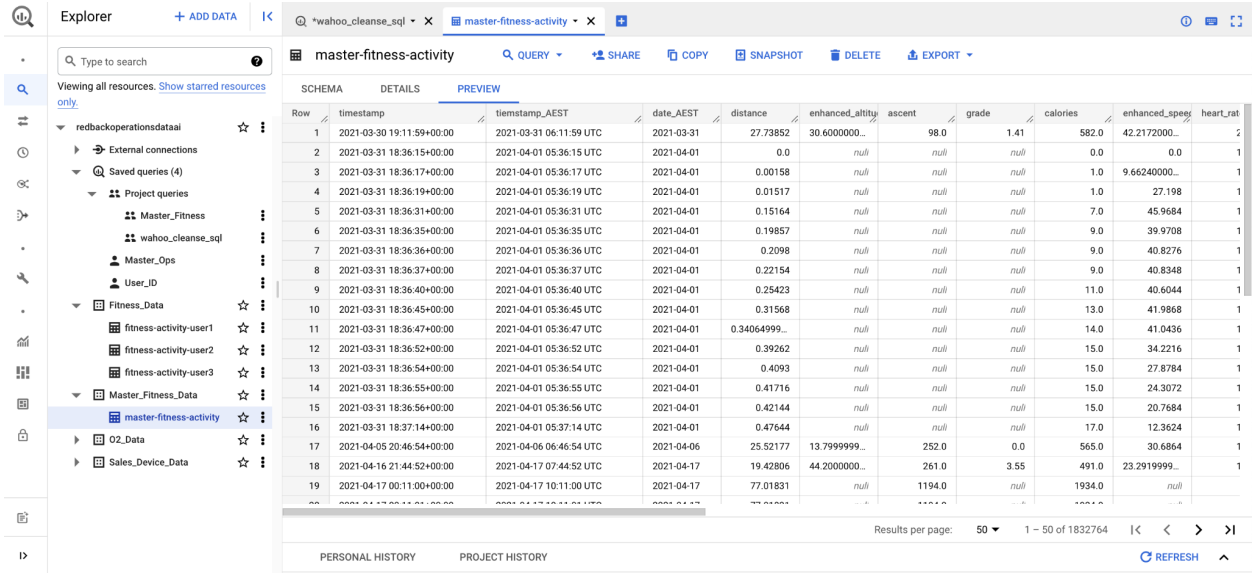
Updated Schema (new data types etc):



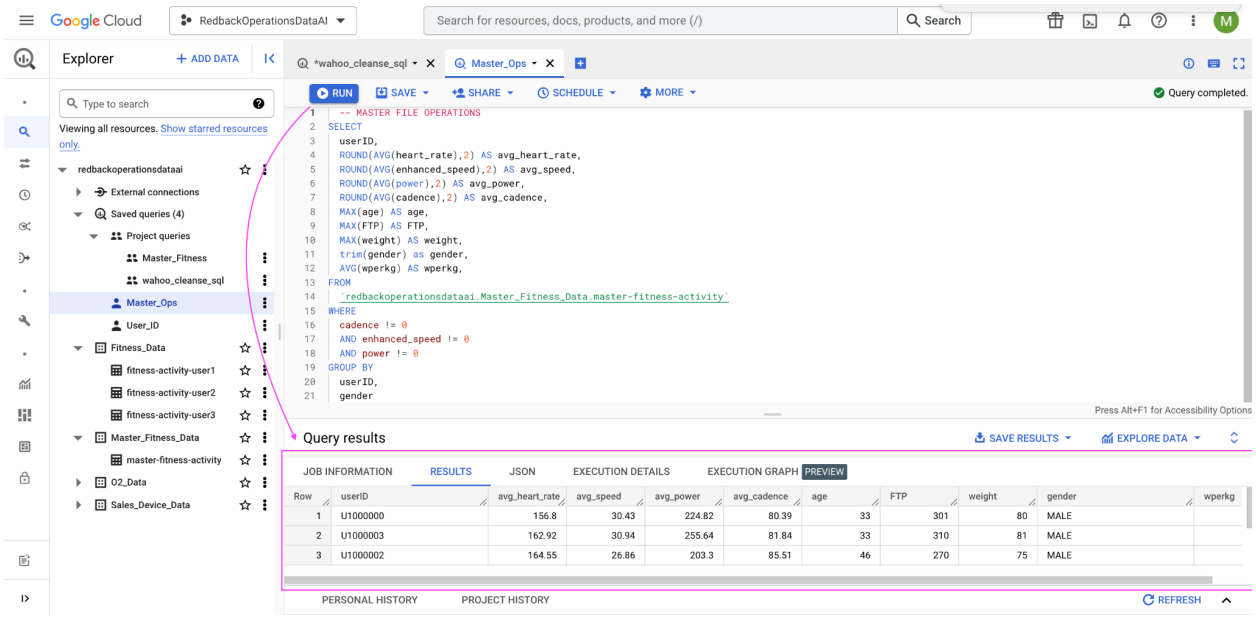
The following query is then scheduled: Given the consistent naming conventions, a wildcard operation can be used to join user tables together:



The mater table is created which houses all user data:



This master table will be used as a key resource; The table can be queried, connected to Tableau etc for data analysis and modelling efforts:



The process was replicated for additional users.

Key Operations (Garmin)

There was a slight change in the set of operations for Garmin-derived data. As a result, certain attributes were omitted i.e., Left/Right balance. To simplify operations and in the interest of time, only one user with Garmin-derived data was included in the final data set. Further work is required to support and update key operations for Garmin-derived data.

Access to Google Console Assets

Google Console Storage

Access to Google Cloud Storage and the relevant project buckets has been restricted due to data privacy requirements and billing controls; I'm using my own personal Google Cloud Storage account to store combined CSV files.

Google Bigquery

Relevant project team members have had access enabled for all datasets - Team members only have viewing privileges. Access details below:

Admin/Owner: mtelley@deakin.edu.au

Access information - watch this Loom video.

Other Notes

Please note, there is availability to query a data table when saving the query results to another data table. This method was not used for a range of reasons, mainly due to the transfer method providing added debugging capabilities should errors occur.

There is an outstanding task that requires some minor SQL development to allocate session IDs to workouts i.e, there can be more than one session in a day, such that each session must have a unique ID. This will be achieved using lag/over operations:

```
WITH
  proto_1 AS (
  SELECT
  timestamp,
    LAG(Distance) OVER(ORDER BY timestamp ASC ) AS PreviousRow,
  CASE
      WHEN Distance<LAG(Distance) OVER(ORDER BY timestamp ASC ) THEN 1
    ELSE
    0
  END
    AS endSession,
  FROM
    `redbackoperationsdataai.Fitness_Data.fitness-activity-user1`
  ORDER BY
    Timestamp ASC)
SELECT
  *,
FROM
  proto_1
WHERE endSession != 0
```

Query results

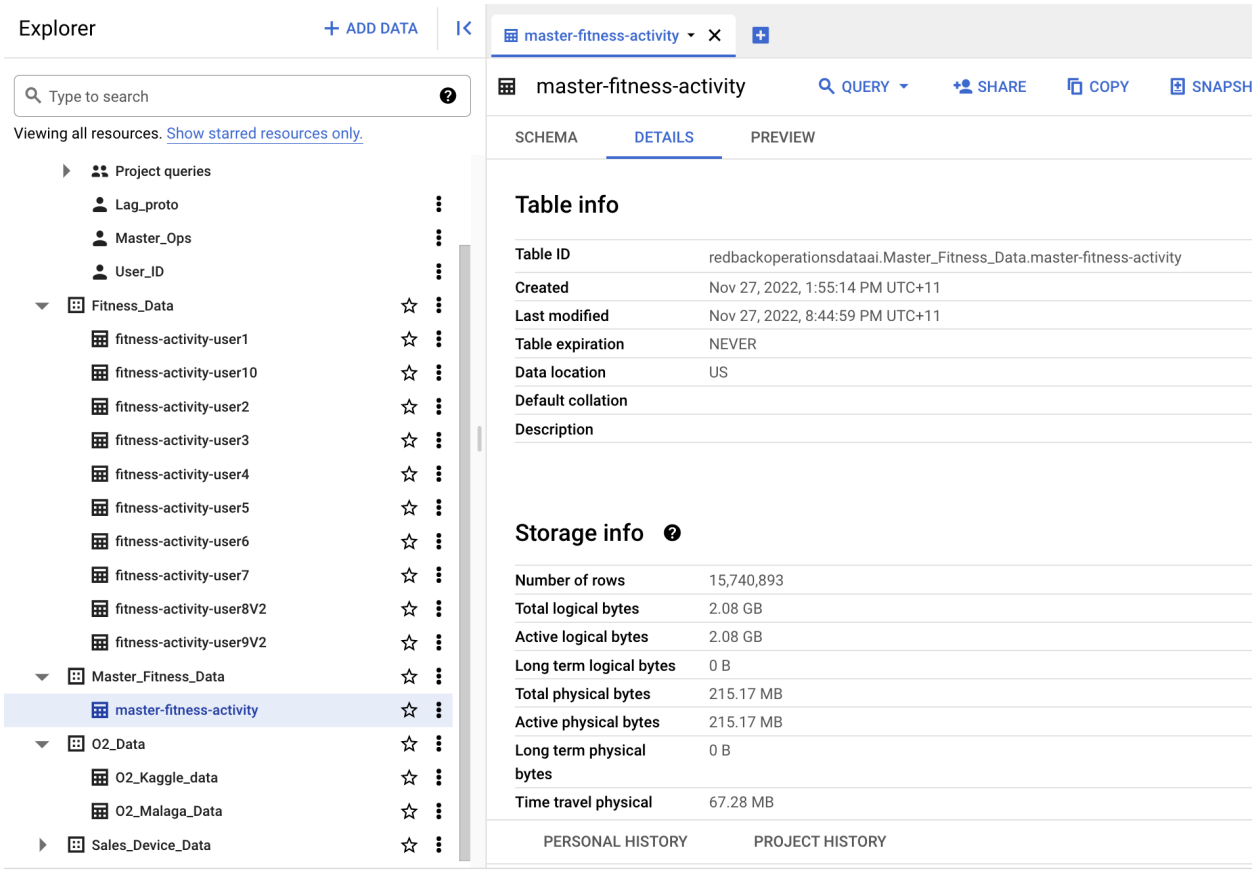
JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	timestamp	PreviousRow	endSession
1	2021-03-09 22:22:07+00:00	62.928	1
2	2021-03-11 22:43:39+00:00	17.13455	1
3	2021-03-13 19:53:22+00:00	10.30559	1
4	2021-03-14 20:54:21+00:00	73.90983	1
5	2021-03-15 07:33:59+00:00	5.18281000…	1

Outcome

Over 15,740,893 seconds (+4300 hours) of exercise data has been collated and stored in a single data set. This data set will be used to create data visualisations, train/test models etc. Moreover, the outcome of this exercise will help inform key data collection efforts for the in-game experience i.e., how to collect data, deploy models and then show the users (at the end of their workout or session) a data visualisation that represents their efforts.

Next page for a summary of the master table:


Explorer
ADD DATA
Type to search
Viewing all resources. Show starred resources only.
Project queries
Lag_proto
Master_Ops
User_ID
Fitness_Data
fitness-activity-user1
fitness-activity-user10
fitness-activity-user2
fitness-activity-user3
fitness-activity-user4
fitness-activity-user5
fitness-activity-user6
fitness-activity-user7
fitness-activity-user8V2
fitness-activity-user9V2
Master_Fitness_Data
master-fitness-activity
O2_Data
O2_Kaggle_data
O2_Malaga_Data
Sales_Device_Data
master-fitness-activity
master-fitness-activity
QUERY
SHARE
COPY
SNAPSH
SCHEMA
DETAILS
PREVIEW
Table info
Table ID
redbackoperationsdataai.Master_Fitness_Data.master-fitness-activity
Created
Nov 27, 2022, 1:55:14 PM UTC+11
Last modified
Nov 27, 2022, 8:44:59 PM UTC+11
Table expiration
NEVER
Data location
US
Default collation
Description
Storage info
Number of rows
15,740,893
Total logical bytes
2.08 GB
Active logical bytes
2.08 GB
Long term logical bytes
0 B
Total physical bytes
215.17 MB
Active physical bytes
215.17 MB
Long term physical bytes
0 B
Time travel physical
67.28 MB
PERSONAL HISTORY
PROJECT HISTORY