Project 10: CovIdentify

Sean Fiscus, Alyssa Shi, Yamil Lopez-Ruiz, Emmanuel Mokel
Duke University BIG IDEAS LAB

Project Background

CovIdentify is an ongoing study at Duke University focusing on using data from wearables to predict and diagnose COVID-19 and the Flu. Studies have shown that biomarkers like heart rate and steps coming from wearable devices such as Fitbit, Garmin, and Apple watches can indicate signs of COVID-19 several days earlier than symptoms arise.[1,2] This project builds on work previously done by the students in the Masters in Interdisciplinary Data Science (MIDS) program, and the BIG IDEAS LAB led by Dr. Jessilyn Dunn.

Duke MIDS
Master in Interdisciplinary Data Science

COVIDENTIFY
A DUKE UNIVERSITY STUDY

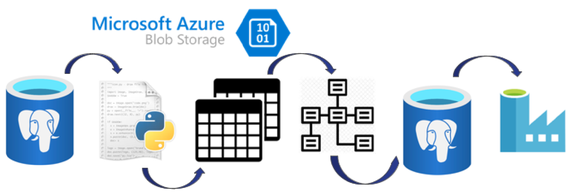
Objectives

1. Improving memory usage of existing pipeline

2. Adapting pipeline to incorporate efficient methods for data storage, remove redundant steps.

3. Implementing all data sources for preprocessing

4. Assessing smart watch user adherence to data collection in data



Methods/Applications

Azure Machine Learning

- A cloud-based service with virtual computers
- Where data was extracted, transformed, then loaded (ETL) to storage

Azure Machine Learning

Azure Data Studio

- Houses SQL databases where data is stored both before and after processing

Results

1. Memory Efficiency Improvements

Problem: After data was read into python from the initial SQL database, it was appended to a quickly-growing Pandas dataframe at two different locations in the pipeline. Using Azure Cloud Services, the team needed to find a way to avoid this large RAM storage.

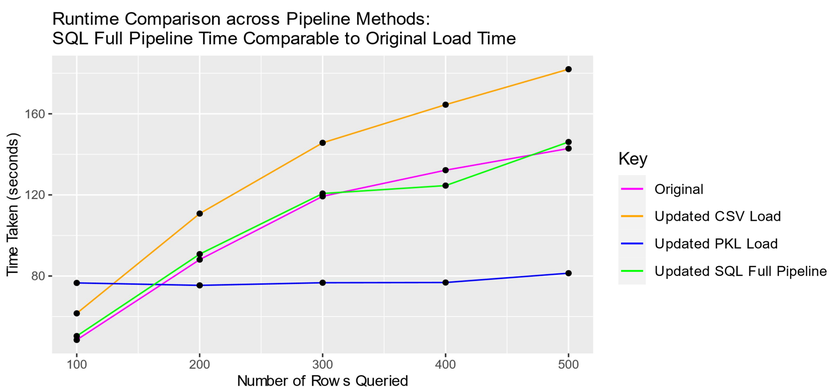
Solution: Implemented batch-pulling method to pull from the database; added a row-by-row normalization and exportation to csv process.

2. Pipeline Speed and Simplification

Problem: Multi-step pipeline with redundant data transfers and stores; data upload speeds now slower with new implementation for improving memory efficiency

Solution: Tested use of pickle files as potentially faster data storage method; created Python scripts to write directly to new SQL database after normalization, cutting out need for csv/pickle.

Simplified pipeline: initial database to python script to final database (removes intermediate Blob Storage step)



Comparison of time taken for query, normalization, and writing to data storage type in original pipeline, modified csv upload and pickle upload with time taken for query, normalization, and writing to database in simplified pipeline.

3. Data Type Expansion

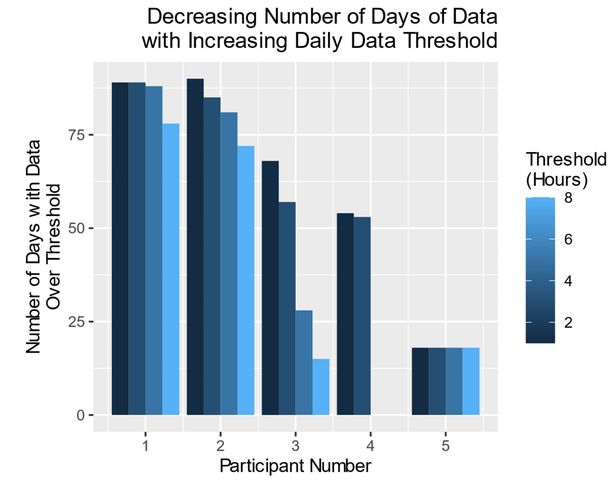
Problem: Data types from users absent from final database

Solution: Added pipeline capability to normalize minute-by-minute Fitbit resting heart rate, steps, and sleep data recorded after user sign-up

4. User Adherence Analysis

Problem: CovIdentify's smartwatch data currently unassessed in terms of users' daily and inter-day adherence to recording data and is thus incomparable to other studies.

Solution: Created SQL scripts to subset data by users' adherence to devices. Can now set and monitor thresholds for minimum recording adherence for daily and inter-day timespans.



For four thresholds (2, 4, 6, and 8 hours), the number of days with data collected for longer than threshold is calculated for five selected participants.

Accomplishments/Next Steps

Key Accomplishments

- Reduced memory usage of the ETL pipeline
- Expanded type and resolution of data in storage
- Created a foundation for future wearable data processing

Next Steps

- Integrating iOS data
- Continue analyzing and applying machine learning algorithms to data

References

1. Mishra, Tejaswini, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, et al. "Early Detection Of COVID-19 Using A Smartwatch." *MedRxiv*, July 7, 2020, 2020.07.06.20147512. https://doi.org/10.1101/2020.07.06.20147512.
2. Natarajan, Aravind, Hao-Wei Su, and Conor Heneghan. "Assessment of Physiological Signs Associated with COVID-19 Measured Using Wearable Devices." *Npj Digital Medicine* 3, no. 1 (November 30, 2020): 1–8. https://doi.org/10.1038/s41746-020-00363-7.