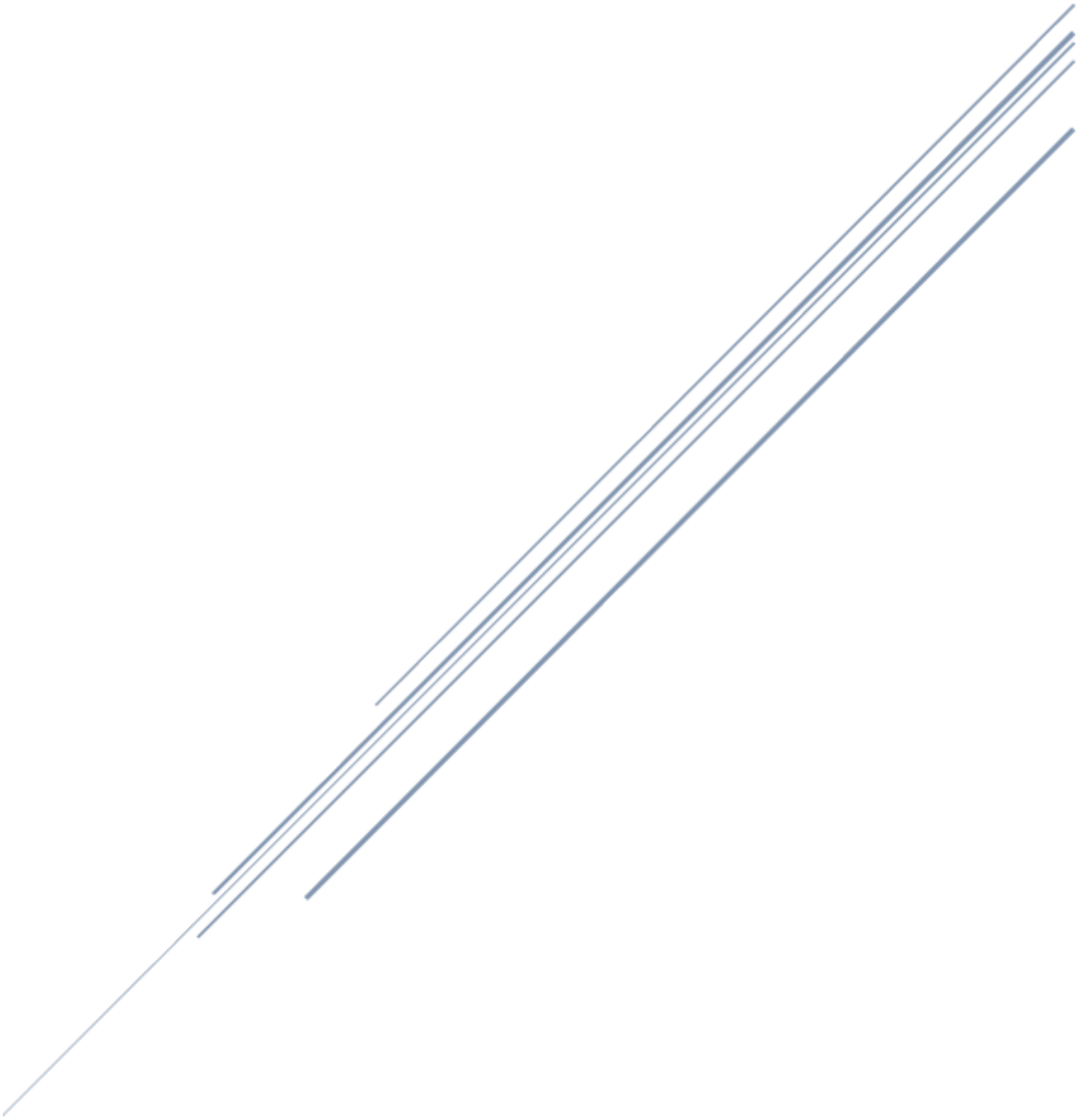
INTERNSHIP REPORT

Priya Majalikar

1927



ZiMetrics Technologies Private Limited

Goa University

Report of Internship done at ZiMetrics Technologies Private Limited

SUBMITTED BY

PRIYA MAJALIKAR

1927

UNDER THE GUIDANCE OF

Mr. Omkar Prabhu
(Center Head - Goa RDC. at ZiMetrics Technologies)

Mr. Aditya Kunkolienkar
(Senior System Analyst at ZiMetrics Technologies)

Internship Certificate

501, Lunkad Sky Station,
Viman Nagar, Pune –411014
Contact No.: 020-41230949
Web: www.zimetrics.com

ZIMETRICS TECHNOLOGIES PRIVATE LIMITED

Date: 31/05/2022

TO WHOM IT MAY CONCERN

This is to certify that Miss. Priya Majalikar, a student of MCA Goa university roll no- 1927.is currently undergoing long internship program at Zimetrics Technologies Pvt Ltd.(10th Jan 2022- Till date)

During the period of her internship program, she is working on Developing ETL for Telematics solutions as an engineering intern.

Her sincere efforts and dedication towards work are greatly appreciated. She is exhibiting overall very good conduct, flexibility and professionalism during this period.

As per the campus drive conducted on 8th November 2021, she will be working with us as a full-time employee from 9th August 2022 onwards.

AshwiniBarve
AshwiniBarve (May 31, 2022 15:58 GMT+5.5)

Sincerely,
For Zimetrics Technologies Pvt Ltd
Ashwini Barve
GM Operations

CIN: U72900PN2015PTC153852, GSTIN 27AAACZ8110B1ZO
Register office: B-3, Konark Campus,Viman Nagar,Pune. 411014

GOA UNIVERSITY

GOA BUSINESS SCHOOL

CERTIFICATE OF EVALUATION

This is to certify that **Ms. Priya Majalikar** has been evaluated for the project work titled **"Report of Internship done at ZiMetrics Technologies Private Limited"** undertaken at **ZiMetrics Technologies Private Limited, Pune** in partial fulfilment for the award of the degree in Master of Computer Applications.

Examiner 1

Examiner 2

Place: Goa University

Date:

Dean, Goa Business School

Acknowledgement

It's great to get an opportunity to work as an intern in a company as you get to learn new technologies, get industry exposure and to put all the knowledge you have into practical work.

I am very grateful to get support from all the seniors initially when I started with the internship and so thereafter. The internship would not be complete without expressing gratitude to everyone who led me throughout the internship period.

I thank Mr. Vikas K Verma (Founder & President of Engineering at ZiMetrics Technologies) for giving me the opportunity to work as an intern at ZiMetrics Technology Private Limited.

I would like to thank Mr. Omkar Prabhu (Center Head - Goa RDC.) for giving me the opportunity to intern at ZiMetrics Technology Private Limited, for making me feel comfortable in the new environment and also to guide me along the way.

I sincerely express gratitude to Mr. Aditya Kunkolienkar (Senior System Analyst at ZiMetrics Technologies) for always being supportive and guiding me in completing the project.

I extend gratitude to Ms. Swati Patil (Head - HR ZiMetrics Technologies Pvt. Ltd.), Mr. Shantanu Waghmare (Associate Manager - Human Resource at ZiMetrics Technologies Pvt Ltd), Yashwanti Patil (HR Manager), Maseera Shaikh (HR Generalist, ZiMetrics) and Mrs. Ashwini Barve (GM Operations at ZiMetrics Technologies) for helping me out whenever needed.

I would also like to thank Ms. Ankita Raul (Office Admin, ZiMetrics) for always being so kind, friendly and caring towards me.

I thank M.S. Dayanand (Dean, Goa Business School, Goa University), Mr. Ramdas Karmali (Prof. and TPO, MCA, Goa Business School, Goa University), Mr. Jarret Stevan Anthony Fernandes (Assistant Prof, MCA, Goa Business School, Goa University) and all the faculty of MCA, Goa University for their constant encouragement and support during the project work.

I would like to thank all my friends for helping me enhance my skills and for encouraging me. Their constant support has been of central importance.

Finally, I would like to express my gratitude towards all my colleagues at ZiMetrics, especially Ashwin Kolgaonkar for being so helpful, kind and friendly in nature.

- **Priya Majalikar**

Table of Contents

Introduction

This report includes a short description of my full-time internship at ZiMetrics Technologies Private Limited.

I joined as an intern in ZiMetrics on 10th January 2022. This report contains all the necessary information about the company, the project I have worked on, the trainings I have received, and some other tasks that I completed in this internship period.

In the following sections I shall include information about the company, the work and culture over here. I shall also include details of the project I worked on, a brief description of the project, the modules I built and the tasks I completed in those modules.

This report emphasizes my learning experience and contribution to the organisation as an intern. This will describe the knowledge that I gained by successfully completing the tasks that were assigned to me.

I'll also be talking about the tools and technologies that were used followed by my internship timeline. I shall conclude by sharing my experience and how it has helped me to grow, both, on the personal and professional front.

Company Profile

ZiMetrics is a niche technology provider and solutions enabler for IoT, Machine data, Big Data analytics, and Data Science.

ZiMetrics is a Confluent & HashiCorp Partner and recognized as the "Top 20 IoT Solutions Providers" by CIO Outlook APAC.

Founded in 2015, ZiMetrics today serves leading global enterprises across Industrial, Oil & Gas, FMCG, MedTech, Internet Advertising, and Retail. Learn of our expertise and transformational value creation stories in Digital Media Analytics, Retail Analytics, Marketing Analytics, Customer Data Analytics, Video Analytics, Big Data, IOT, Machine Vision, Natural language Processing, Machine Learning, and Deep Learning.

Research and Innovation at Core

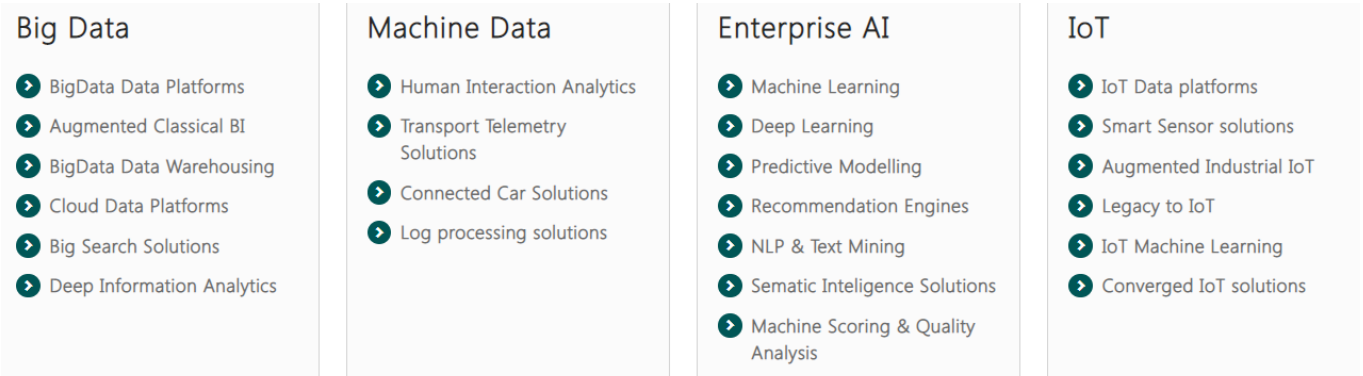
Innovation and research are at the center stage for ZiMetrics. ZiMetrics is heavily open-source-driven and invests heavily in research across data engineering, decision models, computer vision, and embedded technology.

Across Locations & Growing

ZiMetrics is headquartered in India, Pune. ZiMetrics has regional delivery centers in Goa, Delhi, NCR, and Bangalore (upcoming) in India. ZiMetrics has network delivery centers in Canada, British Columbia, and in the USA which helps in delivering a seamless cross-time zone network execution experience.

The fundamentals and aims of ZiMetrics are as follows:

- One that combines data sciences, domain disciplines, and data technologies into crafts that create transformational customer value.
- One that puts the customer's interests first.
- And one that would be recognized as one of the great places to work.

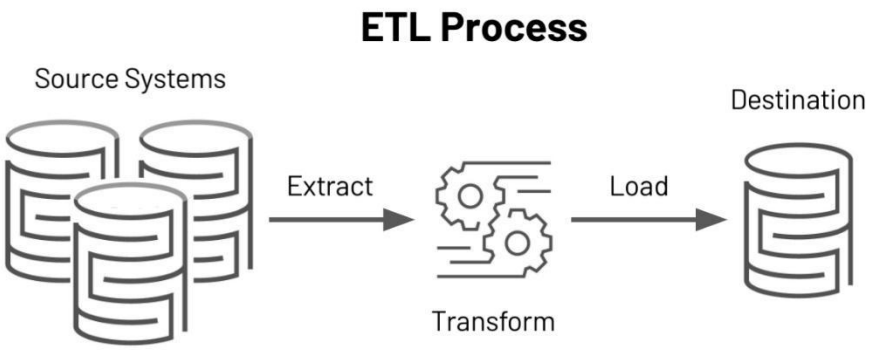


Project – Telematics Data Processing

Problem Statement

Given large amounts of data of EV vehicles being tracked continuously with a specific time gap (every 10 mins), the aim is to clean, transform, enrich and integrate all the data to analyze the performance of the vehicles in different areas such as elevations, mountain pass or mountain range.

Also, to create visualisations from the data to gain better understanding of the vehicle performance.



Overview

The project aim is to perform ETL operations on the given data using Google Cloud Platform. The Google Data Fusion platform is a service that helps users efficiently build and manage ETL/ELT data pipelines. The data pipeline would be scheduled to run at a specific time. The data pipeline would integrate data stored in different files within a bucket on the Google Cloud Storage, cleanse, transform and enrich the data and finally store it in Google BigQuery. The Google Roads API and Google Elevation API is used to enrich the data.

Platforms, Tools and Technologies used

- Google Cloud Platform
    - Compute Engine
    - Google Cloud Storage
    - Google Data Fusion
        - Wrangler
        - Studio
    - BigQuery
- Maven

- Git
- Postman
- Google Data Studio

My Contributions

I was assigned to the Telematics Data Processing project wherein my responsibilities were:

- Exploring the Alooma Platform.
- Exploring Google Data Fusion, create and run data pipelines using Data Fusion on different data sources and different data destinations.
- Creating a VM instance on GCP, scheduling and running program that makes API calls and stores responses in files.
- Exploring BigQuery Data warehouse.
- Understanding Google Roads API and Google Elevation API.
- Developing custom plugin to be used in Data Fusion.
- Exploring PowerBI and Google Data Studio.
- Perform visualisations using Google Data studio.

Details of my contribution are as follows:

Google Cloud Platform:

- Creating and managing a VM Instance with a custom machine type to meet the requirements and such that it is cost-effective.
- Static IP creation and calling XML API and storing the data to Virtual Machine.
- Scheduling a unix cron-job to run a python script after every 10 minutes which makes API calls and stores responses in files on VM instance.

Exploring Google Data Fusion:

- Creating and managing Data Fusion Instance.
- Connecting to different data sources such as REST API, files stored with Google Cloud Storage and data from BigQuery.
- Integrating data from these different data sources, transforming and enriching the data using built-in plugins in Data Fusion.
- Developing custom plugin to be used inside Data fusion and storing the data to BigQuery.
- Triggering one data pipeline to run after another.

BigQuery:

- Creating datasets and tables within them to load data and run queries on the stored data.
- Understanding the pros and cons of using BigQuery over other data storage platforms.

Custom plugin development to be used in Google Data Fusion:

- Creating a custom plugin to perform transformations that were not possible with built-in plugins available in Data Fusion.
- Added more functionality to the custom plugin to add more enrichments to the data (added a few more columns/fields).

Exploring PowerBI and Google Data Studio:

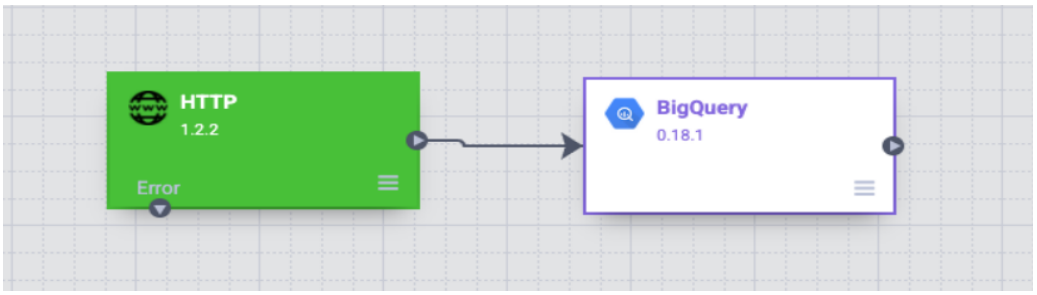
- Plotting different graphs in PowerBI as well as Google Data Fusion.
- Exploring the two platforms and comparing them in terms of their features.

Perform visualisations using Google Data studio:

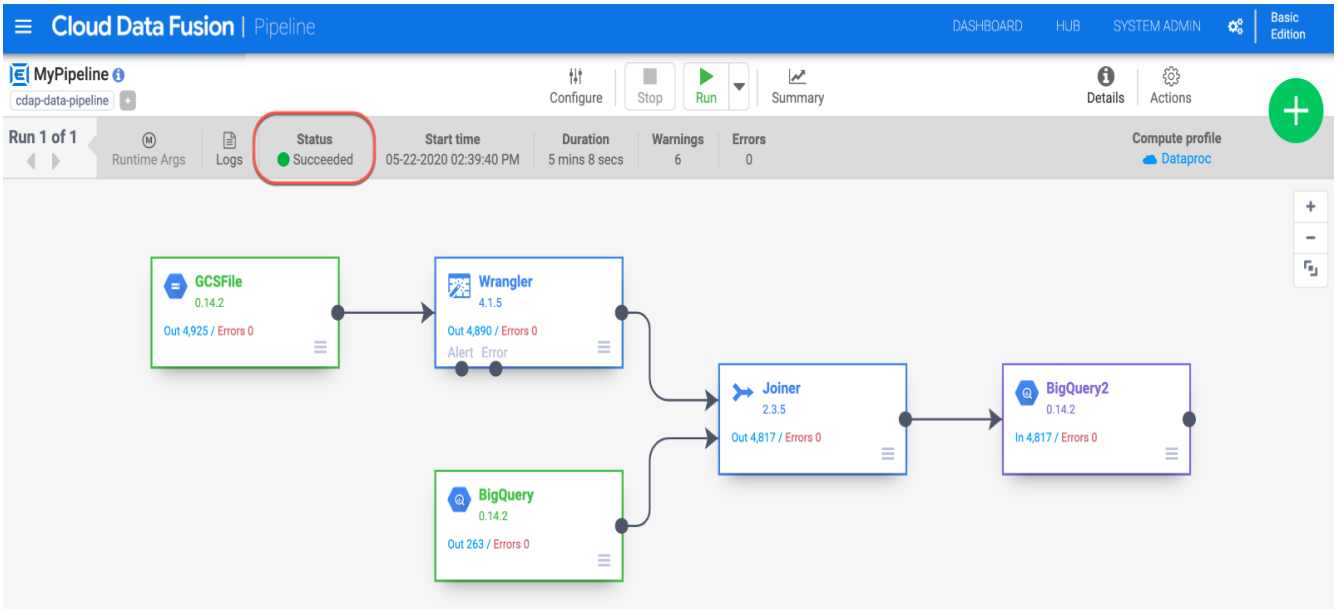
- Connecting with BigQuery data using BigQuery connector and using custom query.
- Creating charts, including line, bar, and pie charts, geo maps and bubble graphs, paginated data tables.
- Plotting the paths taken by the EV vehicles on geo map, using the drill-down feature provided by Data Studio.

Screenshots

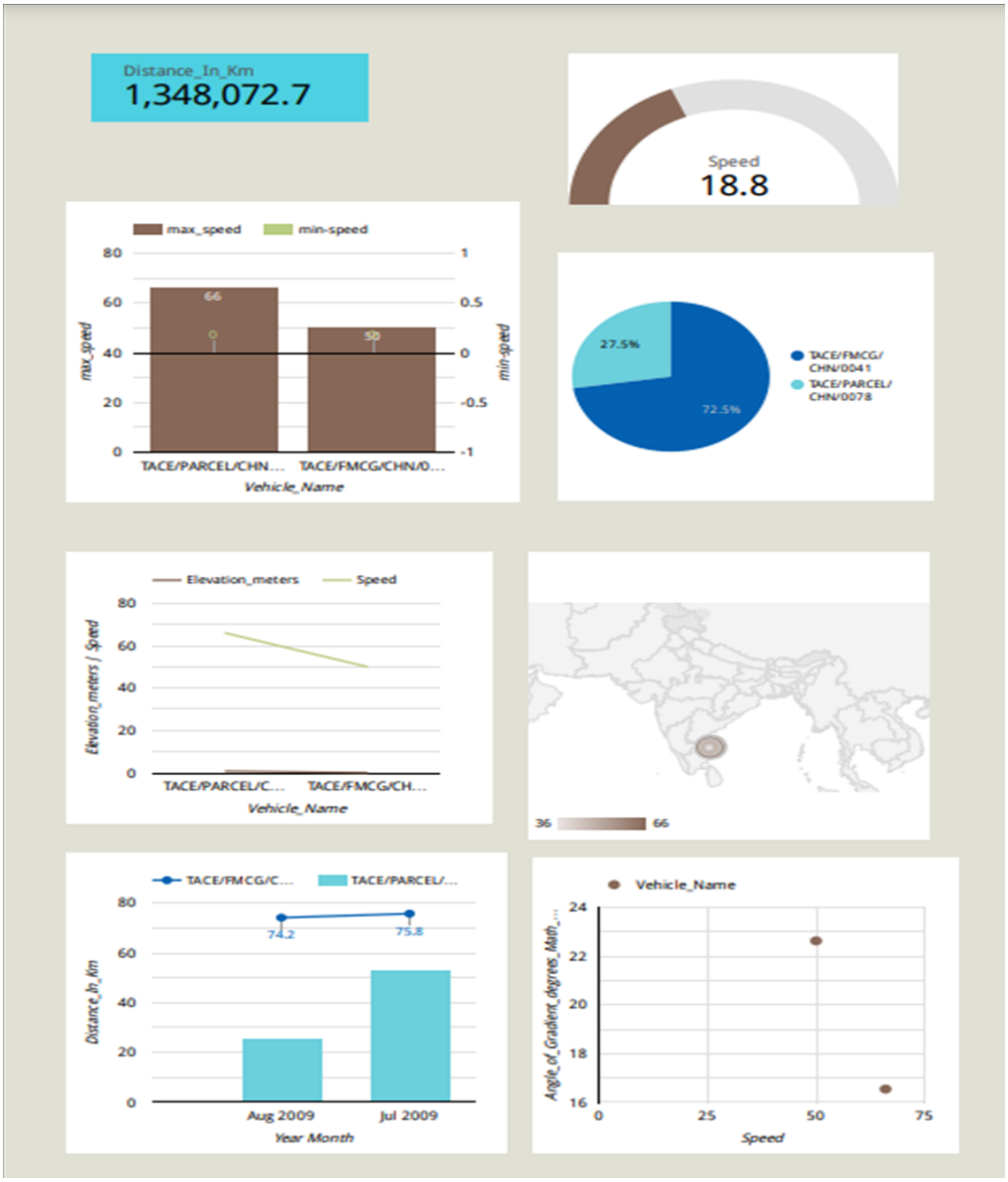
Static API Batch Pipeline:

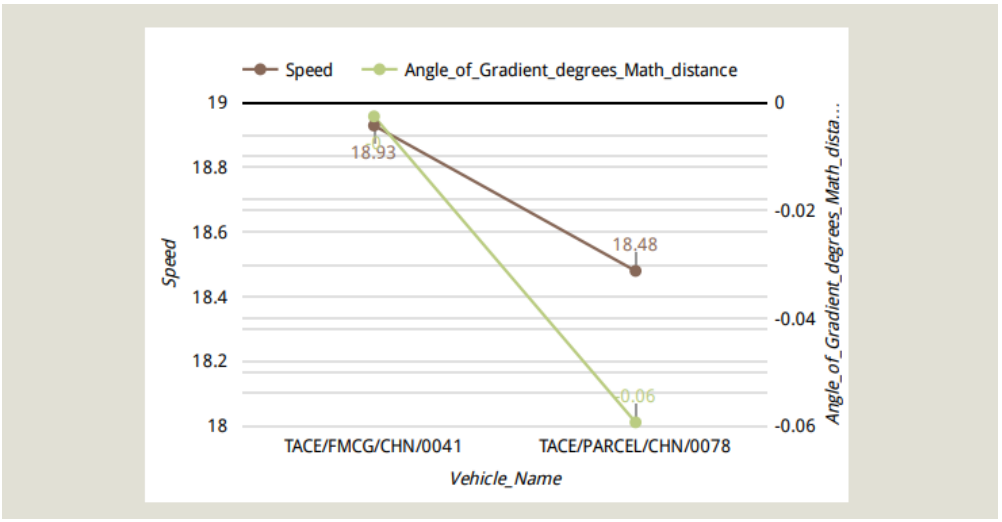


Integrating multiple data sources and storing the data in BigQuery warehouse:



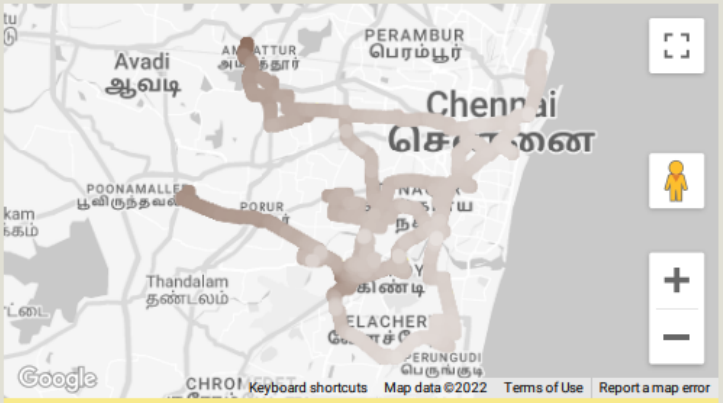
Visualisations using Google Data Studio:





	Vehicle_Name	Final_Elevation_meters
1.	TACE/PARCEL/CHN/0078	17.84
2.	TACE/FMCG/CHN/0041	27.41

1 - 2 / 2



Final_Elevation_meters 1 27.41

	final_location	Vehicle_Name	Trip_ID	Final_Elevation_meters	Speed
1.	13.120205, 80.1495...	TACE/FMCG/CHN/0041	5	26.04	21

1 - 100 / 37000

Other tasks

Consuming Rest Api:

- Consuming REST API in XML format and to get specific data using xpaths.
- Learning flask framework (basic) to call REST API.

Tools and technologies used:

- Postman
- Flask

Certifications and Trainings Completed under Internship and Self Study:

Coursera Course:

- Exploring and preparing your data with BigQuery

Other Self-Study:

- Learning python
- Learning MySQL
- Learning UNIX commands and shell scripting

Trainings Completed Under Internship:

- Java
- Git

HackerRank Preparations:

- Earned 3 stars in Java preparation.
- Earned 1 star in Python preparation.

Platforms, Tools And Technologies Used

Google Cloud Platform:

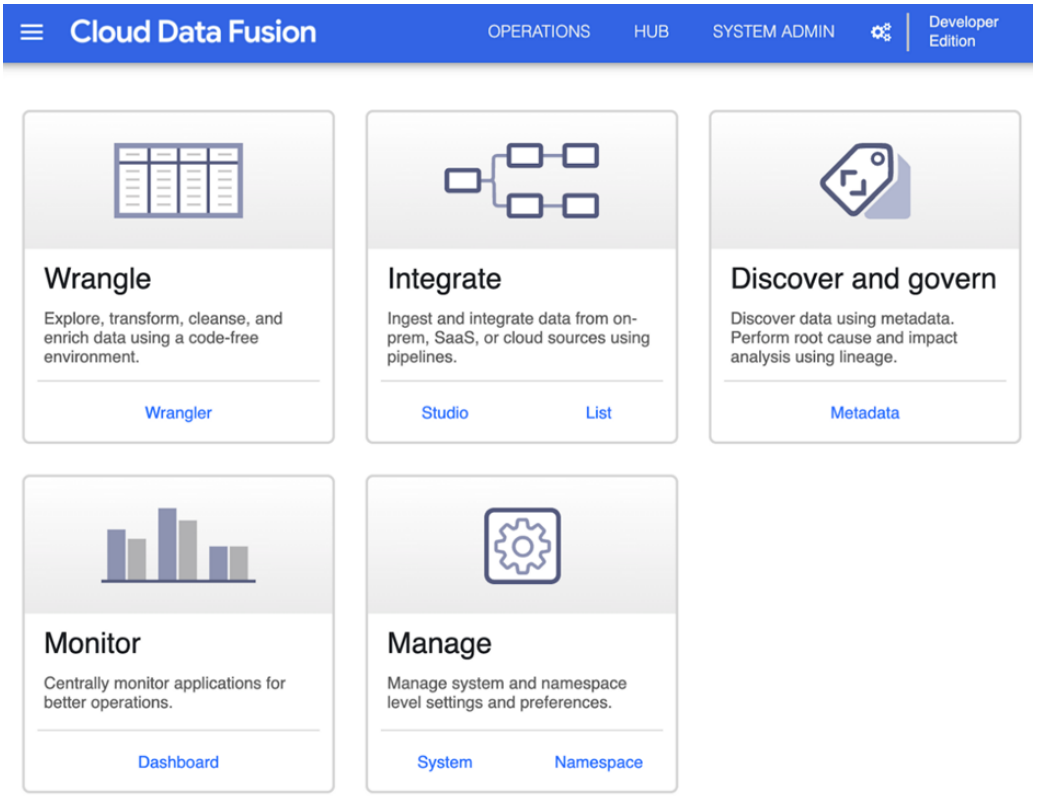
Google Cloud is a suite of public cloud computing services offered by Google. The platform includes a range of hosted services for compute, storage and application development that run on Google hardware.

- Virtual Machine
    - Create a VM with a custom machine type:
        - You can create a VM instance with custom virtualized hardware settings. Specifically, you can create a VM instance with a custom number of vCPUs and amount of memory, effectively using a custom machine type.
        - You can add memory to a machine type with no limitations per vCPU. You can add extended memory up to certain limits based on the machine type.
    - Reserving A Static External IP Address:
        - If a virtual machine (VM) instance requires a fixed external IP address that does not change, you can obtain a static external IP address for that instance.
        - An external IP address can be assigned to an instance or a forwarding rule if you need to communicate with the internet, with resources in another network, or with a resource outside of the Compute Engine.
        - Compute Engine supports two types of external IP addresses:
            - Static external IP addresses:
                - These addresses are assigned to a project long term until they are explicitly released from that assignment, and they remain attached to a resource until they are explicitly detached.
                - For VM instances, static external IP addresses remain attached to stopped instances until they are removed.
            - Ephemeral external IP addresses:
                - Ephemeral external IP addresses remain attached to a VM instance only until the VM is stopped and restarted or the instance is terminated.

Cloud Data Fusion:

- Cloud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines.
- The Cloud Data Fusion web UI allows you to build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.
- Features:
    - **Google Cloud-native:** features like security, reliability, scalability provided by Google.
    - **Delivers hybrid infrastructure:** Cloud data fusion is built using an open-source project called CDAP which ensures data pipeline portability
    - **Multi-cloud integration**
    - **Code free environment**
    - **Seamless operations**: provided by Restful APIs, pipeline state base triggers, logs, metrics and various monitoring dashboards.
- Modules:
    - Cloud Data Fusion instance
        - A Cloud Data Fusion instance is a unique deployment of Cloud Data Fusion.
    - Execution environment
        - Cloud Data Fusion creates ephemeral execution environments to run pipelines when you manually run your pipelines or when pipelines run through a time schedule or a pipeline state trigger.
    - Pipeline
        - A pipeline is a way to visually design data and control flows to extract, transform, blend, aggregate, and load data from various on-premises and cloud data sources.
    - Triggering
        - You can create a trigger on a data pipeline (called the upstream pipeline), to have it run at the completion of one or more different pipelines (called downstream pipelines).
    - Plugin

        - A plugin is a customizable module that can be used to extend the capabilities of Cloud Data Fusion.
        - The various categories of plugins available in Cloud Data Fusion are described below:
            - **Sources**: Sources are connectors to databases, files, or real-time streams from which you obtain your data.
            - **Transforms:** Transforms allow you to manipulate data after the data is ingested.
            - **Analytics:** Analytics plugins are used to perform aggregations such as grouping and joining data from different sources, as well as running analytics and machine learning operations.
            - **Actions:** Action plugins define custom actions that are scheduled to take place during a workflow but don't directly manipulate data in the workflow.
            - **Sinks:** Data must be written to a sink. Cloud Data Fusion contains various sinks, such as Cloud Storage, BigQuery, Spanner, relational databases, file systems, and mainframes.
            - **Error collectors:** When nodes encounter null values, logical errors, or other sources of errors, you can use an error collector plugin to catch errors.
            - **Alert publishers:** Alert Publisher plugins allow you to publish notifications when uncommon events occur.
            - **Conditionals:** Pipelines offer control flow plugins in the form of conditionals. Conditional plugins allow you to branch your pipeline into two separate paths, depending on whether the specified condition predicate evaluates to true or false.
- **Compute profile**
    - A compute profile specifies how and where a pipeline is executed.
    - A profile encapsulates any information required to set up and delete the pipeline's physical execution environment.



- Components
    - Wrangler
        - In the Wrangler UI, you can clean, transform and further prepare your dataset.
    - Integrate
        - Studio: Allows building pipelines
        - List: Displays the deployed pipelines and the pipelines which are saved as drafts
    - Metadata
        - In the Metadata section, you can search and check the data governance by following the lineage of the data.
    - Dashboard
        - Centrally monitor applications for better performance
    - Manage
        - Manage system and namespace level settings and preferences.
    - Hub
        - In the Hub section on the top right corner, you can share your pipelines and wrangling recipes across your entire organization.

    - Operations:
        - lists the various jobs performed

BigQuery:

- BigQuery is a fully-managed, serverless data warehouse that enables scalable analysis over petabytes of data.
- BigQuery is a fully managed enterprise data warehouse that helps you manage and analyze your data with built-in features like machine learning, geospatial analysis, and business intelligence.

Cloud Storage:

- Cloud Storage is a service for storing your objects(files) in Google Cloud. An object is an immutable piece of data consisting of a file of any format. You store objects in containers called buckets.

Google Data Studio:

- Google Data Studio is an online tool for converting data into customizable informative reports and dashboards.

Git:

- Git is a distributed version-control system for tracking changes in source code during software development. It is designed for coordinating work among programmers, but it can be used to track changes in any set of files.

Maven:

- Maven is a build automation tool used primarily for Java projects. Maven can also be used to build and manage projects written in C#, Ruby, Scala, and other languages.

MySQL:

- MySQL is an open-source relational database management system. Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language.

Postman:

- Postman simplifies each step of building an API and streamline collaboration so you can create better APIs—faster.

Project timeline/Project diary

January 2022:

- Week 1:
    - Completed MySQL self-learning training.
- Week 2:
    - Completed Python self-learning training.
- Week 3:
    - Completed Linux self-learning training.

February 2022:

- Week 1:
    - Learnt python flask basics to make REST API calls.
    - Learnt to consume REST APIs using the Postman tool.
- Week 2:
    - Revised Java Concepts.
    - Prepared a short documentation for a medical project.
- Week 3:
    - Research on an ETL platform named Alooma.
    - Started with a course on coursera to explore and prepare data with BigQuery.
- Week 4:
    - Started learning Google Cloud Platform.

March 2022:

- Week 1:
    - Started learning Google Data Fusion tool.
    - Created a simple data pipeline using Data Fusion.
- Week 2:
    - Created a data pipeline to integrate different data sources and store resulting data in different destinations.
- Week 3:
    - Fetched REST API data in XML format using Google Data Fusion plugin.
    - Parsed the XML data and stored it in a csv file.

    - Learnt xpaths to retrieve specific data from xml based data.
- Week 4:
    - Understanding pros and cons of BigQuery compared to other platforms
    - Learnt to use triggers on data pipelines.
- Week 5:
    - Learnt to use Loggers to log errors in the data pipeline.
    - Learnt to clean data using Data Fusion wrangler.

April 2022:

- Week 1:
    - Researched on different machine families and machine types and their costs.
    - Created VM instance with a custom machine type.
    - Reserved a static external IP Address and assigned it to the VM instance.
- Week 2:
    - Scheduled a cron-job on the VM instance to run a python script every 10 minutes.
    - Wrote a python script to make API call.
- Week 3:
    - Learnt powerBI basics.
    - Learnt Google Data Studio basics.
    - Compared powerBI and Data Studio in terms of features
- Week 4:
    - Explored Google Data Studio
    - Connected with BigQuery data using both BigQuery connector and using custom query.
    - Created dashboards with few charts.

May 2022:

- Week 1:
    - Used regular expression to extract part of a column in Data Studio.
    - Added more charts to the dashboard such as geo map and bubble map
    - Used drill-down feature on the map.
- Week 2:

    - Went through python code written by the Pune team to transform data.
    - Prepared pseudo-code for the Python code.
- Week 3:
    - Prepared flowcharts for a few functions in the code.
    - Started learning how to create a custom plugin and use it in Cloud Data Fusion.
    - Created a basic plugin in java.
- Week 4:
    - Created a custom plugin to cleanse and transform data.
    - Added more functionality to the plugin code to add one more column to the data based on values of existing columns(added column for data quality).

JUNE 2022:

- Week 1:
    - Added more functionality to the plugin code to create a hashmap of lists to store data rows based on trips covered by the vehicles, another list to store all non-trip rows and one more list to store all the data rows.
    - Added three more columns to the data.

Reflections/ Experiences of Internship

My experience of working as an intern at ZiMetrics has been wonderful. I started working on Google Cloud Platform which was new to me. It was great learning new technologies. I found myself growing after joining the company.

The work environment at ZiMetrics is very good. There is freedom given to interact, or seek help from seniors.

Apart from work, there are Recreational activities being conducted once every week where all members come together and play some indoor games. Once a month, a meeting is conducted wherein all members talk about their work.

You get appreciated for the work you do successfully which motivates you to do your work and also get help whenever needed.

Overall, I am happy to be a part of ZiMetrics. My best wishes to everyone here.

References

- https://www.javatpoint.com/mysql-tutorial
- https://www.javatpoint.com/linux-tutorial
- https://www.javatpoint.com/python-tutorial
- https://cloud.google.com/compute/docs/instances/creating-instance-with-custom-machine-type#pricing
- https://cloud.google.com/compute/docs/ip-addresses/reserve-static-external-ip-address
- https://cloud.google.com/data-fusion/docs/concepts
- https://datastudio.google.com/overview
- https://www.youtube.com/watch?v=kehG0CJw2wo
- https://medium.com/cdapio/getting-started-with-cdap-plugin-development-bcd21cc7ae66
- https://cdap.atlassian.net/wiki/spaces/DOCS/pages/480313897/Developing+Plugins+Guide
- https://www.restapiexample.com/python/consuming-a-restful-api-with-python-and-flask/
- https://cloud.google.com/data-fusion/docs/create-data-pipeline
- https://www.cloudskillsboost.google/focuses/12363?parent=catalog
- https://jtaras.medium.com/cloud-data-fusion-using-http-plugin-to-batch-api-calls-a5d0d1b32ecd