PhonePe Pulse Project

# Overview

This project entails the following:

1. Load PhonePe **Transaction**, **Insurance** and **User** data from a cloned github data repository into **Pandas** Dataframes
2. Apply required data cleanup and transformations
3. Store the dataframe contents into CSV files
4. Store the CSV contents in a MySQL DB for performing analytics
5. Develop a Streamlit web-app that displays the insights on the gathered data

# Python Notebooks

1. **pp\_domin**: Contains the the domain classes which could be used to load data from github and store data in MySql DB
2. **pp\_dbinit**: Script to create 'phonepe\_pulse' DB and its schema.
3. **pp\_dao**: Code to fetch dashboards and other data insights from the MySQL DB
4. **pp\_webapp**: Streamlit web-app to presents the insight - contains geo map and dashboards

# Configuration Files

1. **pp\_template.json**: A JSON file capturing template for providing DB connection parameters and other GCP config to access the DB
2. **india\_states.geojson**: Geo-JSON file for Indian states that shall be used by the Choropleth mapping library
3. **ppstyle.css**: CSS file for styling the web-app

# CSV Data Files

Data collected from github are stored in CSV files. These files are created by code in **pp\_domain.ipnyb** file. These CSV files are used to populate the tables in the MySQL DB.

## **Geo Location Data:**

1. **states.csv**: Contains records pertaining to all states in India, that are referenced by the PhonePe data
2. **districts.csv**: Contains records pertaining to each district in India, which is referenced in the PhonePe data
3. **Pincodes.csv**: Contains records pertaining to each district in India, which is referenced in the PhonePe data

## Transaction Data:

1. **transaction\_agg.csv:** Contains aggregated records for transactions
2. **transaction\_top.csv:** Contains top N records for transactions
3. **transaction\_hover.csv:** Contains hover records for transactions

## Insurance Data:

1. **insurance\_agg.csv:** Contains aggregated records for insurance
2. **insurance\_top.csv:** Contains top N records for insurance
3. **insurance\_hover.csv:** Contains hover records for insurance

## User Data:

1. **user\_agg.csv:** Contains aggregated records for registered users
2. **user\_top.csv:** Contains top N records for registered users
3. **user\_hover.csv:** Contains hover records for registered users

# Database Tables:

1. **geo**
2. **transaction\_agg**
3. **transaction\_top**
4. **transaction\_hover**
5. **insurance\_agg**
6. **insurance\_top**
7. **insurance\_hover**
8. **user\_agg**
9. **device\_agg**
10. **user\_top**
11. **user\_hover**

# Other Utility Python Libraries:

1. **dbconnect.py**: Code to connect to MySQL DB
2. **tools.py**: Code to load the DB connection config form a JSON file

# Domain Class Design:

1. **Transaction**: Abstraction of transaction record collections. Abstracts how transaction data is loaded from github, converted to CSV and stored in DB
2. **Insurance**: Abstraction of Insurance record collections. Abstracts how insurance data is loaded from github, converted to CSV and stored in DB
3. **User**: Abstraction of User/Device record collections. Abstracts how user and device data is loaded from github, converted to CSV and stored in DB
4. **Geo**: Abstraction of Geo location information. Abstracts how geo location data is extracted from github data, converted to CSV and stored in DB
5. **Base**: Base class for Transaction, Insurance and User. This contains common logic involved in decoding, loading and storing records