# README.md

## Setup and Execution Guide

This document provides a step-by-step guide to setting up the environment and executing the necessary files to process, transform and load data. Follow these instructions carefully to ensure successful execution.

### Prerequisites

* Python 3.x installed on your machine
* Google BigQuery access and credentials set up

### Step 1: Create a Virtual Environment

1. Open your terminal or command prompt.
2. Navigate to your project directory.
3. Create a virtual environment named advisense by running the following command:

* python -m venv advisense

1. Activate the virtual environment:
   * On Windows:
   * .\advisense\Scripts\activate
   * On macOS/Linux:
   * source advisense/bin/activate

### Step 2: Install Dependencies

1. Ensure you are in the root directory of your project.
2. Install the required dependencies by running:

* pip install -r requirements.txt

### Step 3: Extract, Validate, and Clean Data

1. Open and run the **Extract\_validate\_and\_clean\_data.ipynb** notebook.
2. This notebook will execute the data extraction, validation, and cleaning process.
3. Upon completion, a folder named Processed\_file will be created in the root directory with the following subfolders:
   * cleaned\_data: Contains the cleaned data files.
   * invalid\_data: Contains the invalid data files.
   * Original\_data: Contains the original data files.

### Step 4: Transform Data into Data Model and store it into Data warehouse

1. Open and run the **Transform\_into\_data\_model.ipynb** notebook.
2. This notebook will:
   * Take the files from cleaned\_data folder
   * Create the data model.(Dimensions and Facts)
   * Load the dimension and fact tables into Google BigQuery warehouse.
   * Additionally, It will also generate a folder named Data\_warehouse in the root directory containing all the files as CSV.

### Step 5: Incremental Data Loading

After the initial data load into the data model in the data warehouse, subsequent data loads can be managed using the following notebooks:

Both of these notebooks are essentially doing the same, only difference is that one is interacting with the local folder named Data\_warehouse (for testing) where as the other file is interacting directly with Google Big Query on the cloud

1. **Incremental\_transform\_to\_data\_model.ipynb**:
   * Interacts with the local Data\_warehouse folder created earlier.
   * Processes and transforms incremental data for loading into the data model.
2. **Google\_big\_query\_incremental\_transform\_to\_data.ipynb**:
   * Interacts directly with the data warehouse in Google Cloud.
   * Processes and transforms incremental data for loading into the data model