**Incremental Data Loading Pipeline Documentation**

**Goal**

Implement an incremental data loading strategy to process only new records from incoming CSV files, and maintain a cumulative dataset without duplications — based on a unique key (Code).

**Folder Structure**

Create the following directories on your Ubuntu machine:

***mkdir -p ~/Downloads/csv\_incremental/data/new***

***mkdir -p ~/Downloads/csv\_incremental/data/processed***

Where to place files:

Drop new CSV files (e.g. DimenLookupEthnic8277.csv) in:

***~/Downloads/csv\_incremental/data/new/***

The script will maintain a cumulative processed file at:

***~/Downloads/csv\_incremental/data/processed/processed\_dimen\_lookup.csv***

**Script: incremental\_loader.py**

Save the following script to:

***~/Downloads/csv\_incremental/incremental\_loader.py***

***import os***

***import pandas as pd***

***new\_data\_dir = os.path.expanduser("~/Downloads/csv\_incremental/data/new")***

***processed\_file\_path = os.path.expanduser("~/Downloads/csv\_incremental/data/processed/processed\_dimen\_lookup.csv")***

***if os.path.exists(processed\_file\_path):***

***processed\_df = pd.read\_csv(processed\_file\_path, dtype=str)***

***print(f"[INFO] Loaded existing processed file with {len(processed\_df)} rows.")***

***else:***

***processed\_df = pd.DataFrame(columns=["Code", "Description", "SortOrder"])***

***print(f"[INFO] No processed file found. Starting fresh.")***

***for filename in os.listdir(new\_data\_dir):***

***if filename.endswith(".csv"):***

***file\_path = os.path.join(new\_data\_dir, filename)***

***print(f"[INFO] Processing {filename}...")***

***new\_df = pd.read\_csv(file\_path, dtype=str)***

***new\_df = new\_df.dropna(subset=["Code"])***

***new\_rows = new\_df[~new\_df["Code"].isin(processed\_df["Code"])]***

***print(f"[INFO] Found {len(new\_rows)} new rows.")***

***if not new\_rows.empty:***

***processed\_df = pd.concat([processed\_df, new\_rows], ignore\_index=True)***

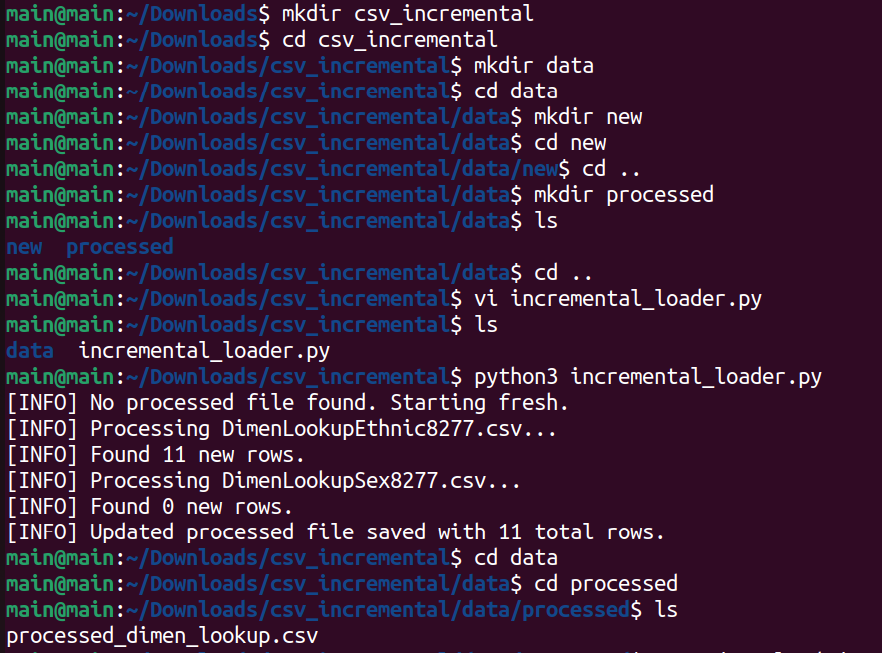
***processed\_df.to\_csv(processed\_file\_path, index=False)***

***print(f"[INFO] Updated processed file saved with {len(processed\_df)} total rows.")***

**How to Run the Pipeline**

***cd ~/Downloads/csv\_incremental***

***python3 incremental\_loader.py***



**Example of Incremental Load**

Initial state:

processed\_dimen\_lookup.csv has 11 rows from DimenLookupEthnic8277.csv.

New File:

Create a test file DimenLookupTestNew.csv in data/new/:

***Code,Description,SortOrder***

***9999,Total people,1***

***1,European,2***

***NEW-001,New Entry 1,1***

***NEW-002,New Entry 2,2***

Run the Script Again:

***python3 incremental\_loader.py***

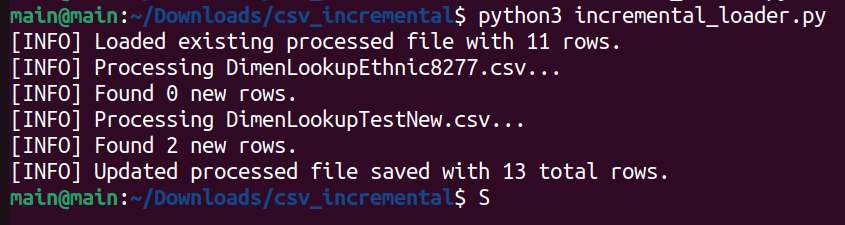
Expected Output:

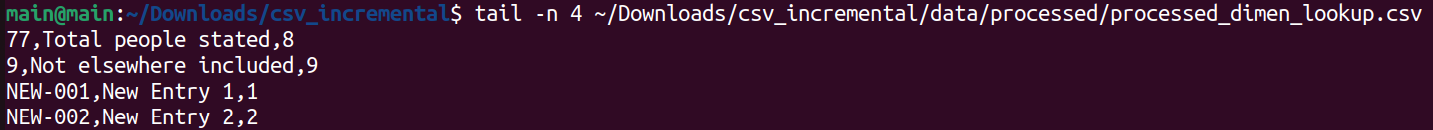
[INFO] Loaded existing processed file with 11 rows.

[INFO] Processing DimenLookupTestNew.csv...

[INFO] Found 2 new rows.

[INFO] Updated processed file saved with 13 total rows.





**Outcome**

This pipeline:

* Ingests new data from CSVs
* Filters out duplicates using Code as a unique identifier
* Appends only new rows to a persistent output file