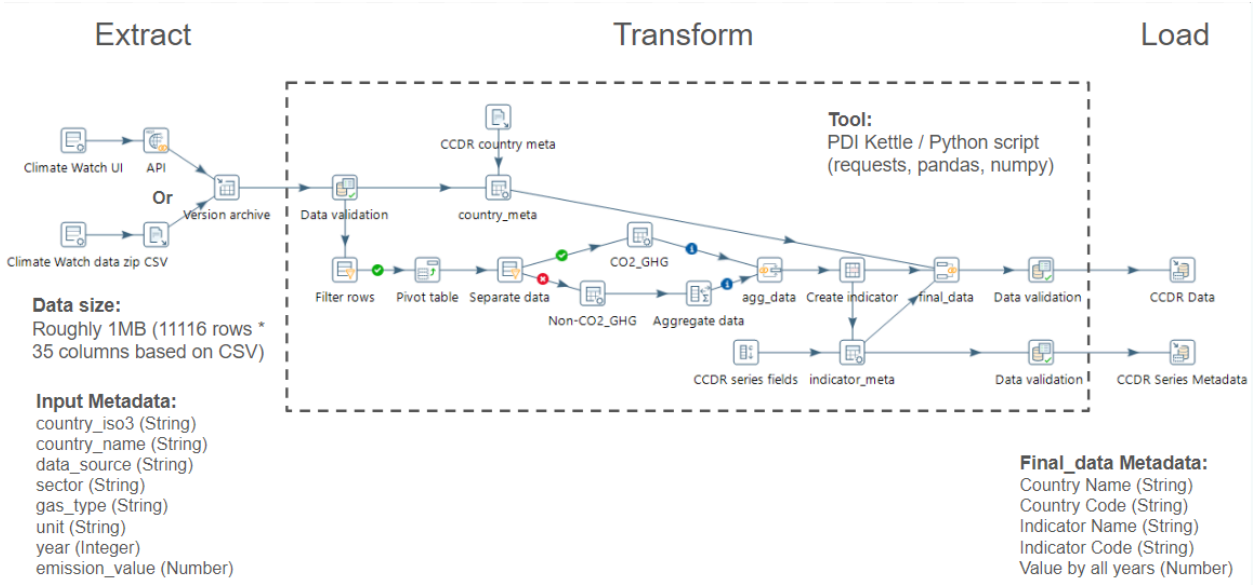

The diagram shows a potential ETL process to migrate Climate Watch (CW) Historical Country Greenhouse Gas Emissions Data into the World Bank Country Climate and Development Reports (CCDR) database. This process, designed in PDI Kettle, can be implemented through a Python script due to the relatively small data size (11,116 rows/1MB). This approach was chosen for its cost-efficiency, but other tools such as AWS, Azure, Hadoop, Apache Airflow, etc. could also be used if this process were to be integrated into a larger, pre-existing ecosystem.

Data extraction from CW can occur through an API or by downloading a zip file, with raw data saved with a version name for version control. The key steps in transformation are data validation, data structure redesign, and indicator metadata creation. Data validation includes checking values (ensuring emissions for All GHG = CO2 + CH4 + N2O + F-Gas, flagging unexpected negative emissions, etc.), date ranges, and emission unit consistency. A country metadata table, using country ISO3 as a key, is created to ensure data consistency between CW and CCDR. Data structure redesign entails pivoting the data table into a wide format by year and reclassifying gas types as: All GHG, CO2, and Non-CO2 GHG (CH4, N2O, and F-Gas). The emissions of Non-CO2 GHG are then aggregated by summing across sectors, countries, and year groups. Indicator metadata creation first requires concatenating gas type, unit, and sector values to generate an indicator field (e.g. Non-CO2 GHG emissions by sector (Mt CO2 eq) - Waste). An indicator metadata table is then created by deduplicating indicators and adding fields (e.g. Series code, Topic, Long definition, etc.) from CCDR series metadata.

The final data table is constructed through an inner join of indicator metadata and country metadata, followed by a left join with the aggregated emission data using the indicator as the key. Subsequent data validation ensures value accuracy (All GHG = CO2 + Non-CO2 GHG etc.), unit consistency, and adherence to CCDR format standards. For the data loading phase, the indicator metadata is loaded to the CCDR Series metadata, and the final dataset is loaded to the CCDR Data. The final dataset can then be loaded to the CCDR database as CSV, SQL, JSON, XML, or other formats.