ETL Pipeline for Market Share Analysis

Technologies Used

1. **Database:** PostgreSQL for structured data storage and efficient querying.
2. **ETL:** Python with pandas for data manipulation and SQLAlchemy for database connectivity.
3. **Visualization:** Power BI for interactive dashboards and insights.

Design

1. ETL Pipeline Architecture:- The ETL pipeline will be built using the following technologies:

1. **Python**: For scripting and handling data transformation.
2. **Pandas**: To manipulate and clean the datasets efficiently.
3. **PostgreSQL**: As the target database to store the cleaned and structured data.
4. **SQLAlchemy**: For seamless database interactions.
5. **Power BI**: For visualizing the final dataset.

Data Flow:

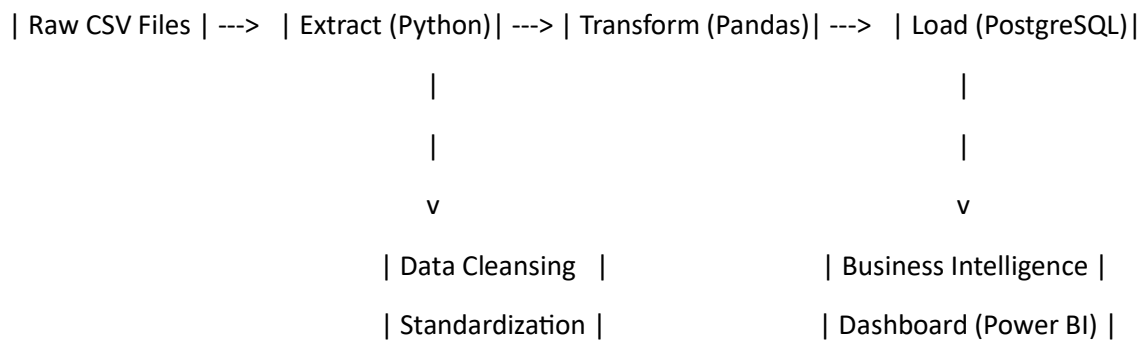
Extract raw data from multiple CSV files into Python.

Perform necessary cleaning, transformations, and aggregations.

Load the transformed data into PostgreSQL tables.

Leverage Power BI for data visualization and reporting.

Data Flow Diagram



3. Data Transformations Required:

Standardization of column names, formats, and data types.

Handling missing values by using appropriate imputation strategies.

Removing duplicate entries to ensure data integrity.

Ensuring revenue data consistency across different years.

Aggregating revenue by Vendor, Region, Country, and Industry Vertical.

Resolving discrepancies in market, country, and vendor definitions over time.

4. Approach to Handling Data Transformation:

Used pandas for data manipulation and cleaning.

Standardized revenue data by adjusting formats and ensuring accuracy.

Automated transformation processes to handle multiple datasets consistently.

Verified data correctness through exploratory data analysis (EDA) and validation checks.

5. Presentation Approach:

Structured data to support Power BI dashboards effectively.

Incorporated interactive filters, drill-down options, and intuitive visual elements.

Ensured key insights are easily accessible for business stakeholders.

Implementation

1. Extract Data

1. Used pandas to read multiple CSV files containing market share data.
2. Established connection with PostgreSQL using SQLAlchemy.
3. Verified database connectivity before proceeding with transformations.

2. Transform Data

1. Renamed columns for consistency and ease of analysis.
2. Converted YR column into Year for uniform representation.
3. Addressed missing values in Super Region, Vendor Name, and HQ Country using appropriate techniques.
4. Removed duplicate records to avoid redundancy.
5. Merged datasets where necessary to create a comprehensive dataset.
6. Aggregated revenue data by Vendor, Region, Country, and Industry Vertical.
7. Created summary tables to facilitate efficient querying and reporting.
8. Validated transformations using descriptive statistics and data visualizations.

3. Load Data

1. Created PostgreSQL tables to store transformed data.
2. Used SQLAlchemy to insert cleaned data efficiently.
3. Verified correctness of loaded data using sample queries and cross-checking with raw files.

Performance Optimization

1. Used pandas' vectorized operations for efficient data transformations.
2. Optimized PostgreSQL queries by implementing indexing.
3. Pre-aggregated data in summary tables to improve dashboard performance.
4. Minimized memory usage by converting data types where applicable.
5. Used batch loading techniques instead of row-by-row insertion to enhance performance.
6. Normalized database schema to reduce redundancy and improve query efficiency.

Error Handling and Logging

1. Implemented robust try-except blocks to catch database connection and query errors.
2. Logged missing values, duplicates, and transformation anomalies for auditing purposes.
3. Validated data integrity at each stage of the ETL pipeline.
4. Ensured logs captured successful and failed transactions for debugging and monitoring.
5. Implemented fail-safe mechanisms to prevent incorrect data from being loaded into PostgreSQL.

Documentation and Communication

1. Provided detailed comments in Python scripts to explain each step of the ETL pipeline.
2. Created structured documentation explaining ETL architecture, data flow, and implementation choices.
3. Ensured the final dataset and dashboards are intuitive for non-technical stakeholders.
4. Developed an executive summary highlighting key findings and insights.
5. Presented results using charts, tables, and summaries to facilitate data-driven decision-making.

Visualizations in Power BI

1. Market Share Trend (2019-2023)

1. Line chart displaying revenue trends over the years.
2. Interactive filters for selecting specific years and comparing trends.

2. Top Vendors by Revenue

1. Bar chart showcasing leading vendors by revenue.
2. Filters for selecting specific time periods and vendors.
3. Limited display to top vendors for better clarity.

3. Revenue by Region and Country

1. Treemap visualization allowing drill-down from Region to Country.

2. Color-coded revenue distribution to highlight key regions.

4. Revenue by Industry Vertical

1. Bar chart comparing revenue distribution across various industry verticals.
2. Sorting and filtering enabled to focus on specific industries.