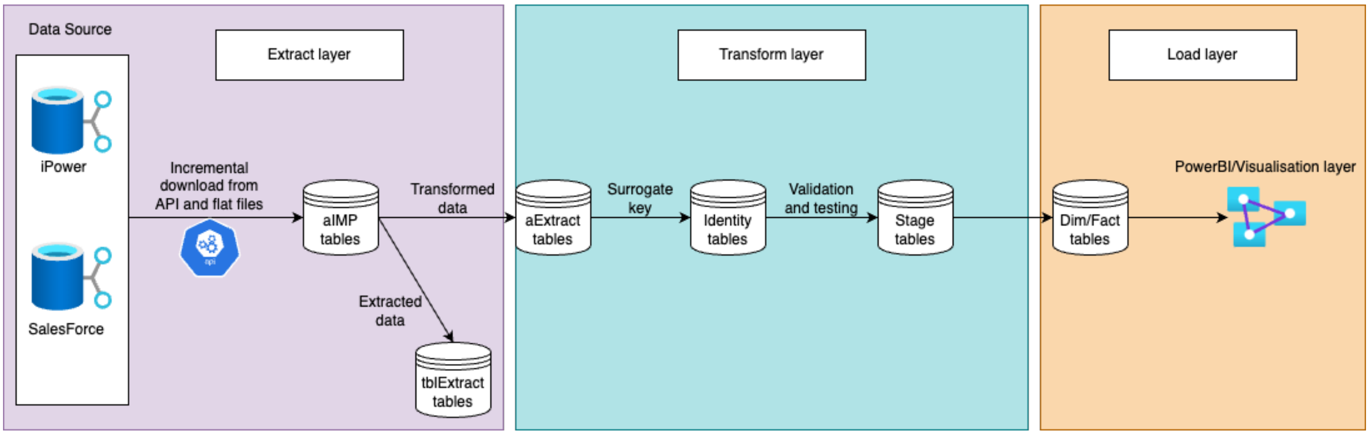
# Data Processing and ETL Workflow Overview

## My Contribution

My role in this project is to maintain incremental loading, monitor the process, add new metrics and calculations and improve it, reducing data loading time. The most important point is to keep the process running so that users have access to the data. **What I can create for your team is a similar automated data warehouse so that there is no need to duplicate the analysis, but only the already created analyzes are filled only with new data.** Thanks to this extensive project, I can create advanced queries in SQL, but also create automatic processes, optimize queries, create visualizations in PowerBI and think like an architect of database solutions.

## Data warehouse model



## System Structure

The system structure revolves around a series of stored procedures in SQL, which work to incrementally load and transform data from source systems into a final, usable format. Here's a step-by-step overview:

1. **Data Import (aIMP):** A stored procedure retrieves data from source tables, concatenates SQL queries into a single statement, and inserts the data into the aIMP table based on the most recent 'lastModifiedDate'.
2. **Data Extraction (tblExtract):** Another stored procedure moves data from aIMP to tblExtract, which serves as a repository for historical data without any transformations.
3. **Data Aggregation (aExtract):** Data from tblExtract is moved to aExtract, where joins and initial transformations occur.
4. **Data Transformation and Identity Assignment:** Data is then processed in Identity tables, where database keys are assigned.
5. **Staging Area (Stage):** Transformed data is moved to Stage for further processing.
6. **Dimension and Fact Tables (Dim and Fact):** The final steps in the pipeline involve transferring data to dimension (Dim) and fact (Fact) tables.

7. **View Creation:** In T-SQL, views are created from the Dim and Fact tables. These views are then consumed by an OLAP cube.
8. **Data Access:** Data is exposed to clients via the OLAP cube, allowing access through various tools such as Excel, Google Sheets, or Power BI reports.

## Data Sources

The system relies on two primary data sources:

- **iPower:** An internal HR tool used for tracking project hours, employee utilization, business travel, and expenses. Data is accessed through a linked server, with incremental data loading from tables and views.
- **Salesforce:** A popular platform for sales and accounting data. Data is extracted via a custom C# API connector.

## ETL Process Workflow

The ETL process follows this sequence:

- **aIMP → tblExtract → aExtract → Identity → Stage → Dim → Fact → Views**

## Tasks Performed in Each Step

- **Data Import and Extraction:** Retrieve data incrementally and store it in tblExtract without transformation.
- **Data Transformation and Aggregation:** Join and transform data in aExtract. Assign database keys in ‘Identity’.
- **Data Staging:** Perform additional transformations in Stage.
- **Final Storage:** Transfer processed data to Dim and Fact.
- **View Creation:** Create views for the OLAP cube.

## Technologies and Tools Used

The system uses a variety of technologies and tools, including:

- T-SQL for stored procedures and query writing
- Tabular Editor and Power BI for data visualization and reporting
- C# for the Salesforce API connector.

## Problems Solved by the Process

This ETL process provides a streamlined and efficient method for managing data pipelines. It addresses:

- System maintenance and error handling
- Timeout resolution
- Data integrity and accuracy through auditing
- Optimization and improvement of existing processes
- Handling user requests, troubleshooting data issues, and answering questions about data calculations

## Main Results and Outcomes

The main outcomes of this process are:

- Robust and reliable data available for reporting and analysis
- A set of comprehensive reports produced by your team
- User-friendly OLAP cubes accessible via Excel, allowing users to create their own analyses

Overall, this ETL system provides a reliable framework for data processing and analysis, supporting various business needs and ensuring data accuracy and integrity.