DATA ENGINEERING PROCESS METHODOLOGY

RAW: Ingestion of raw source data to the staging area. Tasks:

- Check if received data is what you need
- Perform a data profiling of all tables

PREPROC: Preprocessing of raw tables individually (no joins). Tasks:

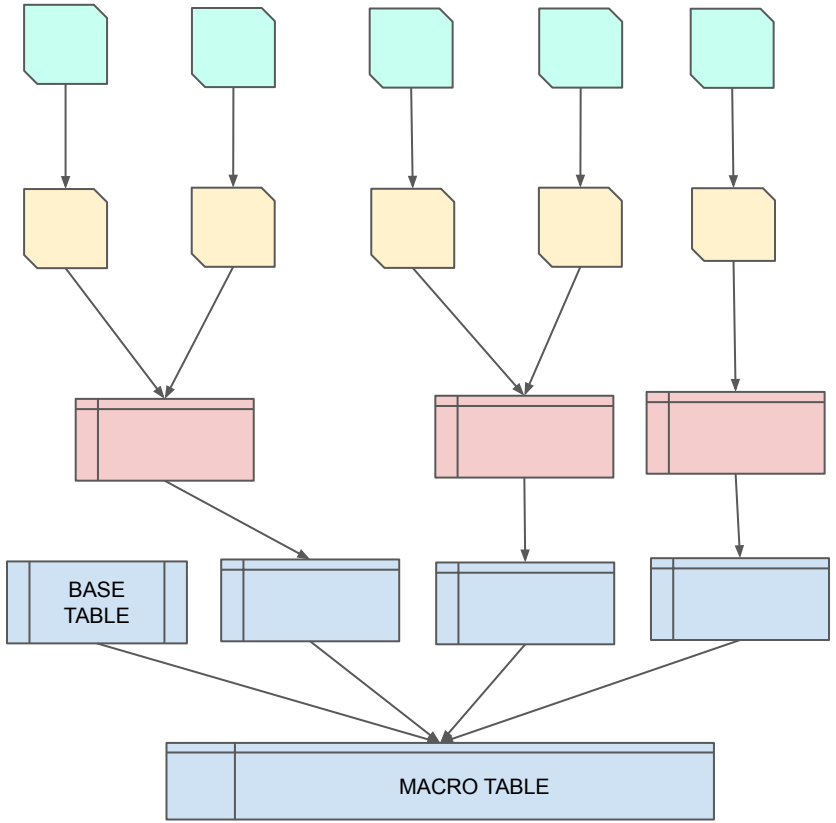
- Fix wrong values *(ex. Spian -> Spain)*
- Pivot tables to the desired format *(ex. when time series are in columns)*
- Filter rows that makes no sense. *(ex. a row without primary key)*
- Remove unuseful columns
- Change bad column names *(ex. col34a_7 -> SalePrice)*

ENRICH: Preprocessing tables jointly without aggregations.

- Join tables directly by the index *(ex. include geolocation data from other separate table)*.
- Calculate simple derived variables from current columns *(ex. if we want to calculate a variable before aggregating)*

AGGREGATE: Phase in which the process of obtaining the Macro Table is carried out following the following steps:

1. Construction of a **Base Table** that contains only the aggregation keys of all the tables, that is, all the possible indexes that exist in the data set.
2. Individually aggregate each table based on the index defined in the Base Table
3. Generate **Macro Table** by joining (left join) the Base Table with all other sources by the corresponding index
4. Solve missings and generate more derived variables



The “E” in the ETL

- The “E” part in this methodology is oversimplified. Bear in mind that when working with real data in real projects for real companies, **data is not stored in someone’s computer in CSV format**

- Many companies are now fully integrated with the “Digital Transformation” mindset, and they do care about their data with proper **Data Governance strategies**, but not all companies are at the same level. Data Governance is the approach to manage data, from the acquisition to its usage in all different projects in a company. DG has to do with **data stewardship, data quality and storage**. Transforming the data acquisition processes in big companies can be very expensive, so this goes slow.

- So, the **access to source files can be very heterogeneous**, from Excel files to distributed Hive metastores in AWS. But, n**o matter the way to access, that this methodology can be applied independently**. Just be sure you have access to your own, separated environment.

SOME COMMENTS

This is **not a flawless methodology**, and probably is not suitable for all cases and I'm sure it'll struggle sometimes. Anyway, **having a methodology, this one or other one, is crucial**. Some of the benefits of following a methodology are:

- To save time. Data Engineering is about the 70%-80% of the time we spend with a data science work, so saving time on this task is important
- It clarifies the intra-team communication, establishing what are the inputs and the outputs of the data manipulation and data analytics parts.
- It **provides some guidance** and avoid the "blank page syndrome"
- It reduces the number of mistakes, and avoid redoing the same operations again an again.
- It paves the way to the future orchestration of automatized ETL workloads.

Although we're working with Python, this methodology can be used with any other data engineering tool like NiFi, AWS Data Pileline, Azure Data Factory, and others.