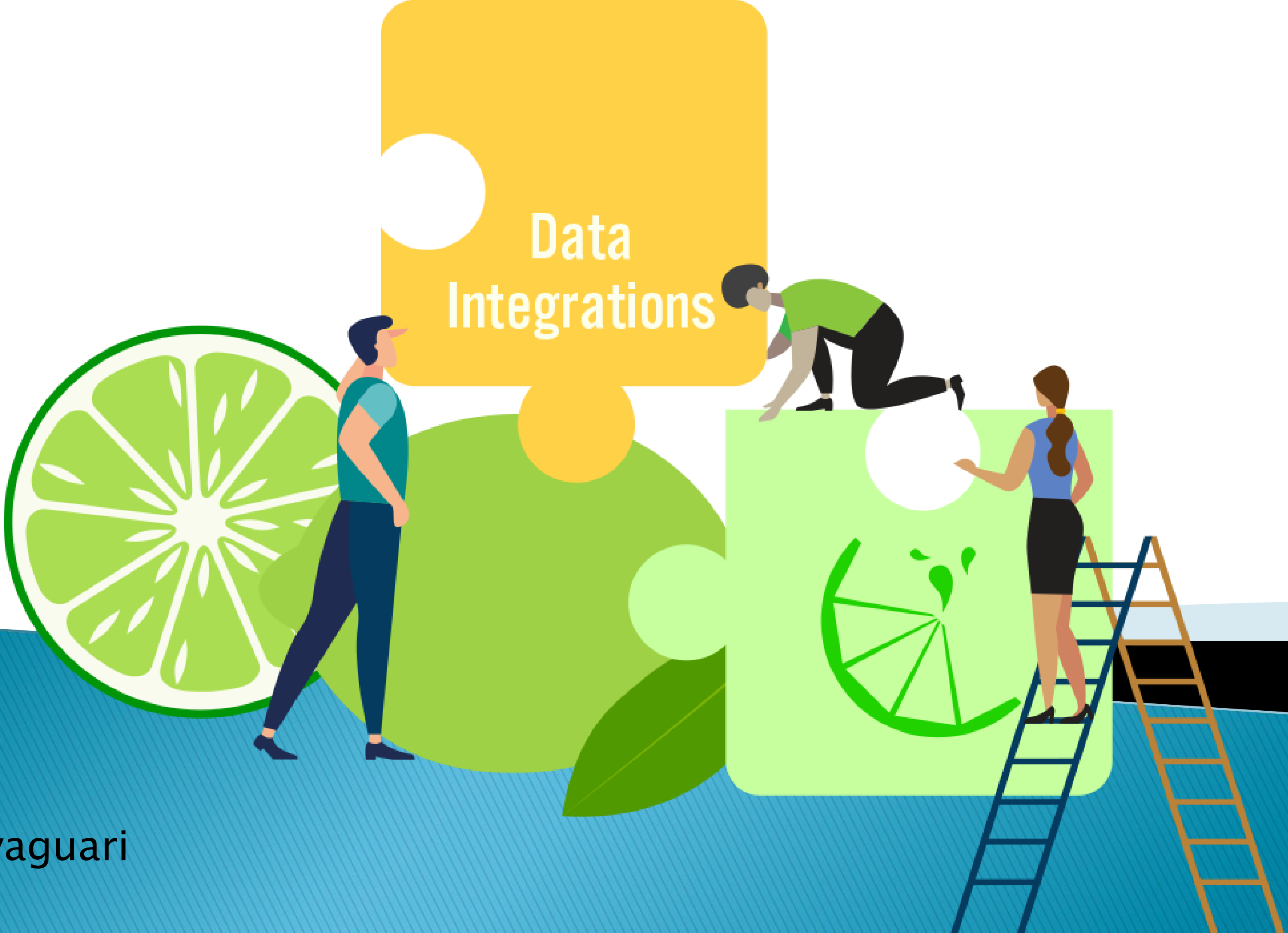
Data Integration Phase

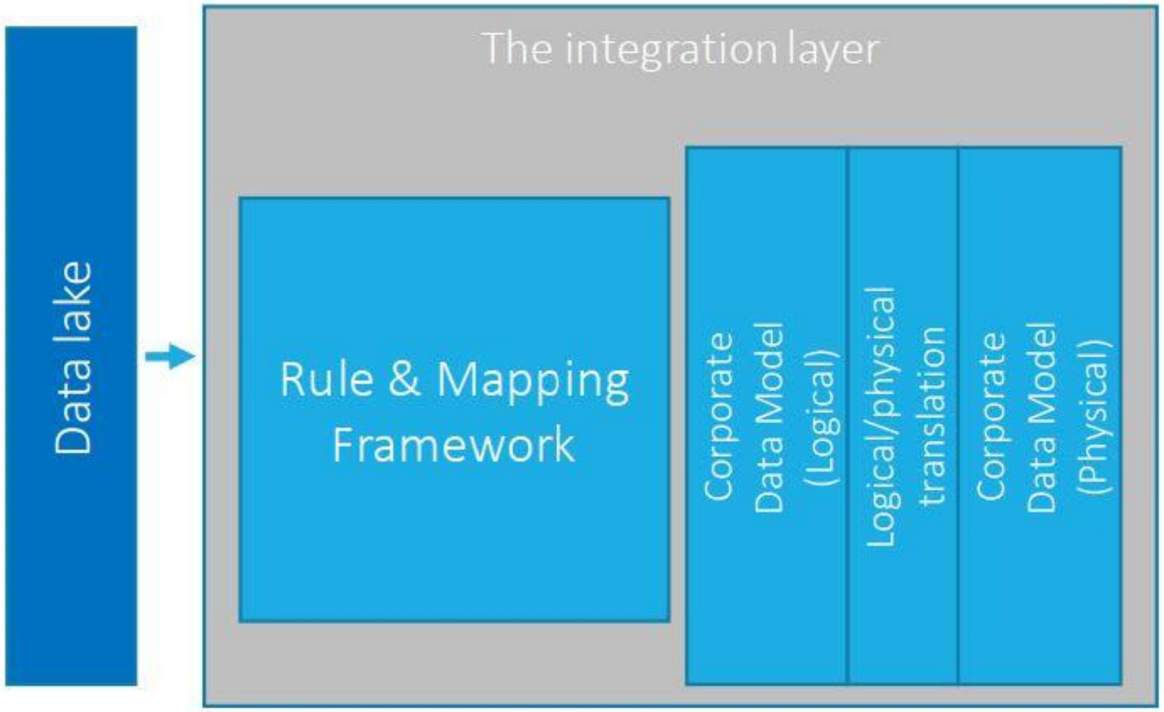


G. Fernando Lojano Mayaguari
Miguel Moraga Muñoz

What is Data Integration?

Data integration involves combining data residing in different sources and providing users with a unified view of them.

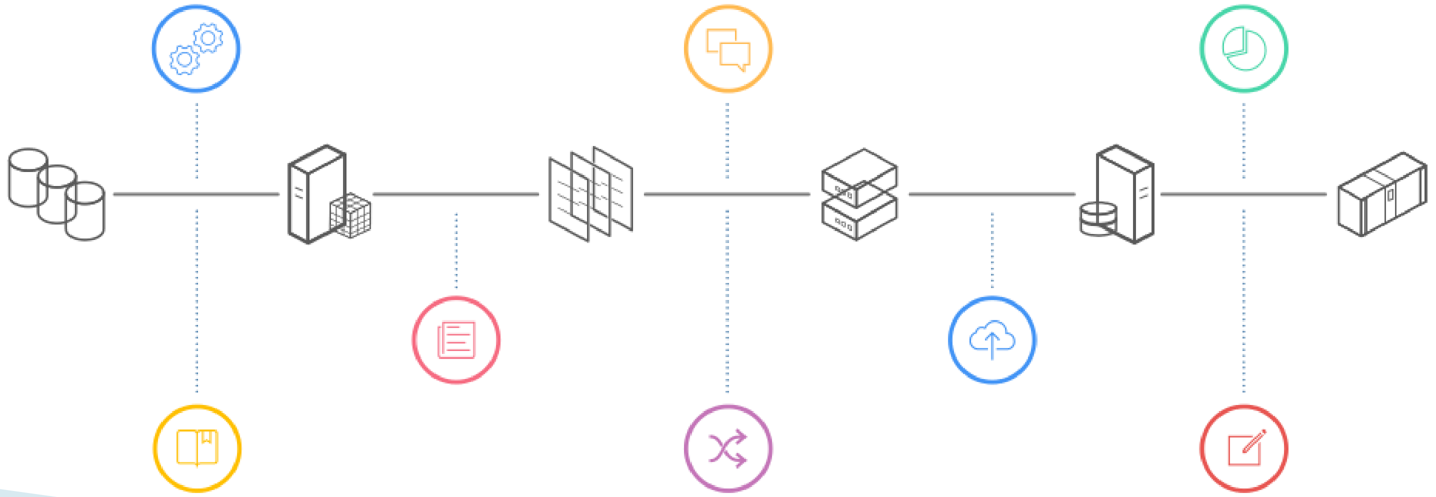
In our case, the different sources comprise the data lake, and the end result is a physical corporate data model.


Data lake
The integration layer
Rule & Mapping Framework
Corporate Data Model (Logical)
Logical/physical translation
Corporate Data Model (Physical)

Source-to-target mapping

Set of data transformation instructions that determine how to convert the structure and content of data in the source system to the structure and content needed in the target system.

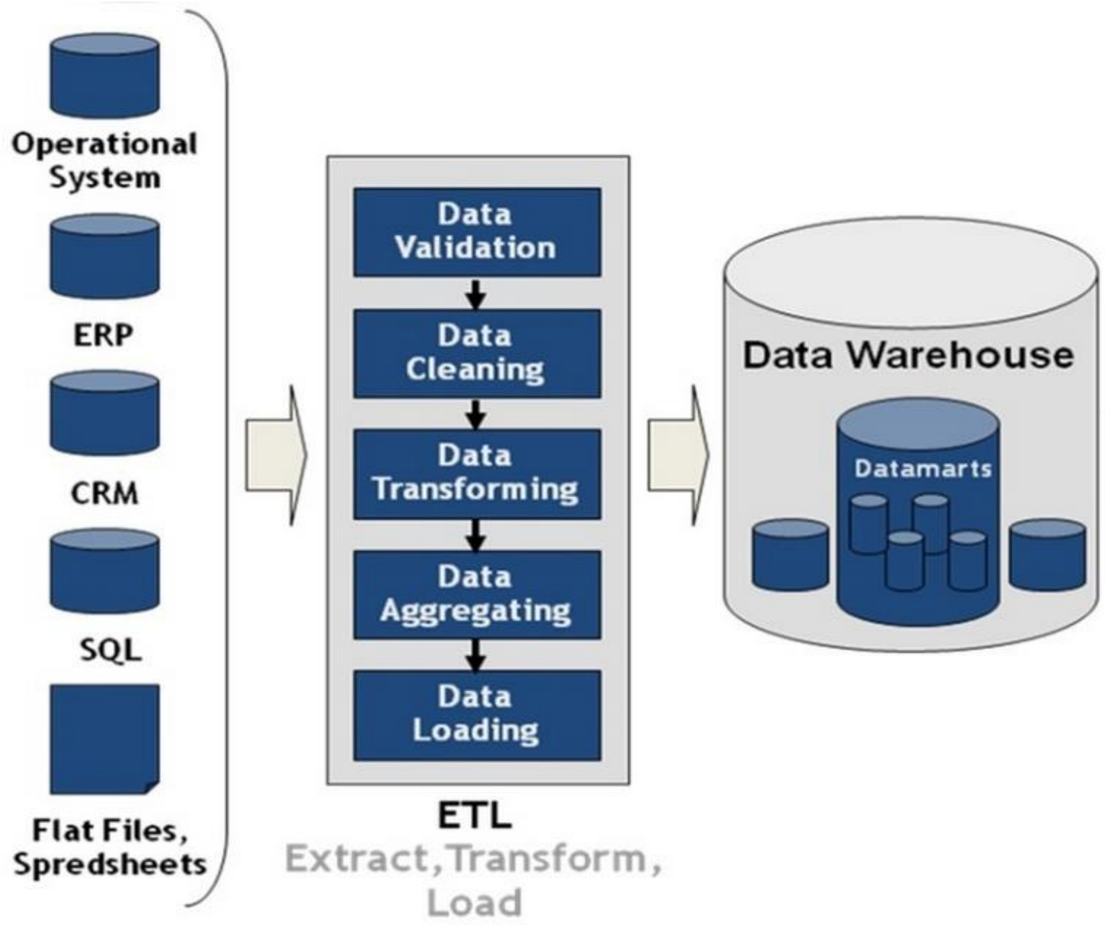
When you create a mapping, you use operators to define the Extraction, Transformation, and Loading (ETL) operations that move data from a source object to a data warehouse target object.



ETL: What is it?

- Data Integration Model System
- Often used to build data warehouse
- 3 main phases:
  - Extract
  - Transform
  - Load

- Benefits
- Challenges & Difficulties



Logical Data Model & Physical Data Model

Logical data model

- Describes data needs for a single project but could integrate with other logical data models based on the scope of the project.
- Designed and developed independently from the DBMS.
- Data attributes will have datatypes with exact precisions and length.
- Normalization processes to the model is applied typically till 3NF.

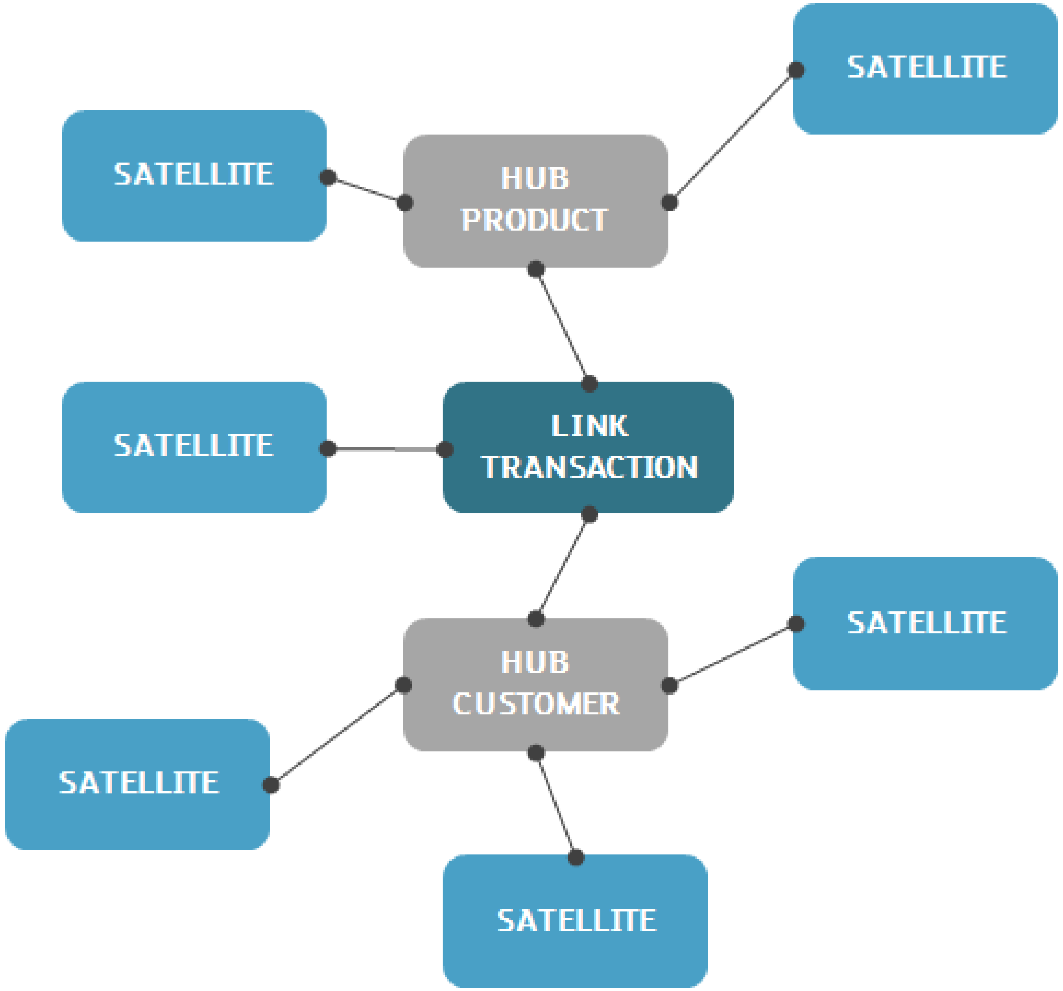
Physical data model

- The physical data model describes data need for a single project or application though it maybe integrated with other physical data models based on project scope.
- Data Model contains relationships between tables that which addresses cardinality and nullability of the relationships.
- Developed for a specific version of a DBMS, location, data storage or technology to be used in the project.
- Columns should have exact datatypes, lengths assigned and default values.
- Primary and Foreign keys, views, indexes, access profiles, and authorizations, etc. are defined.

About Data Vault

- Hybrid Data Modelling Technology
- 4 main principles:
  - Traceability
  - Non relevant data selection
  - Tolerance to change
  - Load speed

- 3 different table components
- Benefits
- Challenges & Disadvantages



Open Source Data Integration Options

Open Source Tool: Apatar

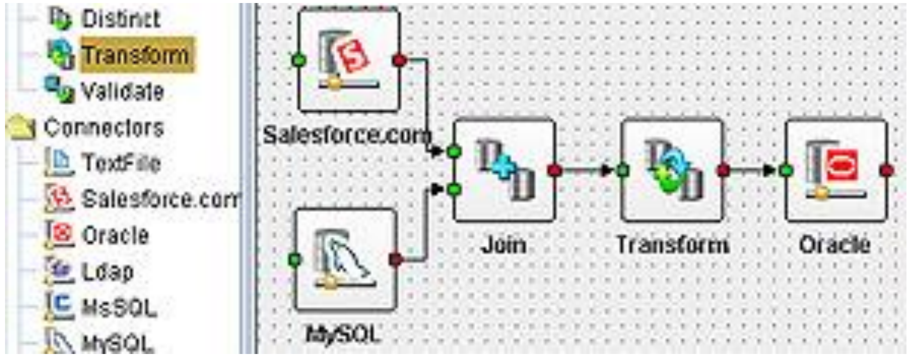
- No coding, just visual job designer
- Flexible deployment Options
- As there is no coding, non developers can easily use it.

BUT…..

Only ETL approach, it not supports the new approaches as ELT or ETLT

Quite outdated, last update made in 2013.

apatar
connecting data



Open Source Data Integration Options

Open Source Framework: Spark

- Provides a number of inter-connected platforms, systems and standards for Big Data projects
- A fast open source data processing engine
- Makes distributed programming very accessible to data scientists

We will use Spark to easily implement the ETL mapping from the data lake to the logical model (3NF)

Then, we will map the logical model into the physical data model through the data vault modelling method