1. Data Integration

- What is data integration?

Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.

Data integration involves combining data residing in different sources and providing users with a unified view of them. Data integration appears with increasing frequency as the volume (that is, big data) and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved. Data integration encourages collaboration between internal as well as external users.

2. Source-to-target mapping

- What is source-to-target mapping?

Source-to-target mapping is a 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. Source-to-target mapping solutions enable their users to identify columns or keys in the source system and point them to columns or keys in the target systems. Additionally, users can map data values in the source system to the range of values in the target system.

A mapping describes a series of operations that pulls data from sources, transforms it, and loads it into targets. 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. Mappings provide a visual representation of the flow of the data from sources to targets and the operations performed on the data.

3. ETL

- What is ETL?

ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources. It's often used to build a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse or other system. Extract, load, transform (ELT) is an alternate but related approach designed to push processing down to the database for improved performance.

ETL can be implemented with scripts (custom DIY code) or with a dedicated ETL tool.

The 3 main phases of a ETL process are:

- **Extraction** of the data from one or multiple source.
- **Transformation** of the extracted data with the possibility of reformatting and cleaning this data whenever is needed to.
- **Load** the data into a data base, a data mart or a data Warehouse, so they can be analysed or used.

- Main benefits of ETL process:
  -The possibility of creating a Master Data Management, which is a central repository that contains all the data of a certain organization or a business.
  Example: We have a client object in a credit data base and another client object in a credits card database. The master repository will define a single client register with all the necessary information for the whole organization.

  -It's useful for integrating systems. EXAMPLE: As all the organizations are growing, they need to aggregate more source data. As a consequence, there are more necessities surging, like integrating a banking online data with old data of a legacy system.

  - Disadvantages:

  -They aren't really good at near-real time or on-demand data access. They are more for a batch model of working. It should be used for well established, slow changing data tranformation for data warehouse.

  ETL tools aren't the kind of tools which are really aimed at data analysts or business users. These users typically understand data and some SQL query language, but not necessarily data infrastructure technology like ETL and data warehouses.

4. Logical Data Model & Physical Data Model:

- Logical data model:

A logical data model is a data of a particular database management product or storage technology expressed in terms of data structures such as relational tables and columns, object-oriented classes, or XML tags.

Its main characteristics are:

- 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:

A physical data model (or database design) is a representation of a data design as implemented in a database management system. In the lifecycle of a project it typically derives from a logical data model, though it may be reverse-engineered from a given database implementation. Its main characteristics are:

- 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 on this model.

5. Data Vault

- Data Vault definition:

The Data Vault is a hybrid data modeling methodology providing historical data representation from multiple sources designed to be resilient to environmental changes.

Data Vault method is used for modelling data warehouses. His main design principle is to separate the business keys, the context and the relations in different tables known as Hub, Satellite and links.

- HUB: containing a list of unique business keys having its own surrogate key. Metadata describing the origin of the business key.
- LNK: establishing relationships between business keys. Essentially describing a many-to-many relationship. Links are often used to deal with changes in data reducing the impact of adding a new business key to a linked Hub.
- SAT: This table holds descriptive attributes that can change over time. As Hubs and Links form the structure of the data model, Satellites contain temporal and descriptive attributes including metadata linking them to their parent Hub or Link tables.

- Data Vault advantages:

  There are several advantages to the Data Vault approach:

  - Simplifies the data ingestion process.
  - Terabytes to Petabytes of information (Big Data).
  - Dynamic Model Adaptation – self healing.
  - Removes the cleansing requirement of a Star Schema.
  - Puts the focus on the real problem instead of programming around it.
  - Easily allows for the addition of new data sources without disruption to existing schema.
  - Created models are highly scalable.

- Challenges of the data vault method

As a result of separating all the information in different tables, there is a high number of data objects (tables, columns). As a main consequence, it produces a longer modelation performance time.

6. Our solution

The huge data volume, the diverse sources and the complexity of them makes the transformation phase slows down drastically turning it in a truly bottleneck. With the increase in the data needs and the 'Big Data' system apparition, the future of the traditional ETL method is getting unclear. ETL tools works really well on batch, but their performance decrease on real time.

There are a lot of opinions that confirms that Big Data era will bring the ETL tool's end. Although, for the moment, due to the fact of the beginning of this era, hybrid solutions are being invented and developed. One of them, the most extended one is Hadoop.

This is one of the most common tool used for data processing since is capable of manage and analyse big volumes of information that may be useful on the future. It's essential for the data analysis on real time and as it was the first platform that came out as a possible solution, it's used like a model for all the other possible solutions.

However, for the case that we have in here, a better approach should be working with Apache Spark, as it is an open source data processing engine that actually works pretty fast. Another big advantage about using it is that is considered the first open source software easy-to-use and accessible not just for the IT team but also for data scientists.

IT team may now a lot about technical information but when it's about the data that may be useful for analytics, researches… data scientists know better which data is actually useful or which data can be discarded. Because of that, they should manage the cleansing phase of Extraction phase of the ETL approach.

Apache Spark technology offers an efficient and already implemented way for doing the Extraction and Load phases, and also includes a lot of functionalities for doing the Transform step. One of the typical approach is the MapReduce method, as a result of this approach, we'll obtain a 3NF data model Format.

After ending the ETL process, we'll have to map our model in a physical data model. We'll use the Data Vault Modelling method. This is designed for offer a long-term historical storage of the data coming from different sources. As we explained before the main challenge of this method is that we have to do a lot of mechanical but easy tasks for modelling the data, but this approach is still affordable because as a result of this, we obtain easy access to the data for analytics, researches…