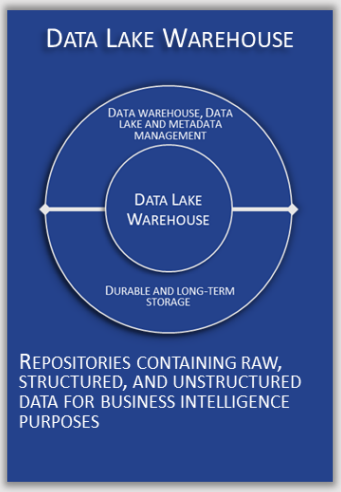
DATA LAKE WAREHOUSE

Repositories contain raw, structured, and unstructured data for business intelligence and data analytics purposes. In TRAIDA, the term 'Data lake warehouse' encompasses data warehouse, data lake, and data lakehouse. The term 'Business intelligence' includes data reporting and OLAP. The term 'data analytics' refers to data science.



1. CONDITIONS OF SUCCESS

When "big data" solutions do not fully meet expectations, most decision-makers believe that AI and knowledge graphs are the solution to better address data analysis needs. However, successfully integrating transformative AI at the decision-making system level requires clarifying the architecture. With TRAIDA, the effort made at the semantic platform level and with shared ontologies facilitates this integration. We will explain how in this TRAIDA card, but first, we need to clarify the meaning of the term "big data" by reducing it to the identification of multimedia databases. Since this term does not impose specific technologies or use cases, it becomes a commodity that is not structurally important for architectural choices.

We need to move beyond the term big data and return to the company's objectives in these two classic realms of decision-making IT, which we group under the generic term "Data Lake Warehouse":

- **Business Intelligence**: Focuses on reporting needs and structured data analysis. These data are described using metadata that provide their structures, definitions, and quality control rules. The technologies used are SQL-type databases and OLAP (Online Analytical Processing), including meta-schema and NoCode approaches. They are grouped under the generic term data warehouse.
- **Data Analytics**: Refers to the domain of data science, which works on more or less extensive multimedia data sets, with or without metadata. The goal is trend calculation, data discovery, detection of atypical cases, general classification, etc. The technologies used are NoSQL and schema-free. They are grouped under the generic term data lake.

AI's power is expressed in each of these two realms separately. However, it brings more potential when applied to a data repository that unifies the data warehouse and the data lake. This is the promise of new data lakehouse solutions. At the time of writing this TRAIDA card, the feedback from such solutions is still recent, making it difficult to assess their maturity. Nevertheless, it is certain that the convergence of data warehouse and data lake will be realized through such mechanisms:

- The ability to extend OLAP technologies to include multimedia data.
- Adding metadata management in the data lake to enhance query power and quality controls. These metadata must be shared with the OLAP part of the unified solution.
- Standardizing mass data storage solutions for both structured (enriched with their OLAP dimensions) and unstructured (multimedia) data inherent to the data lake.
- Unifying data manipulation languages between the data warehouse and the data lake necessary for injections, cleaning, aggregations, etc.

- Sharing a universal data access layer (OLAP, SQL, data lake) usable by data visualization tools.
- The ability to export vectorized databases from all data sources, both OLAP and data lake. This vectorization is necessary for enriching AI conversations with RAG (Retrieval Augmented Generation) technique and training AI assistants.
- Advances in transaction management (ACID) for both the data warehouse and the data lake will allow the implementation of data update processes directly at the decision-making IT architecture level. It will be possible to build integrated business intelligence and data analytics solutions in operational application systems. For example, a data lakehouse could modify a customer's data as part of an end-to-end process with a CRM application. Transactional management will then encompass both decision-making and operational systems.

To prepare for this kind of evolution, a good practice is to build the necessary ontologies for your MDM, ODS, and EKG repositories (see the respective TRAIDA cards). Indeed, it is from these shared ontologies that metadata are developed and implemented. They bring to life the semantic platform highlighted by the TRAIDA framework. This metadata is necessary to enhance the power of your data lake and to configure the OLAP dimensions in your data warehouse. They will be used for the unification of the two solutions through the evolving data lakehouse technology on the market. All of these MDM, ODS, and particularly EKG repositories are also the necessary data sources to train your AIs and enrich them on demand (RAG).

In parallel with this technological landscape, it is also possible to use the EKG repository as a data analysis solution (see TRAIDA EKG card). Indeed, the technology of knowledge graph-oriented databases offers specific benefits that OLAP and data lake solutions do not:

- The OLAP approach stores data to aggregate them according to multiple axes of analysis.
- The data lake stores data in a raw manner without imposing strong structuring.
- The knowledge graph stores data in the form of a meta-structure in triplets (starting object, relationship, ending object).

Therefore, the EKG is a complementary opportunity for value creation in the field of decision-making IT. It is unlikely that graph technology will replace OLAP and data lake solutions in a reasonable timeframe. However, it can handle certain data analysis cases that avoid the use of OLAP or data lake, and offer others that are impossible to implement without graph management, such as inference algorithms.

Finally, a powerful way to combine the EKG with OLAP and the data lake is to consider knowledge graphs as a layer above the data lakehouse to manage metadata and enrich data analysis systems. In choosing a technical solution for the data lakehouse, it is important to understand the mode of mass data storage (OLAP, SQL, multimedia) but also to evaluate the availability of a knowledge graph-oriented database used as a supervisory layer. If the barycenter of the IT system is the data lakehouse, EKG technology can then come with the data analysis solution. However, as mentioned earlier, it is still too early to validate the technical maturity of such an apparatus. A more reasonable approach is to choose EKG technology that does not depend on the future choice of a data lakehouse. If the latter has a semantic layer based on a knowledge graph, it will need to be integrated with your EKG.

2. IMPORTANCE OF THIS CARD FOR YOUR TRANSFORMATIVE AI

The TRAIDA "Data Lake Warehouse" card is not necessary for successfully deploying AI and its associated data at scale. Indeed, the semantic platform recommended by TRAIDA relies on the MDM, ODS, and EKG repositories described in specific cards.

However, companies need to conduct data analyses, and the repositories of the semantic platform are insufficient when it comes to multidimensional analysis (OLAP) or the exploitation of large amounts of minimally or unstructured information (data lake).

The contributions of TRAIDA to decision-making IT are at two levels:

1. First, the ontologies modeled in the semantic platform can be reused to improve data warehouse and data lake solutions. These ontologies provide the metadata necessary for enriching analyses and enhancing data quality.
2. Second, the MDM, ODS, and EKG repositories are key data sources for the data warehouse and data lake.

DATA WAREHOUSE, DATA LAKE AND METADATA MANAGEMENT

The following architectural principles are proposed:

- The ODS is the preferred repository for feeding data warehouses with operational data. The MDM is the source of reference and master data, which are used to build data hierarchies and analysis dimensions (OLAP).
- The EKG is the preferred repository for feeding data lakes with already accumulated knowledge for AI needs. Additional sources of multimedia data can be added to complement the content from the EKG.
- In all cases, the shared ontologies at the semantic platform level serve as a reference for the metadata handled in data warehouses and data lakes.
- In the case of a data lakehouse equipped with a semantic layer based on a knowledge graph, integration with the EKG is even more natural. The knowledge graph at the EKG level should then be considered the master repository.

According to needs, AI can be used at all levels of the data warehouse and data lake. For example, an AI can be trained at the EKG level to be used on data from the data lake, or directly trained at the data lake level without attempting to reuse it at the EKG level. It is not possible to define generic governance rules that would apply to all companies. It is preferable to adopt a pragmatic approach and frame the training and use of AIs according to their operational or decision-making scope. In other words, a distinction should be made between AIs that act on operational systems (MDM, ODS, and EKG) and AIs that operate at the data analysis level (data warehouse, data lake).

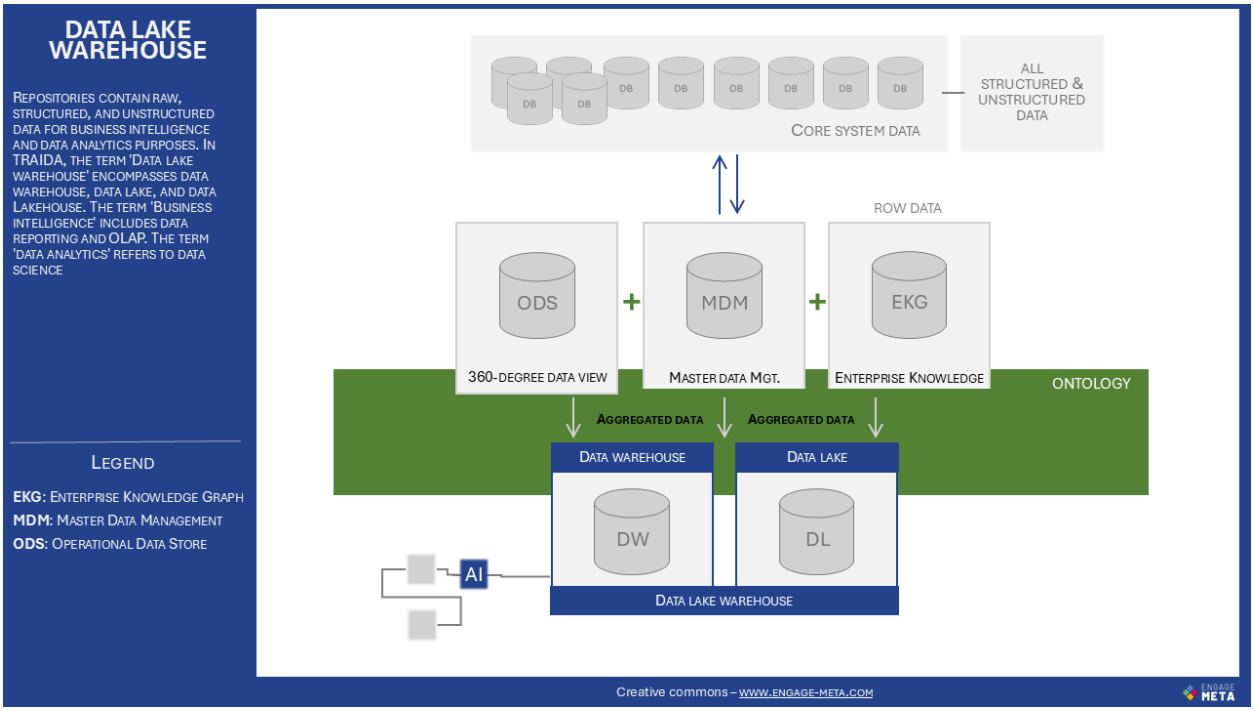
DURABLE AND LONG-TERM STORAGE

In the field of repositories for data analysis, the volumes handled are generally large. It is therefore necessary to specify the means used to ensure their storage by distinguishing between two conservation horizons:

- **Durable Storage**: This corresponds to the ability to provide data on demand, including large volumes, with redundancy mechanisms, multiple backups, real-time access security guarantees, and durability of access APIs. The physical storage media (disks, memories) must be able to change with technological advancements without altering the access mechanisms. Some well-known solutions include: Amazon S3, Google Cloud Storage, Azure Blob Storage.
- **Long-term Storage**: This corresponds to the archiving and recycling of data over long time horizons (several years, decades) without allowing real-time access to the data. When an archive needs to be loaded, a process lasting several hours or days is initiated in accordance with a service contract. The physical storage media must be maintained or even transferred from an old technology to a more recent one transparently for the archive user. Over a long period of several years, it is essential to ensure that the physical storage media does not degrade and remains readable with the most recent technologies. Therefore, regular maintenance is required. Some well-known solutions include: Amazon Glacier, Google Cloud Archive Storage, Azure Archive Storage.

AI solutions contribute to better management of these storage systems by monitoring access to detect potential fraud, anticipating failures, optimizing the transition between durable and long-term storage (hybrid storage), eliminating redundant data sets, and more.

3. BLUEPRINT



4. YOUR SITUATION & OBJECTIVES