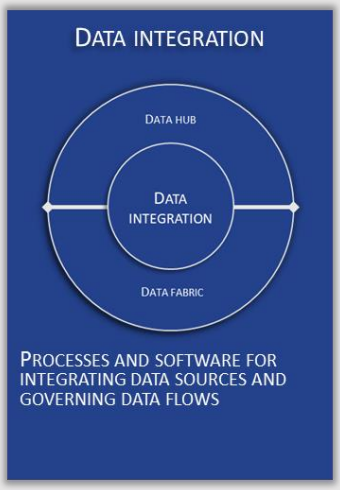
DATA INTEGRATION

Processes and software for integrating data sources and governing data flows. The data hub might compete with the ODS (Operational Data Store) of the semantic platform; and the data fabric might compete with the EKG (Enterprise Knowledge Graph). Therefore, a choice must be made to either use the data fabric as a component of the semantic platform or integrate it with more transversal MDM (Master Data Management), ODS, and EKG.



1. CONDITIONS OF SUCCESS

Data integration synchronizes and transforms multiple sources of information to provide a standardized data flow to consumers. These consumers can be repositories like MDM (Master Data Management), ODS (Operational Data Store), EKG (Enterprise Knowledge Graph), data warehouses, data lakes or application systems and AI systems for training.

Historically, this need has been covered by ETL (Extract, Transform, Load) and EAI (Enterprise Application Integration). However, to handle the complexity of integration processes, specific developments are often necessary to adapt them. These implementations become a significant technical debt and create a high rigidity in data flow integration. This rigidity is incompatible with agile governance. For instance, a simple change in data type requiring several days of maintenance would be unacceptable in a business emergency.

To address this rigidity of ETL-EAI, data hub and data fabric solutions have emerged.

Although the boundaries of these solutions vary depending on software vendors, their value proposition is based on greater agility in data flow integration. To achieve this, they use metadata and repositories for information storage that contribute to flow management. Consequently, they not only integrate data flows but also manage repositories. As vendors of these solutions ride technological and marketing waves, defining a solid architectural framework is not straightforward.

In this difficult-to-decipher marketing context, TRAIDA approaches the choice of data hub and data fabric by considering that unified data repositories like MDM, ODS, and EKG (see respective TRAIDA cards) must be preserved. They form the foundation of the semantic platform for AI.

Therefore, when considering a data hub or data fabric solution, it is essential to evaluate its ability to provide robust MDM, ODS, EKG repositories or to integrate with those of the semantic platform. For example, if the data hub establishes a metadata catalog, its integration with the shared ontologies in the semantic platform must be carefully examined. Neglecting this issue would result in managing two metadata catalogs: one at the global level housed in the semantic platform and the other accompanying data flow integration in the data hub. These two catalogs should share the same ontologies to avoid creating silos, which could lead to poor data quality and high maintenance costs.

To help you build the best data flow integration solution with the semantic platform for AI, here are the criteria to consider; for the data hub :

- The origin of the data hub is primarily technical and resembles an ETL-EAI solution enhanced with metadata management. This enhancement promotes better governance of data flows and quality control rules. The functions of data transformation, mapping, and flow propagation then rely on metadata. Consequently, the need for specific software development is reduced, making room for configuration and a NoCode approach.
- Some data hubs integrate data repository management, such as CDI (Customer Data Integration). In this case, the data hub is no longer limited to just data integration; it also provides a new repository. It then resembles a specialized ODS focused on a functional domain (here, customer management), competing with the generalist ODS of the semantic platform. Over time, some companies find themselves with a CDI data hub coexisting with other data hubs like HR, Supplier, Marketing, etc. Unfortunately, the creation of these siloed ODS repositories does not support unified data governance and generates maintenance difficulties. In the TRAIDA vision, it is preferable to build a unified ODS independent of the data hub. This would limit the data hub's scope to only managing data flow integration.

For the data fabric :

- A data fabric is a multi-technology framework that offers unified data management, with AI governance functions (management of data spaces for AI training, vectorized databases, prompt management, a unified user interface over different LLMs, etc.). It integrates data preparation and transformation functions identical to those of the data hub, based on metadata management.
- Most data fabrics incorporate a knowledge graph-oriented database technology, raising the question of integration with the EKG of the semantic platform. Unlike the data hub, which offers a solution that might compete with the ODS, the data fabric extends into the EKG repository.

How to make the right choice?

To avoid creating technological silos, the choice of a data hub or a data fabric must align with the unified repositories of the semantic platform, namely the ODS and EKG, followed by the MDM for metadata management. These repositories have a scope of action that transcends silos. They should not create new silos during data integration, as this would harm the information system's governance and data quality. These issues can arise if a data hub or data fabric imposes its own ODS and EKG repositories without sufficient integration capability with the semantic platform. To avoid this risk, follow these three recommendations:

1. **Specify your ODS and EKG needs independently**: It is necessary to specify your ODS and EKG needs independently of the study of data hub and data fabric solutions. The TRAIDA framework offers a sufficiently generic knowledge to achieve this (see the MDM, ODS, and EKG cards).

2. **Metadata management synchronization**: Data hub and data fabric solutions rely on metadata management. They impose a sort of verticalized MDM on the metadata that should be synchronized with the MDM of the semantic platform. Without such synchronization, different ontologies would be maintained between the operational management level (semantic platform) and flow integration (data hub or data fabric). These divergences degrade data quality and increase maintenance costs.

3. **Event-driven architecture**: To avoid point-to-point exchanges between systems, data flow integration relies on an asynchronous architecture, meaning event-driven management. For example, the ODS listens to a data channel for updates without being directly connected to the source application providing the flow. Modern data hubs offer this type of automation. The decoupling between repositories (MDM, ODS, EKG) and the systems that provide and consume data is crucial for architectural robustness. The alternative solution of exposing each system's access to all others resembles point-to-point exchanges. The negative consequences of this architecture are well-known and often illustrated by the metaphor of "spaghetti architecture."

By following these recommendations, you can ensure a cohesive and well-governed data integration strategy that leverages the strengths of your semantic platform while maintaining flexibility and data quality.

2. IMPORTANCE OF THIS CARD FOR YOUR TRANSFORMATIVE AI

The domain of data integration is strategic for successfully scaling AI. The MDM, ODS, and EKG repositories of the semantic platform must be synchronized with the upstream systems that provide them with data. They are also synchronized with the downstream systems that utilize them, including AI systems for training and real-time conversation enrichment (Retrieval Augmented Generation). In the realm of business intelligence, data warehouse, data lake, and data lakehouse repositories must also be fed from the semantic platform's repositories. These repositories then expose their data flows to information visualization tools and other AI systems.

Governance and quality control of data flows

There is a need for governance of data flows, including a description of data producers, consumers, transformation and normalization rules, as well as quality control, enrichment, security, traceability, version management rules, and more.

Choosing a data hub

When choosing a data hub, the decision to use its ODS component is relatively straightforward. It depends on the transactional quality of the database. For an SME (Small and Medium-sized Enterprise) and provided that the data hub's ODS is not verticalized on a specific functional domain, it is possible to make it a transversal solution for the entire company, thus becoming a component of the semantic platform. Regarding metadata management, the architectural principles presented for selecting a data fabric should also be applied (see below).

Choosing a data fabric

Selecting a data fabric is more complex than the choice of a datahub. Indeed, it must be coordinated with the EKG repository and the MDM. Depending on the size of the company and the complexity of its technical infrastructure, several scenarios are possible; the following two are the most significant:

- For an SME: A solution centered on a data fabric with a knowledge graph-oriented data repository should also be usable as the EKG repository of the semantic platform. Metadata management could also be handled within the data fabric, replacing a more general MDM, at least during the initial deployment phase and while waiting for a more robust semantic platform. It's important to note that the MDM repository provides business governance functions similar to a complete application system, which typically do not exist in the data fabric. The ODS needs to be addressed with a transactional framework (OLTP), which is not always scalable with knowledge graph technology.

- For a Large Enterprise: Implementing MDM, ODS, and EKG repositories independently of the data fabric seems essential. The repositories in the data fabric should be considered tactical and synchronized with the master repositories in the semantic platform. If this integration is too burdensome, reconsidering the choice of data fabric is advisable; otherwise, there is a significant risk of creating silos at the level of strategic repositories. In a multi-year deployment program, it is also possible to consider some data fabric repositories as master during a limited time, awaiting reversibility into the semantic platform.

In all scenarios, ensuring the coherence of decisions hinges on sharing ontologies. No technical solution should impose specific ontologies that diverge from those shared at the enterprise level.

DATA HUB

As previously mentioned in this document, the data hub provides two main functions:

1. Integration of data flows.
2. A data repository resembling a verticalized ODS, such as a Customer Data Integration (CDI).

Deploying verticalized ODS within a data hub exacerbates the problems caused by silos, which unnecessarily duplicate data. Therefore, it is better to limit the data hub's scope to the integration of data flows, essentially an ETL-EAI augmented with a metadata catalog. This approach reduces the need for specific development in favor of flow configuration. It is essential to study the synchronization of this catalog with the MDM of the semantic platform. Finally, the data hub must offer asynchronous mechanisms for flow integration, meaning an event-driven architecture.

Ideal architecture

In summary, the ideal architecture relies on a data hub that handles data flow integration by integrating with the MDM of the semantic platform to unify metadata management (structure of flows, list of data-producing and consuming systems, service contracts, etc.). The ODS needs are not addressed at the data hub level and remain the responsibility of the semantic platform.

Contribution to AI scaling

The data hub does not directly contribute to scaling AI. However, it is essential for industrializing data flow integration both upstream and downstream of the semantic platform.

DATA FABRIC

The data fabric is a technological assembly based on data flow integration, similar to a data hub. Beyond this integration, the innovative aspect of the data fabric lies in the provision of a knowledge graph-oriented data repository akin to the EKG of the semantic platform. The technical quality of this repository and the functions offered by the data fabric determine its proximity to the semantic platform. This technical quality can be assessed through the following main criteria:

- Ability to handle increased data volume.
- Compliance with transactional management (ACID) even under intense multi-user demands.
- Ability to synchronize the metadata catalog and ontologies with third-party tools (unified MDM).
- Availability of governance functions for version management, including comparison and branch merging.

Key functions of the data fabric

The main functions offered by the data fabric include:

- Metadata manager: Some solutions provide a semantic modeling tool for building ontologies.
- Metadata and ontology use: For data flow integration, reducing specific development in favor of configuration.
- Integration with AI systems: Configuring data training flows (datasets), managing fine-tuning processes, and prompt catalogs.
- Version management: Managing versions of metadata, ontologies, and integration processes to control changes over time.
- User and access management: Security based on profiles.
- Interfaces for data visualization Tools Integration.

Data Fabric deployment scenarios

For an SME or as a first tactical deployment in a large enterprise, a data fabric with powerful knowledge graph technology can serve as a component of the semantic platform as described in TRAIDA. After the initial implementation effort, the sustainability of the technology within the semantic platform should be evaluated. This deployment mode quickly provides AI governance (training flow configuration, prompt catalog, version management, etc.).

If a tactical project is not desired for deciding the target solution, consider integrating the EKG with the knowledge graph-oriented database of the data fabric. The benefits of this choice include:

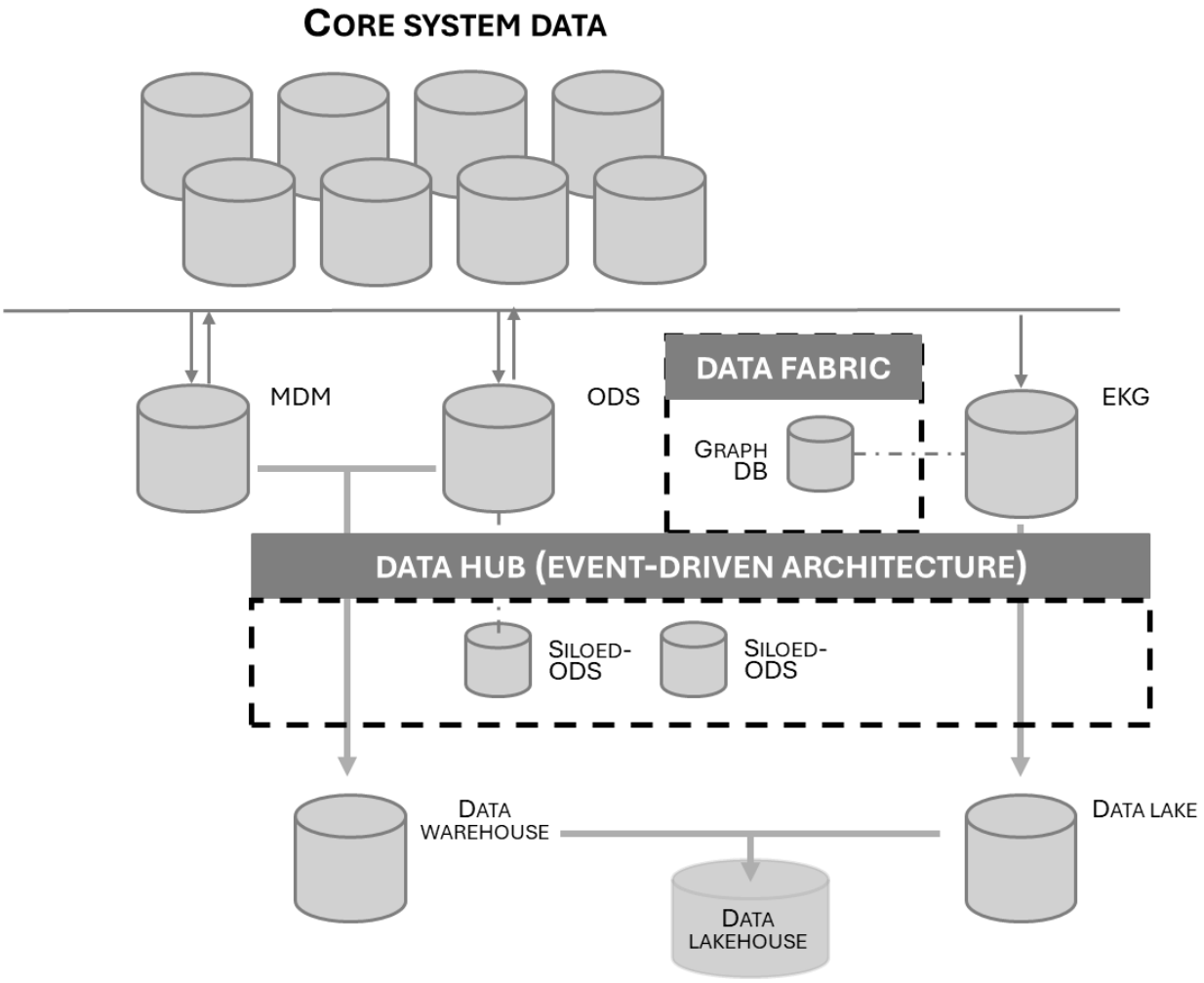
- Reduced dependency: Maintaining an EKG repository in addition to the data fabric technology reduces dependency on the latter. Changing the data fabric can leverage the data capitalized in the EKG.
- Selective data storage: The EKG can contain information that doesn't need to be copied into the data fabric. Without it, there's a risk of treating the data fabric's repository as a universal storage solution, increasing technological dependency. For example, a knowledge graph that imports regulatory documentation can be managed at the EKG level without using the data fabric's storage repository.

Key considerations for AI governance

Regardless of the path chosen for analyzing and deploying your data fabric solution, it's crucial to specify your AI governance needs. Refer to the governance cards in the TRAIDA framework for more information (green cards). The advantage of the data fabric lies in its ability to manage AI system training data flows (datasets). However, it's not just about creating prompts and uploading information into AI assistants. It's also about connecting these elements with the business concepts (ontologies) the company uses in its operations and managing versions. For example, consider how to ensure an AI forgets certain data when necessary.

Therefore, the choice of a data fabric also depends on your AI governance needs. If no existing solution meets your expectations, consider developing specific enhancements around the EKG of the semantic platform.

3. BLUEPRINT



4. YOUR SITUATION & OBJECTIVES