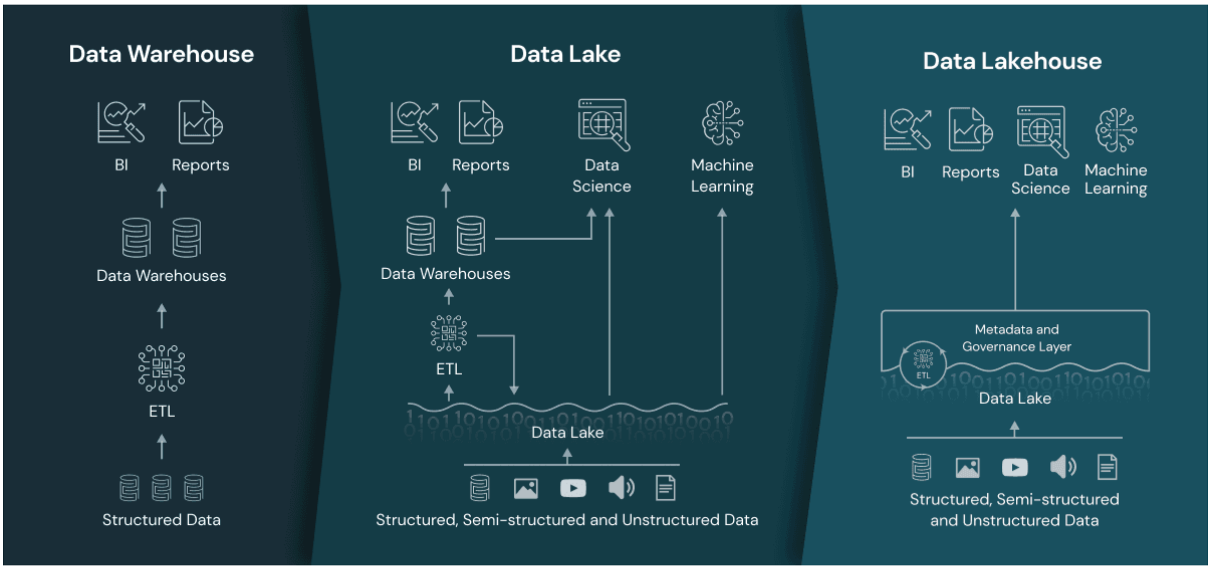
Proposal of meta management and integrated query technology using AGE for DW

Data Warehouse (DW) was introduced by IBM in the mid-1980s as an information warehouse. Later, in the late 1980s, Inmon began using the data warehouse concept as a data access strategy. The DW system has grown into a representative system in the field of information System. Recently, Data LakeHouse has been proposed as a data architecture that combines Data Lake, which stores raw data, and Data Warehouse, which manages refined data. The Data Lakehouse concept emphasizes simplicity, flexibility, and low cost. In particular, the flexibility required to solve and respond to many problems in the IT business is being emphasized more.



<Source: Lakehouse, databricks.com>

In addition, cloud-based DW is linked to data collection, BI, and machine learning, making it easy for people without special technical capabilities to use it, and reducing trial and error and risk of introducing a solution by using and returning the service only when necessary. As such, DW must be an important system that must be used by private companies and government agencies even if the IT environment changes.

However, these DW, Data Lake, and Data Lakehouses also have disadvantages, as shown in the table below.

Traditional DW & Problems with existing big data	**[Problems with traditional DW]** 1. The more data you have, the more expensive the cost of storing and processing your data 2. Lack of flexibility in collected data 3. High redundancy data and ETL or ELT redundancy costs **[DW problems in big data environment]** 1. In Data Lake, it is not easy to apply data security for personal information protection, and there is a problem of cost due to the up-to-date data - Since the data in the data lake is data before it is refined, it is difficult to select personal information and apply access control, masking, and encryption. - Difficult to ensure up-to-date due to different data collection cycles and levels 2. Data Lake is huge and expensive to find and refine the data you want or store unnecessary data 3. Data Lake has difficulty managing data. - Because you don't know who and how to use it from the perspective of managers who need to manage integrated management

In the end, the contents requested by data consumers can be summarized as follows.

- Data consumers must be able to quickly find and utilize the high-quality data they want.
- It must be flexible to adapt to rapid environmental changes.
- The costs between supply and demand must also be considered.
- The backbone system should not be loaded.

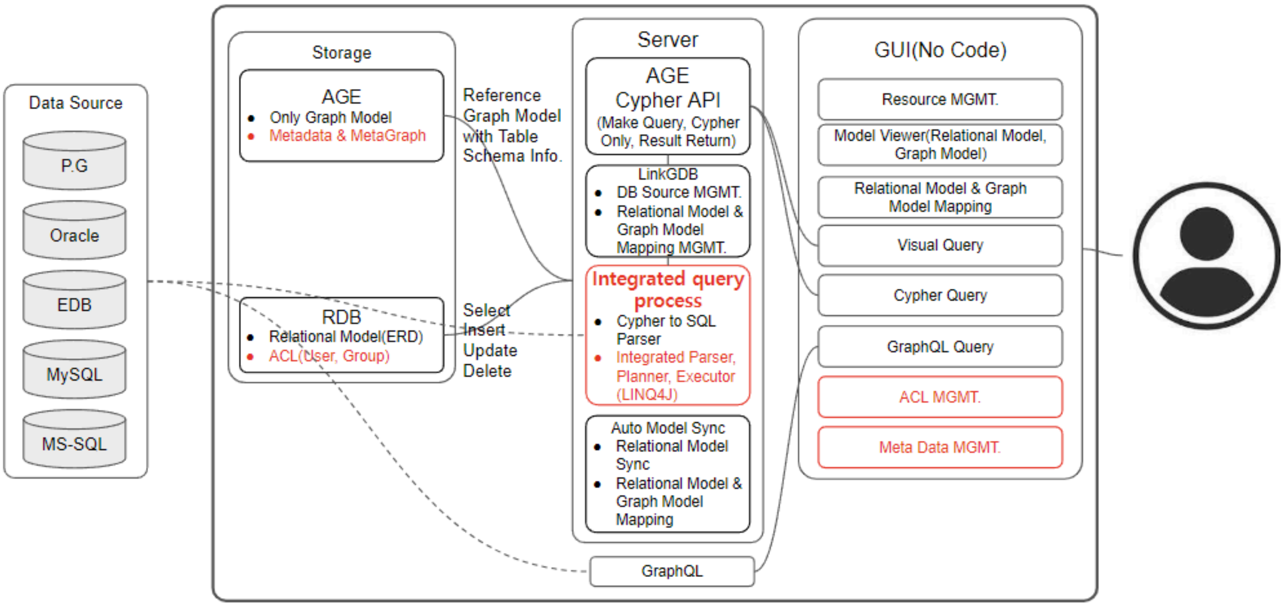
The DW that data consumers want requires an environment that is flexible in responding to the environment, does not load existing systems, and can use high-quality data quickly and inexpensively. It is difficult to find an ideal technology that meets all of the requirements. However, there are platforms that provide technologies accordingly, and they want to share related architectures and directions.

[Build a flexible DW environment in an environment where data consumers can easily find and quickly utilize the high-quality data they want.]

I intend to propose an architecture and direction to provide high-quality data to data consumers at a low cost through a flexible integrated query system considering data Warehouse data virtualization technology and data security.

The following architectures and technologies are expected to be required to build a DW environment such as this title.

* Legend: AGE YELLOW (Black text), Proposal (Red text)



<**Integrated query concept**, Integrated query server and data sources architecture>

technology summary	**Technical Message:** Dataeta & MetaGraph for quality integration in real-time replication or distributed environments for large PostgreSQL based Data Warehouse, integrated quality processing technology ● **Data Source**: Data replication through various replicas (If only used PostgreSQL) ○ Perfect Sync-based streaming replication technology ○ Read/Write & Read only ● Access security and data source management ○ ACL (Group, User, Object) Manager ○ Integrated meta (Metadata & MetaGraph & ERD) ● Data virtualization technology ○ Distributed data - **Metadata storage/processing technology**

	○ **Logical MetaGraph technology for flexible data connectivity** ● **Integrated query processing technology** ○ Ansi SQL & Graph Grammar ○ Parser and ACL referencing Meta ○ Heterogeneous Data Source Integrated Query Planner (Optimizer) ○ Integrated query executor ○ Merge integrated query results

The advantages of the above architecture are as follows.

- By managing the data meta of the DB on the DW cluster or data source, unnecessary ETL and storage space are saved, lowering costs and increasing flexibility.
- MetaGraph enables data users to quickly find the data they want and ask an Integrated query.
- Secure data exists in an encrypted state on the data source, and access control through ACL is possible even when querying.