Factored

DATA WAREHOUSE OR DATA LAKEHOUSE?

TECH SERIES

DECIDING WHICH PLATFORM IS **RIGHT FOR YOUR ANALYTICS APPS**

Information is without a doubt the most valuable asset for modern businesses. It allows them to make good decisions and drive strategies. For decades, there was this concept of a data warehouse, where a company could store all their data and leverage it to deliver insights. Then, a much newer concept arrived, the data lakehouse. Which architecture is right for you? Is there a reference for how to use them?

In this article, we define and explore data analytics patterns with these two options. The data lake is not introduced here as it lacks a data governance layer, making it difficult to deliver efficient data analytics.

A CLOSER LOOK AT THE DATA WAREHOUSE

A data warehouse is a central repository of structured data that powers business intelligence applications. **Why is it important?** Intuition and instinct are not enough to be successful. That is why we need data. A data warehouse enforces standardization and accuracy. According to a Gartner study, organizations believe that poor data quality is responsible for $15 million a year on average in losses[1]. Ensuring data is accurate and in a central location also boosts efficiency as it eliminates the need to ask for validations and assistance from support staff or other analysts. It also enables a myriad of ways to automate new processes.

This architecture has a governance layer that makes data secure, manageable, and scalable. It includes[2]:

- **Metadata:** To locate, tag, and discover the structure of the data
- **Data model:** How is the data abstracted and organized?
- **Data lineage:** The history of the data including its origin and all its transformations
- **KPIs:** Key Performance Indicators that drive the design of the warehouse
- **ETL:** The technology that makes the data usable.

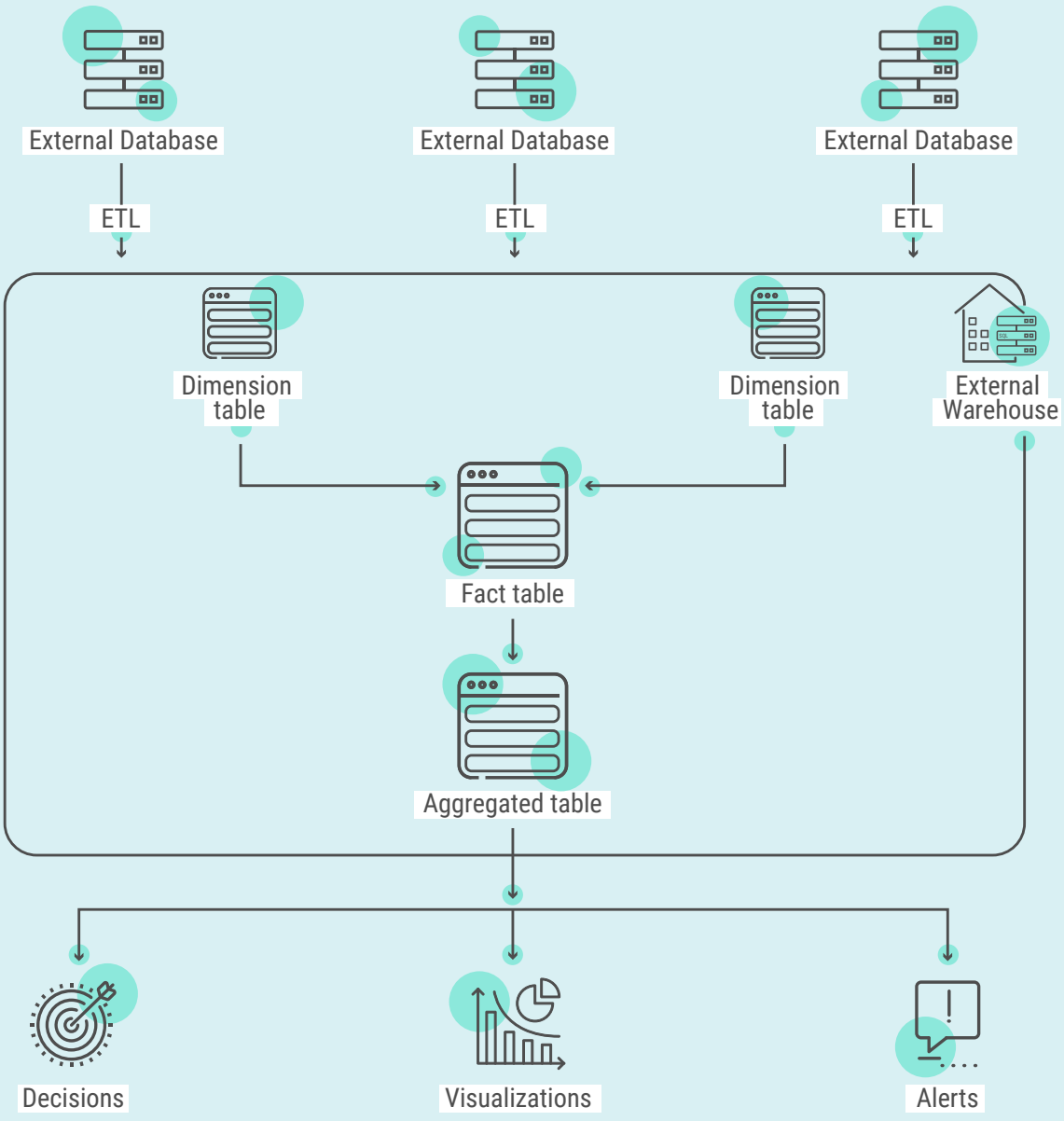
According to a Gartner study, organizations believe that poor data quality is responsible for $15 million a year on average in losses[1].

Furthermore, this concept was created to work with structured data, stored in relational tables inside a database. There is a common design pattern called the star schema. It consists of creating fact tables containing measurements from the common operation of the business, and dimensional tables that give context to the facts. Naturally, dimensions contain data that does not change often.

As the data is structured, it is meant for businesses with a concrete use case, where the shape of the data does not change over the short term. The most common example is a bank. First, you start asking questions like:

- Which location has the highest monthly transactions?
- Which bank loaned the highest amount of money this year?

The facts and dimensions would be joined, then aggregated using functions like summation, average, or median, over the field where the data will be grouped. In this case, the location of the banks. Finally, this aggregated data is transformed into visualizations that answer business questions quickly and intuitively.



Data warehouse reference. Created by Factored.

THE INS AND OUTS OF THE DATA LAKEHOUSE

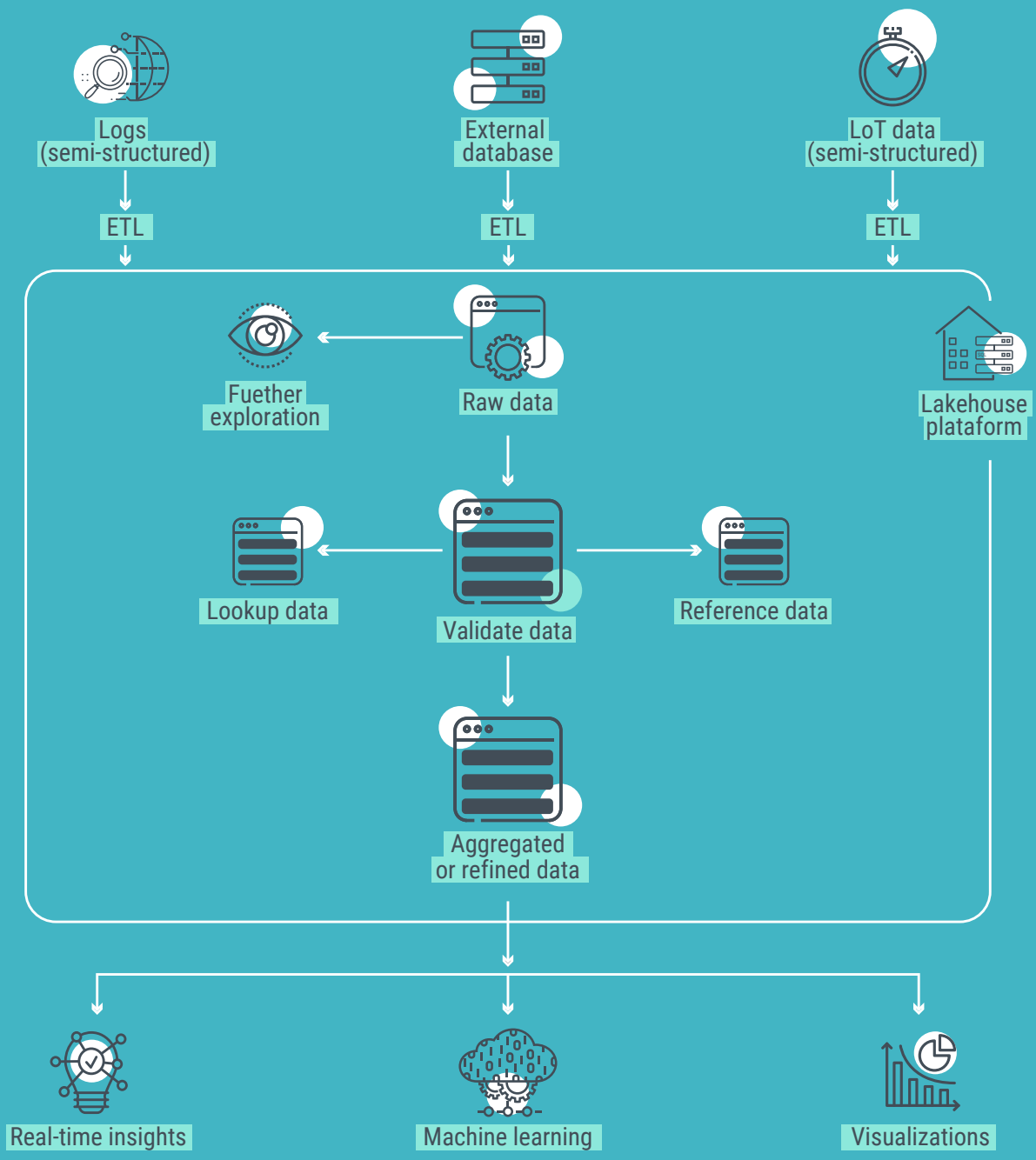
A **data lakehouse is also** a repository of data. The difference is that lakehouses can also ingest semi-structured and unstructured data as well, to develop both batch and streaming applications. **Why is it important?** Nowadays, data comes from a large variety of sources and shapes. According to IBM, unstructured data accounts for over 80% of enterprise data[3].

Moreover, this architecture is built to work with open file formats like parquet and ORC, much like a data lake, allowing us to easily train machine learning models. **These advancements enable advanced analytics applications that can be tailor-made for modern businesses.**

Consider the example of an IoT Consider the example of an IoT company, which has thousands of sensors placed on a farm. It needs to monitor the properties of the soil in near real-time to optimize the use of water and other products. With the data lakehouse, we can develop a multi-layer data model.

- The first layer ingests the data in its raw format.
- The second layer transforms and validates the data. Reference data can also be joined at this stage.
- The third layer aggregates the data for visualization.

> According to IBM, unstructured data accounts for over 80% of enterprise data[3].



Data lakehouse reference. Created by Factored.

The raw layer is possible because of the cheap, open storage system of the lakehouse. Further batch analytics can be performed from this layer to gain deeper insights that might have been missed in the real-time application and can follow the same multi-layer pattern explained above.

On top of all these capabilities, there is the same governance layer that warehouses provide on top of the data, making it usable, easily discoverable, traceable, and secure.

SO SHOULD WE **ALL OPT FOR DATA LAKEHOUSES?**

There is no silver bullet when it comes to data. One has to analyze the requirements for the use case and pick the tool that best suits their needs. If most of your data comes from relational tables, and the nature of the business does not rapidly change, then a data warehouse would be a great option. There are very mature platforms like Snowflake, Amazon Redshift, and Azure Synapse.

On the other hand, if you want to extract value from various types of data sources and your data models are likely to evolve, then a data lakehouse might be a good choice. There is also a two-tier architecture that uses a data lake, followed by a warehouse, but it is more costly and involves replicating data from one repository to the other. In addition, if you need to perform machine learning modeling, or other advanced analytics in a single platform, then consider a data lakehouse.

REFERENCES

1. Moore, S. (2018). How to Create a Business Case for Data Quality Improvement. Https://Www.Gartner.Com/Smarterwithgartner/How-to-Create-a-Business-Case-for-Data-Quality-Improvement.

2.Inmon Bill, Levins, M., & Srivastava, R. (2021). Building the Data Lakehouse (J. Hoberman, Ed.). Technics Publications.

3.IBM Cloud Education. (2021). Structured vs. Unstructured Data: What's the Difference? Https://Www.Ibm.Com/Cloud/Blog/Structured-vs-Unstructured-Data.