Objectives

In this Lecture you will learn:

- What is Data Warehouse.
- The core idea behind datawarehousing.
- The basic layer of Enterprise Data Wrehouse.
- Staging area, ETL and Data marts
- Various application of data warehousing.

1. Introduction

Throughout the day we make many decisions relying on previous experience. Our brains store trillions of bits of data about past events and leverage those memories each time we face the need to *make a decision*. Like people, companies generate and collect tons of data about the past. And this data can be used to *make better decisions*.

In today's business environment, an organization must have *reliable reporting and analysis of large amounts of data*. Businesses need their data collected and integrated for different levels of aggregation, from customer service to partner integration to top-level executive business decisions. This is where data warehousing comes in to make reporting and analysis easier. This rise in data, in turn, increases the use of data warehouses to manage business data.

2. Enterprise data warehouse components EDW

There are a lot of instruments used to set up an enterprise data warehousing platform. Let's have a bird's eye view of the purpose of each component and its functions.

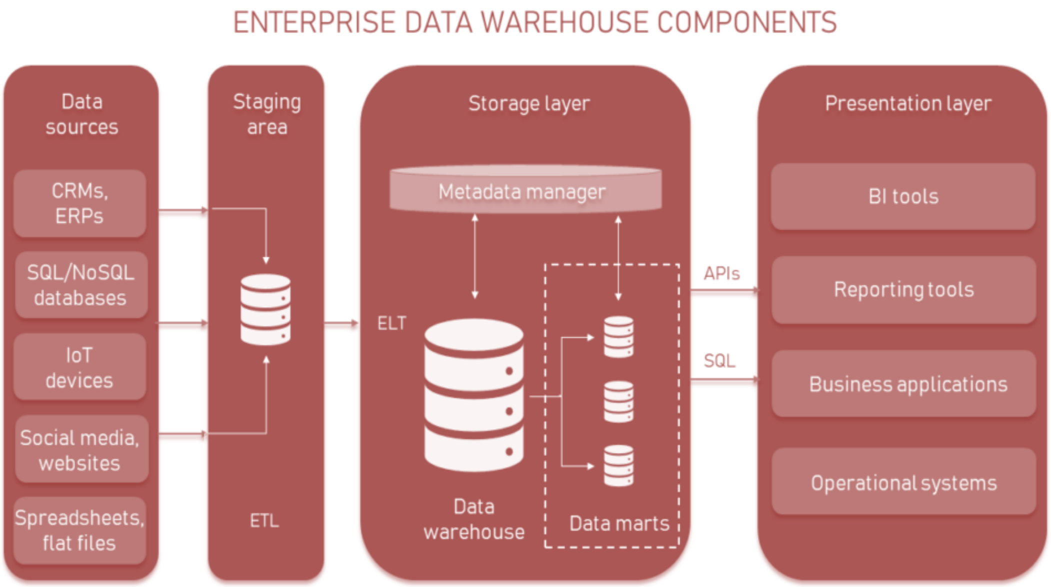

Fig1. EDW components.

2.1 Data sources

These are all the data sources where raw data originates and/or is stored. They can range from simple *spreadsheets* to flat files relational, *SQL databases*, *IoT systems*, and more.

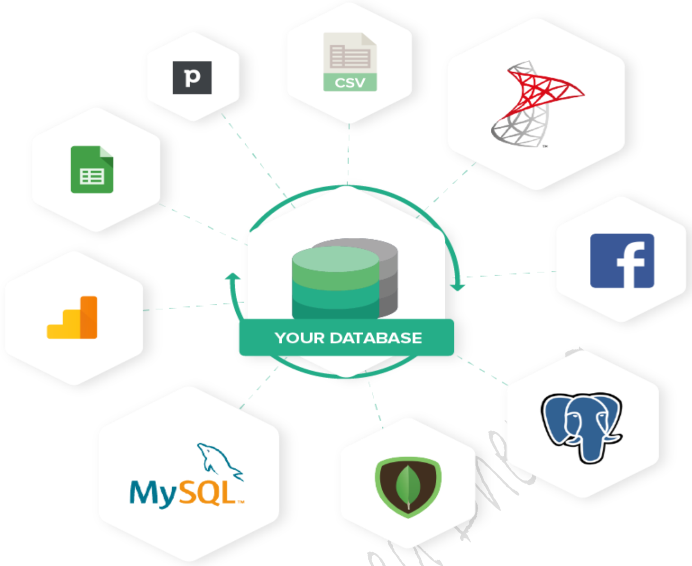

Fig2. Data Source.

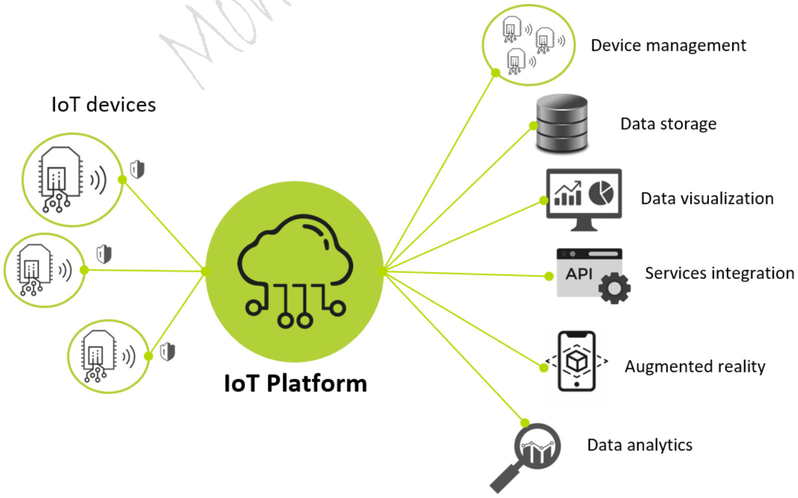

Fig3. IoT Data Source.

2.2 Staging area

ETL stands for **Extract**, **Transform**, and **Load**, and it is the most common method used by Data Extraction Tools and Business Intelligence Tools to extract data from a data source, transform it into a common format suitable for further analysis, and then load it into a common storage location, usually a Data Warehouse.

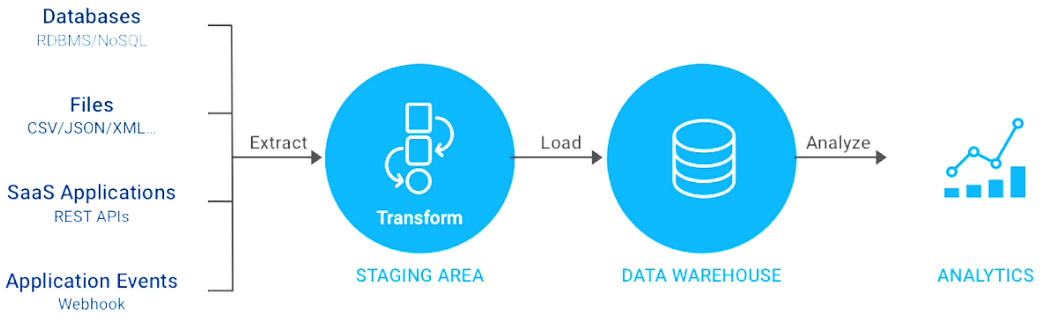

Fig4. ETL Process.

Ever tried juggling data from different sources, each with its own quirks? That's where the data staging area swoops in, cleaning up the mess and making sure everything speaks the same language before it joins the party.

Thus, the *data staging area* in a data warehouse is a critical intermediate space where data is **processed** and **prepared** before being transferred to the warehouse. It serves as a buffer zone for **cleansing**, **transforming**, and **consolidating** data from various sources. This area ensures data **quality** and **consistency**, making it ready for analysis and reporting in the data warehouse.

In the case of ETL, the staging area is the place where data is transformed before EDW. Here, it will get **cleansed**, **de-duplicated**, **split**, **joined**, and **converted** into a **unified format** to fit a given data model of a warehouse.

2.3 Storage layer

The data is finally **loaded into the storage space**. As we mentioned, data warehouses are most often relational databases. DW will also include a database management system and additional storage for metadata.

2.3.1 Metadata module

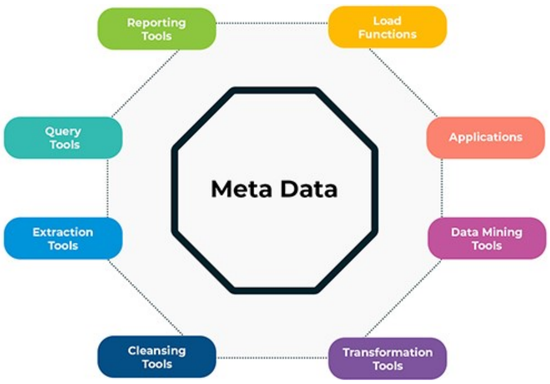

Fig. 5. Metadata Tools.

Put simply, metadata is data about data. These are the explanations that give hints for users/administrators of what subject/domain this information relates to. This data can be **technical meta** (e.g., initial source), or **business meta** (e.g., region of sales).

All the metadata is stored in a **separate module** of EDW and is managed by a **metadata manager**. In some cases, there might be an additional layer built on top of the whole infrastructure to curate metadata like a **data virtualization layer**.

2.3.2 Data marts (optional)

In some cases, an EDW can have a set of **smaller subsections called data marts** that are built specifically for a particular subject area, business function, or group of users. For example, there can be a separate data mart for marketing purposes and a data mart for a financial department.

Data Marts **are tailored storage solutions for specific business units or lines of business**, while Data Warehouses serve as a **central repository** for the entire organization.

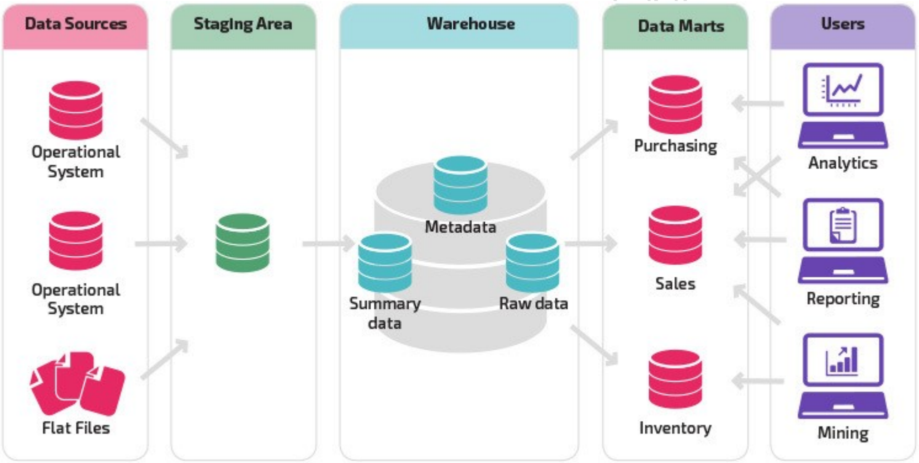

Fig. 5. EDW with Data Marts.

CheckPoint: What is the difference between metadata, raw, and summary data?

2.4 Presentation layer

The final building block of an EDW comprises tools that give **end users access to data**. Also called the **BI interface**, this layer will serve as a dashboard for data visualization, business reporting, and pulling out separate pieces of information for such tasks as **machine learning on data mining**.

- Business intelligence tools (BI tools) or reporting tools are all about helping you understand trends and derive insights from your data so that you can make tactical and strategic business decisions. Examples are SAP, Datapine, and MicroStrategy.

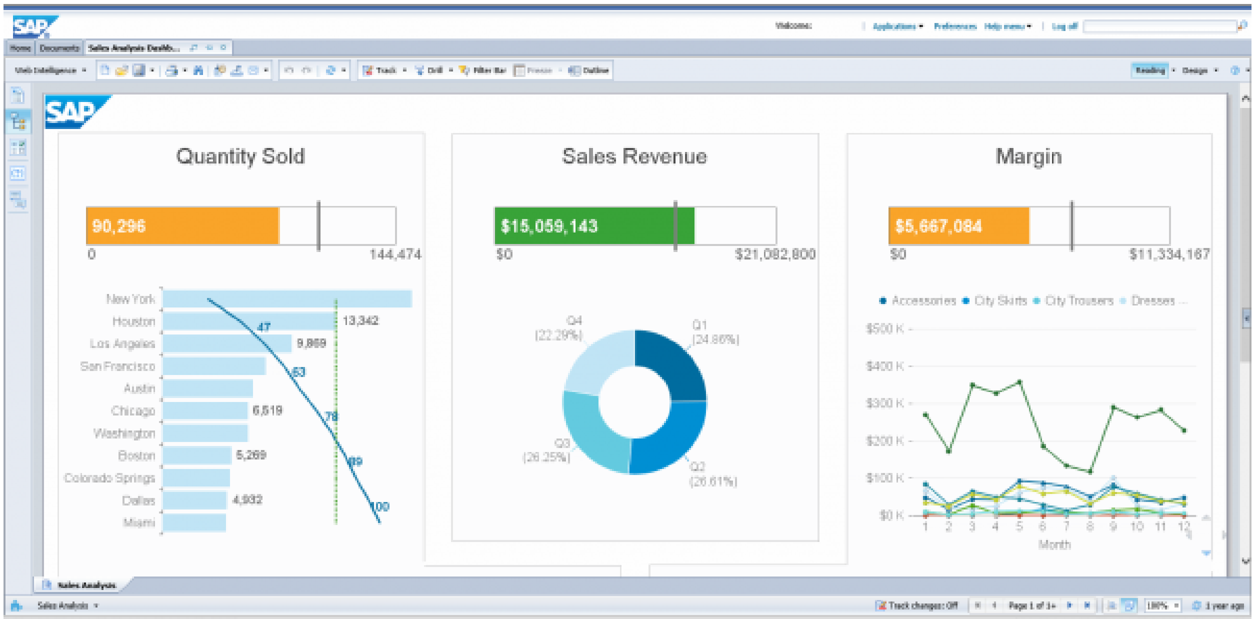

Fig. 6. BI Tools.

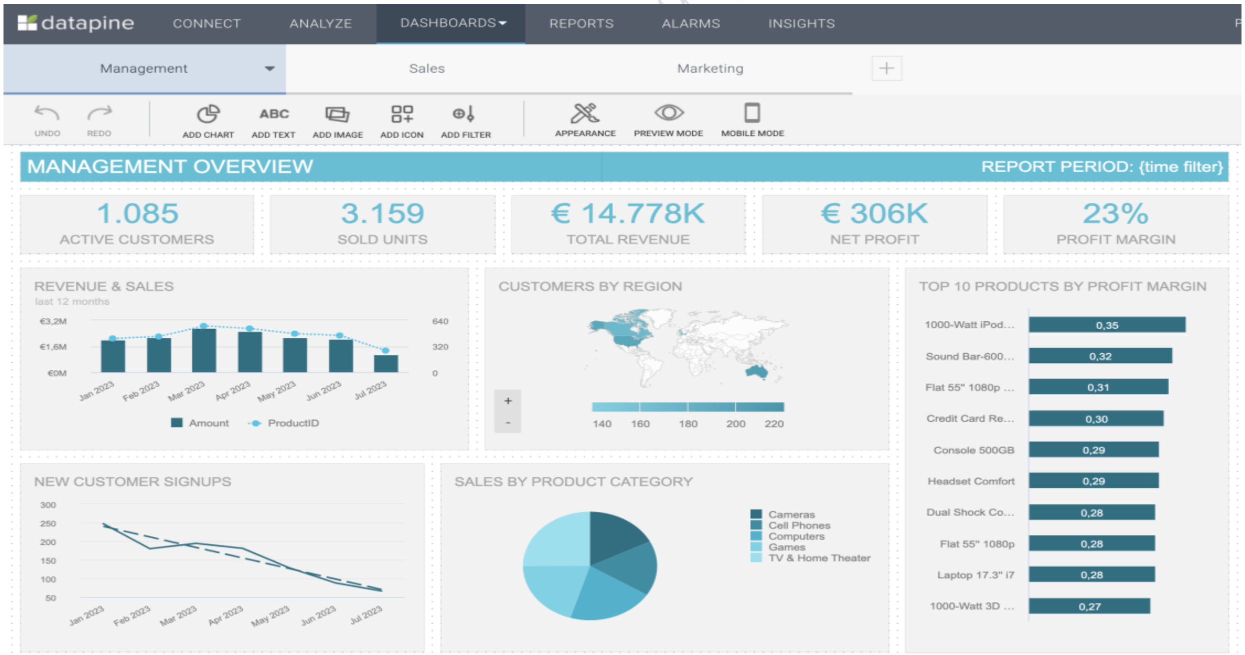

Fig. 7. BI Tools.

3. The best applications of Data Warehousing

3.1 E-commerce:

E-commerce platforms need to gather key marketing metrics (such as clicks, impressions, website visitors, etc.) from marketing tools and use that to approach their customers in a better way. An example is the story of a pregnant mother in the USA.

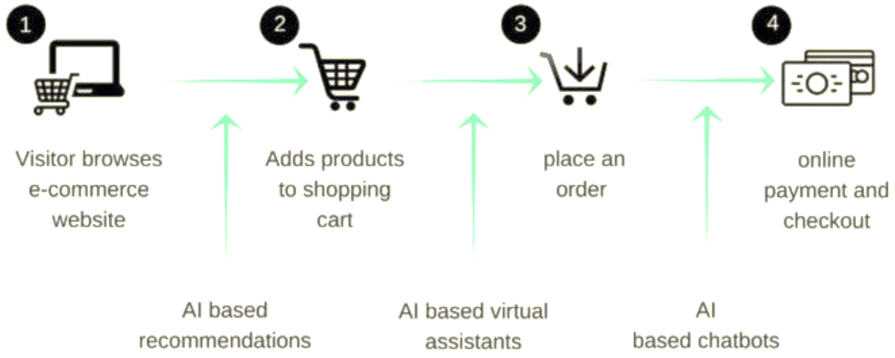
1
Visitor browses e-commerce website
2
Adds products to shopping cart
3
place an order
4
online payment and checkout
AI based recommendations
AI based virtual assistants
AI based chatbots

Fig. 8. E-commerce.

3.2 Retail:

Data warehouses can be used by retailers to easily identify **products with high demand** and the **fastest selling demand**. The data can then be used to react to a rise or fall in consumer demand quickly, which can ultimately be used to gain a competitive advantage.

3.3 Artificial Intelligence/Machine Learning (AI/ML):

With many companies embracing AI for their data journey, it's critical to get a reliable data warehouse now.

AI enables data maturity, which is intertwined with the flexibility, scalability, and agility that a warehouse offers.

On the other hand, machine learning is used on the data after the data has been replicated and transformed in the warehouse, to help newer business models emerge and advance digital disruption. Examples are recommendation engines on YouTube, Chatbots, speech **recognition**-based commercials on Facebook, or racism, terrorism, and hate speech of X **detection**.

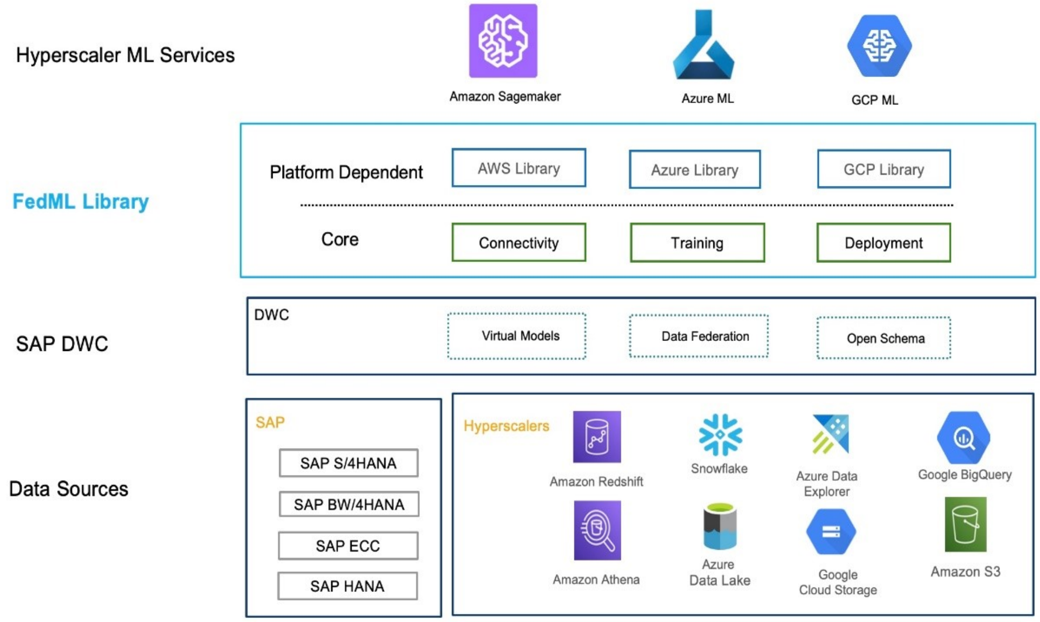

Fig. 9. SAP Fed-ML.

3.4 Agritech

Data storage is a must when it comes to the new age of farming. With advanced analytics, engineers and business analysts can figure out inefficiencies in the ecosystem, such as problems in the soil quality, unnecessary use of pesticides, etc., and iron them out.

Checkpoint: you can list more real-world data warehouse applications.