Mestrado em Engenharia Informática
KE

2022/2023

BigData

Programa

Large Scale Data Handling

1. Characteristics
2. Modeling and Management
3. MapReduce
4. Hadoop

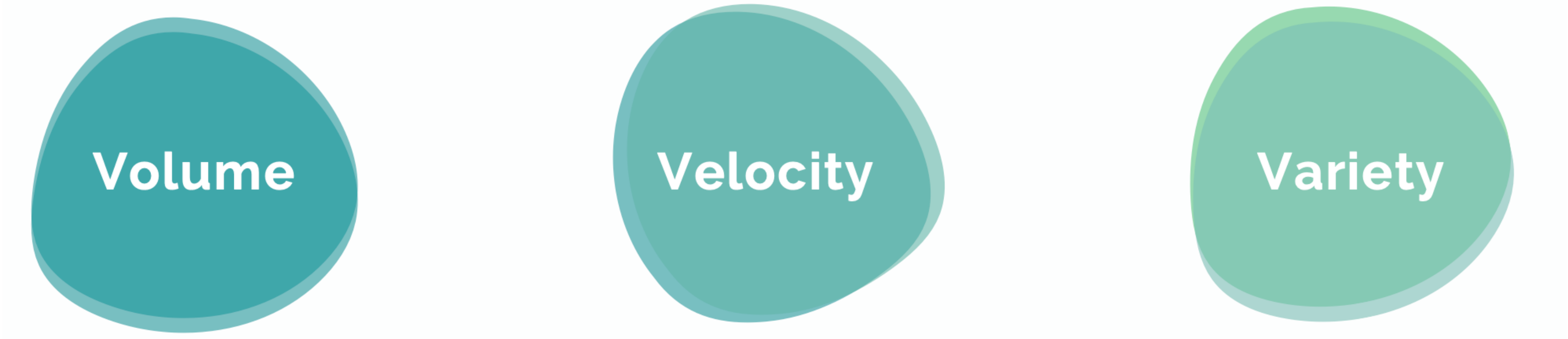
BIG DATA CONCEPT

The concept of Big Data remains so far a relative term with regard to the boundary between what is and is not considered Big Data.

For a company such as Google, the concept and size of Big Data is much different from that assumed for a medium-sized company.

BIG DATA CONCEPT

The most accepted definition was given by Douglas Laney. Laney observed that Big Data grew in three different dimensions:



Volume

Velocity

Variety

BIG DATA CONCEPT

However, other authors have crossed these characteristics by adding several other V's to this definition, such as:

- Value,
- Veracity,
- Visualization,
- Viscosity,
- Virality,
- among others.

The 4th most consensual V's is undoubtedly the **veracity**.

BIG DATA CHARACTERISTICS

VOLUME

The volume of data gives the large amount of data, mostly described in several petabytes or even more. However, not even this definition is consensual among the authors, since the definition depends on the type of data being analyzed.

BIG DATA CHARACTERISTICS

VELOCITY

The velocity concerns both the rate of data generation and the speed of analysis they require. Big Data Velocity deals with the speed at which data flows in from sources.

BIG DATA CHARACTERISTICS

VARIETY

The variety of data has increased exponentially due to the diversity of collection sources. Data can have several organizations and reach the collection point in a structured, semi-structured or even unstructured way. In addition, data formats must be taken into account.

BIG DATA CHARACTERISTICS

VERACITY

Veracity encompasses the reliability inherent in some sources of data collection. For example, information taken from a social network cannot be given the same relevance as information taken from hospital software.

BIG DATA MODELING

Why Is Data Modeling Necessary?

BIG DATA MODELING

Large amounts of data imply a system or method to keep everything in order.

The process of sorting and storing data is called "data modeling".

A data model is a method by which we can organize and store data.

BIG DATA MODELING

Proper models and storage environments offer the following benefits to large data:

- **Performance,**
- **Cost**,
- **Efficiency**, and
- **Quality**.

BIG DATA MODELING

Performance: Ensures fast query and reduces I/O output.

BIG DATA MODELING

Cost: Significantly reduces data redundancy, reducing storage and computing costs for the
large data system.

BIG DATA MODELING

Efficiency: They greatly improve the user experience as well as the efficiency of data use.

BIG DATA MODELING

Quality: They make data statistics more consistent and reduce the possibility of
computing errors.

6 TIPS FOR MODELING BIG DATA

01 DON'T IMPOSE TRADITIONAL MODELING

02 DESIGN A SYSTEM, NOT A SCHEMA

03 LOOK FOR BIG DATA MODELING TOOLS

04 FOCUS ON DATA THAT IS CORE TO YOUR BUSINESS

05 DELIVER QUALITY DATA

06 LOOK FOR KEY INROADS INTO THE DATA

BIG DATA MANAGEMENT

Big Data Management is a set of practices that promotes the

- **collection,**
- **organization,**
- **administration and**
- **interpretation**

of large volumes of data.

BIG DATA MANAGEMENT

- **Adequacy**

Ability to analyze a large amount of information, structured or not, allows the detection and correction of errors in stored information

BIG DATA MANAGEMENT

- **Integration**

Ability to filter and classify data so that it can later be handled assuming a standardized structure.

BIG DATA MANAGEMENT

- **Migration**

Ability to move data from one environment to another quickly and conveniently.

BIG DATA MANAGEMENT

- **Management**

Ensure the availability and security of data, ensuring that it follows all the organization's policies and standards.

BIG DATA MANAGEMENT: ADVANTAGES

Increase in company revenue

More accurate decision making

Strategy improvement

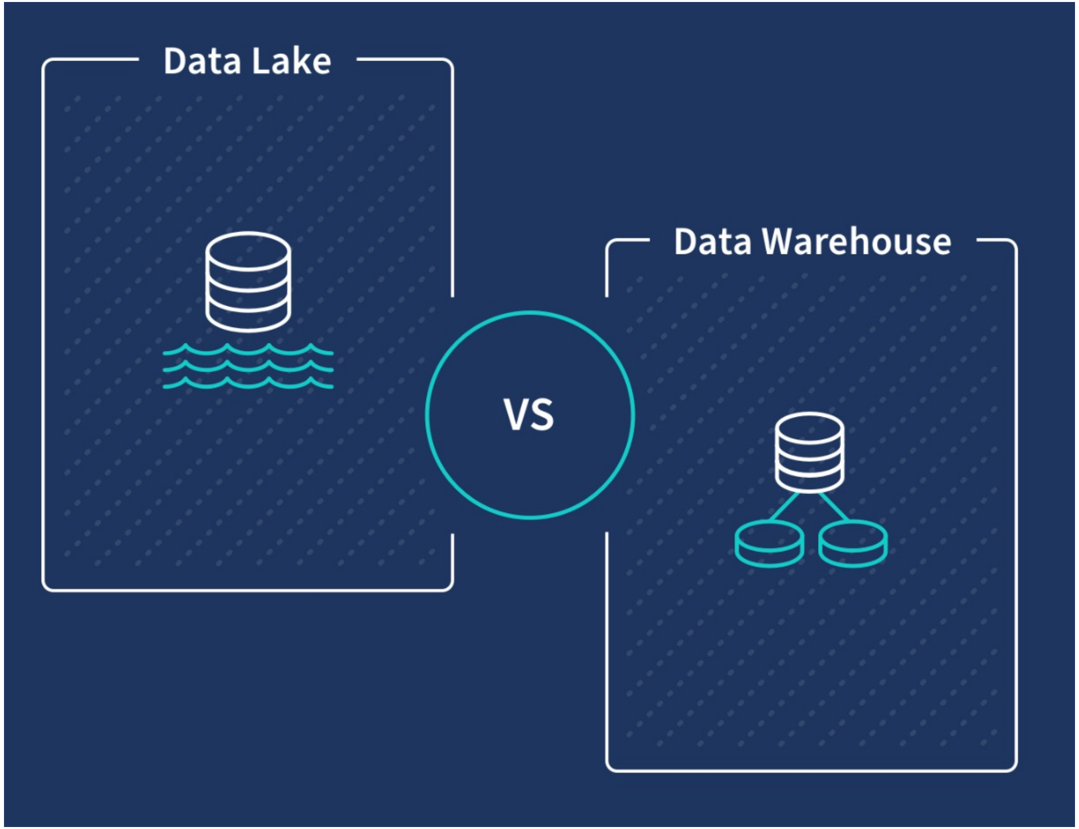
Team productivity and efficiency

BIG DATA MANAGEMENT

Choosing a data lake or data warehouse

Data lakes and **data warehouses** are fundamentally very different storage solutions, each with their own pros and cons.

DATA WAREHOUSE VS DATA LAKE



DATA WAREHOUSE VS DATA LAKE

A data lake is a massive repository of structured and unstructured data, and the *purpose for this data has not been defined*.

A data warehouse is a repository of highly structured historical data which has been processed *for a defined purpose*.

DATA LAKE

What is a Data Lake?

A data lake is a repository that stores all of organization's data — both structured and unstructured. Think of it as a massive storage pool for data in its natural, raw state (like a lake).

DATA LAKE

What is a Data Lake?

A **data lake architecture** can handle the huge volumes of data that most organizations produce without the need to structure it first.

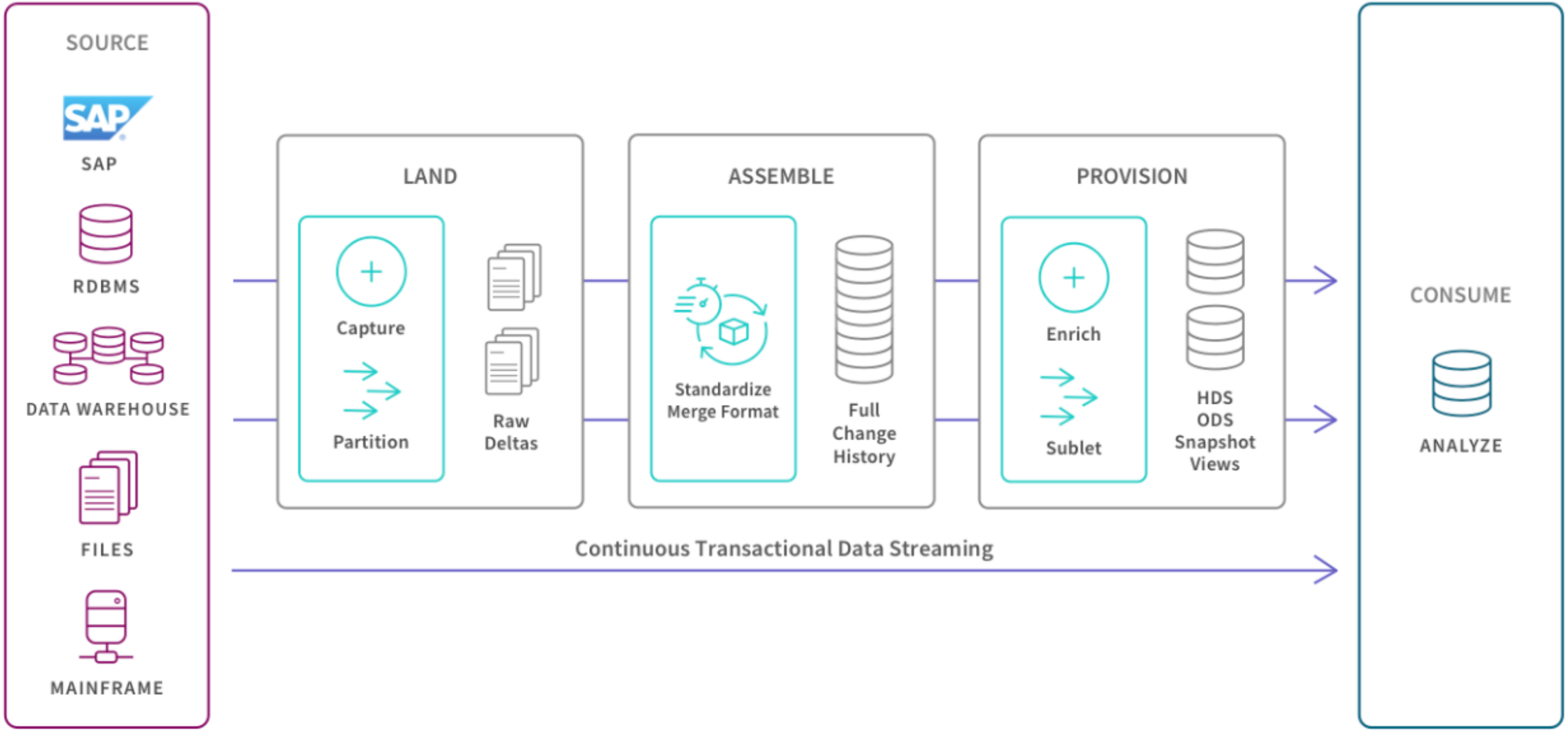
Data stored in a data lake can be used to build data pipelines to make it available for data analytics tools to find insights that inform key business decisions.

DATA LAKE

Benefits

Because the large volumes of data in a data lake are not structured before being stored, skilled data scientists or end-to-end self-service-bi tools can gain access to a broader range of data far faster than in a data warehouse.

DATA LAKE


SOURCE
SAP
RDBMS
DATA WAREHOUSE
FILES
MAINFRAME
LAND
Capture
Partition
Raw Deltas
ASSEMBLE
Standardize Merge Format
Full Change History
PROVISION
Enrich
Sublet
HDS ODS Snapshot Views
Continuous Transactional Data Streaming
CONSUME
ANALYZE

DATA LAKE

Benefits

1. Massive volumes of structured and unstructured data like ERP transactions and call logs can be stored cost effectively.

2. Data is available for use far faster by keeping it in a raw state.

3. A broader range of data can be analyzed in new ways to gain unexpected and previously unavailable insights.

DATA WAREHOUSE

Similar to a data lake, a data warehouse is a repository for business data. However, unlike a data lake, only highly structured and unified data lives in a data warehouse to support specific business intelligence and analytics needs.

Think of it like an actual warehouse, where contents are first processed, then organized into sections and onto shelves (called data marts). Data from a warehouse is ready for use to support historical analysis and reporting to inform decision making across an organization’s lines of business.

DATA WAREHOUSE

A **cloud data warehouse** is a database stored as a managed service in a public cloud and optimized for scalable BI and analytics. It removes the constraint of physical data centers and lets you rapidly grow or shrink your data warehouses to meet changing business budgets and needs.

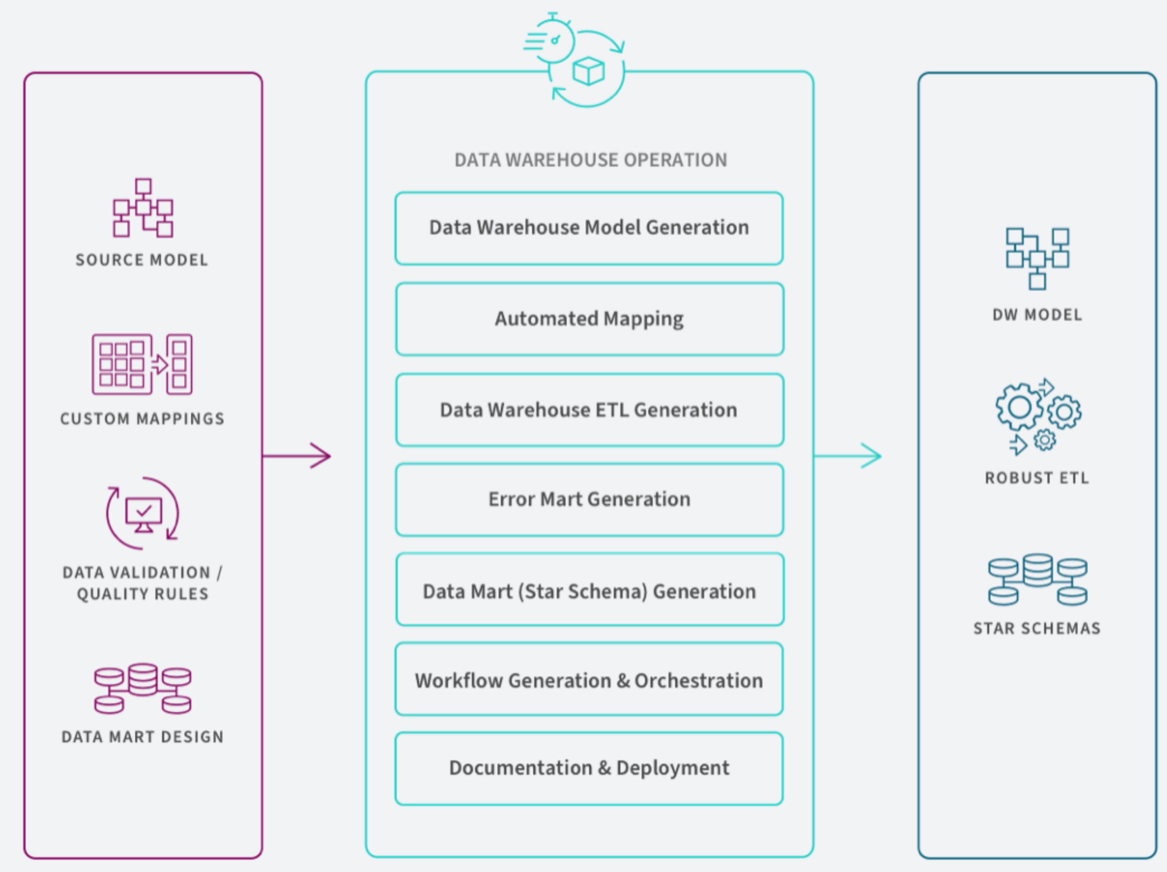
DATA WAREHOUSE

Benefits

A data warehouse offers enormous benefits to organizations, especially as it relates to BI and analytics.

After the initial work of cleansing and processing, data stored in a warehouse serves as a consistent "single source of truth" which is invaluable to business data analysis, collaboration, and better insights.

DATA WAREHOUSE



SOURCE MODEL
CUSTOM MAPPINGS
DATA VALIDATION / QUALITY RULES
DATA MART DESIGN
DATA WAREHOUSE OPERATION
Data Warehouse Model Generation
Automated Mapping
Data Warehouse ETL Generation
Error Mart Generation
Data Mart (Star Schema) Generation
Workflow Generation & Orchestration
Documentation & Deployment
DW MODEL
ROBUST ETL
STAR SCHEMAS

DATA WAREHOUSE

Three major advantages of a data warehouse include:

1.Little or no data prep needed, making it far easier for analysts and business users to access and analyze this data.

2.Accurate, complete data is available more quickly, so businesses can turn information into insight faster.

3.Unified, harmonized data offers a single source of truth, building trust in data insights and decision-making across business lines.

DATA WAREHOUSE -> Data Mart

What is a Data Mart?

A data mart is a structured data repository purpose-built to support the analytical needs of a particular department, line of business, or geographic region within an enterprise.

Data marts are typically created as partitioned segments of an enterprise data warehouse, with each being relevant to a specific subject or department in your organization such as finance or sales.

Data marts help you perform analysis faster given that you’re working with a smaller, more applicable dataset.

DATA WAREHOUSE -> Data Mart

Benefits

More trustworthy data. Data marts create a “single source of truth” regarding a certain subject or department. This gives your teams a collective view of the data and allows them to focus on finding insights, making decisions, and taking action rather than sharing spreadsheets and wondering which data is accurate.

DATA WAREHOUSE -> Data Mart

Benefits

Easier access to data. Since data marts hold a subset of data, you can access the data you need with less effort than dealing with a cluttered data warehouse. Plus, by establishing connections to the appropriate data sources, you can access live data anytime without waiting for IT to perform periodic extracts.

•DATA WAREHOUSE -> Data Mart

Benefits

Faster insights & decisions. The focused nature of a data mart also allows you to more quickly leverage your analytics and business intelligence tools because you’re only working with a relevant, frequently needed data set.

Lower cost. Data marts typically cost far less to set up than establishing a full data warehouse.

•DATA WAREHOUSE -> Data Mart

Benefits

Easier implementation & maintenance. Unlike data warehouses, which require integration with a wide variety of internal and external data sources, data marts only contain data essential to the particular business unit or department. This makes for faster and easier implementation and maintenance because you're serving the needs of a specific business team rather than your entire organization.

•DATA WAREHOUSE -> Data Mart

Benefits

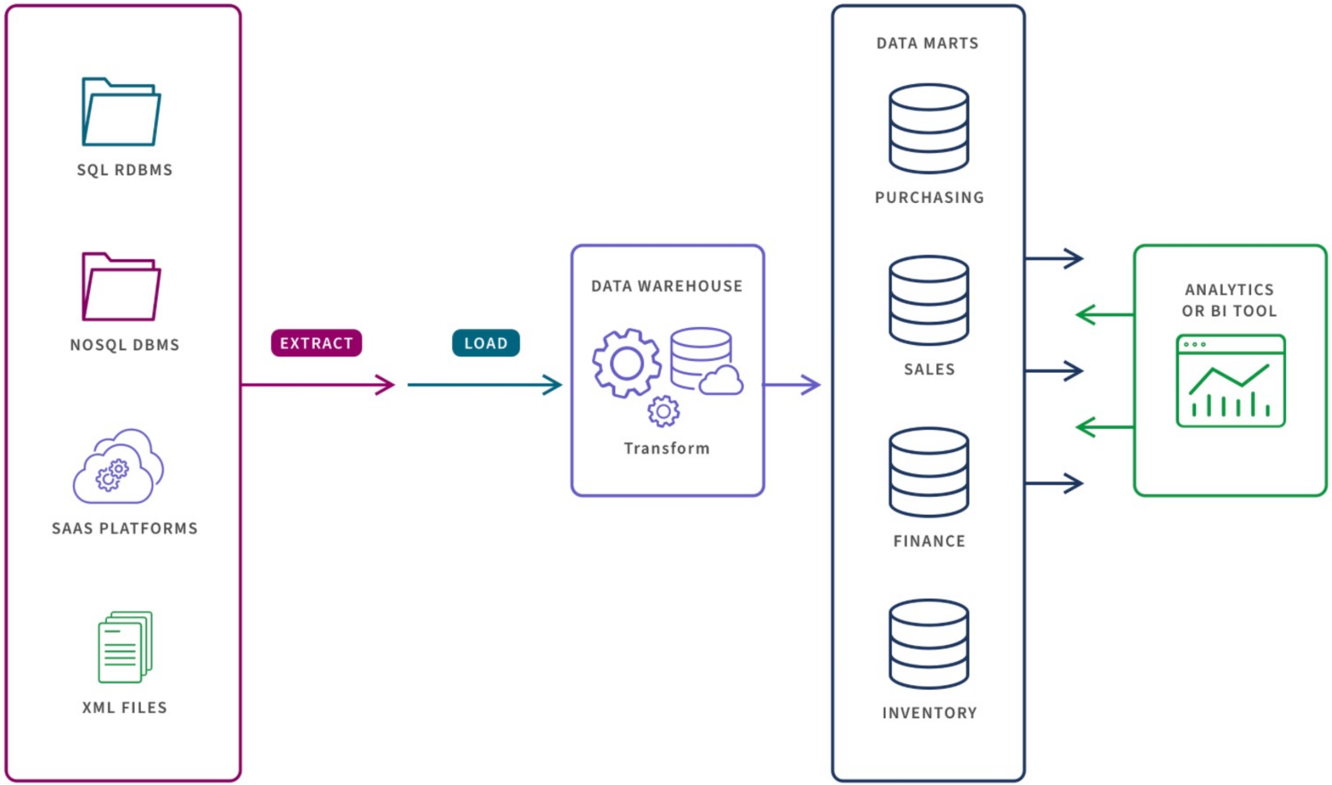
Better support short-term projects. As noted above, you can quickly and cost-effectively establish a data mart, so they are well-suited for short-term data analysis projects such as determining the effectiveness of an advertising campaign.

Better data access control. Data in your mart is partitioned from the broader data warehouse. This gives you the ability to control data access privileges at a granular level.

Data Lake vs Data Warehouse vs Data Mart

The terms data lake, data warehouse, and data mart should not be used interchangeably. They each serve different needs in your organization and here we describe key differences between them.

Data Warehouse vs Data Mart



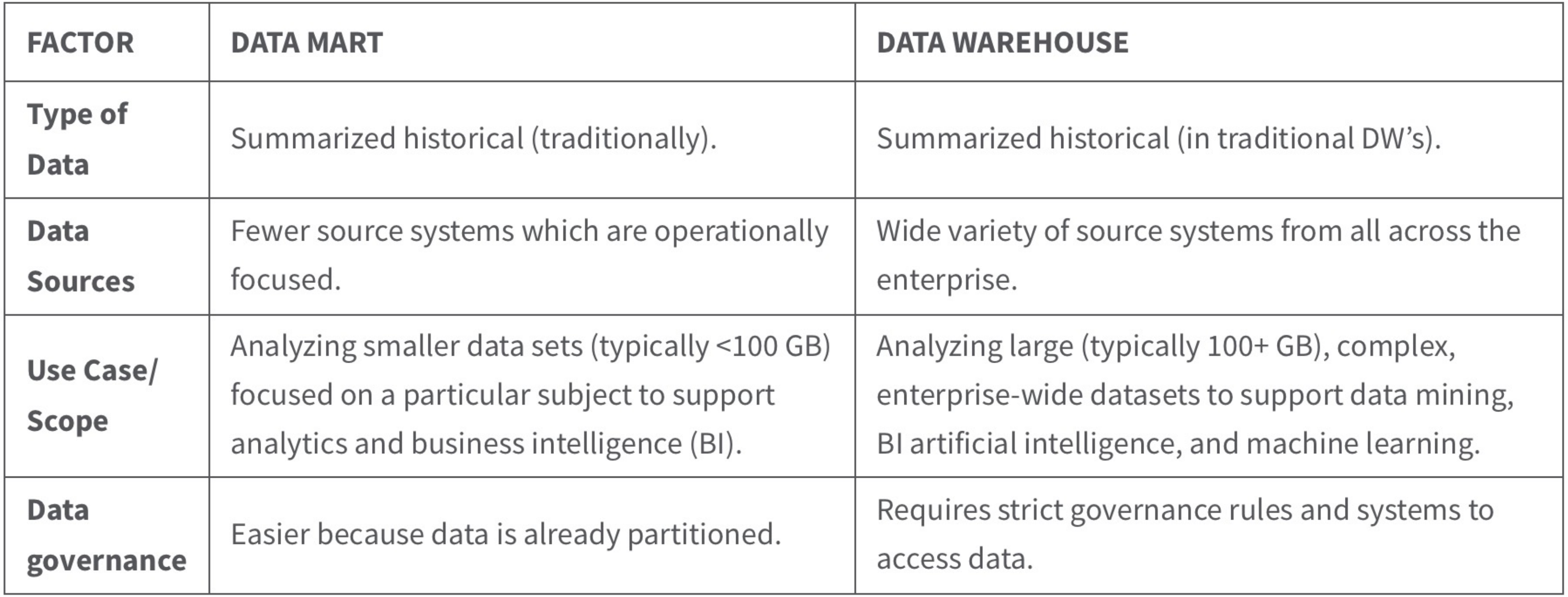
Data Warehouse vs Data Mart

Marts and warehouses are both read-only, structured data repositories of transactional data. But they differ in the scope of data which is stored.

Data warehouses aggregate large volumes of data from multiple sources such as transactional applications and application log files into a single repository of highly structured and unified historical data.

Data marts consist of a subset of this warehouse data which is relevant to a specific subject or department in your organization. As shown below, they're added between the warehouse and the analytics tools.

Data Warehouse vs Data Mart



FACTOR	DATA MART	DATA WAREHOUSE
Type of Data	Summarized historical (traditionally).	Summarized historical (in traditional DW's).
Data Sources	Fewer source systems which are operationally focused.	Wide variety of source systems from all across the enterprise.
Use Case/ Scope	Analyzing smaller data sets (typically <100 GB) focused on a particular subject to support analytics and business intelligence (BI).	Analyzing large (typically 100+ GB), complex, enterprise-wide datasets to support data mining, BI artificial intelligence, and machine learning.
Data governance	Easier because data is already partitioned.	Requires strict governance rules and systems to access data.

Data mart vs data lake

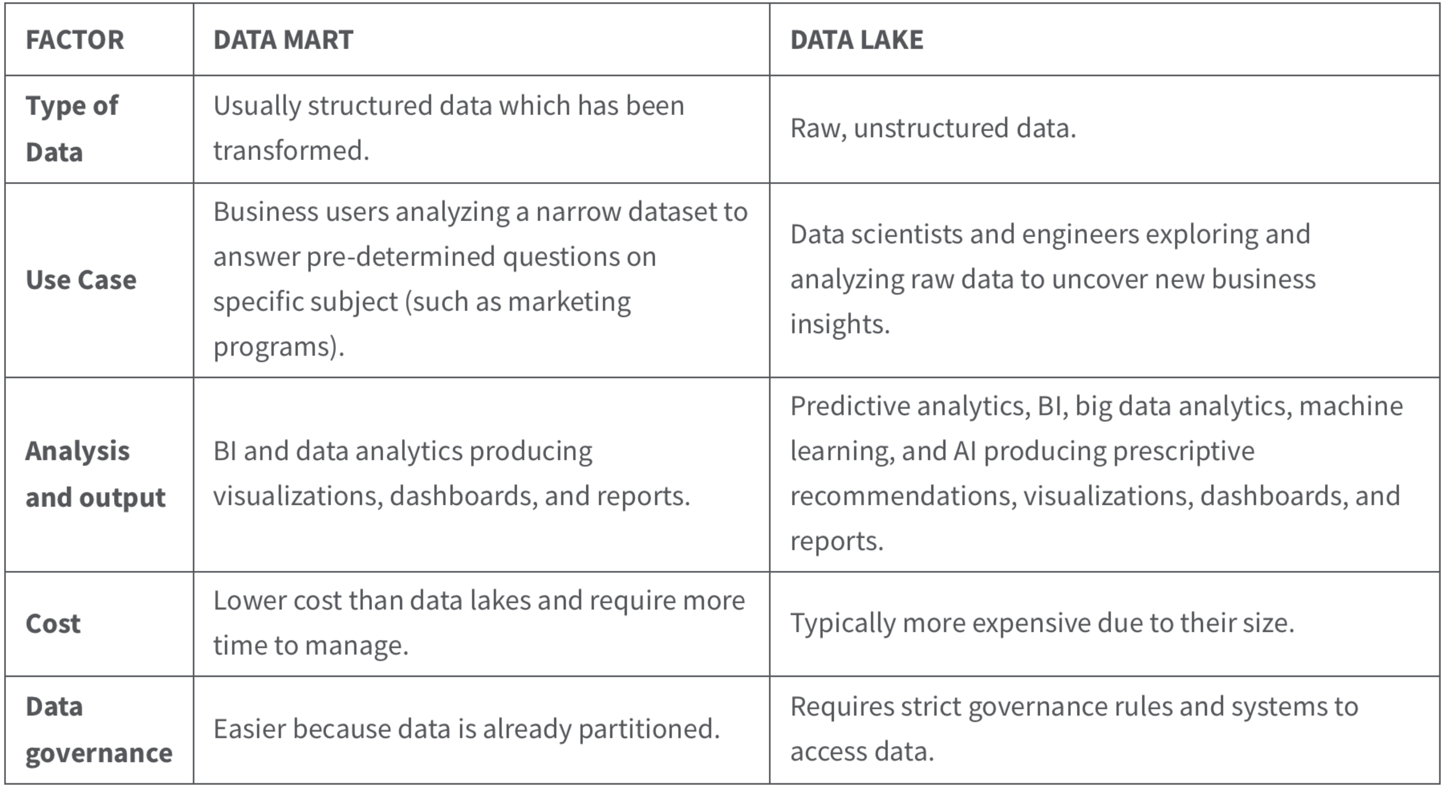
The main difference in data mart vs data lake is the type and volume of data stored.

Marts typically hold smaller amounts of structured data which has been transformed whereas data lakes consist of massive amounts of raw, unstructured data.

Another key difference is that the data in marts has been selected to serve a well-defined need whereas the purpose of data in data lakes has not necessarily been defined.

Many organizations use both systems to accommodate their range of storage needs.

Data mart vs data lake



FACTOR	DATA MART	DATA LAKE
Type of Data	Usually structured data which has been transformed.	Raw, unstructured data.
Use Case	Business users analyzing a narrow dataset to answer pre-determined questions on specific subject (such as marketing programs).	Data scientists and engineers exploring and analyzing raw data to uncover new business insights.
Analysis and output	BI and data analytics producing visualizations, dashboards, and reports.	Predictive analytics, BI, big data analytics, machine learning, and AI producing prescriptive recommendations, visualizations, dashboards, and reports.
Cost	Lower cost than data lakes and require more time to manage.	Typically more expensive due to their size.
Data governance	Easier because data is already partitioned.	Requires strict governance rules and systems to access data.

Data lake

What is a Data Lake?

A data lake is a data storage strategy whereby a centralized repository holds all of your organization's structured and unstructured data. It employs a flat architecture which allows you to store raw data at any scale without the need to structure it first.

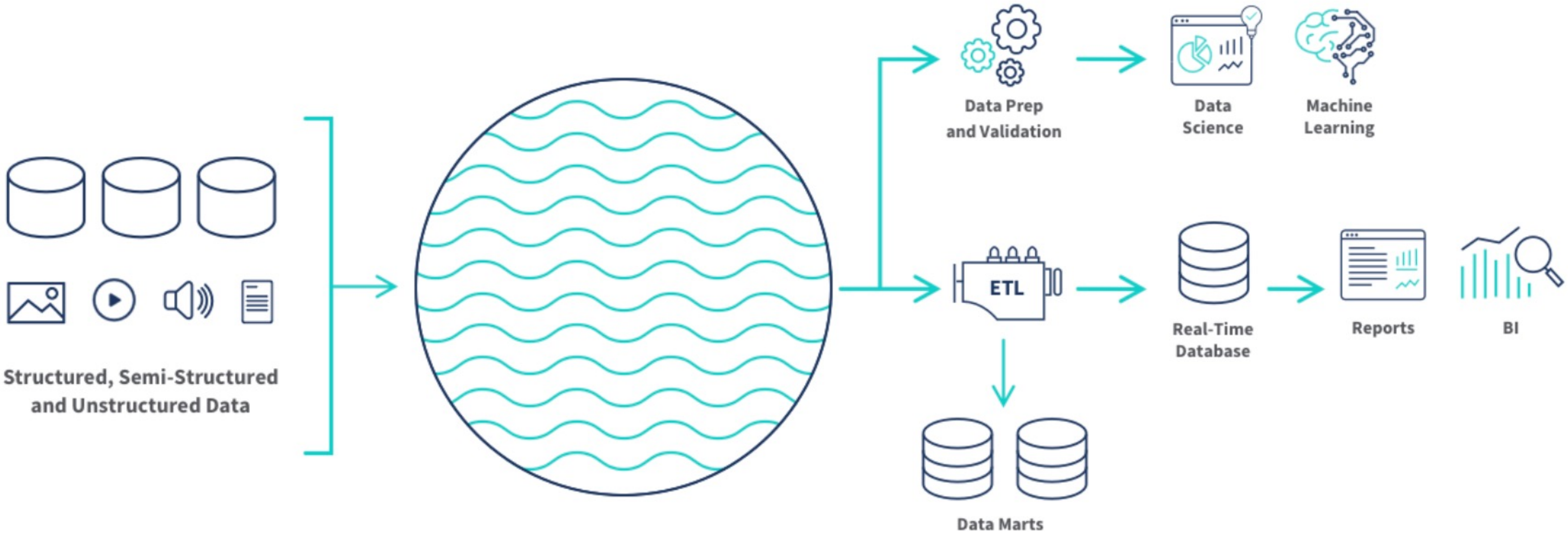
Instead of pre-defining the schema and data requirements, you use tools to assign unique identifiers and tags to data elements so that only a subset of relevant data is queried to analyze a given business question. **This analysis can include real-time analytics, big data analytics, machine learning, dashboards and data visualizations** to help you uncover insights that lead to better decisions.

Data lake

Architecture

There are a number of different tools you can use to build and manage your data lake, such as Azure, Amazon S3 and **Hadoop**.
Therefore, the detailed physical structure of your system will depend on which tool you select. Still, you can see below how it can fit into your overall data integration strategy.

Data lake


Data Prep
and Validation
Data
Science
Machine
Learning
Structured, Semi-Structured
and Unstructured Data
ETL
Real-Time
Database
Reports
BI
Data Marts

Data lake

Data teams can build ETL data pipelines and schema-on-read transformations to make data stored in a data lake available for data science and machine learning and for analytics and business intelligence tools.

As we discuss below, managed data lake creation tools help you overcome the limitations of slow, hand-coded scripts and scarce engineering resources.

Data lake

Benefits

Because the large volumes of data are not structured before being stored, skilled data scientists or end-to-end self-service BI tools can provide you access to a broader range of data far faster than in a data warehouse.

Data lake

Benefits

1.Agility. You can easily configure queries, data models, or applications without the need for pre-planning. In addition to SQL queries, the data lake strategy is well suited to support real-time analytics, **big data analytics**, and **machine learning**.

2.Real-time. You can import data in its original format from multiple sources in real-time. This allows you to perform real-time analytics and machine learning and trigger actions in other applications.

Data lake

Benefits

3. Scale. Because of its lack of structure, data lakes can handle massive volumes of structured and unstructured data such as ERP transactions and call logs.
4. **Speed.** Keeping data in a raw state also makes it available for use far faster since you don’t have to perform time-intensive tasks such as transforming the data and developing schemas until you define the business question(s) that need to be addressed.

Data lake

Benefits

5. Better insights. You can gain unexpected and previously unavailable insights by analyzing a broader range of data in new ways.
6. Cost savings. Data lakes have lower operational costs since they’re less time-consuming to manage. Also, storage costs are less expensive than traditional data warehouses because most of the tools you use to manage them are open source and run on low-cost hardware.

	Data Lake	Data Warehouse
1. Processing	ELT (Extract, Load, Transform). Data is extracted from its source(s), loaded into the lake, and is structured and transformed only when needed.	ETL (Extract, Transform, Load). Data is extracted from its source(s) and then scrubbed and structured before loading into a repository.
2. Storage	Contains all of your organization's data in both a structured and raw, unstructured form.	Contains only structured data which has been cleaned and processed based on predefined business needs.
3. Schema	Schema is defined after the data is stored. This makes the process of capturing and storing the data faster.	You have to define schema before the data is stored. This lengthens the time it takes to process the data, but once complete, the data is available for immediate use.
4. Users	Data is typically used by data scientists and engineers who prefer to study data in its raw form.	Data is typically accessed by managers and business-end users looking to answer pre-determined questions.
5. Analysis	Predictive analytics, machine learning, data visualization, dashboards, BI, big data analytics.	Data visualization, dashboards, BI, data analytics.
6. Expense	Storage costs are typically lower than a data warehouse. Plus, operational costs are lower since data lakes take less time to manage.	Data warehouses cost more and also require more time to manage, resulting in additional operational costs.

ETL

What is ETL?

ETL stands for **“Extract, Transform, and Load”** and describes the set of processes to extract data from one system, transform it, and load it into a target repository. An ETL pipeline is a traditional type of data pipeline for cleaning, enriching, and transforming data from a variety of sources before integrating it for use in data analytics, business intelligence and data science.

ETL

Benefits

Using an ETL pipeline to transform raw data to match the target system, allows for systematic and accurate data analysis to take place in the target repository.

ETL

Benefits

- More stable and faster data analysis on a single, pre-defined use case. This is because the data set has already been structured and transformed.
- Easier compliance with data protections standards . This is because users can omit any sensitive data prior to loading in the target system.
- Identify and capture changes made to a database via the change data capture (CDC) process or technology. These changes can then be applied to another data repository or made available in a format consumable by ETL.

ETL vs ELT

The key difference between the two processes is when the transformation of the data occurs. Many organizations use both processes to cover their wide range of data pipeline needs.

The ETL process is most appropriate for small data sets which require complex transformations. For larger, unstructured data sets and when timeliness is important, the ELT process is more appropriate.

ETL vs ELT

Extract > Transform > Load (ETL)

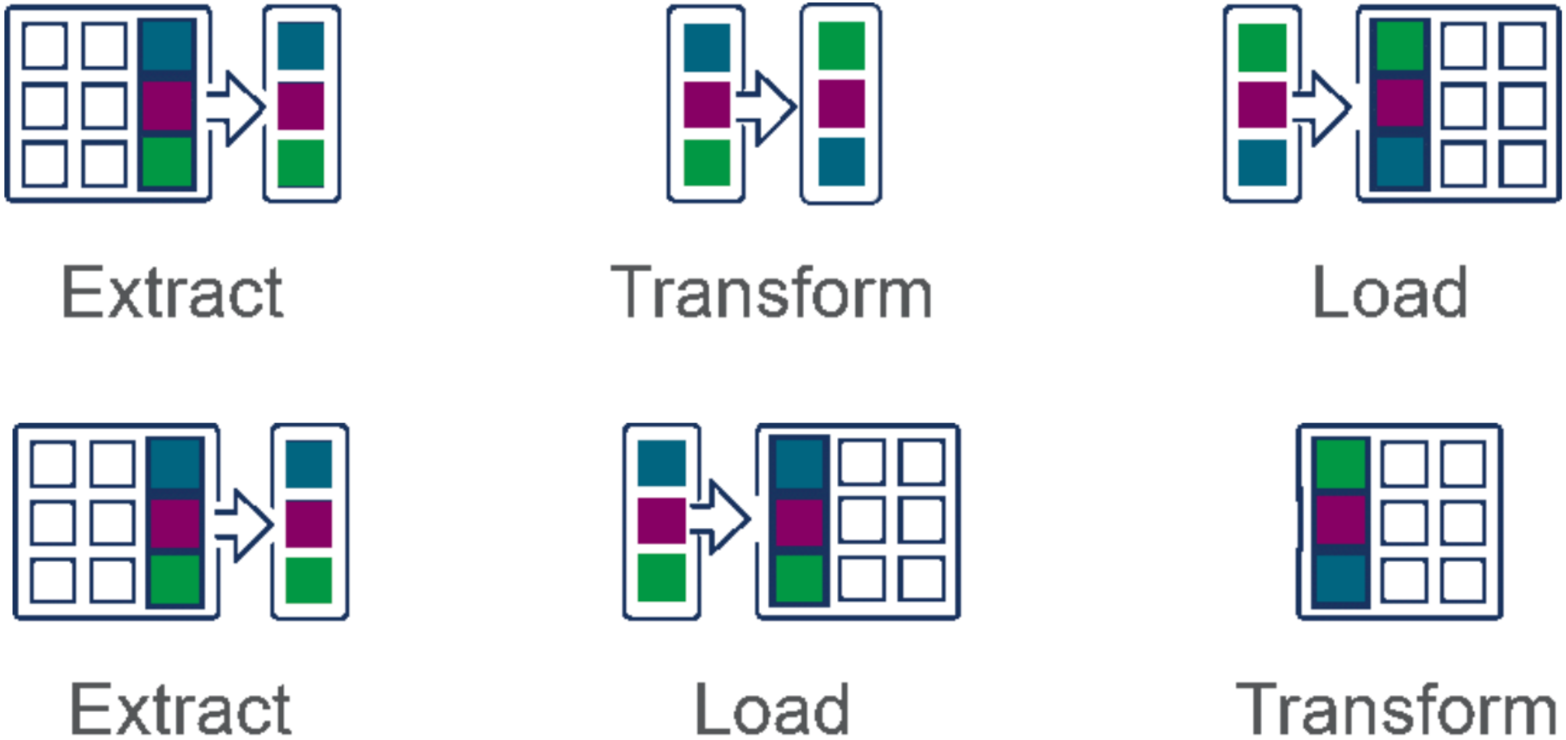
In the ETL process, transformation is performed in a staging area outside of the data warehouse and before loading it into the data warehouse.
The entire data set must be transformed before loading, so transforming large data sets can take a lot of time up front. The benefit is that analysis can take place immediately once the data is loaded. This is why this process is appropriate for small data sets which require complex transformations.

ETL vs ELT

Extract > Load > Transform (ELT)

In the ELT process, data transformation is performed on an as-needed basis within the target system.
This means that the ELT process takes less time. But if there is not sufficient processing power in the cloud solution, transformation can slow down the querying and analysis processes. This is why the ELT process is more appropriate for larger, structured and unstructured data sets and when timeliness is important.

ETL vs ELT



Extract Transform Load

Extract Load Transform

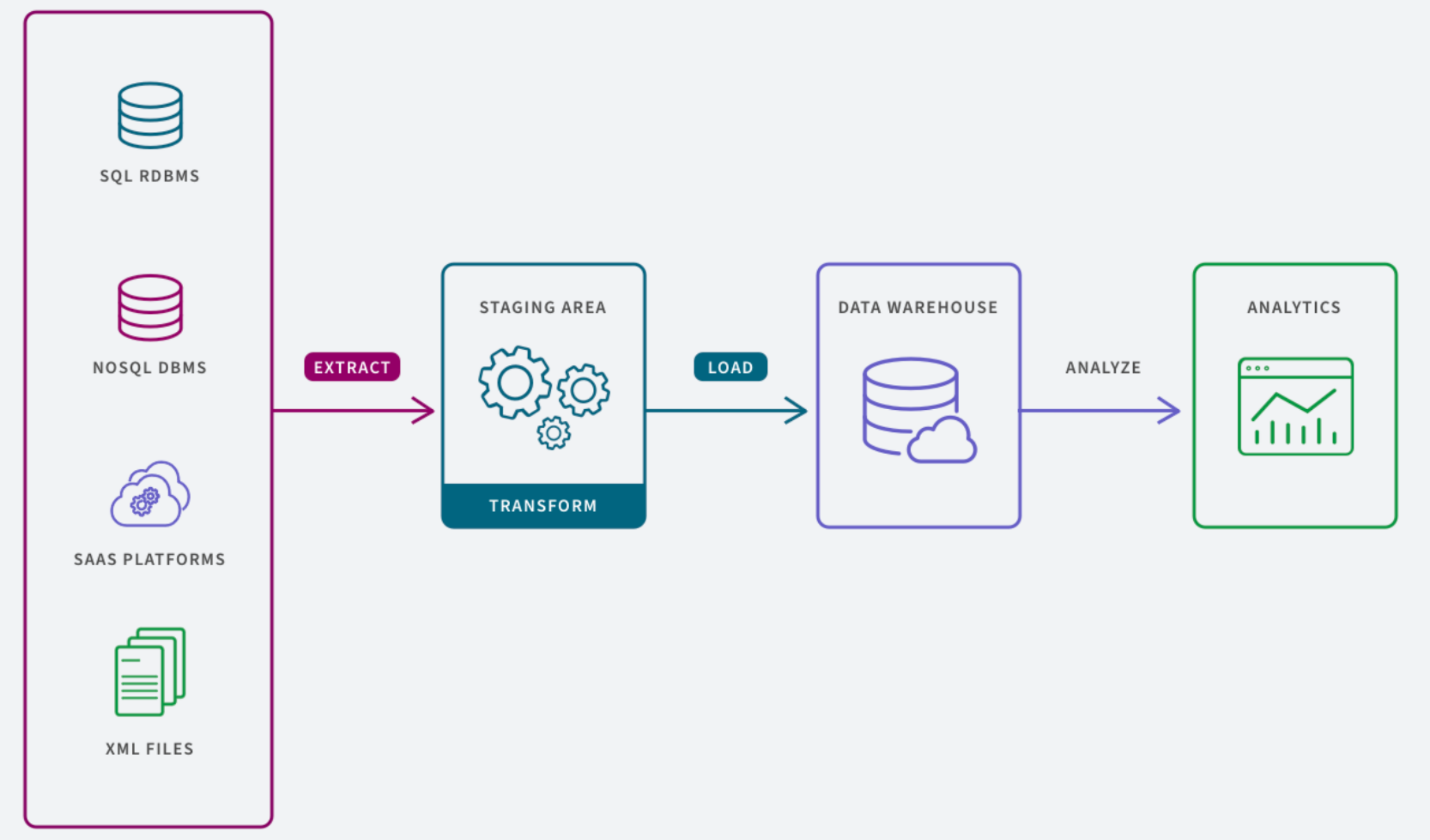
ETL

- **Extract:** the process of pulling data from a source such as an SQL or NoSQL database, an XML file or a cloud platform holding data for systems such as marketing tools, CRM systems, or transactional systems.

- **Transform:** the process of converting the format or structure of the data set to match the target system.

- **Load:** the process of placing the data set into the target system which can be a database, data warehouse, an application, such as CRM platform or a cloud data warehouse, data lake from providers such as Snowflake, Amazon RedShift, and Google BigQuery.

ETL


SQL RDBMS
NOSQL DBMS
SAAS PLATFORMS
XML FILES
EXTRACT
STAGING AREA
TRANSFORM
LOAD
DATA WAREHOUSE
ANALYZE
ANALYTICS

ETL

ETL Use Cases

The ETL process helps eliminate data errors, bottlenecks, and latency to provide for a smooth flow of data from one system to the other. Here are some of the key use cases:

- Migrating data from a legacy system to a new repository.
- Centralizing data sources to gain a consolidated version of the data.
- Enriching data in one system with data from another system.
- Providing a stable dataset for data analytics tools to quickly access a single, pre-defined analytics use case given that the data set has already been structured and transformed.
- Complying with data protection standards.

Mestrado em Engenharia Informática
KE

2022/2023