Big Data and Data Mining

Labbi — Oct 2025 (Lecture)

HI, I AM LABBI

Labbi Karmacharya (she/her)

Data Scientist, Big data and Data Minining Lecturer

Contact:

www.labbi.com.np

linkedin.com/in/labbi-karmacharya

Recap:

- Overview of the Module
- The 5Vs of Big Data:
- Database Systems:
- Hands-On SQL Server

Expectations and Goals (Instructor Expecations)

- **Engagement and Curiosity:** Participate in class discussions.
- **Commitment:** 76% of independent guided study.
- **Collaboration:** Work together in groups when assigned
- **Diligence**: Regularly check MST
- **Patience:** Understand that I might not know answer to all
- **Respect** and Professionalism

SQL Overview

- SQL (Structured Query Language) is the standard language for managing and manipulating databases.
- While the core syntax of SQL is fairly consistent, different database systems have their extensions and specific features.

SQL Dialects

- Despite its universal application, SQL manifests in different dialects, each tailored to the specificities of individual database management systems (DBMS)
- Prominent SQL dialects: **MySQL, PostgreSQL, SQL Server, Oracle, and SQLite**, offering insights into their unique features, applications, and syntax variations.

SQL Dialects

- SQL dialects are essentially variations of the SQL language, each incorporating unique syntax, functions, and features that align with the peculiarities of a specific database system.

SQL Dialects

- MySQL: Widely used for web applications due to its simplicity and speed. Good replication features and a broad community.
- PostgreSQL: Supports advanced data types and a wide range of SQL functions.
- SQL Server: Deep integration with other Microsoft products, excellent transactional support, comprehensive business intelligence and analytics tools.

Medium article on the same

Why SQL Server?

Medium article on the same

WHY SQL Server?

SQL Server is one of the leading database management systems used globally.

- Integration with Microsoft Products:
- Comprehensive Tools for BI and Analytics:
- Strong Transactional Support:
- Scalability and Performance:

SQL Command Types:

DDL

- CREATE
- DROP
- ALTER
- TRUNCATE
- RENAME

DQL

- SELECT

DML

- INSERT
- UPDATE
- DELETE
- MERGE

DCL

- GRANT
- REVOKE

TCL

- COMMIT
- ROLLBACK
- SAVEPOINT

Self work: Learn the syntax of the commands along with its use..

Example:

ALTER

- **Syntax**: ALTER TABLE table_name ADD column_name datatype;
- **Use**: Modifies an existing table, such as adding a new column.
- **Example**: ALTER TABLE Employees ADD Email VARCHAR(100);

Complex SQL queries

- Multi-table Joins:
- Subqueries:
- Conditional Expressions and Case Statements:
- Window Functions:
- Recursive Queries:

Multi-table Joins

Queries that retrieve data by joining multiple tables. These can include various types of joins like INNER, LEFT, RIGHT, and FULL OUTER JOIN, often combined in a single query.

- Example Usage: Creating a comprehensive report that pulls together customer data, order details, and shipment status from separate tables.

Subqueries

A subquery is a query nested inside another query. A correlated subquery is a type of subquery that depends on information from the outer query.

- Example Usage: Finding products that are priced above the average price of products in their category, where the average is calculated on the fly for each category.

Conditional Expressions and Case Statements:

These allow SQL queries to execute different logic paths based on conditions, similar to if-else statements in programming.

- Example Usage: Adjusting the results of SQL queries dynamically based on certain criteria, like applying discounts to orders based on order date or quantity.

Window Functions:

Functions that perform calculations across sets of rows related to the current row, allowing for running totals, moving averages, and ranking without needing to group the entire query.

- Example Usage: Example Usage: Calculating a running total of sales, ranking users based on their activity, or segmenting transaction data into quantiles.

Recursive Queries:

Queries that refer to themselves to retrieve hierarchical or iterative data. Commonly implemented with Common Table Expressions (CTEs).

Exploring the SQL Server Suite

SQL Server isn't just a database management system; it's a comprehensive suite designed for data integration, reporting, and analysis

Exploring the SQL Server Suite

SQL Server isn't just a database management system; it's a comprehensive suite designed for data integration, reporting, and analysis

One of the Modules Aims:

- provide students with an understanding of key data mining concepts, techniques and process for **business Intelligence.**

Have you heard about Business Intelligence?

BI

RELEVANT & RELIABLE INFORMATION
TO THE RIGHT PEOPLE
AT THE RIGHT TIME

With the goal of achieving

BETTER
DECISIONS
FASTER

To do this, BI requires

METHODS & PROGRAMS
COLLECT & STRUCTURE DATA

ing INNOVATE NEPAL GROUP

Converts into

INFORMATION

islington college (इस्लिङ्टन कलेज)

informatics college • pokhara

ITAHARI INTERNATIONAL COLLEGE

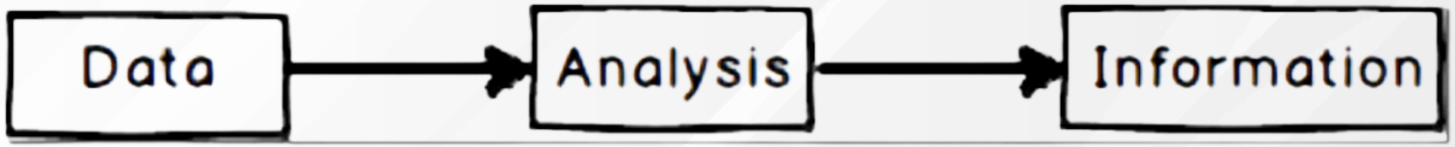
In simple words, BI is

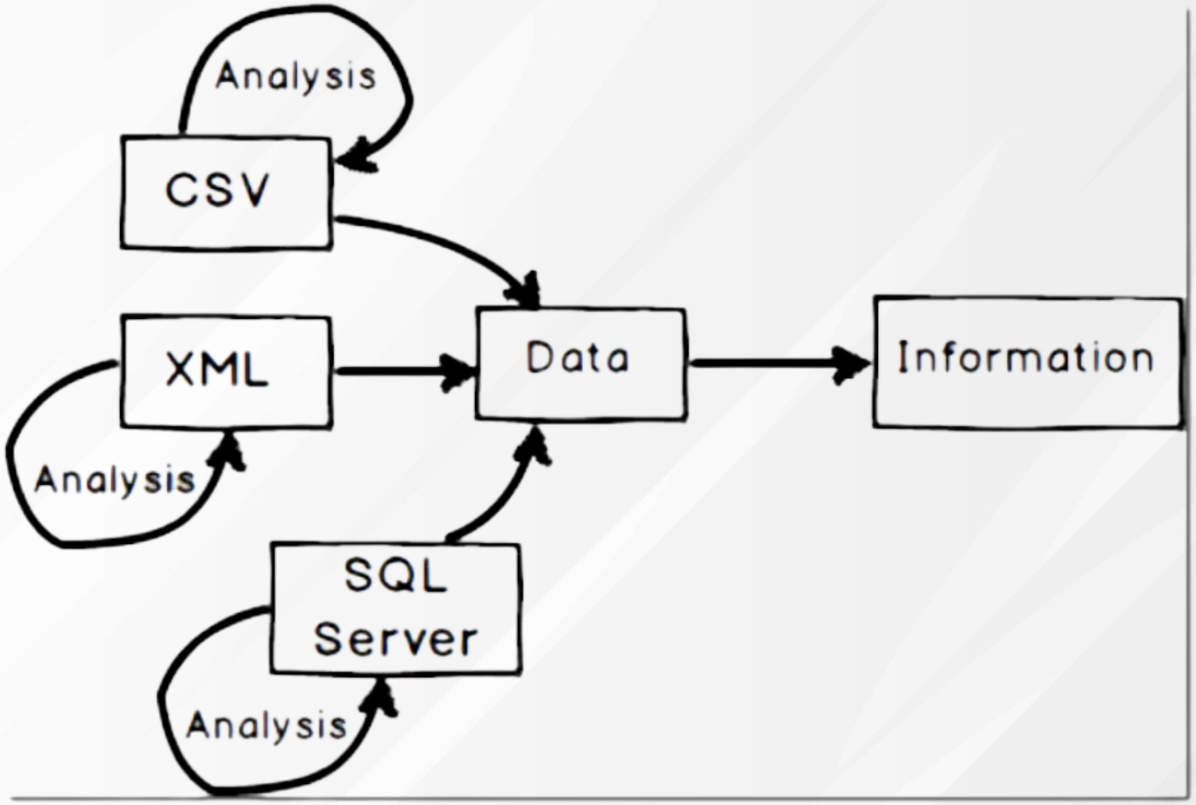
Data consists of raw facts and figures—unprocessed, unorganized elements that alone might not carry meaning or context.

Information is data that has been processed, organized, or structured to provide context and meaning.

Data can be in different format

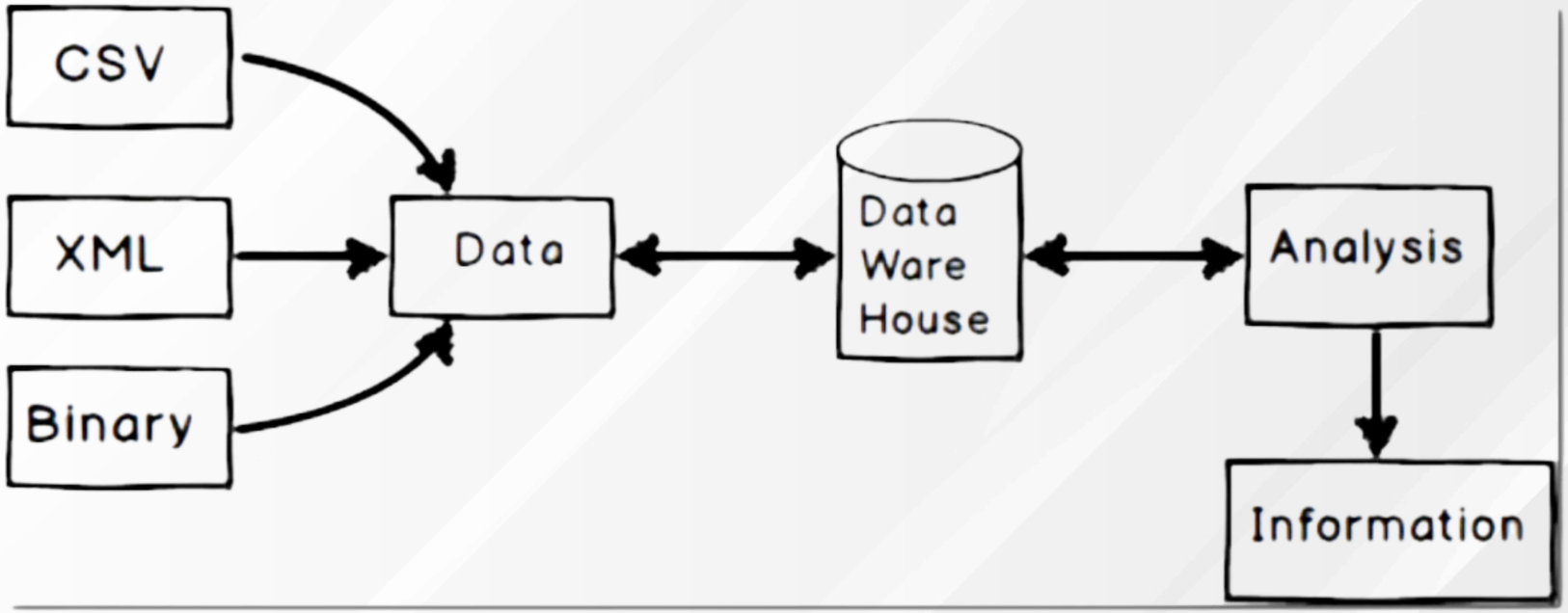
Now this journey, involves



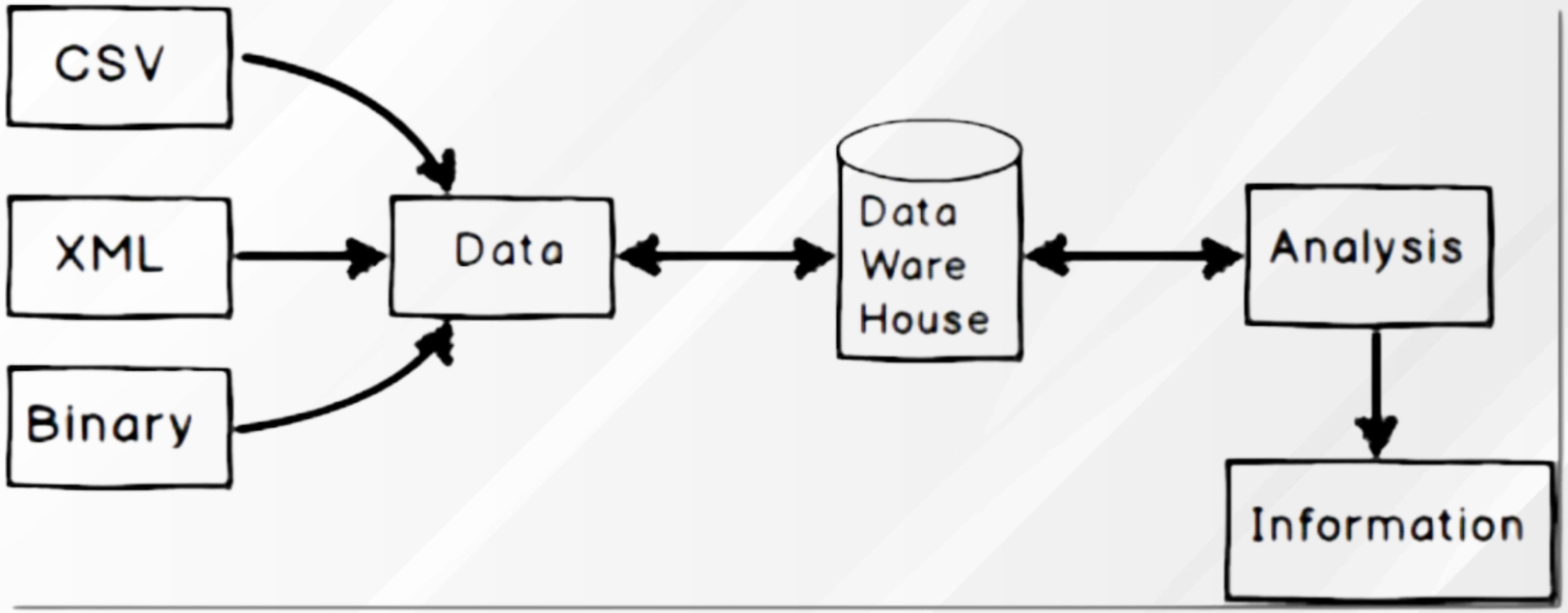


Running the analysis in these individually is tedious and illogical

Hence,



Hence,



One of the Modules Aims:

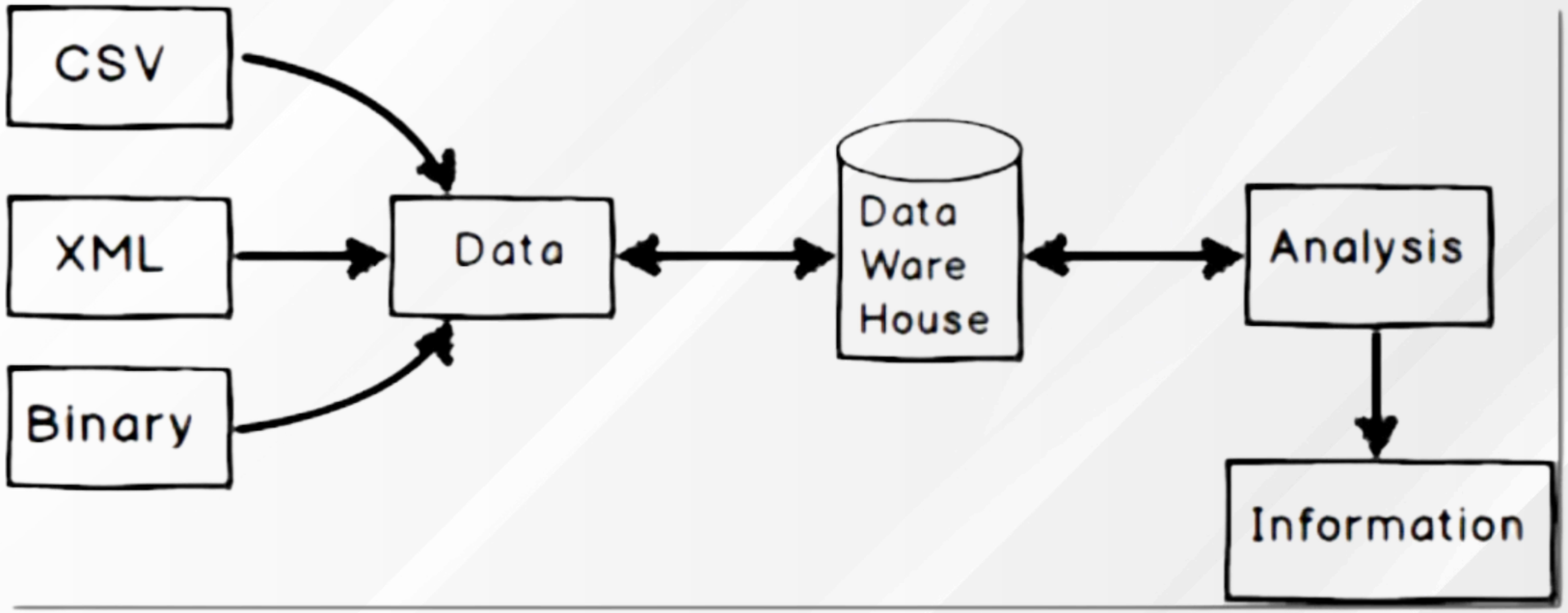
- provide students with an understanding of key data mining concepts, techniques and process for **business Intelligence.**

What is a Data Warehouse?

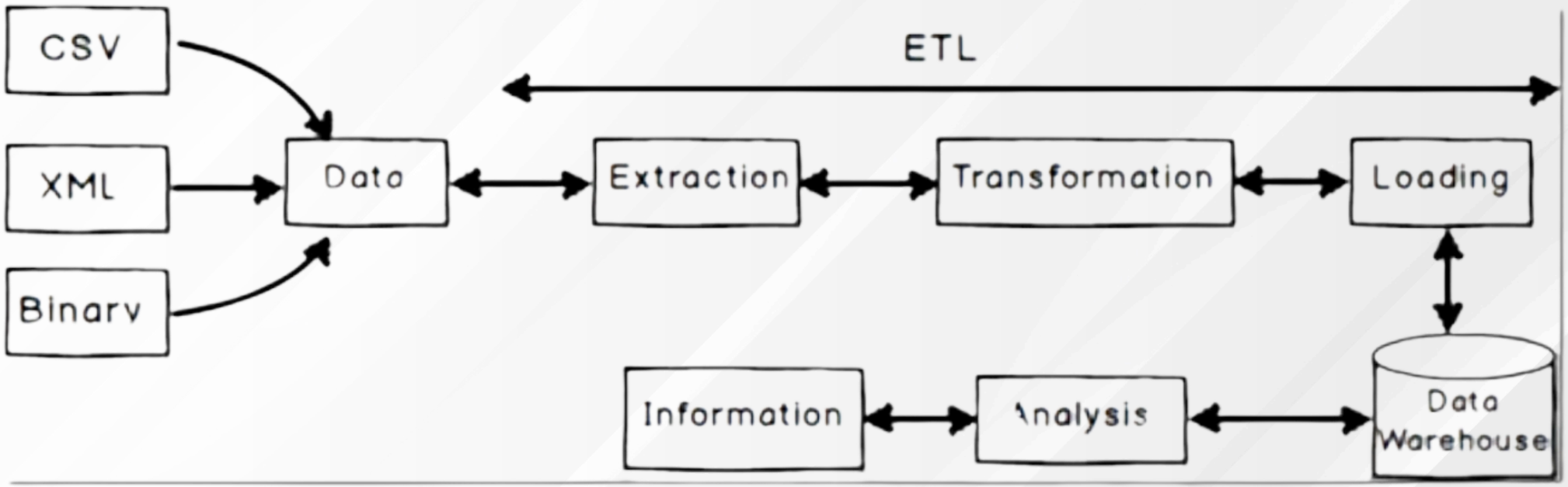
A data warehouse is a system used for reporting and data analysis, acting as a central repository of integrated data from one or more disparate sources.

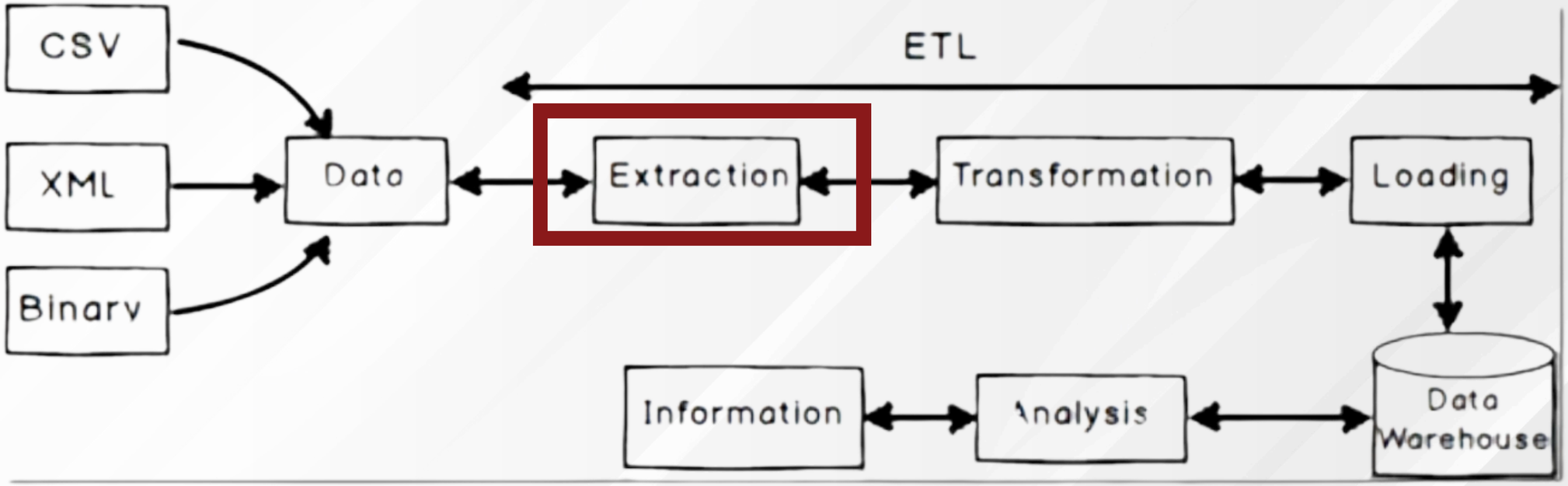
Examples of Data warehouse tools: Amazon Redshift, Microsoft SQL Server Analysis Services, Snowflake, Google BigQuery, Teradata

Hence,



The journey from Data to Dataware house is not easy...

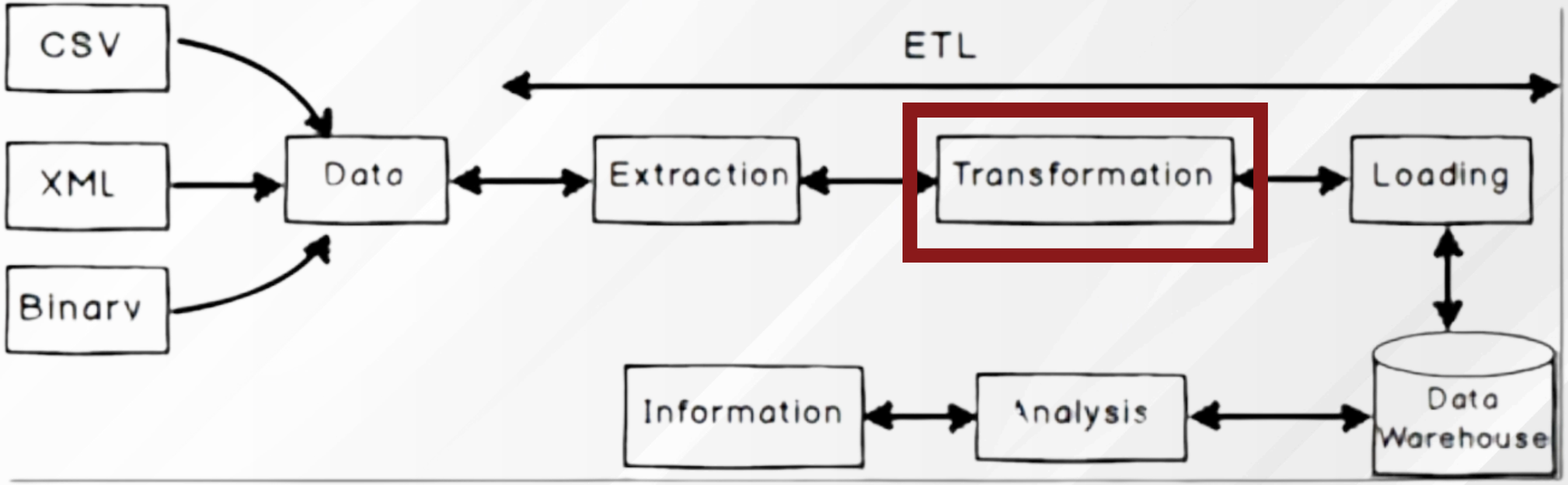

CSV
XML
Binary
Data
ETL
Extraction
Transformation
Loading
Information
Analysis
Data
Warehouse


CSV
XML
Binary
Data
ETL
Extraction
Transformation
Loading
Information
Analysis
Data Warehouse

Extract

The first step in the ETL process involves pulling data from various source systems. These sources might include databases, CRM systems, flat files, web services, and more.

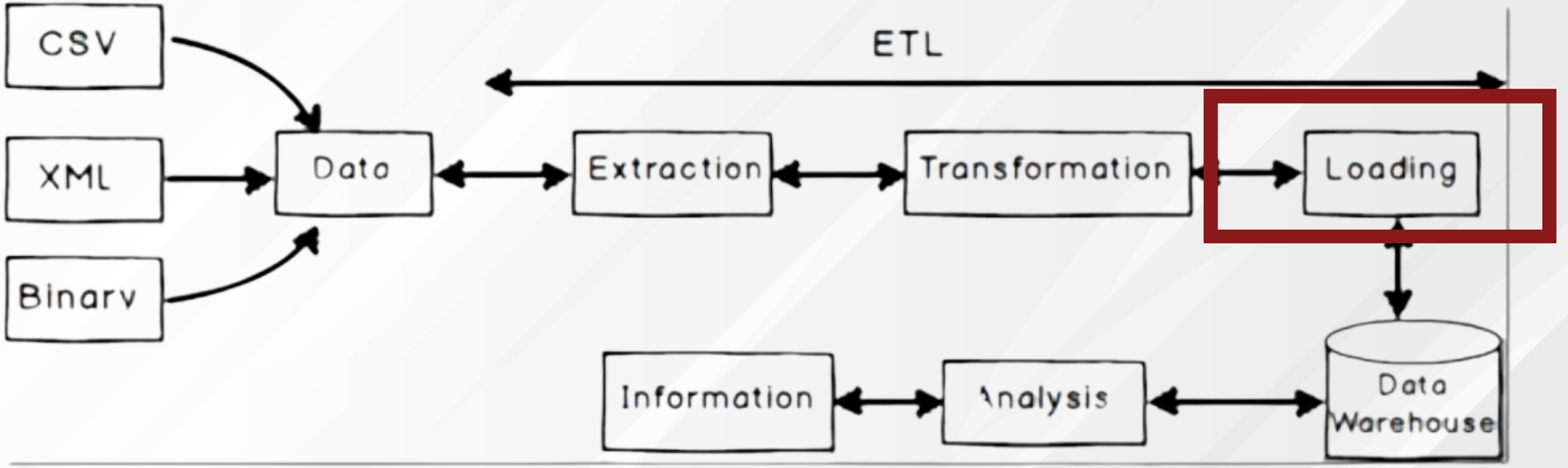
The main challenge here is dealing with different data formats and ensuring the integrity of data during extraction.


CSV
XML
Binary
Data
ETL
Extraction
Transformation
Loading
Information
Analysis
Data Warehouse

Transform

Once data is extracted, it must be cleansed, formatted, and modified to fit the needs of the data warehouse schema. This might include converting data types, normalizing text fields, deduplicating records, and applying business rules.

Transformation is often the most complex part of ETL, as it requires robust processing to ensure the data is accurate and useful.


CSV
XML
Binary
Data
ETL
Extraction
Transformation
Loading
Information
Analysis
Data Warehouse

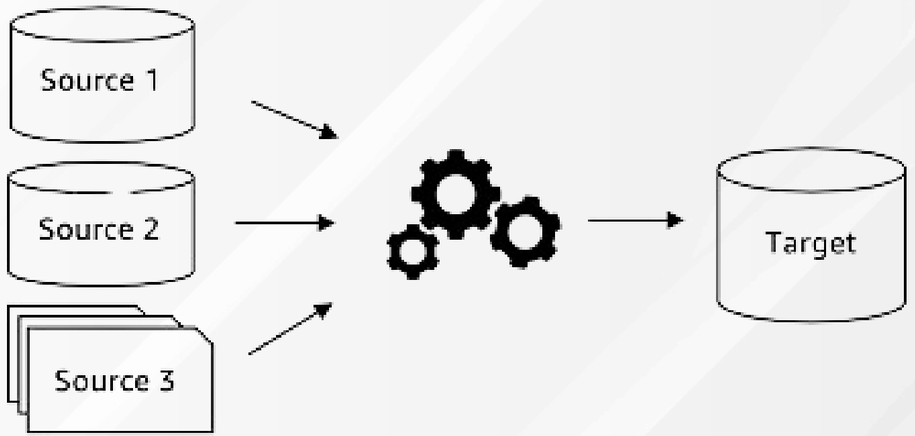
Load

The final step is loading the transformed data into the data warehouse. Depending on the requirements, this could be a simple bulk load or a more complex incremental load where only new or changed data is added.

Ensuring the load process does not disrupt the performance of the data warehouse and that data integrity is maintained.

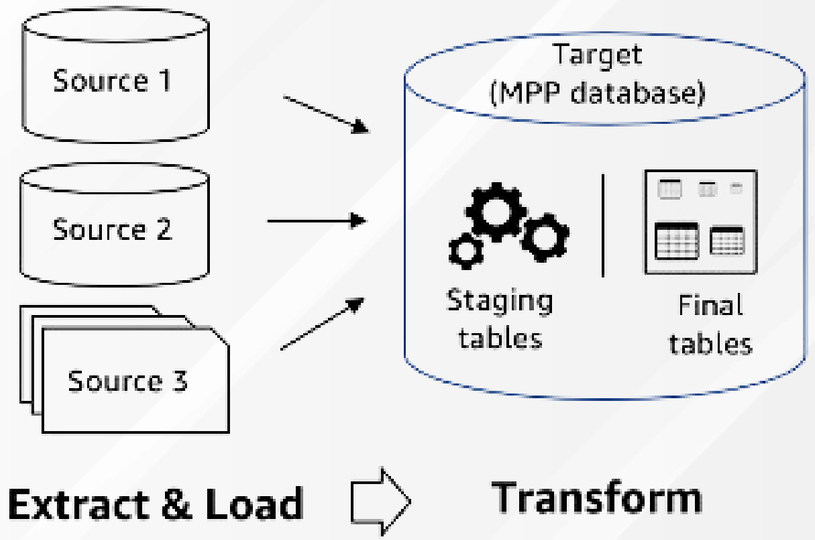
ETL Process

- You extract raw data from various sources
- You use a secondary processing server to transform that data
- You load that data into a target database



Another process: ELT

- You extract raw data from various sources
- You load it in its natural state into a data warehouse or data lake
- You transform it as needed while in the target system



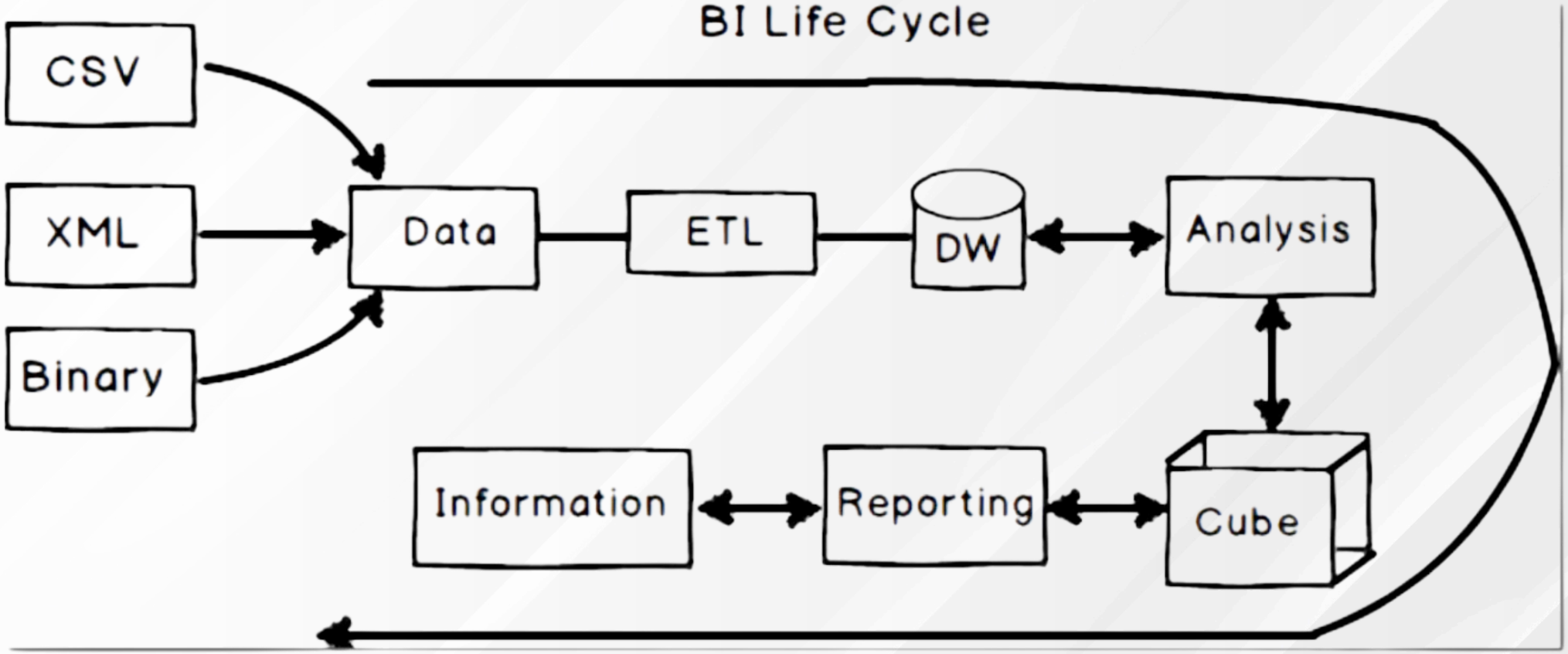
E -> L -> T

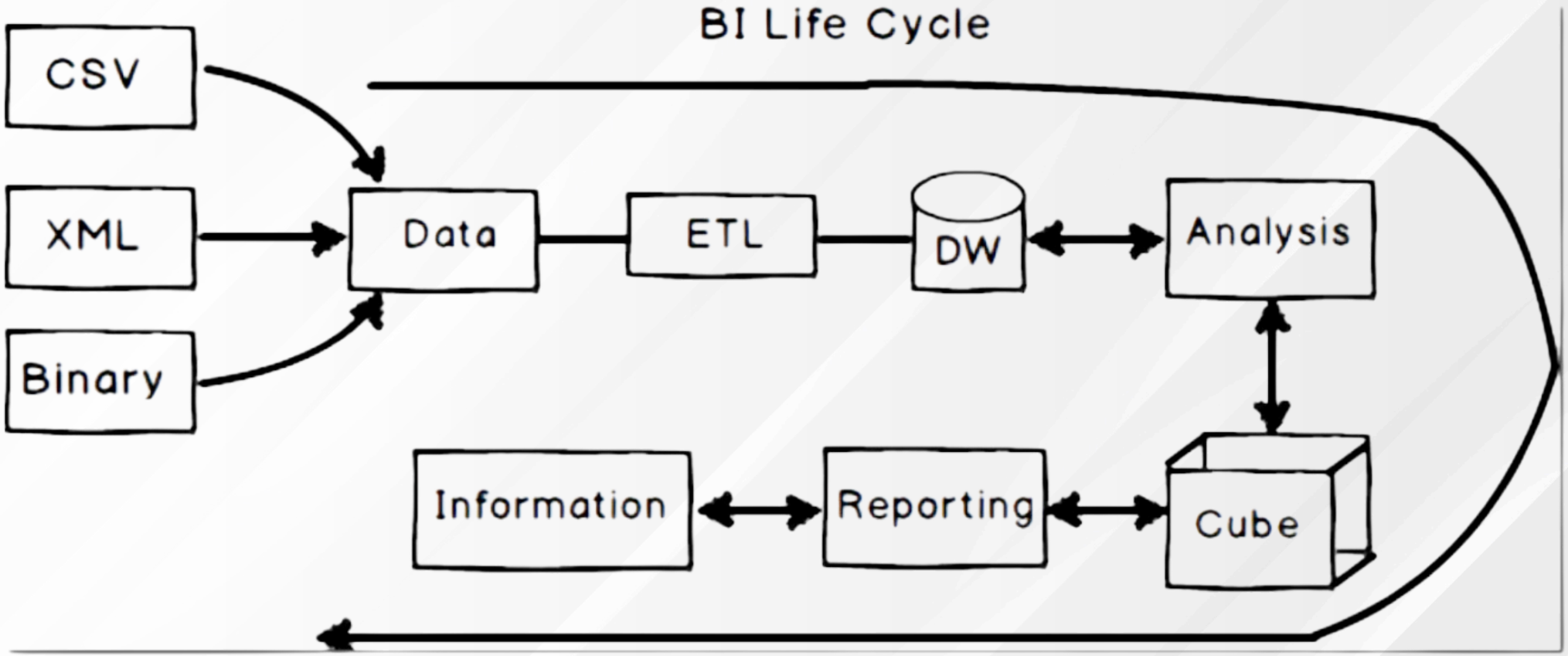
ETL VS ELT

- ETL : Developed in the 1970s as data warehousing started to gain traction.
- Use Cases: Best for systems where data quality and consistency are critical, and the computational resources are limited in the target environment

ETL VS ELT

- **ELT:** Gained popularity with the rise of big data technologies and cloud computing, which provide scalable compute resources.
- Use Cases: Ideal for environments with massive data sets where scalability and speed are priorities, such as big data applications in cloud environments.


BI Life Cycle
CSV
XML
Binary
Data
ETL
DW
Analysis
Information
Reporting
Cube

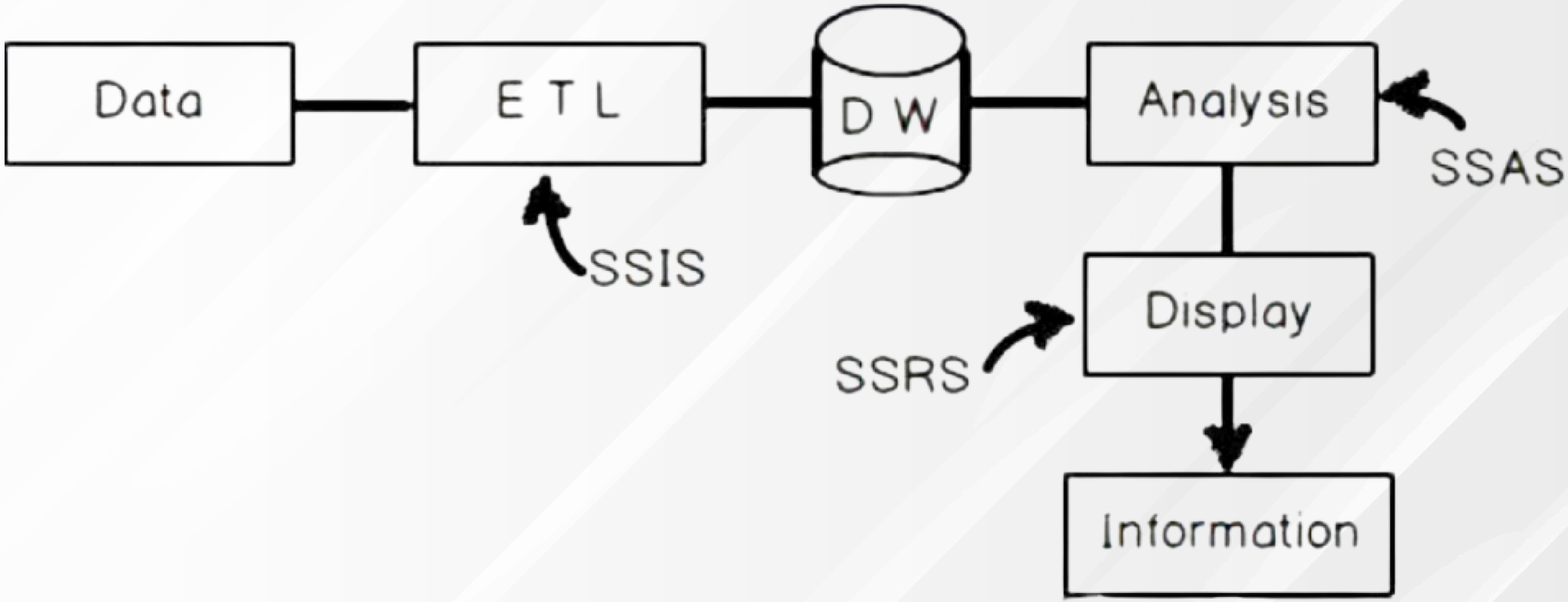


A cube refers to a multidimensional database that is optimized for data analysis and reporting.

MS BI tools

- SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), and SQL Server Integration Services (SSIS) are Microsoft business intelligence (BI) tools that work together to help organizations use their data to make better decisions:

MS BI tools



S. No.	Factor	Data Science	Business Intelligence
1.	Concept	It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.	It is basically a set of technologies, applications and processes that are used by the enterprises for business data analysis.
2.	Focus	It focuses on the future.	It focuses on the past and present.
3.	Data	It deals with both structured as well as unstructured data.	It mainly deals only with structured data.
4.	Flexibility	Data science is much more flexible as data sources can be added as per requirement.	It is less flexible as in case of business intelligence data sources need to be pre-planned.

Thank you so much!