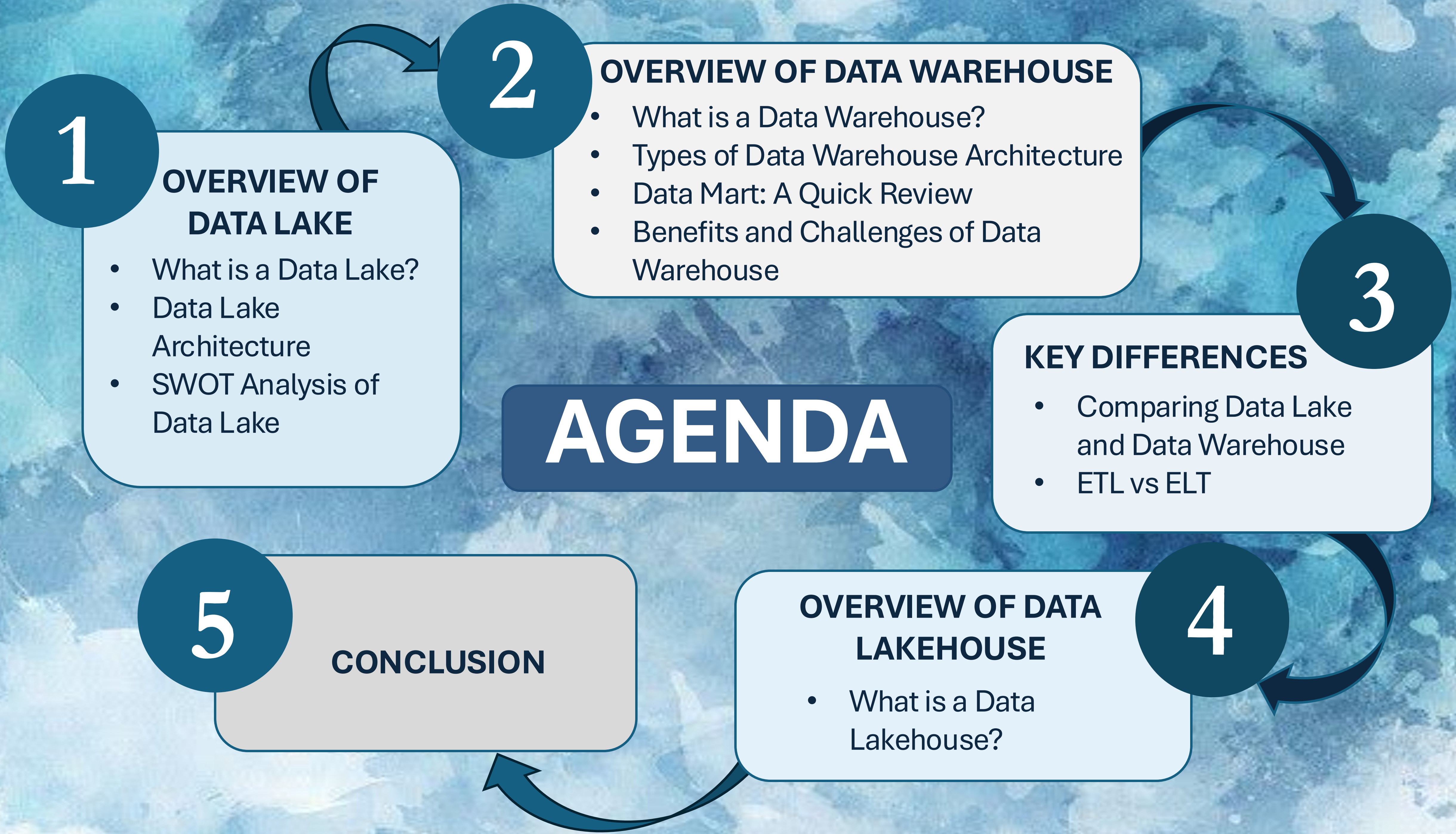
Data Lake & Warehouse

Understanding Modern Data Architectures

AGENDA



1 OVERVIEW OF DATA LAKE

- What is a Data Lake?
- Data Lake Architecture
- SWOT Analysis of Data Lake

2 OVERVIEW OF DATA WAREHOUSE

- What is a Data Warehouse?
- Types of Data Warehouse Architecture
- Data Mart: A Quick Review
- Benefits and Challenges of Data Warehouse

3 KEY DIFFERENCES

- Comparing Data Lake and Data Warehouse
- ETL vs ELT

4 OVERVIEW OF DATA LAKEHOUSE

- What is a Data Lakehouse?

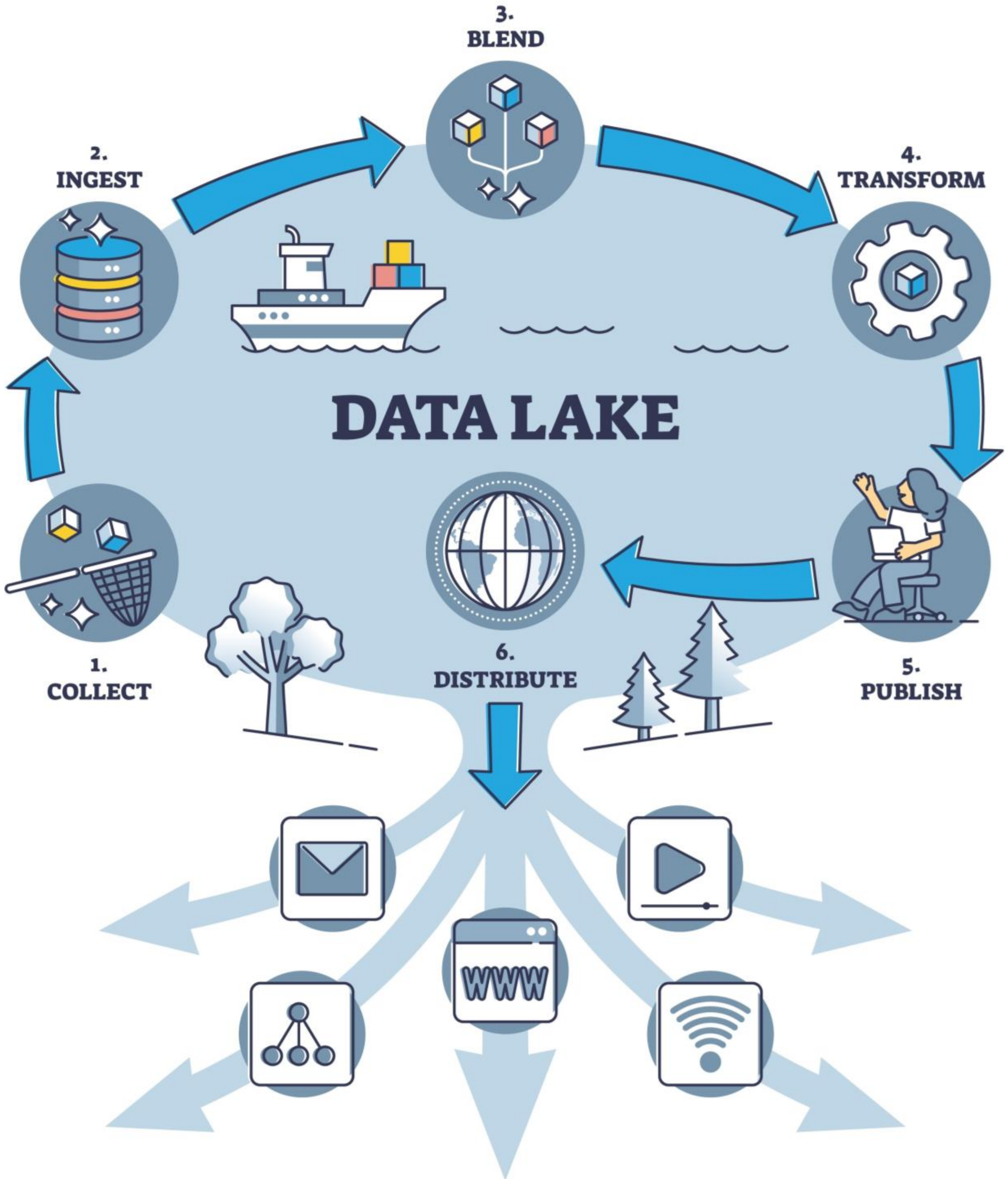
5 CONCLUSION

What is a Data Lake ?

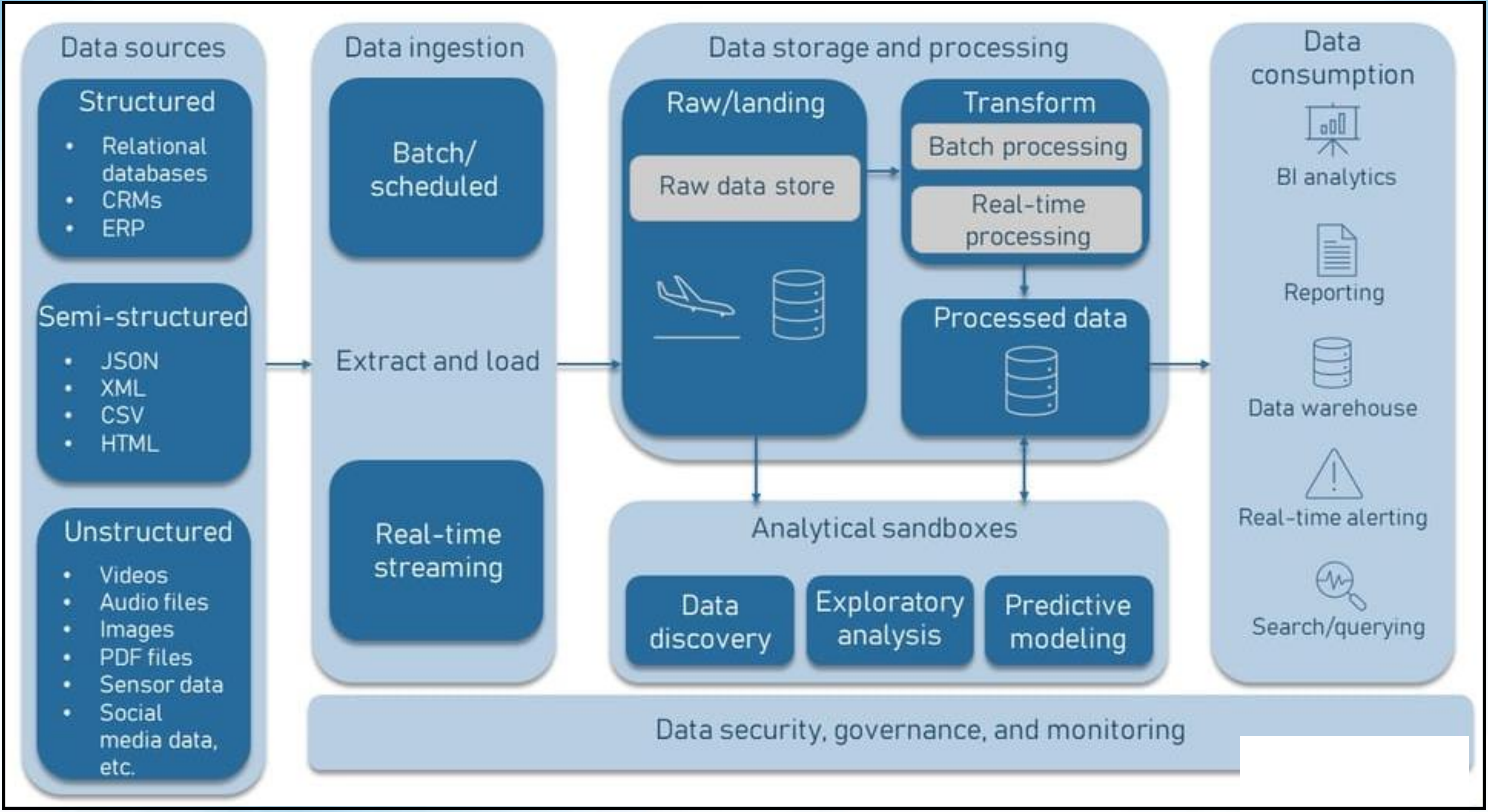
A data lake is a centralized storehouse that enables organizations to store and manage enormous amounts of data in its native format— whether it's structured, semi-structured, or unstructured—in a single, secure, managed environment that's easily accessible across the enterprise.

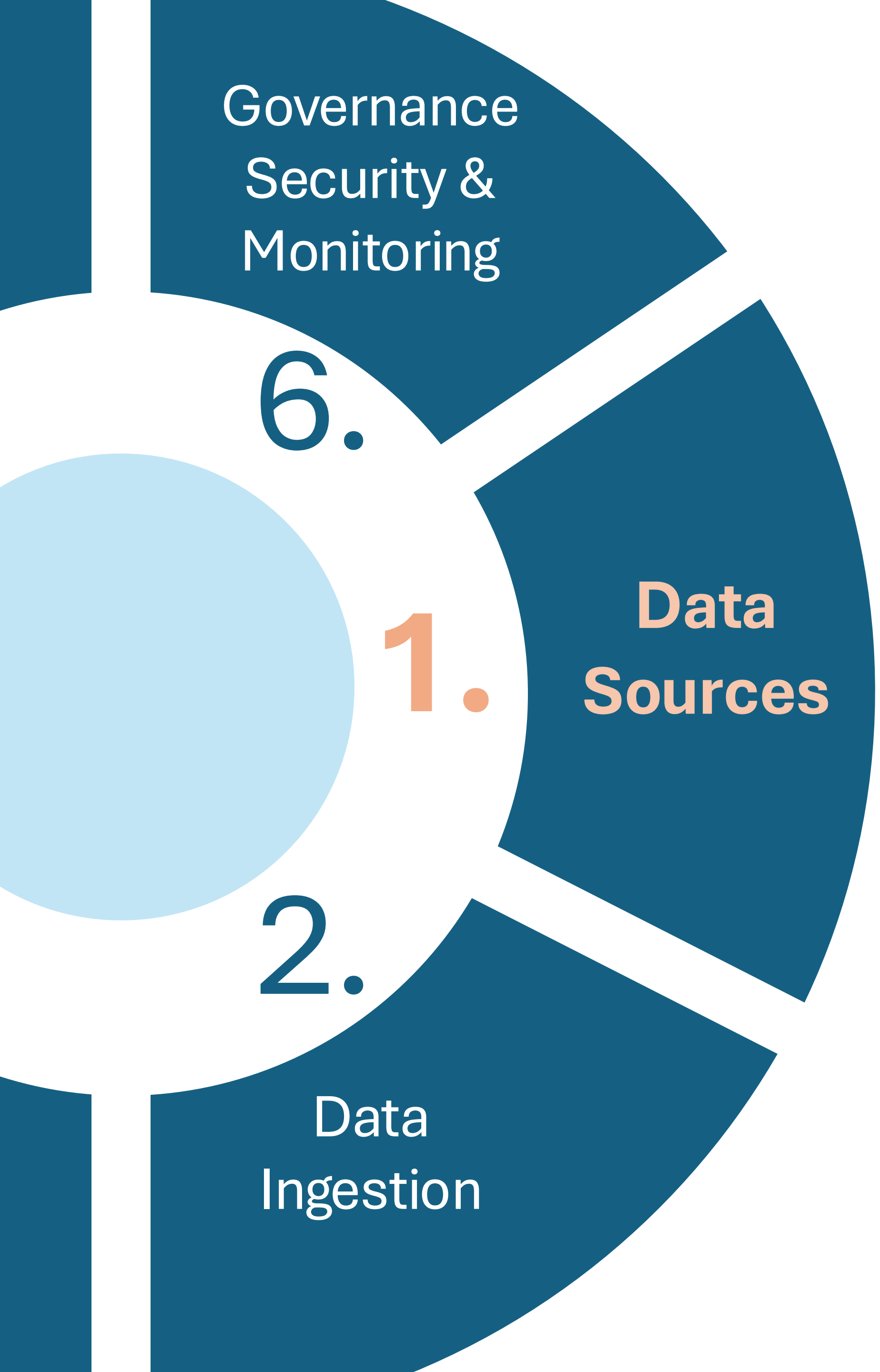
TOP 7 DATA LAKE TOOLS:

1. Amazon S3 & AWS Lake Formation
2. Databricks Lakehouse (Delta Lake)
3. Microsoft Azure Data Lake Storage (ADLS)
4. Google Cloud Storage & BigLake
5. Snowflake Data Cloud
6. Apache Iceberg
7. Dremio Lakehouse Platform



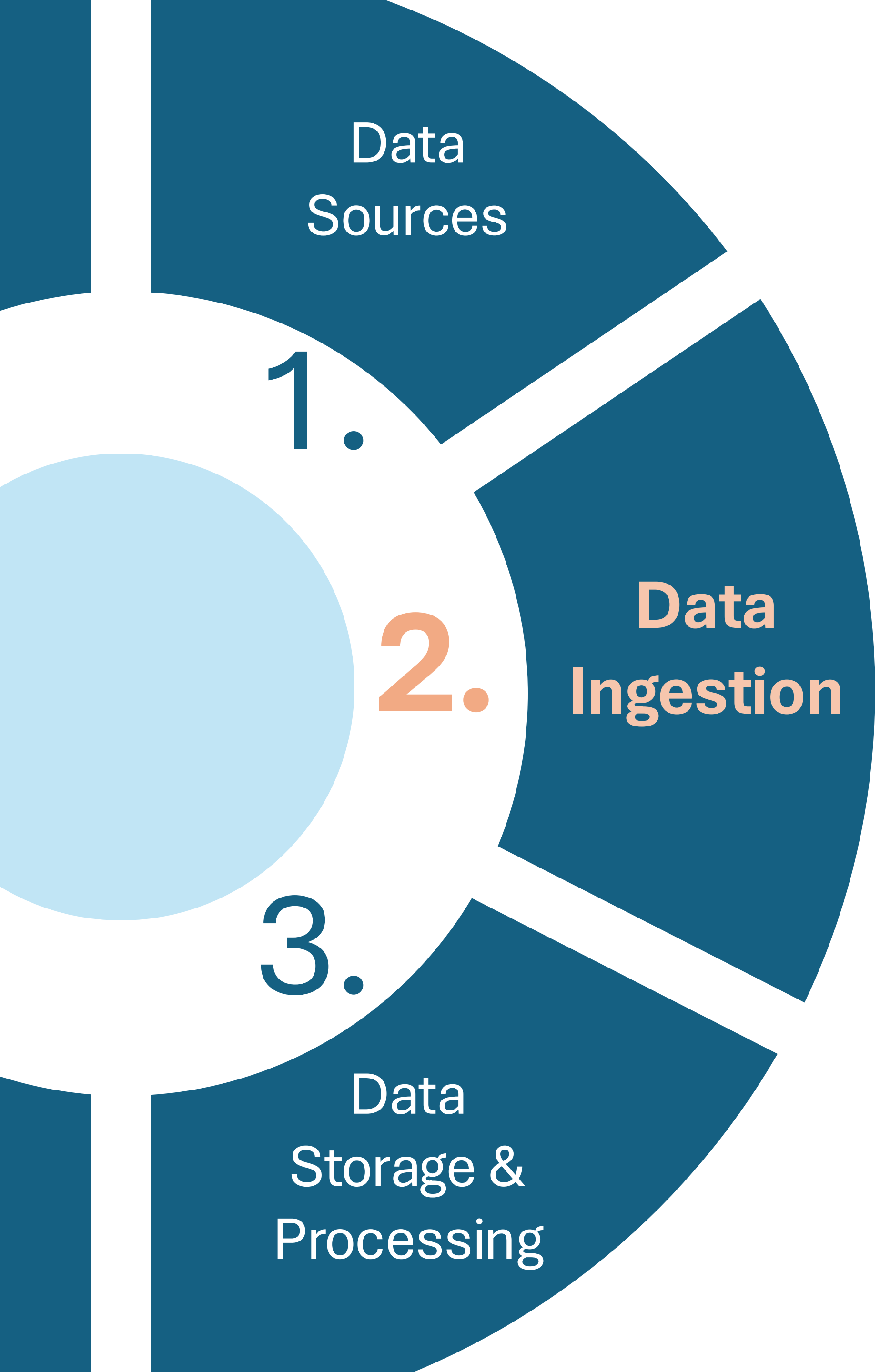
Data Lake Architecture





Data Lake Architecture

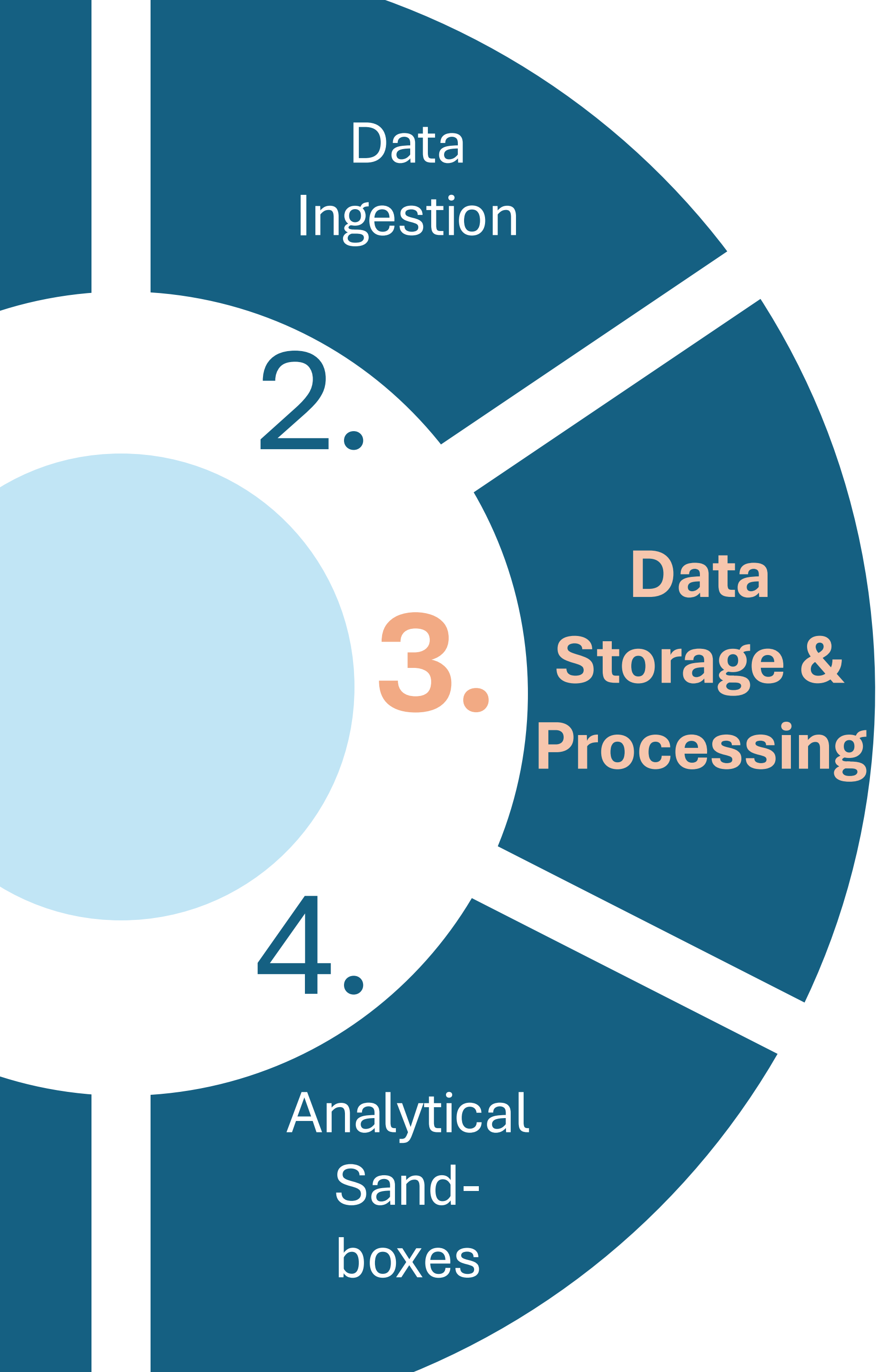
- **Structured:** Well-defined formats (e.g., SQL databases like MySQL, Oracle).
- **Semi-Structured:** Partial structure (e.g., JSON, XML, HTML).
- **Unstructured:** No fixed format (e.g., videos, IoT sensor data, social media).



Data Lake Architecture

- **Batch Ingestion:** Scheduled, interval-based method of data importation. Tools often used for batch ingestion include Apache NiFi, Flume, and traditional ETL tools like Talend and Microsoft SSIS.

- **Real-Time Ingestion:** Immediately brings data into the data lake as it is generated. Apache Kafka and AWS Kinesis are popular tools for handling real-time data ingestion.



2. Data Ingestion

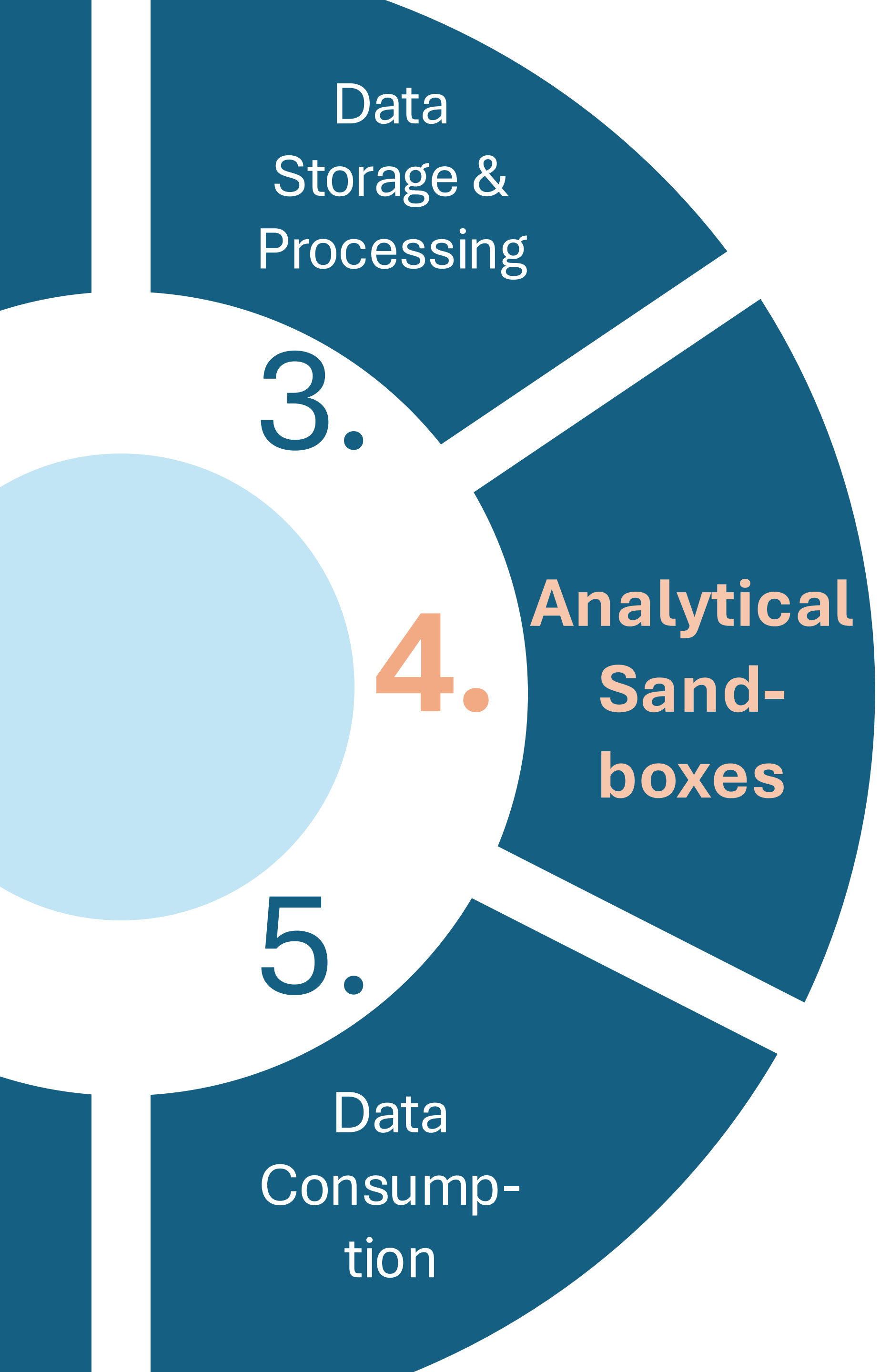
3. Data Storage & Processing

4. Analytical Sand-boxes

Data Lake Architecture

- **Raw Zone:** Stores data in its native form. Utilizes storage solutions like Hadoop HDFS, Amazon S3.
- **Transformation section:**
  - **Data cleansing** involves removing or correcting inaccurate records, inconsistencies in the data.
  - **Data enrichment** adds value to the original data.
  - **Normalization** modifies the data into a common format, ensuring consistency.
  - **Structuring** often involves breaking down data into a structured form suitable for analysis.
- **Processed data section:** Additional transformation and structuring. Tools like Dremio or Presto may be used for querying this refined data.

Data Lake Architecture



3. Data Storage & Processing

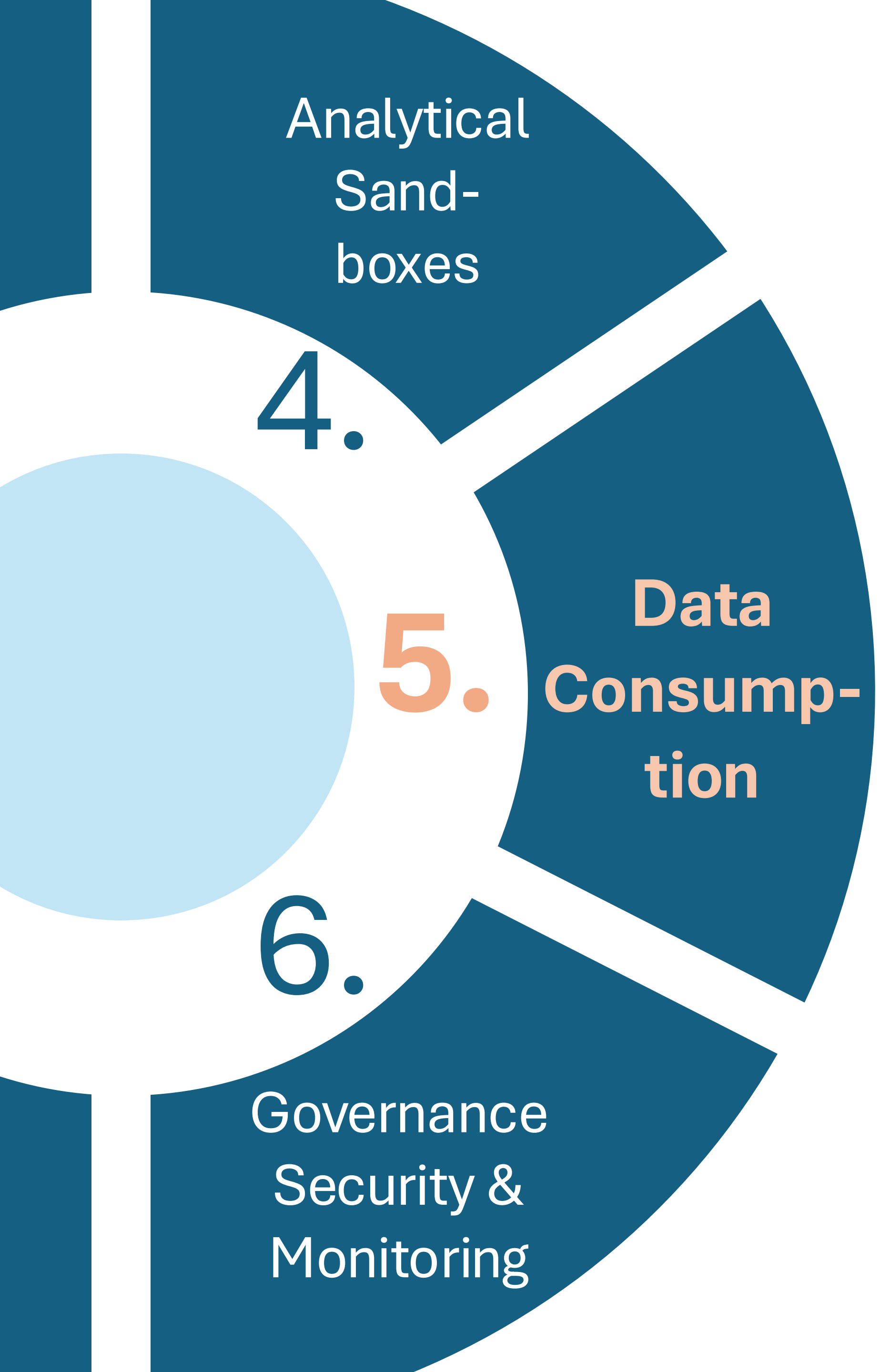
4. Analytical Sandboxes

5. Data Consumption

- **Data discovery:** Explore the data to understand its structure & quality by statistics and data visualization.
- **Machine learning and predictive modelling:** Involve a range of ML libraries like TensorFlow, PyTorch, or Scikit-learn to create predictive or classification models.
- **Exploratory data analysis (EDA):** Statistical graphics, plots, and information tables are employed to analyze the data and understand the variables' relationships, patterns, or anomalies without making any assumptions.

Tools like Jupyter Notebooks, RStudio, or specialized software like Dataiku or Knime are often used within these sandboxes.

Data Lake Architecture



4. Analytical Sandboxes

5. Data Consumption

6. Governance Security & Monitoring

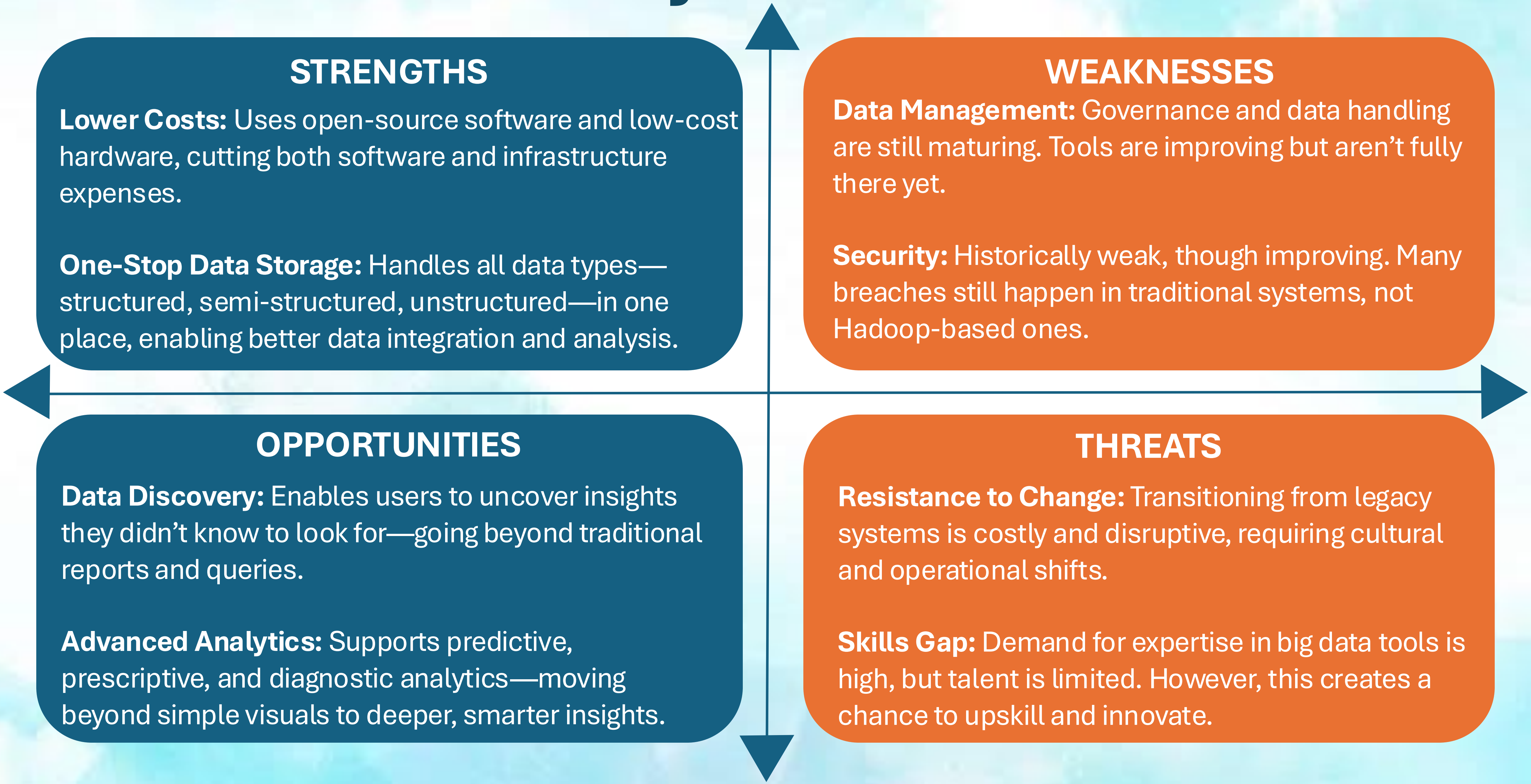
- Final output layer of polished and reliable data.
- Data is exposed via Business Intelligence tools like Tableau, Power BI.
- Consumers are data analysts, decision makers, business teams and executives.

Data Consump-tion
5.
6.
Governance Security & Monitoring
1.
Data Sources

Data Lake Architecture

- **Governance** defines rules and policies for data access and quality. (e.g. Apache Atlas, Collibra)
- **Security protocols** protects data from unauthorized access. (e.g. Varonis, McAfee Total Protection for Data Loss Prevention)
- **Monitoring** handles the oversight and flow of data from its raw form into more usable formats. (e.g. Talend, Apache NiFi)
- **Stewardship** involves active data management and oversight. (e.g. Alation, Waterline Data assist)

SWOT Analysis of Data Lake



STRENGTHS

Lower Costs: Uses open-source software and low-cost hardware, cutting both software and infrastructure expenses.

One-Stop Data Storage: Handles all data types—structured, semi-structured, unstructured—in one place, enabling better data integration and analysis.

WEAKNESSES

Data Management: Governance and data handling are still maturing. Tools are improving but aren't fully there yet.

Security: Historically weak, though improving. Many breaches still happen in traditional systems, not Hadoop-based ones.

OPPORTUNITIES

Data Discovery: Enables users to uncover insights they didn't know to look for—going beyond traditional reports and queries.

Advanced Analytics: Supports predictive, prescriptive, and diagnostic analytics—moving beyond simple visuals to deeper, smarter insights.

THREATS

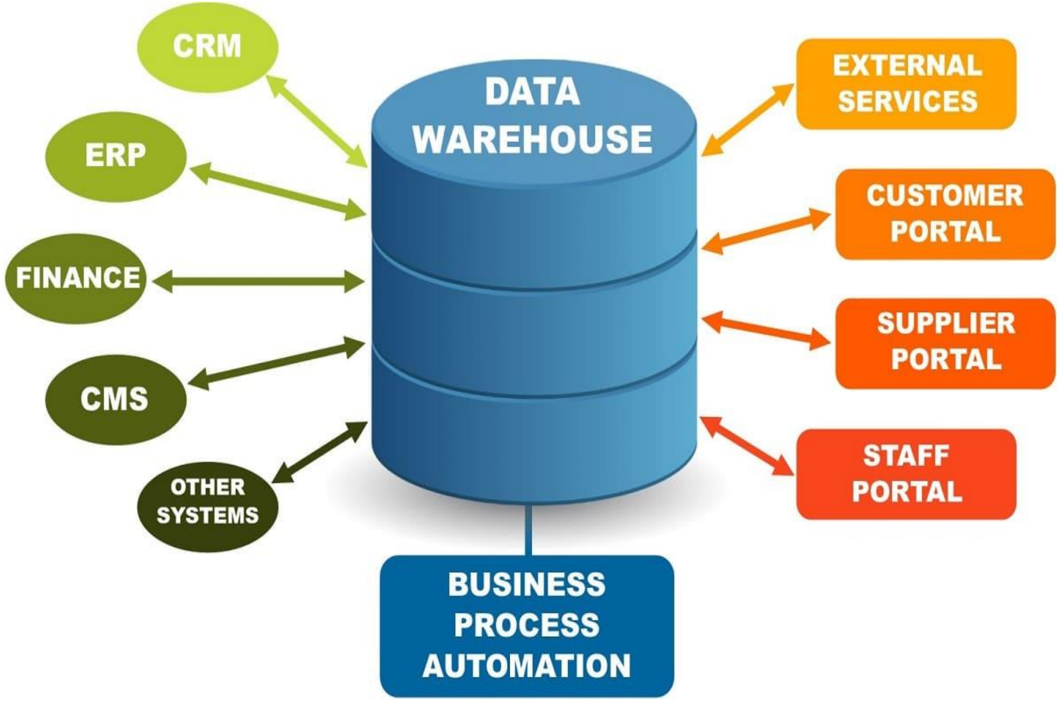
Resistance to Change: Transitioning from legacy systems is costly and disruptive, requiring cultural and operational shifts.

Skills Gap: Demand for expertise in big data tools is high, but talent is limited. However, this creates a chance to upskill and innovate.

What is a Data Warehouse ?

A data warehouse is a **centralized repository** optimized for storing structured data from various sources, where the data is cleaned, transformed, and organized into a **consistent format** to support high-performance querying, reporting, and business intelligence—unlike a data lake, which stores raw data in its native format.

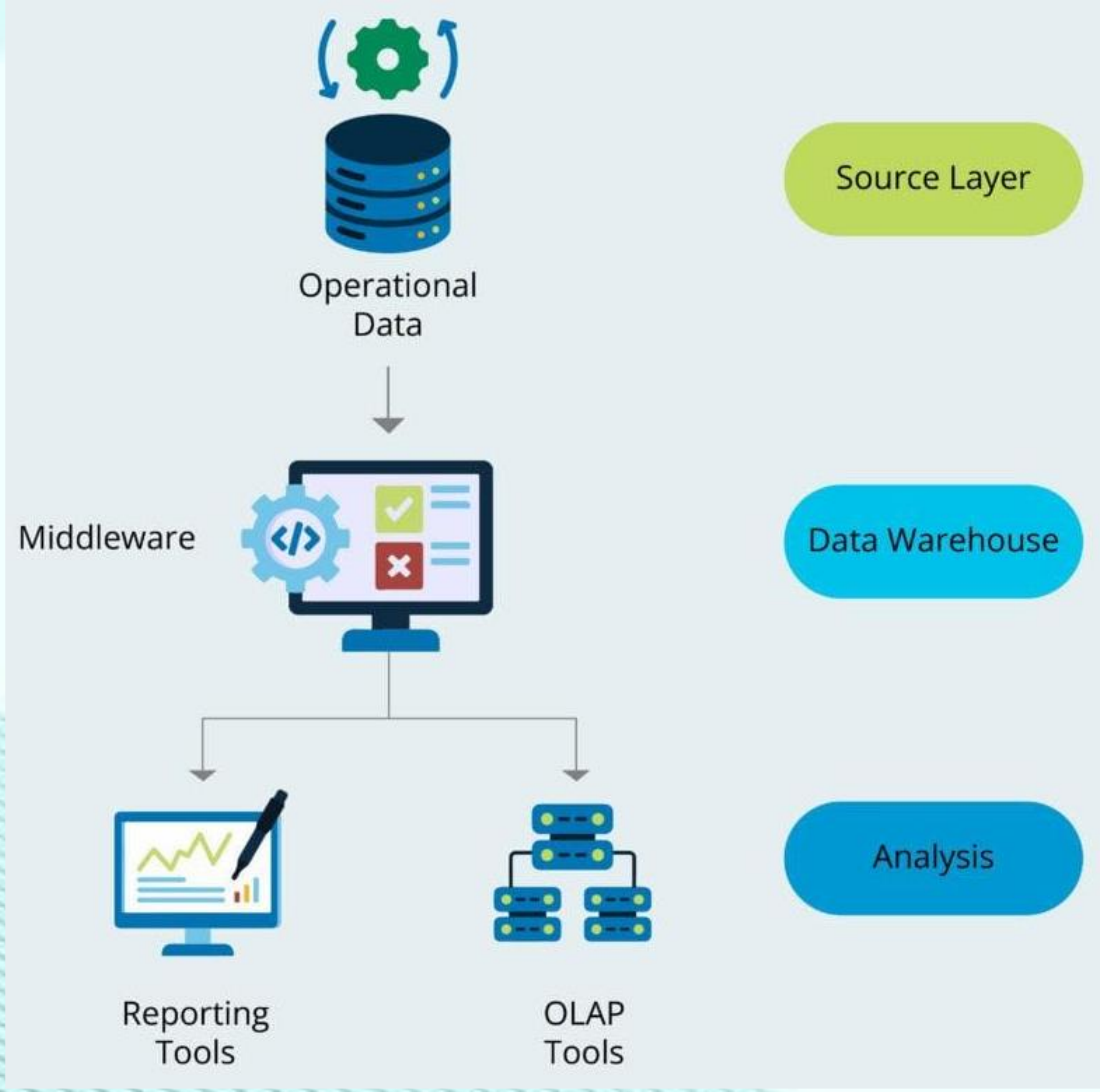
It help businesses quickly analyze big data and make smart decisions.



TOP 7 DATA WAREHOUSE TOOLS:

1. Snowflake
2. Azure Synapse Analytics
3. Google BigQuery
4. Amazon Redshift
5. IBM Db2 Warehouse
6. Oracle Autonomous Data Warehouse
7. Firebolt Cloud Data Warehouse

Types of Data Warehouse Architecture



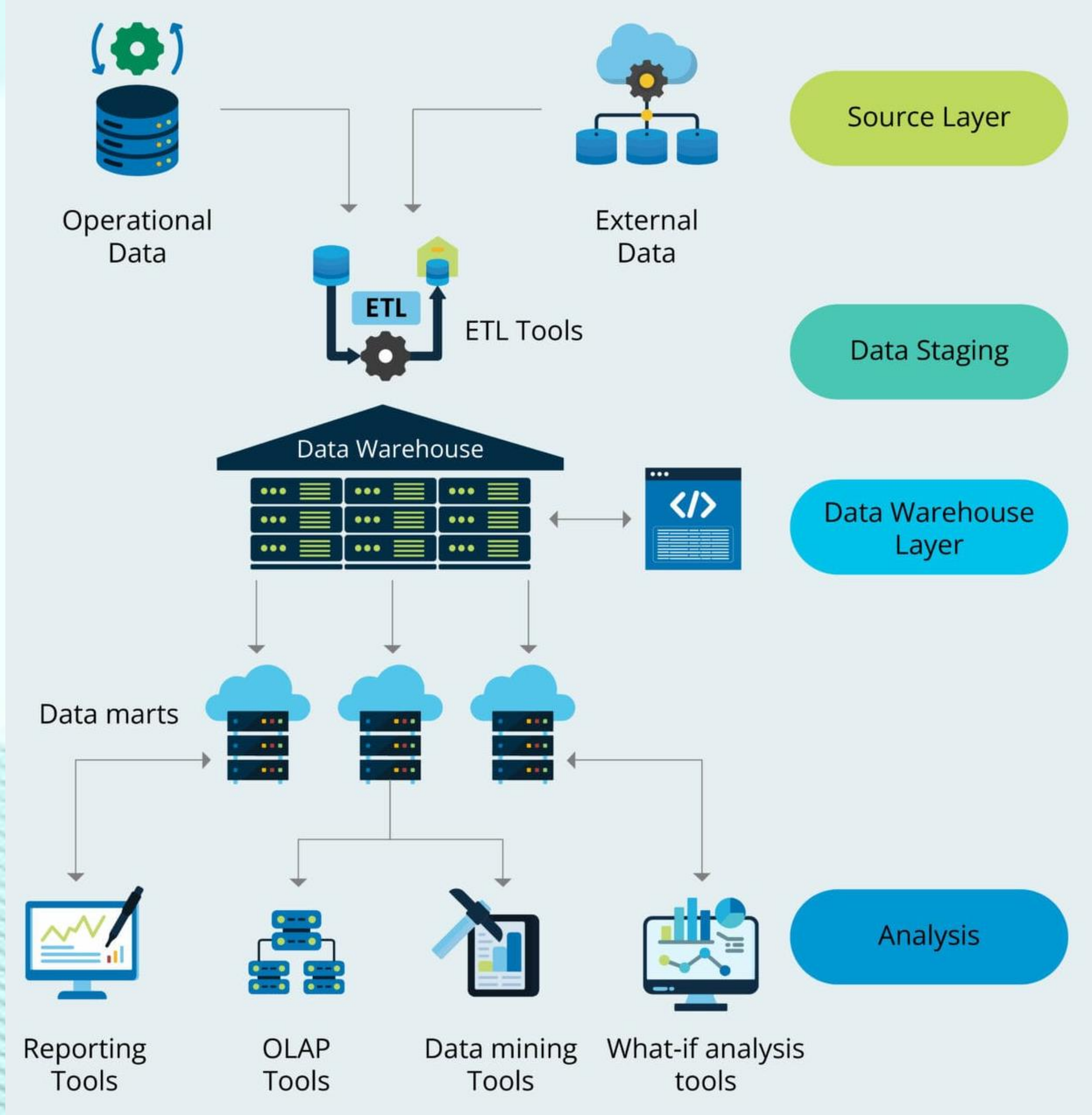
Single-tier Data Warehouse Architecture

Designed for small-scale environments. Combines data storage and processing in a single layer.

- **Source Layer:** Collects raw, real time data (operational data) from internal systems (Enterprise Resource Planning, Customer Relationship Management, spreadsheets etc.)
- **Data Warehouse:** Stores, transforms, and processes data in one unified layer. ETL tasks (Extract, Transform, Load) are embedded here.
- **Analysis Layer:** Performs both transactional and analytical processing on the same system. Can lead to performance bottlenecks due to shared resources.

Small businesses with limited data and basic reporting needs.

Types of Data Warehouse Architecture



Two-tier Data Warehouse Architecture

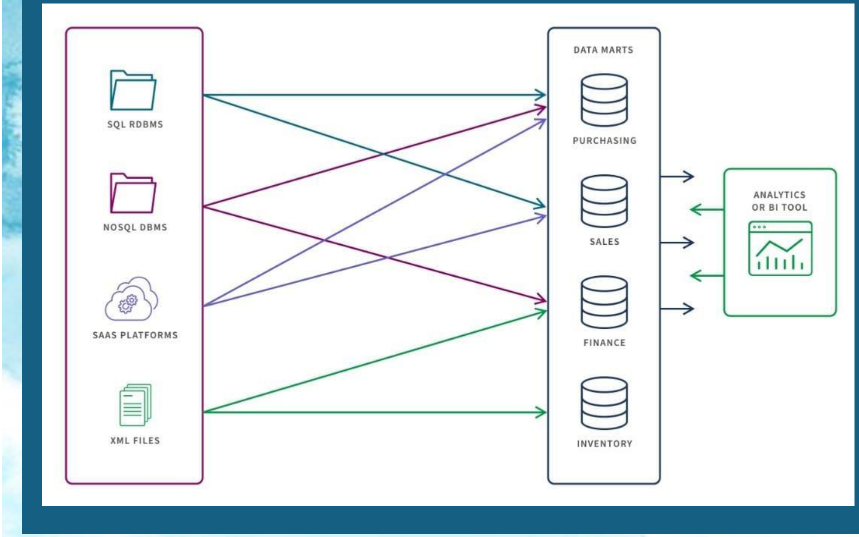
Adds a dedicated staging area and separates analytical from transactional systems. Offers improved performance over single-tier but with limited scalability.

- **Source Layer:** Collects data from multiple sources — e.g., CRM, ERP, flat files, spreadsheets.
- **Staging Layer:** Temporary holding area for data. Performs ETL operations like cleaning, deduplication, formatting.
- **Warehouse Layer:** Stores transformed, structured data. Organized into schemas optimized for querying.
- **Analysis Layer:** Provides user access through BI tools and reporting interfaces. Supports dashboards, ad-hoc queries, and summary reports.

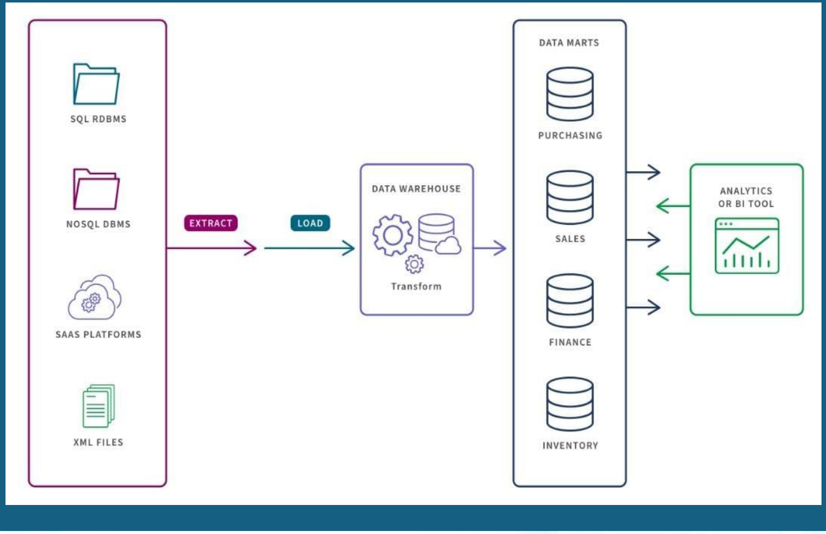
Medium-sized businesses requiring clean data and moderate reporting needs.

Data Mart: A Quick Review

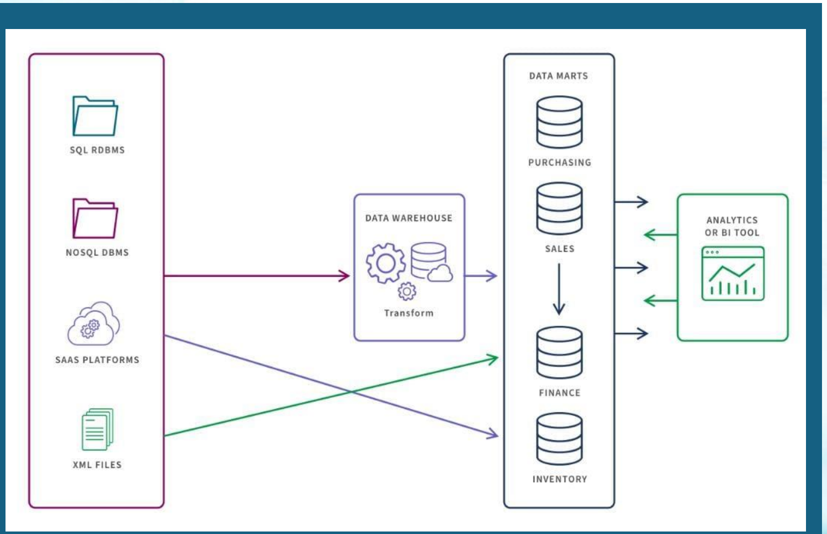
A data mart is a specialized subset of a data warehouse focused on a specific functional area or department within an organization. It provides a simplified and targeted view of data, addressing specific reporting and analytical needs.



Independent Data Mart
Built separately for a specific department without relying on a central warehouse.

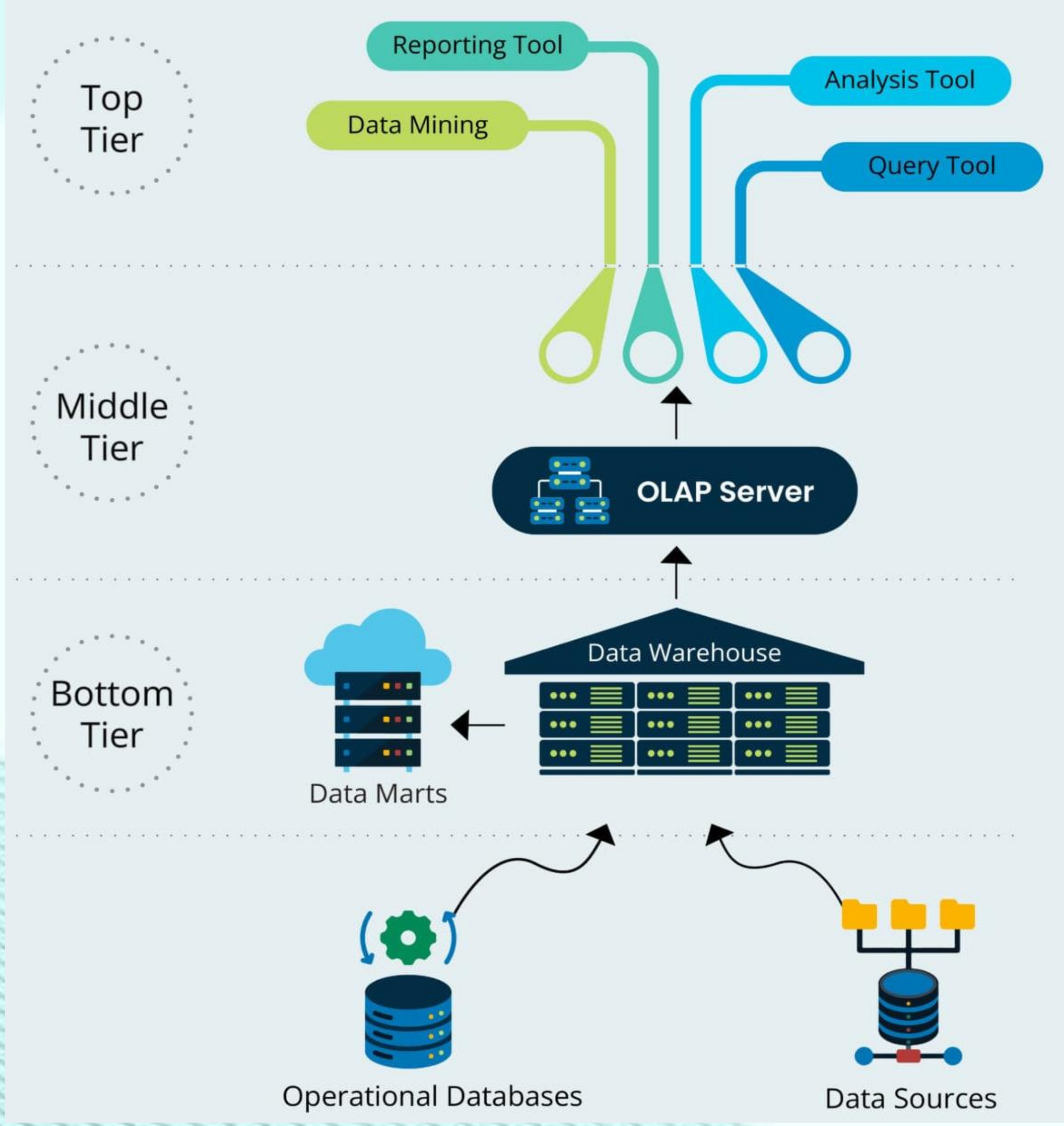


Dependent Data Mart
Created from a central data warehouse to ensure consistent and integrated data.



Hybrid Data Mart
Combines warehouse data with local sources to balance consistency and flexibility.

Types of Data Warehouse Architecture



Three-tier Data Warehouse Architecture

Most widely adopted. Separates storage, processing, and user interaction layers.

Bottom Tier (Data Warehouse Layer): Stores large volumes of cleansed, transformed data. Ensures consistency and quality.
Middle Tier (OLAP / Reconciled Layer): Uses OLAP servers to organize data into a user-friendly format. Enables multidimensional analysis and fast querying.
Top Tier (Presentation Layer): User access layer with dashboards, reports, BI tools. Provides business insights and visualizations.

Large enterprises with high data volumes and complex analytics needs.

Benefits & Challenges of Data Warehousing

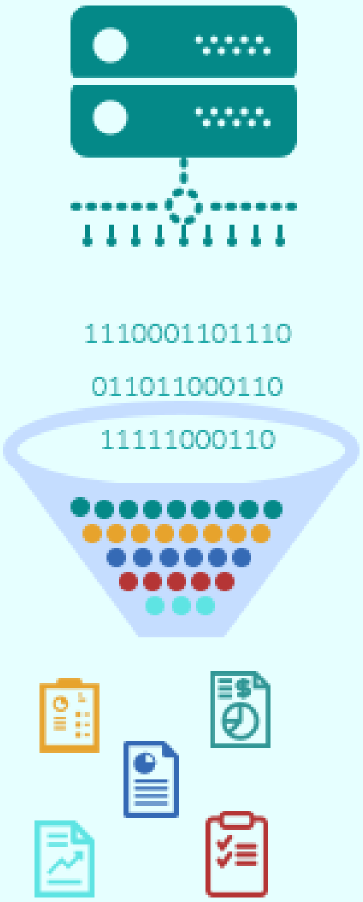
Benefits

- **Improved data quality:** Data is cleaned, transformed, and organized, ensuring high accuracy and consistency.
- **Competitive advantage:** Enables strategic insights by uncovering hidden customer and market trends.
- **Productivity of decision makers:** Enhances decision-making efficiency through consistent and integrated historical data.
- **Cost-effective decision making:** Reduces IT dependency and external data needs by centralizing data storage.
- **Faster data access:** High-speed servers and structured storage allow rapid querying and retrieval of data.

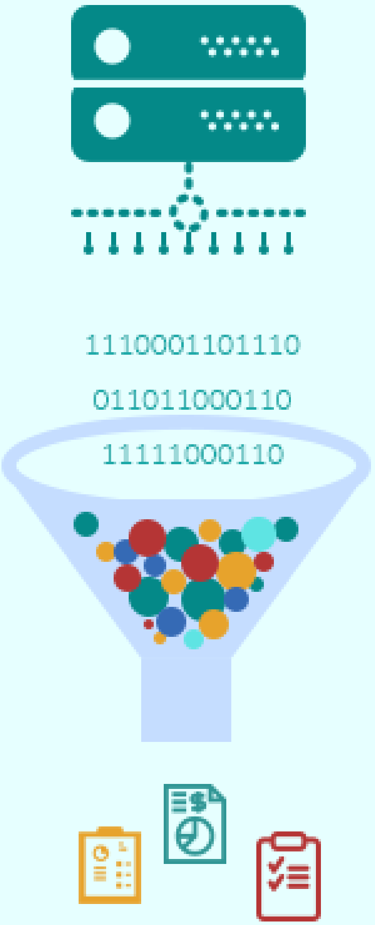
Challenges

- **Underestimation of data loading resources:** Data cleaning and loading often take more time than anticipated.
- **Hidden problems in source systems:** Data quality issues in source systems may surface only after integration.
- **Data homogenization:** Standardizing data formats can lead to the loss of nuanced or valuable data.
- **Complex setup and training:** Needs expert professionals and thorough staff training to avoid mismanagement or data loss.
- **Real-time delays:** Real-time data must be processed before storage, causing delays in availability.

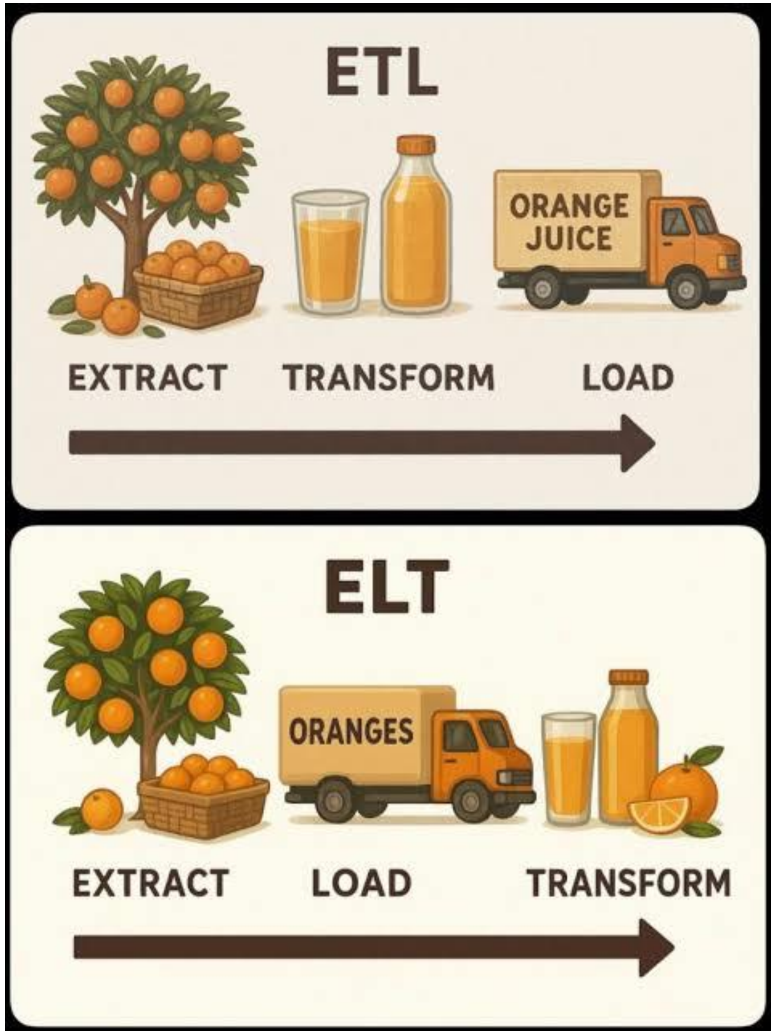
Comparing Data Warehouse and Data Lake



Data Warehouse	vs	Data Lake
structured, processed	DATA	structured/semi-structured /unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.



ETL vs ELT

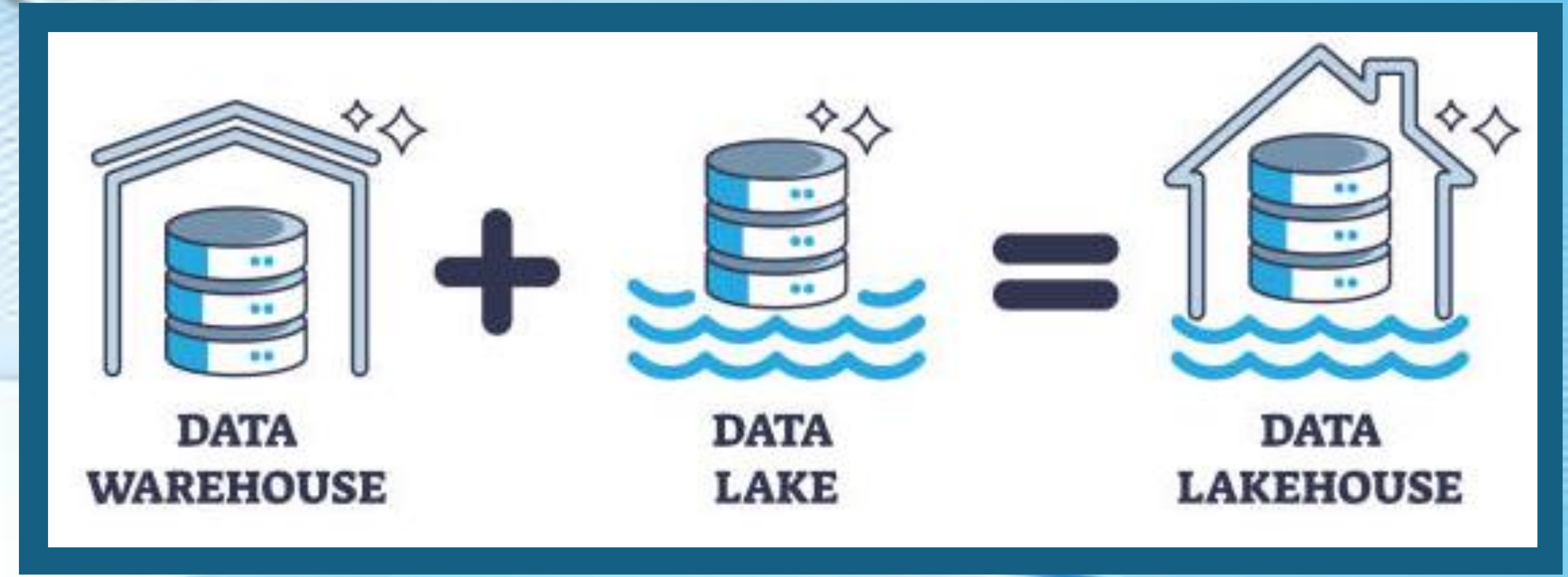


Category	ETL	ELT
Purpose	ETL extracts data, transforms it using a secondary processor, and loads it.	ELT extracts data, loads it, and transforms it within the system.
Expense	Using multiple servers is expensive.	The simplified process cuts costs.
Speed	Transferring transformed data takes a lot of time.	ELT is faster because the data is loaded and transformed in the same system.
Maturity	ETL has been around for over 20 years.	ELT is a new integration method.
Privacy	The pre-loaded transformation helps you meet government regulations.	Loading raw data directly requires extra protection.
Maintenance	Managing a second processor is a hassle.	The fewer systems make maintenance easier.

What is a Data Lakehouse ?

- A Data Lakehouse is a modern data architecture that combines the best features of a data warehouse and a data lake.

- It offers the scalability and flexibility of a data lake for storing all types of data, with the reliability and performance of a data warehouse for analytics.



Conclusion

- The future of data management is **hybrid** and **unified**
- Data Warehouses are best for structured data and traditional BI reporting.
- Data Lakes are great for storing large, raw, and diverse data types.
- Data Lakehouses combine the strengths of both — flexible, scalable, and analytics-ready.
- A **strategic blend of data lake, warehouse, and lakehouse** can drive smarter decisions and long-term innovation.

Thank you