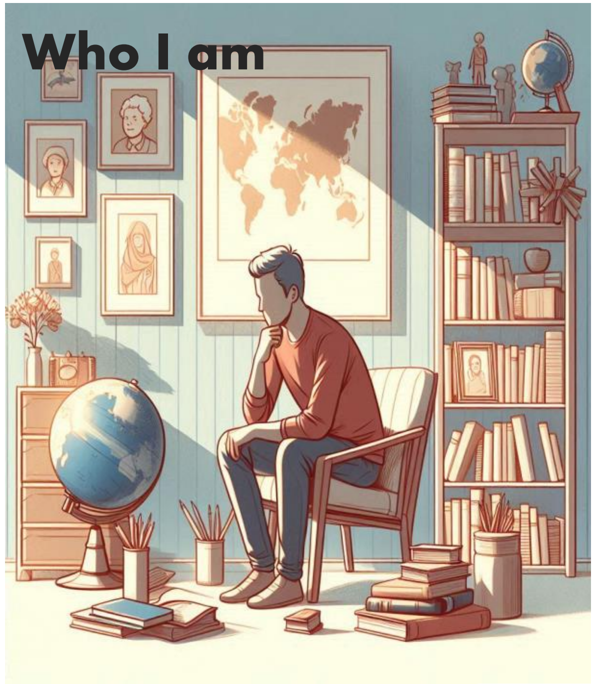
THE LIFE CYCLE OF DATA AND THE STATE OF THE SECTOR

An in-depth exploration of data management frameworks and industry trends

Who I am



High School
- Hatay Dörtyol Anatolian High School
- Konya Ereğli Atatürk High School

University
- Yıldız Teknik – Mathematical Engineering

Master Degree
- Bahçeşehir – Engineering Management

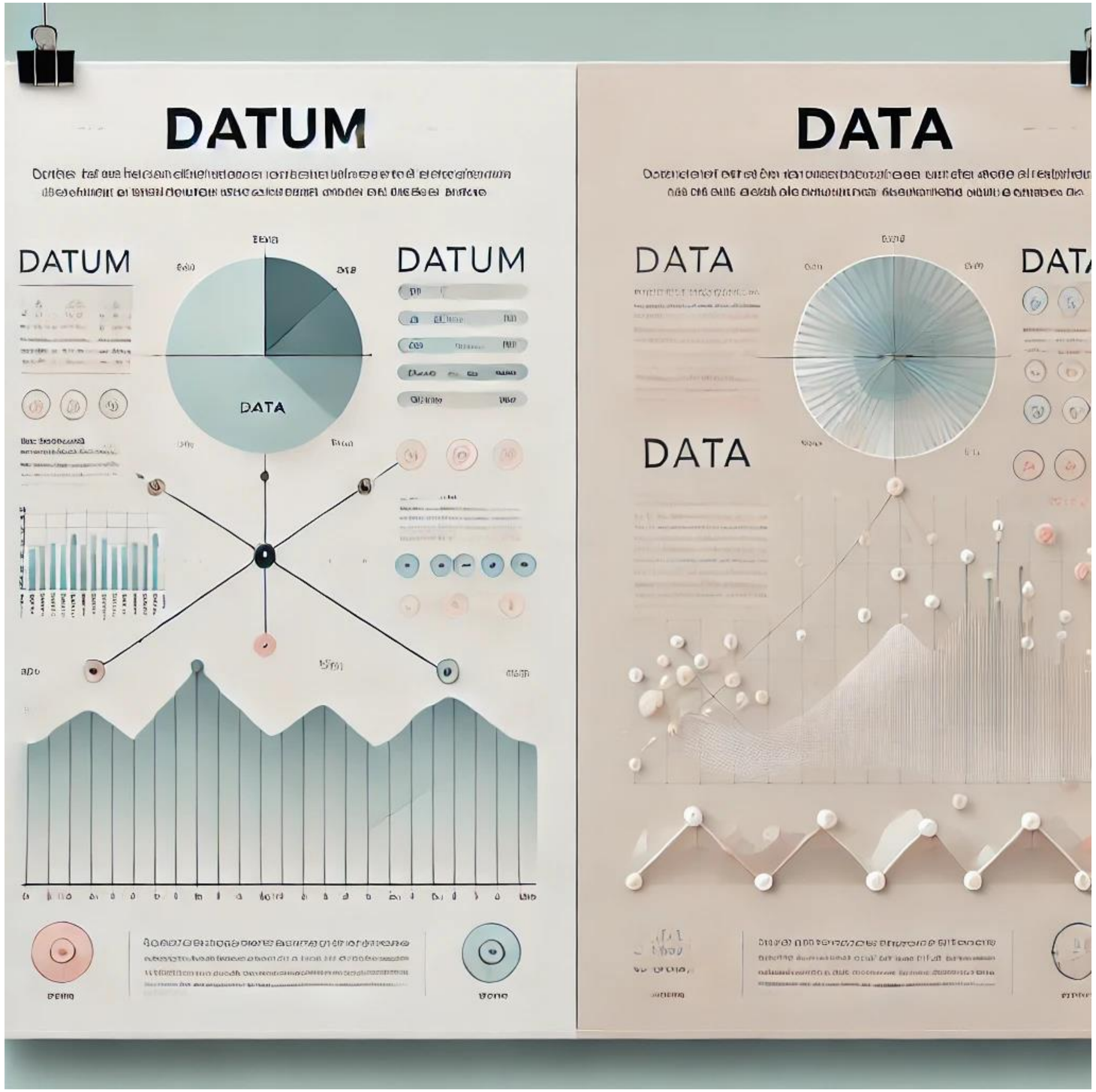
Teknosa (2004-2007) → Smartiks (2007-2009)
- Pfizer
- Deva
- Avea

→ Avea (2009 -2013) → TEB (2013-2013) → KKB(Kredi Kayıt Bürosu) (2013-2016) → Vodafone (2016-2018) → YKB (2018-)

Company - Project Experience

Teknosa	Smartiks	Avea	TEB	KKB	Vodafone	YKB
• Sales Reporting	• BI (Business Intelligence) Developer	• Telco Data Warehouse Project • Real Time Data Transfer and Reporting • Campaign Management	• Financial Dwh Project	• National Financial Reporting Project • Factoring Financial Reporting Project	• Telco DWH Project • Campaign Management • Establishment of Bigdata Environment	• Warehouse Transformation • Clikview to PowerBI Dashboarding • Big Data Analytical Environments

How Describe Datum & Data



Datum is the singular form of "data" and refers to a single piece of factual information or a single measurement.

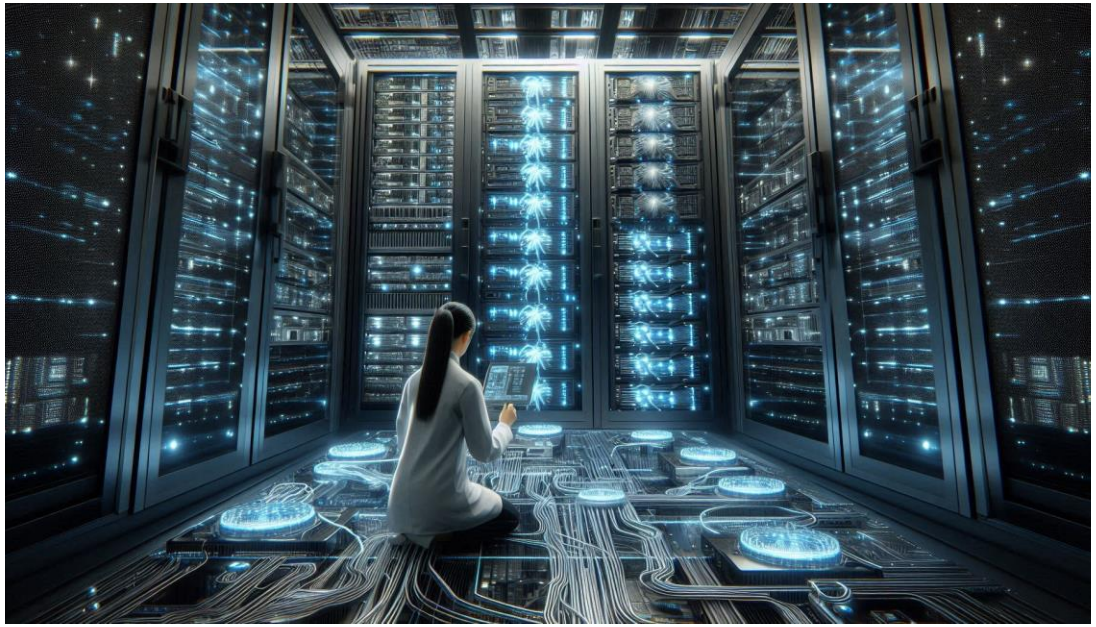
Data is the plural form of datum, although it is commonly treated as a singular mass noun (e.g., "data is" rather than "data are") in everyday language.

Datum = a single fact or measurement.
Data = multiple pieces of information collectively used for analysis and interpretation

i.Data Life Cycle Stages

1. **Data Generation & Collection**
   Data is continuously generated from a wide range of sources,
   - Customer interactions,
   - Business transactions,
   - IoT devices, and
   - Digital platforms.

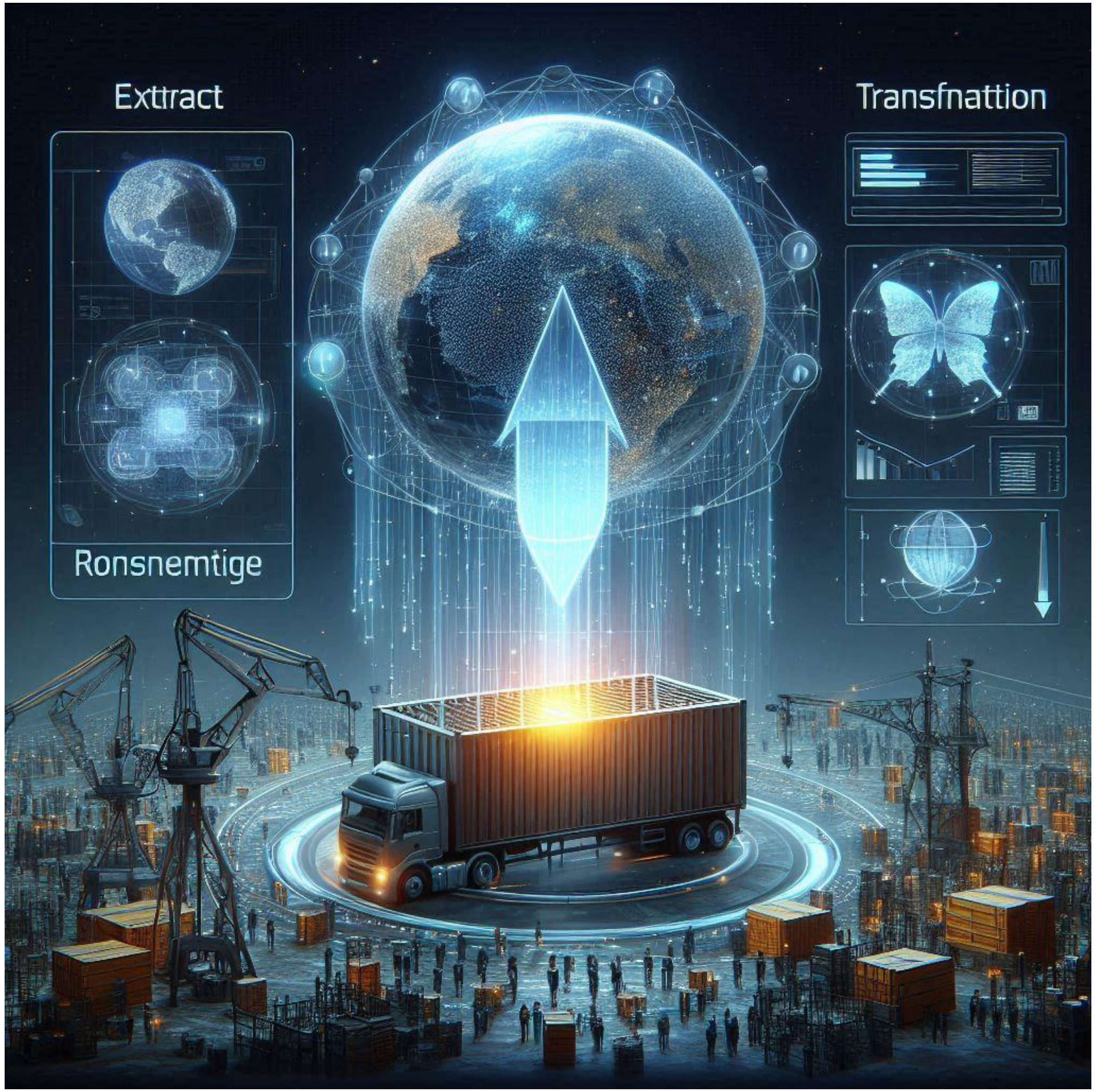
   Effective collection strategies are essential to ensure data quality, accuracy, and compliance from the outset.



2. Data Storage & Retention
Once collected, data needs to be stored securely and in a way that supports accessibility and scalability. Traditional databases are often supplemented or replaced by data lakes, data warehouses, or cloud storage solutions. Advances in cloud technology have made data storage more flexible and cost-effective, while increasing storage capacities for handling vast amounts of information.

3. Data Processing & Transformation

Raw data is rarely useful in its collected form. It typically requires cleaning, filtering, and transforming to make it ready for analysis. This stage often includes **ETL** (Extract, Transform, Load) processes, where data is extracted from different sources, transformed for quality and consistency, and then loaded into data warehouses or analytics platforms. Automation and machine learning are playing larger roles in these processes, reducing the time and labor involved.

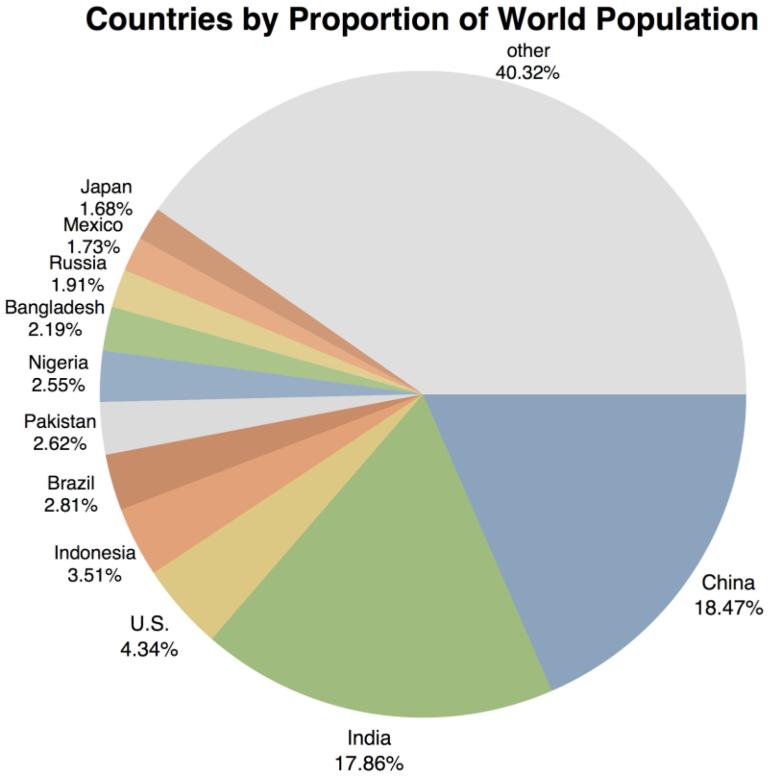


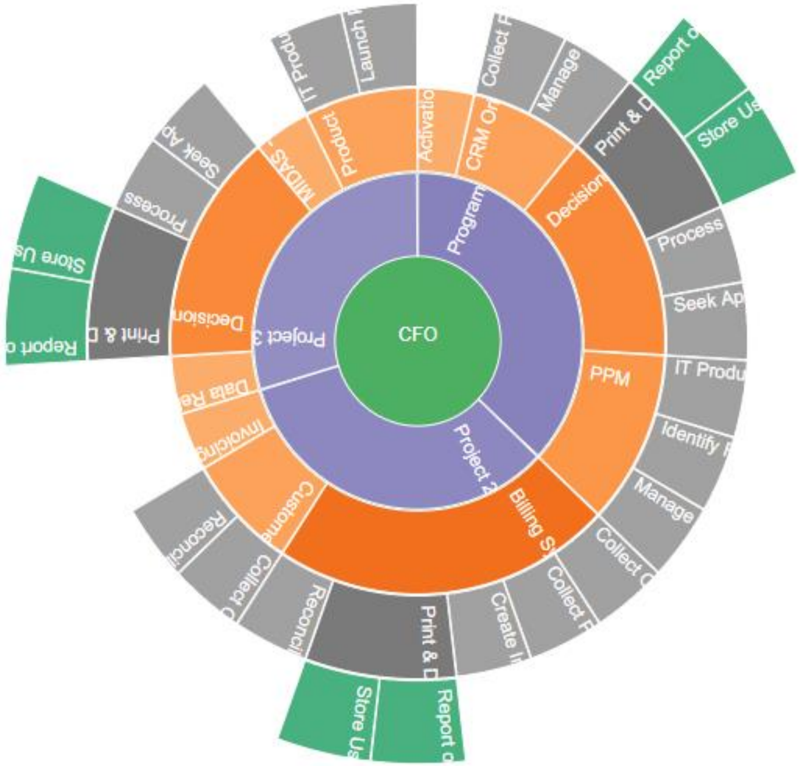
4. Data Analysis

This stage focuses on extracting insights from processed data. Advanced analytics tools, powered by machine learning and AI, enable businesses to uncover trends, patterns, and anomalies in their data. Visualization tools also help transform raw data into meaningful insights that can inform decision-making across an organization.

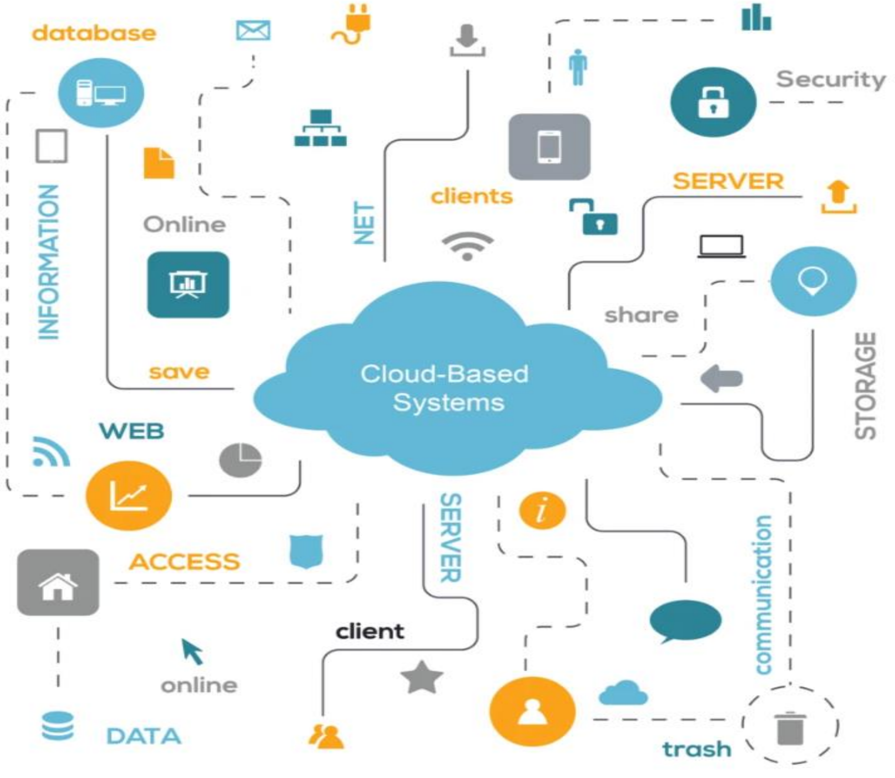
5. Data Sharing & Usage

Insights generated from data analysis are shared across departments to aid in strategic decision-making. Ensuring data is accessible to the right users while maintaining security is a primary challenge here. Modern data management practices emphasize data democratization, allowing non-technical users to leverage data for decision-making.



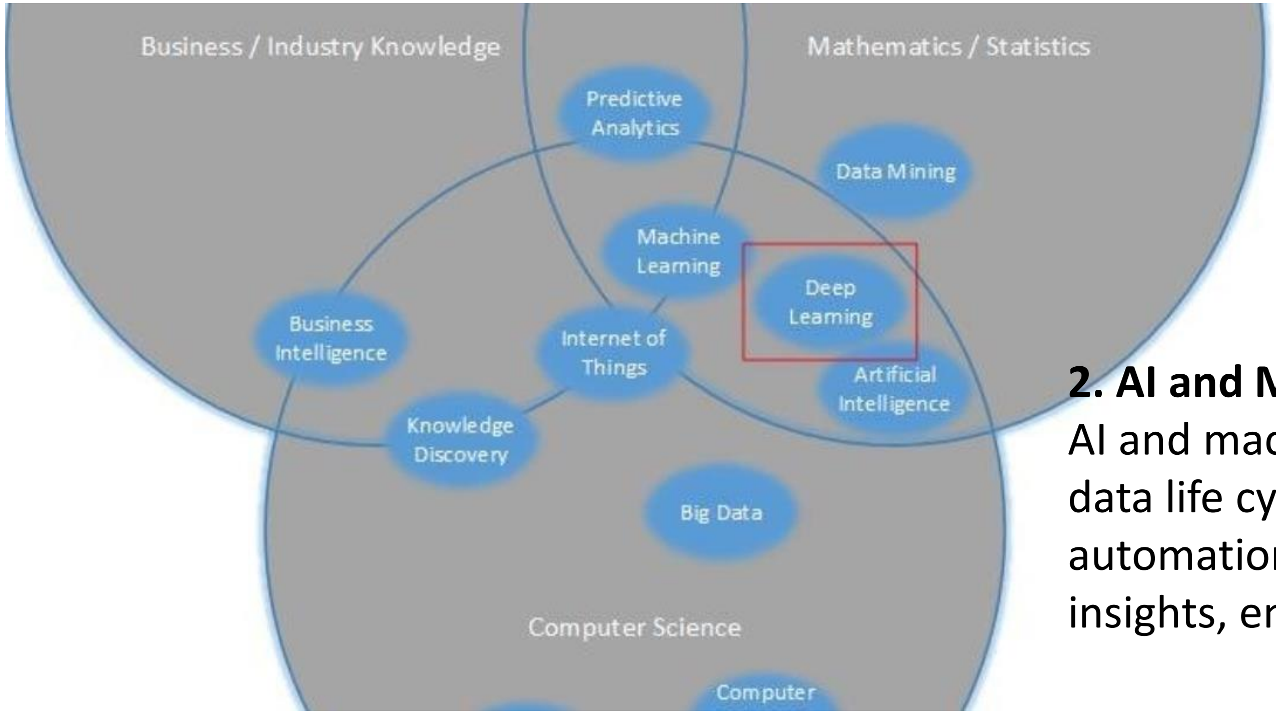


ii. State of the Data Management Sector



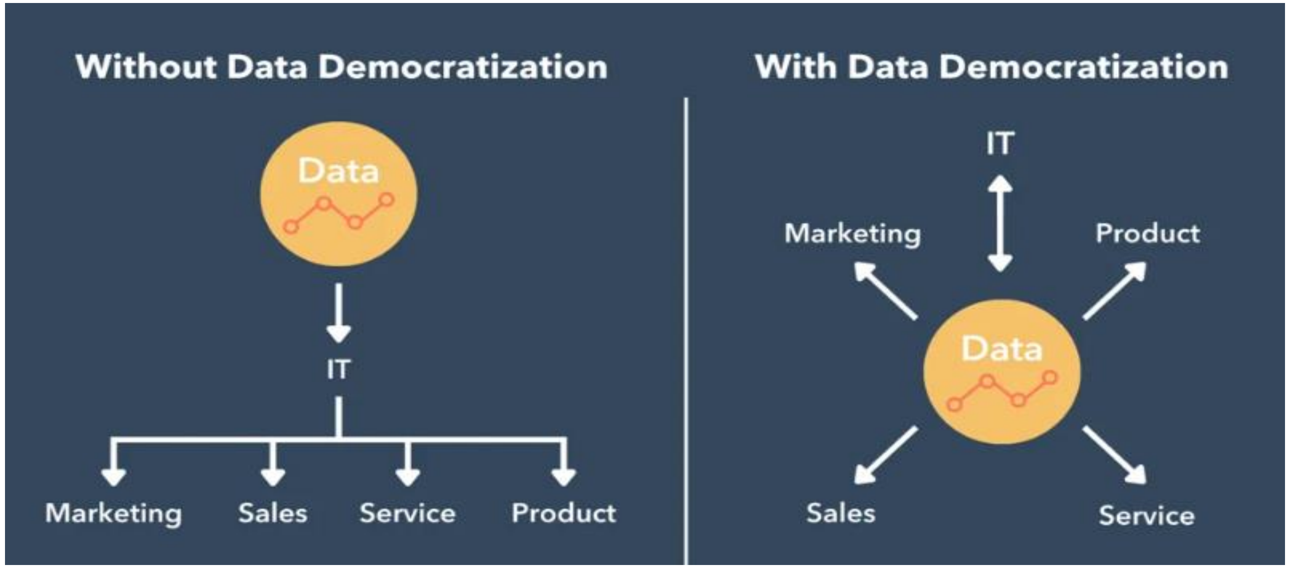
1. Cloud-Based Data Management

Cloud adoption continues to rise as companies seek scalable, flexible, and cost-effective solutions for data storage, processing, and analytics. Cloud platforms offer integrated data management tools that streamline the data life cycle, especially for organizations with distributed teams or data sources.



2. AI and Machine Learning Integration

AI and machine learning are becoming integral to all stages of the data life cycle, especially in data processing, analysis, and automation. These technologies enable faster and more accurate insights, empowering businesses to be more proactive and adaptive.



4. Data Democratization and Self-Service Analytics

There is a growing push to make data more accessible across organizational levels, allowing non-technical employees to leverage data insights through self-service tools. This trend promotes a data-driven culture and encourages collaborative, insight-driven decision-making.

5. Real-Time Data Processing

With an increasing demand for real-time insights, companies are prioritizing technologies that enable rapid data processing. Streaming data platforms and edge computing are allowing businesses to make time-sensitive decisions with minimal latency, especially in sectors like finance, healthcare, and manufacturing.

My Team Tool Experience Chart

Role	Total	Informatica	Business Object	Power BI	JAVA	Big Data	Python	React	SAS
Architect	**2**	2	0	0	0	0	1	0	0
BI Developer	**12**	1	12	12	1	0	0	0	0
Data Analyst	**32**	10	3	4	0	0	0	0	1
ETL Developer	**36**	36	1	1	0	0	2	0	4
Full Stack Developer	**10**	0	0	0	2	8	10	10	0
Total	**92**	**49**	**16**	**17**	**3**	**8**	**13**	**10**	**5**

100
Ekip Üyesi

58
Developer

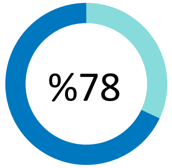
40
Analyst

2
Architect

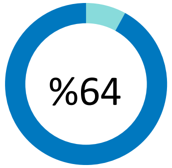
DWH Architecture

We use an Oracle database located at the centre of the DWH architecture and Informatica tool for Oracle-to-Oracle data flow.

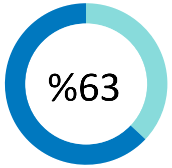
Resource Utilisation Statistics:



Storage
585 TB Usage
750 TB Capacity



Memory
970 GB Usage
1.5 TB Capacity



CPU
68 Core
CPU usage

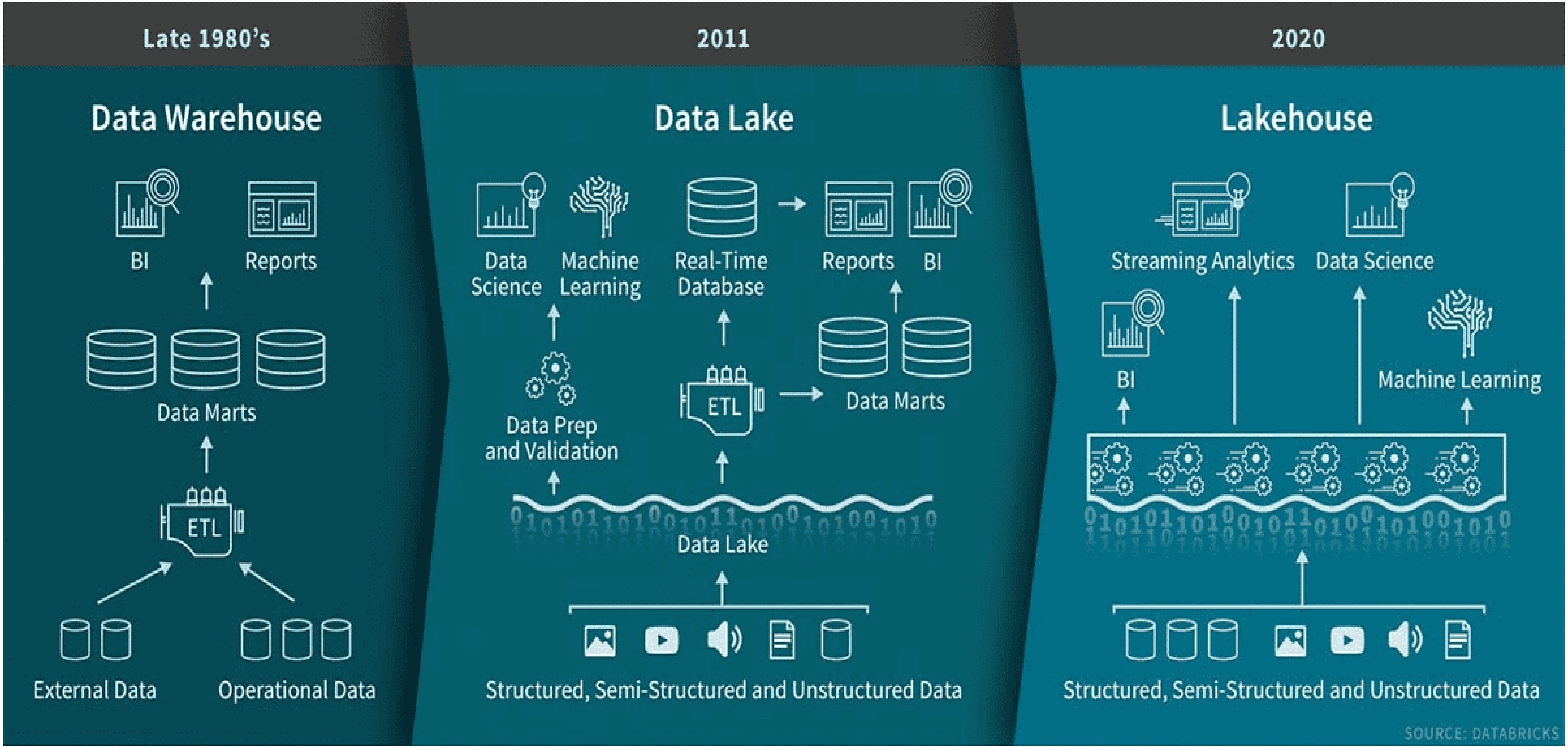
12 TB
Monthly Storage Growth

17/20 TB
Daily Average Processed Data

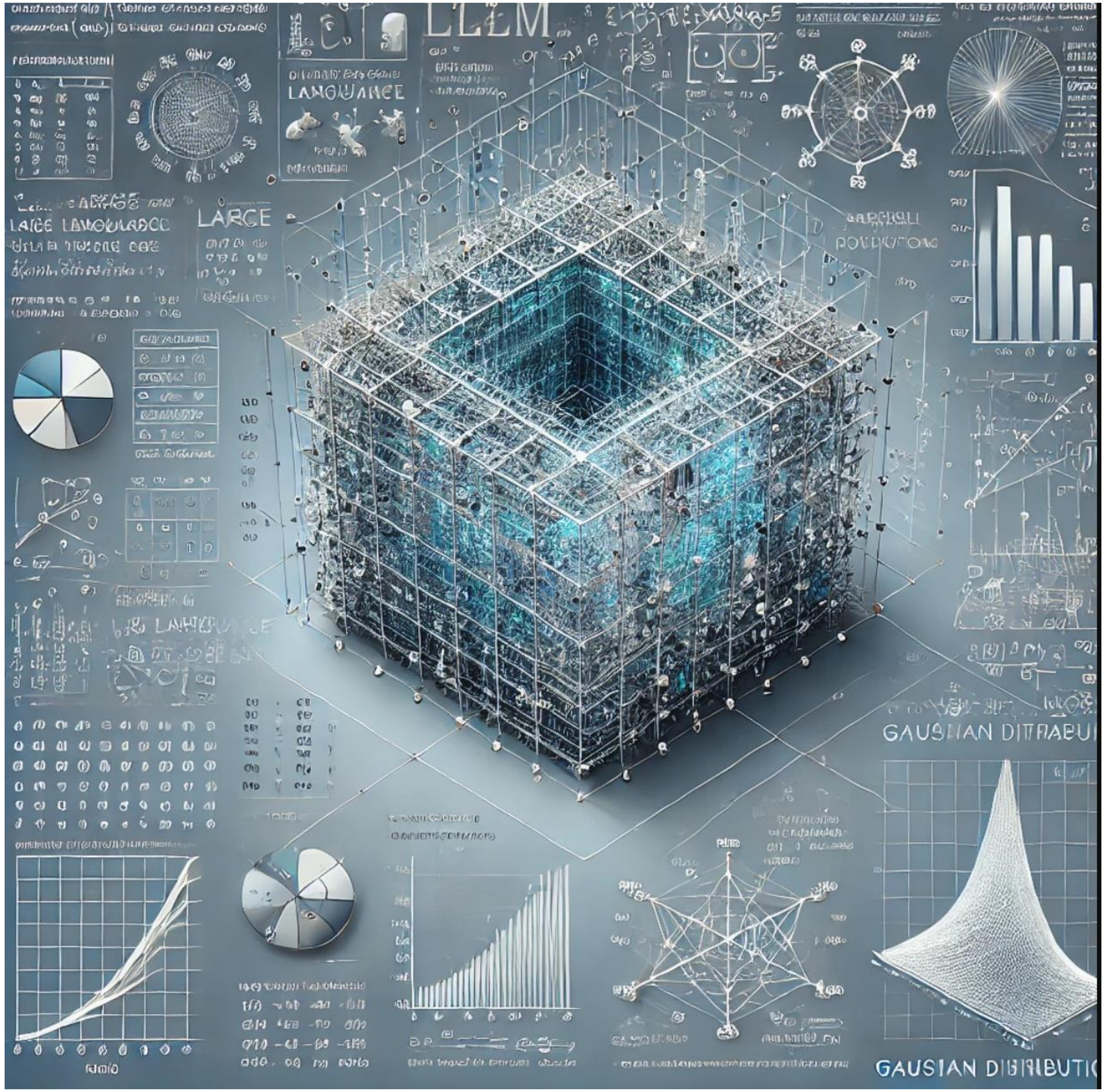
~ 30 K
DWH ETL | End of day select count

~ 100 K
DWH | Number of selects per day

Data Warehouse Approach



Vector Databases and Matrix Connections in LLMs



- **In the context of Large Language Models (LLMs), vector databases store embedding vectors, which constitute the numerical representations of textual data. Each text segment is encoded as a vector situated within a high-dimensional semantic space.**
- **These vectors are maintained in matrix form, and query operations are executed through mathematical manipulations applied to these matrices. Similarity assessments are most commonly performed using cosine similarity or the dot product, enabling efficient comparison between vector representations.**
- **Consequently, documents that exhibit the highest semantic proximity to a given query are retrieved rapidly through matrix-based computational procedures.**

Mathematical Foundations of Large Language Models

- **Large Language Models (LLMs) are fundamentally grounded in linear algebra, probability theory, and calculus.**
- • **Linear Algebra:** Matrix multiplication, tensor operations, and the updating of model parameters.
- • **Probability Theory:** The softmax function and the distribution of attention weights within attention mechanisms.
- • **Calculus:** Optimization and learning through gradient descent and related gradient-based techniques.
- **These mathematical foundations enable the model to learn semantic relationships and to generate coherent textual outputs.**


TOGETHER
EVERYONE
ACHIEVES
MORE