DATA LAKE OR DATA WAREHOUSE?
DATA CLEANING OR DATA WRANGLING?
HOW TO ENSURE THE QUALITY OF YOUR DATA?

Anastasija Nikiforova
Assistant Professor of Information Systems, Faculty of Science and Technology,
Institute of Computer Science, Chair of Software Engineering, University of Tartu
European Open Science CLoud (EOSC) Task Force “FAIR metrics and data quality”

Data Science Seminar: ***When, why and How? The important of Business Intelligence***

9.11.2022, Tartu, Estonia

BIO

PhD in Computer Science – Data Processing Systems and Data Networking

Research interests include but are not limited to data management with a focus on data quality, open government data, Smart City, Society 5.0, sustainable development, IoT, HCI, digitization.

Most recent experience:

- ✓ **Assistant professor at the University of Tartu, Faculty of Science and Technology, Institute of Computer Science, Chair of Software Engineering**
- ✓ **European Open Science Cloud Task Force "*FAIR Metrics and Data Quality*"**
- ✓ **associate member of the Latvian Open Technology Association.**
- ✓ **expert of the Latvian Council of Sciences in (1) *Natural Sciences – Computer Science & Informatics*, (2) *Engineering and Technology-Electrical Engineering, Electronics, ICT*, (3) *Social Sciences – Economics and Business***
- ✓ **expert of the *COST – European Cooperation in Science & Technology***
- ✓ **visiting researcher at the Delft University of Tehnology, Faculty Technology Policy and Management**
- ✓ **assistant professor at the Faculty of Computing, University of Latvia**
- ✓ **researcher in the Innovation Laboratory, Faculty of Computing, University of Latvia**
- ✓ **IT-expert at the Latvian Biomedical Research and Study Centre, BBMRI-ERIC LV National Node**
- ✓ **advisor for the Institute for Social and Political Studies, University of Latvia**

TARTU ÜLIKOOL · UNIVERSITY OF TARTU · 1632

EUROPEAN OPEN SCIENCE CLOUD

TU Delft
Delft University of Technology

LATA

DATA & POLICY

CAMBRIDGE
UNIVERSITY PRESS

JeDEM · eJOURNAL OF DEMOCRACY & OPEN GOVERNMENT

INTERNATIONAL JOURNAL ON
Semantic Web and Information Systems

Editorial Board Member for
BMC Research Notes
bmcresnotes.biomedcentral.com
BMC
Part of Springer Nature

Latvian Council of Science

BEFORE WE START…

Data Quality Management

- ✓ **Towards automating data quality specification by extracting data quality requirements from data features**
- ✓ **Data Deduplication (A Record Linkage-Based Data Deduplication Framework)**
- ✓ **Data Quality-aware Software Development**
- ✓ **User-oriented data object-driven approach to data quality assessment ⇒ DQMBT (Data Quality Model-Based Testing of IS)**
- ✓ **Data Lake and Data Wrangling for Ensuring Data Quality in CRIS**

Integrating artificial intelligence technologies into customer service: Improving the Actionability of Customer Feedback Analysis Using Machine Learning

BEFORE WE START…

Data Quality Management

- ✓ Towards automating data quality specification by extracting data quality requirements from data features
- ✓ Data Deduplication (A Record Linkage-Based Data Deduplication Framework)

- ✓ Data Quality-aware Software Development
- ✓ User-oriented data object-driven approach to data quality assessment ⇒ DQMBT (Data Quality Model-Based Testing of IS)
- ✓ Data Lake and Data Wrangling for Ensuring Data Quality in CRIS ➔

  Azeroual, O., Schöpfel, J., Ivanovic, D., & Nikiforova, A. (2022, May). Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS. In CRIS2022: 15th International Conference on Current Research Information Systems

Integrating artificial intelligence technologies into customer service: Improving the Actionability of Customer Feedback Analysis Using Machine Learning

BACKGROUND

Today, billions of data sources continuously generate, collect, process, and exchange data ⇒ with the rapid increase in the number of devices and IS, the amount and variety of data are increasing.

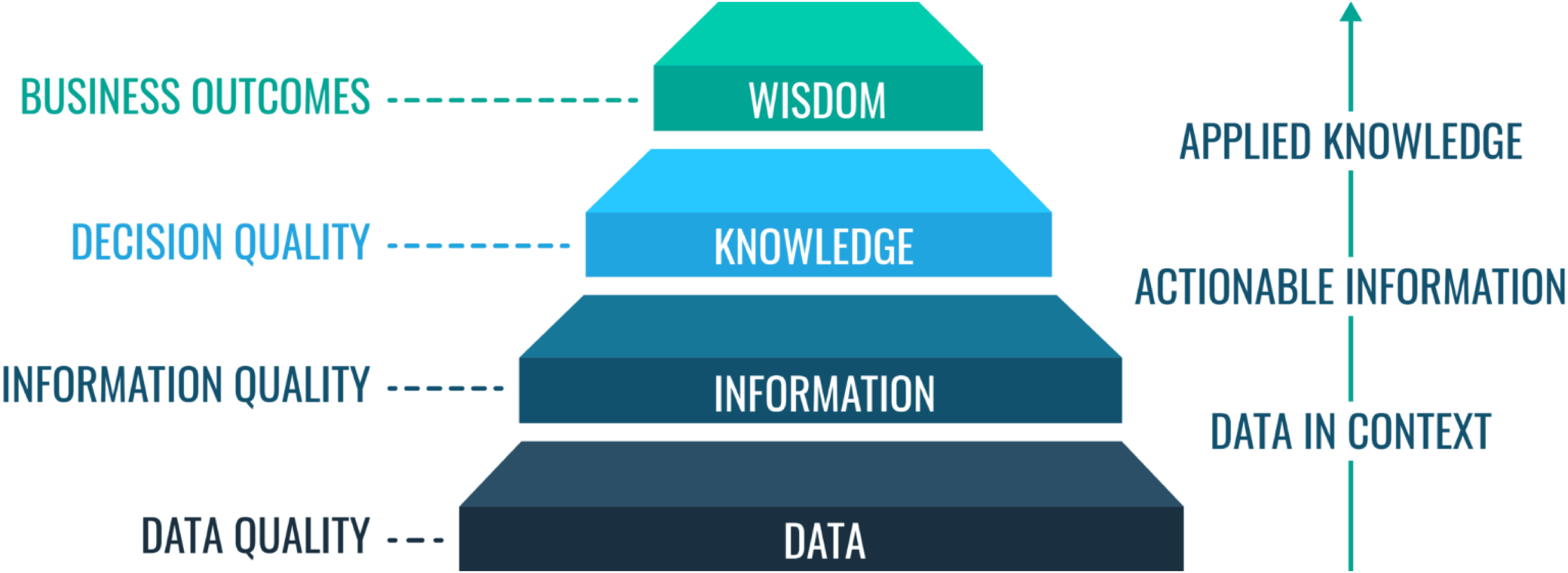
There is a need to integrate ever-increasing volumes of data, regardless of the source, format or amount, where the data quality, flexibility and scalability in connecting and processing different data sources are crucial.

⬇

an effective mechanism should be employed to ensure faster value creation from these data

DATA QUALITY

Why data quality? Again? and still?



BUSINESS OUTCOMES - - - - - - - - - - - - WISDOM

DECISION QUALITY - - - - - - - - - - KNOWLEDGE

INFORMATION QUALITY - - - - - - INFORMATION

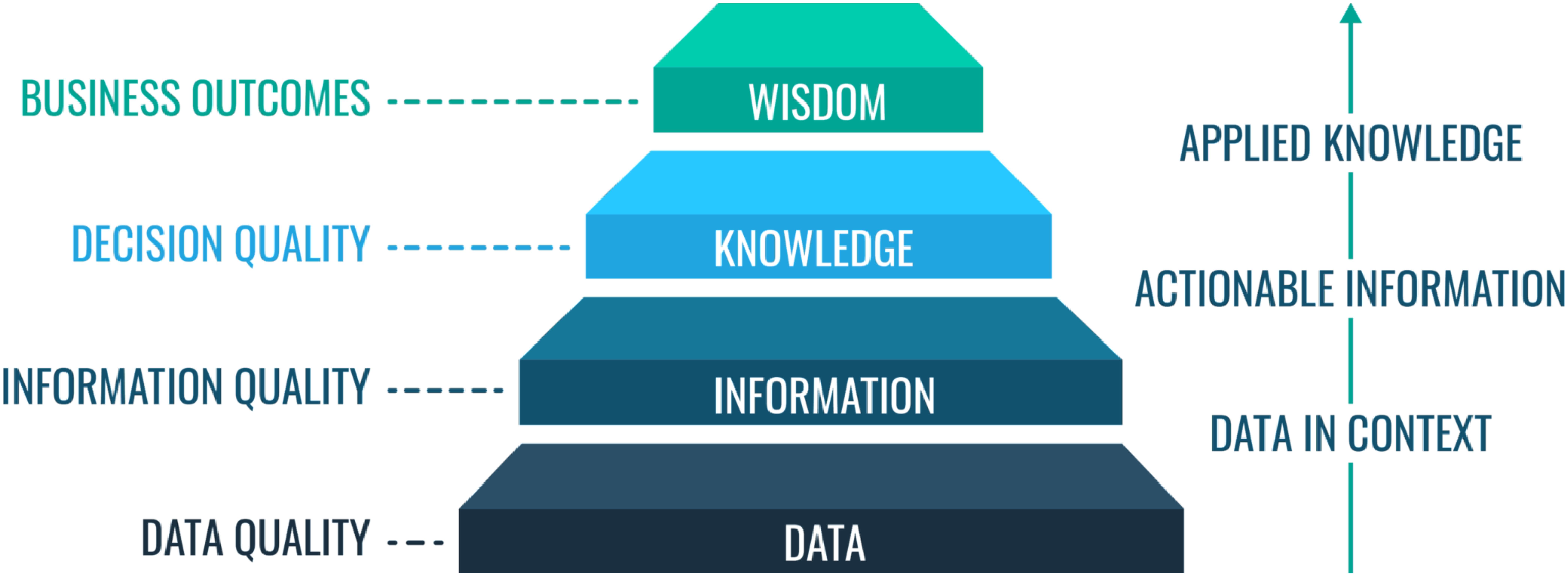
DATA QUALITY - - - DATA

APPLIED KNOWLEDGE

ACTIONABLE INFORMATION

DATA IN CONTEXT

[DATA QUALITY - WHAT, WHY, HOW, 10 BEST PRACTICES & MORE - Enterprise Master Data Management • Profisee]

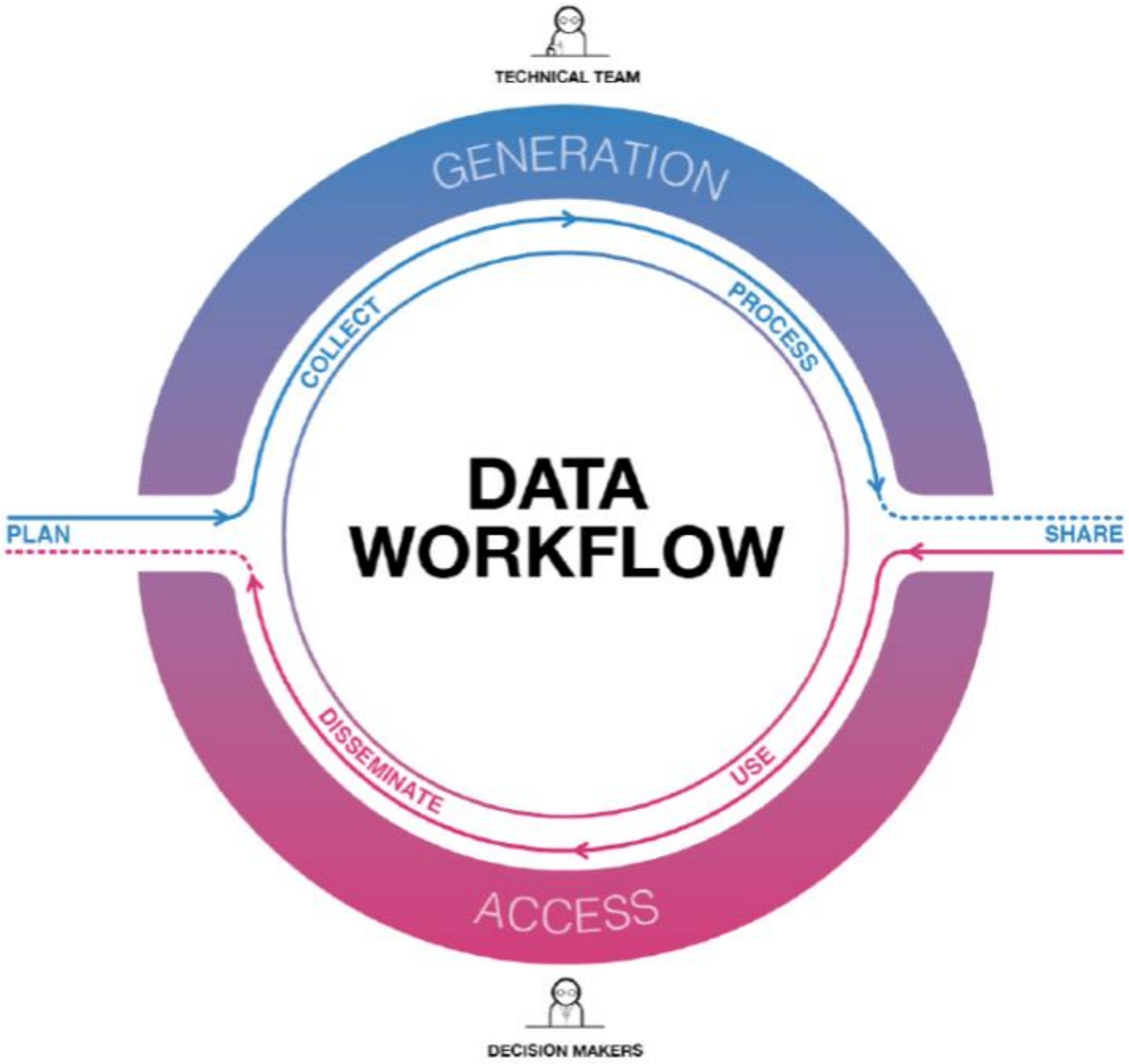


Among other "nuances", data quality is use-case dependent and dynamic (as well as relative) in nature!

***** "absolute data quality" == a level of data quality at which the data would satisfy all possible use cases - is not possible to achieve, but this is the objective to be pursued**

DATA QUALITY - WHAT, WHY, HOW, 10 BEST PRACTICES & MORE - Enterprise Master Data Management • Profisee

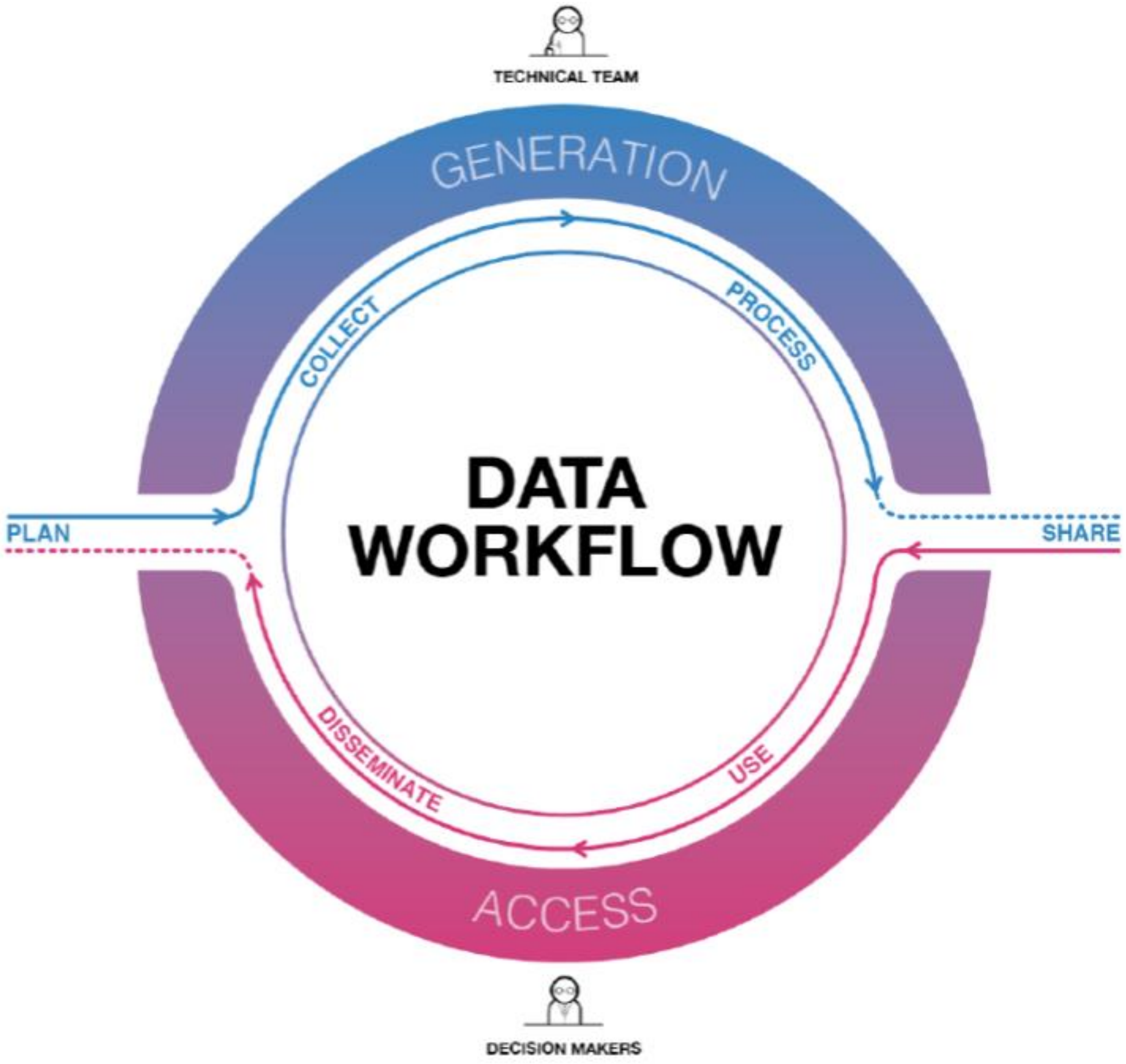
DATA REPOSITORY

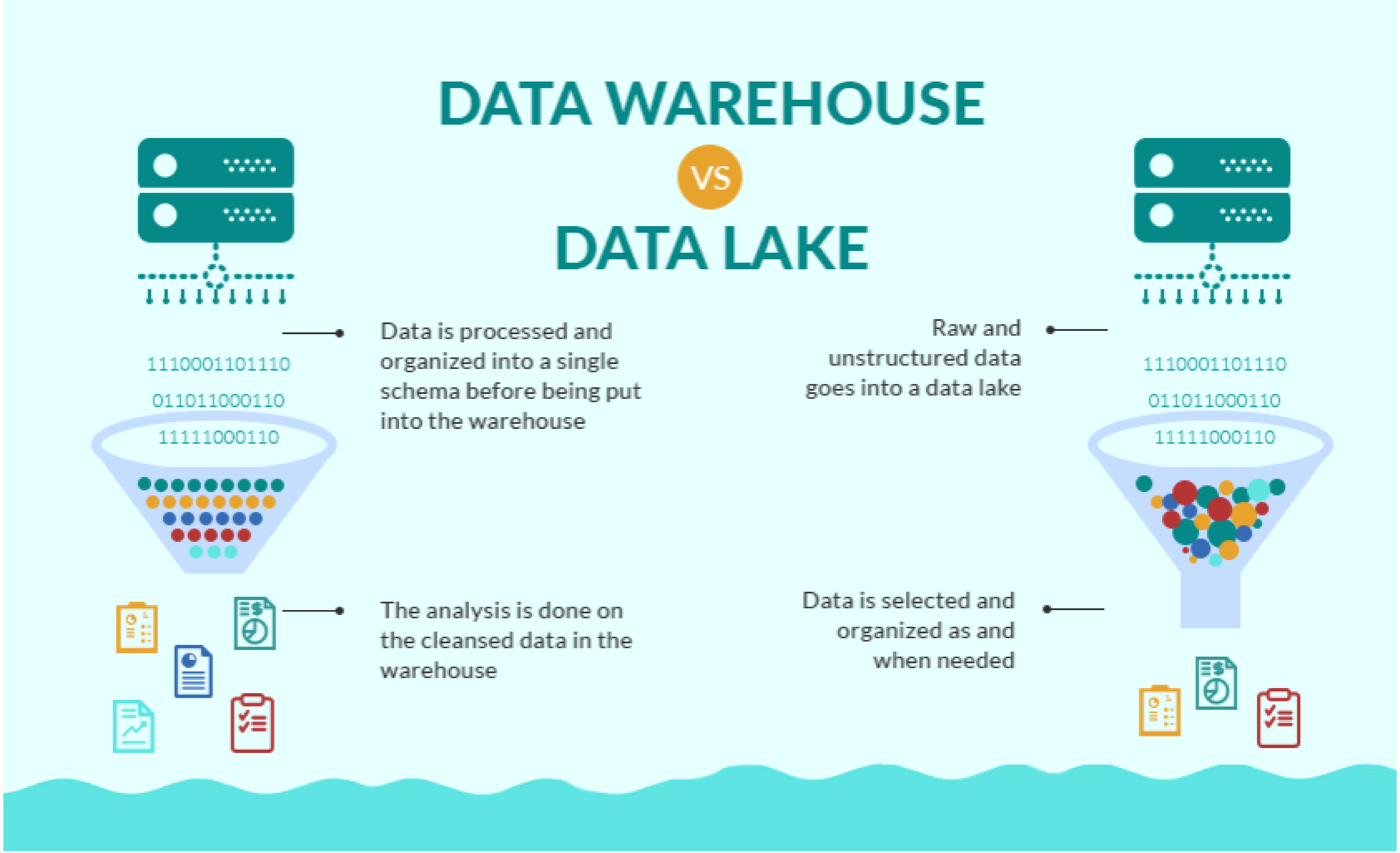


DATA WAREHOUSE

DATA LAKE?

Maybe even something more?





DATA WAREHOUSE VS DATA LAKE

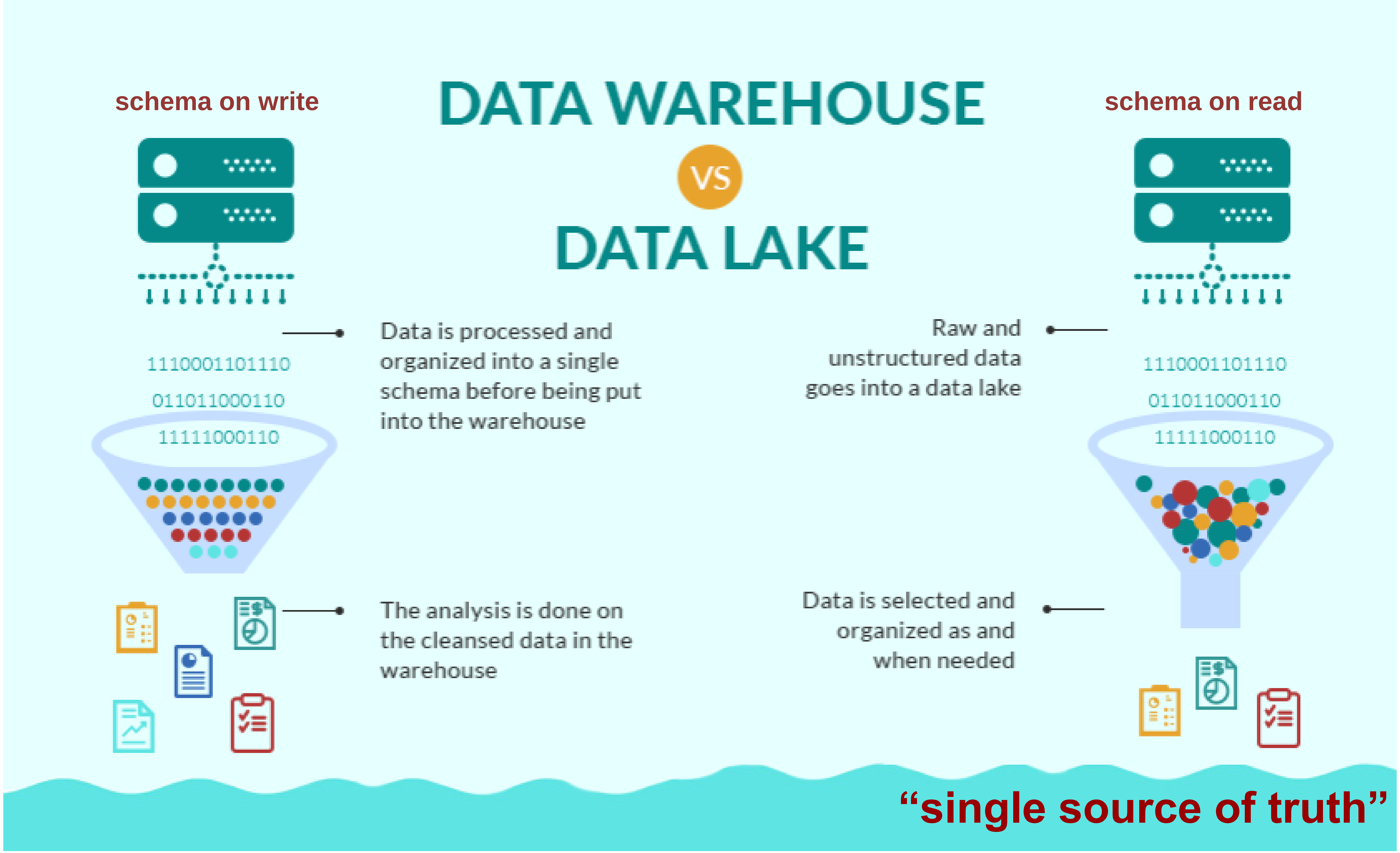
Data is processed and organized into a single schema before being put into the warehouse

The analysis is done on the cleansed data in the warehouse

Raw and unstructured data goes into a data lake

Data is selected and organized as and when needed

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/



DATA WAREHOUSE VS DATA LAKE

schema on write

Data is processed and organized into a single schema before being put into the warehouse

The analysis is done on the cleansed data in the warehouse

schema on read

Raw and unstructured data goes into a data lake

Data is selected and organized as and when needed

“single source of truth”

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/

DATA LAKE vs DATA WAREHOUSE

DATA LAKE

Data	Users	Use cases
unstructured	Data Scientists, Data Analysts	Stream Processing, Machine Learning, Real time analysis

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

DATA WAREHOUSE

Data	Users	Use cases
Structured	Business Analysts	Batch Processing, BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

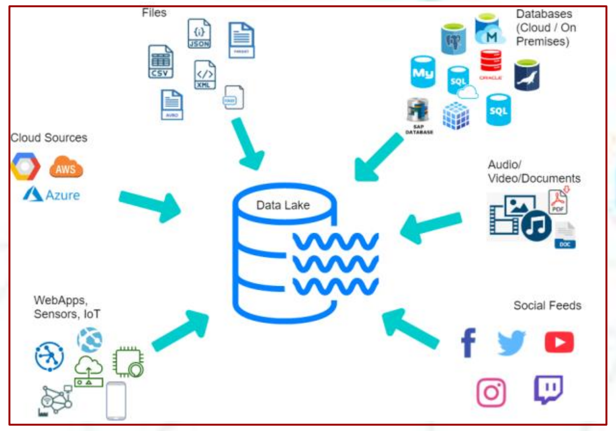
Data Warehouses contain historic and relational data, such as transaction systems, operations etc

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/

DATA LAKE vs DATA WAREHOUSE

DATA LAKE

Data



Users

Data Scientists,
Data Analysts

Use cases

Stream Processing,
Machine Learning,
Real time analysis

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

DATA WAREHOUSE

Data

Structured

Users

Business Analysts

Use cases

Batch Processing,
BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

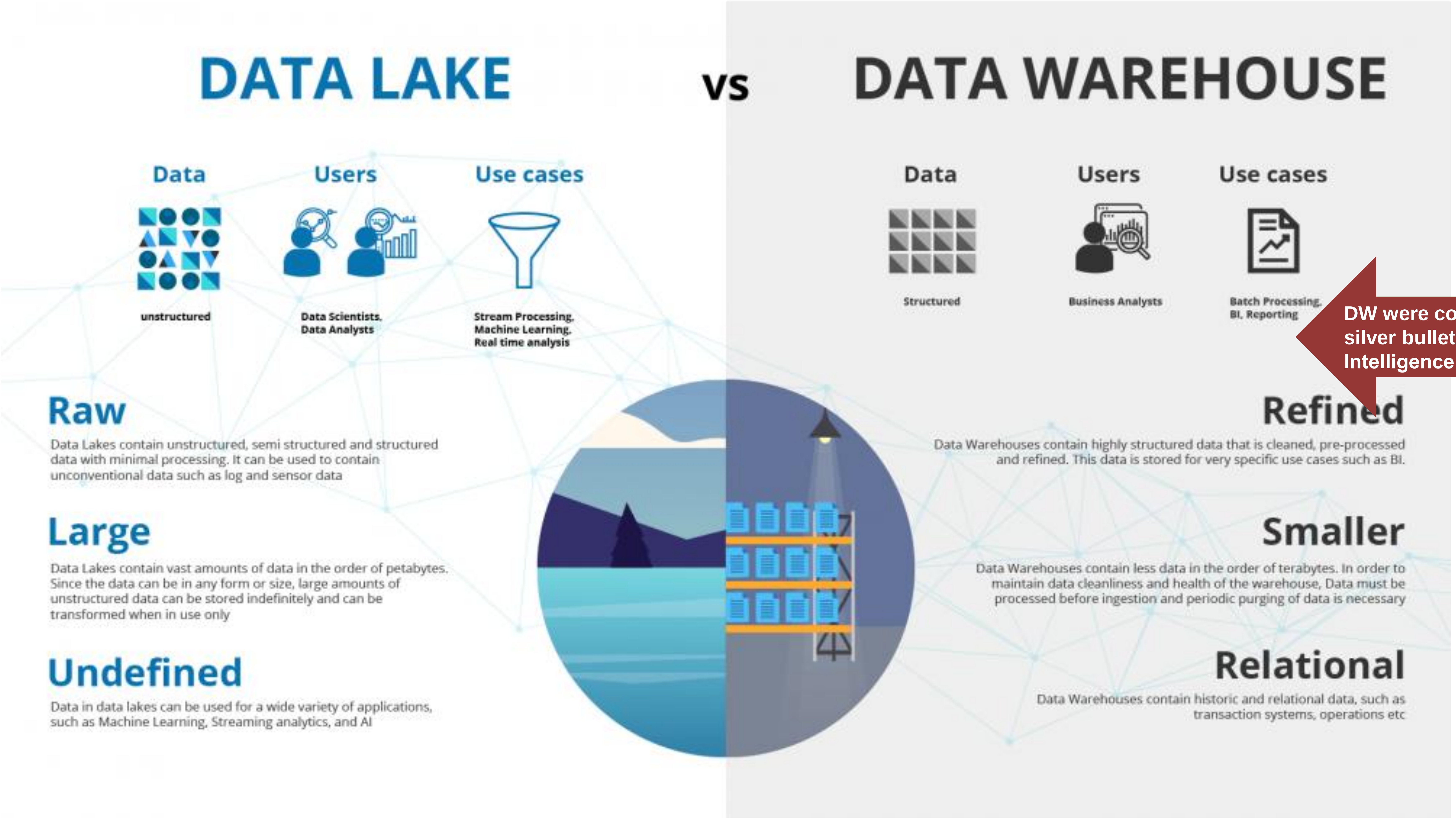
Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/



DATA LAKE vs DATA WAREHOUSE

DATA LAKE

Data	Users	Use cases
unstructured	Data Scientists, Data Analysts	Stream Processing, Machine Learning, Real time analysis

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

DATA WAREHOUSE

Data	Users	Use cases
Structured	Business Analysts	Batch Processing, BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

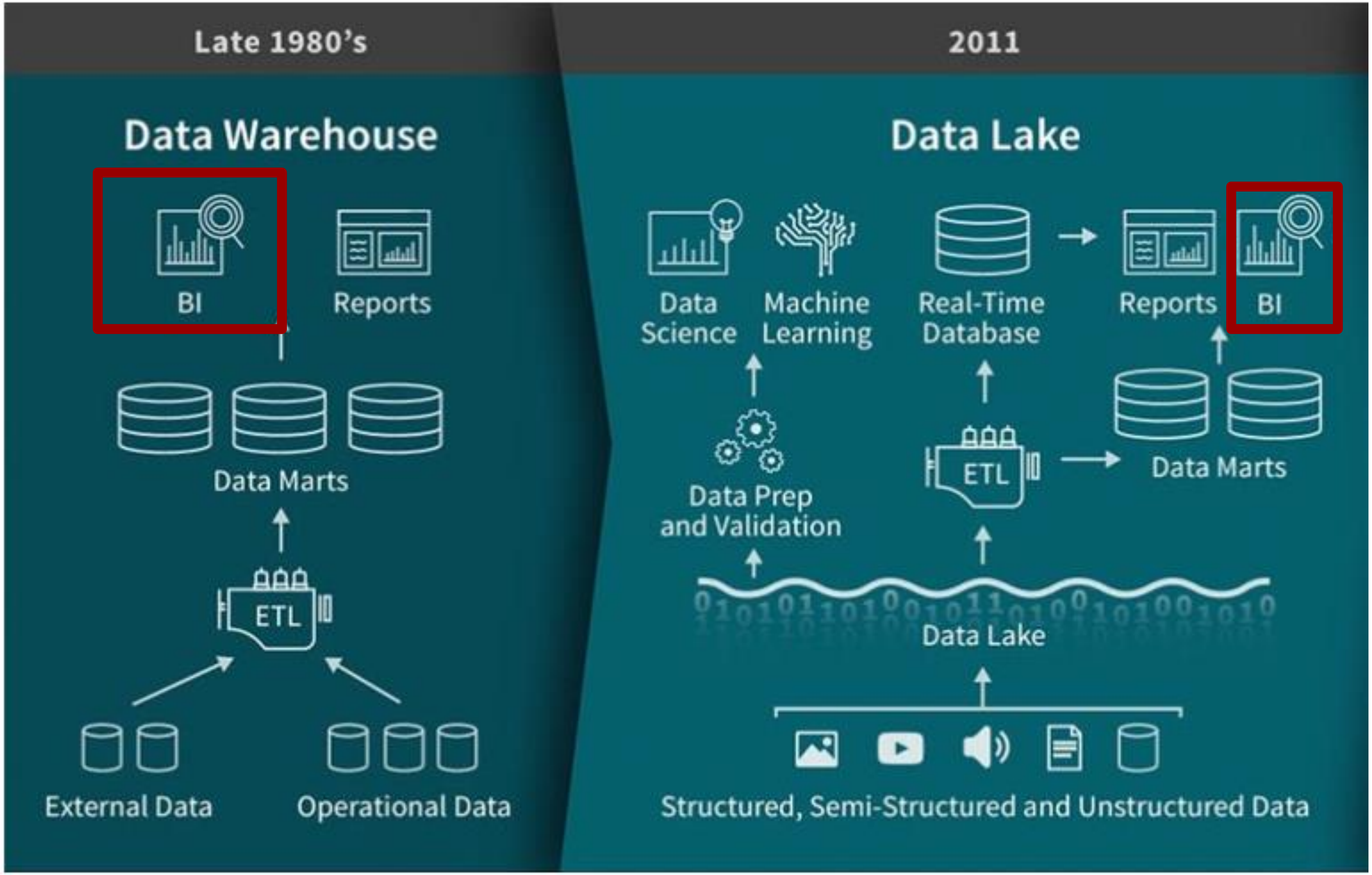
Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

DW were considered to be «a silver bullet» for Business Intelligence …

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/



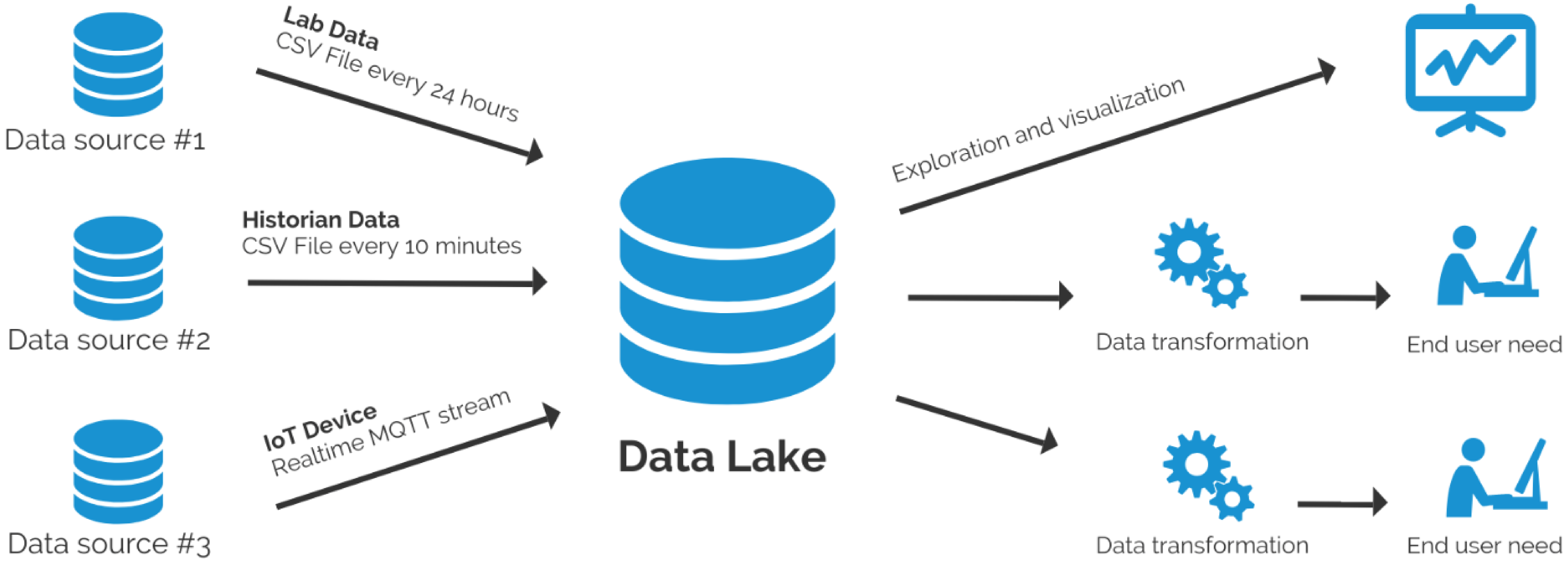
Implementing a Data Lake or Data Warehouse Architecture for Business Intelligence? | by Lan Chu | Towards Data Science

Data Swamp

- ✗ No metadata
- ✗ Broken metadata management
- ✗ No data governance
- ✗ Broken ingestion process

Data Lake

- ✓ Metadata
- ✓ Information is in rows and columns
- ✓ Easily ordered and processed with data mining tools
- ✓ Has data context
- ✓ Contains a data set for running analytics
- ✓ Has directories and sub-directories



So how to get its benefits?

Image source: https://towardsdatascience.com/augment-your-data-lake-analytics-with-snowflake-b417f1186615

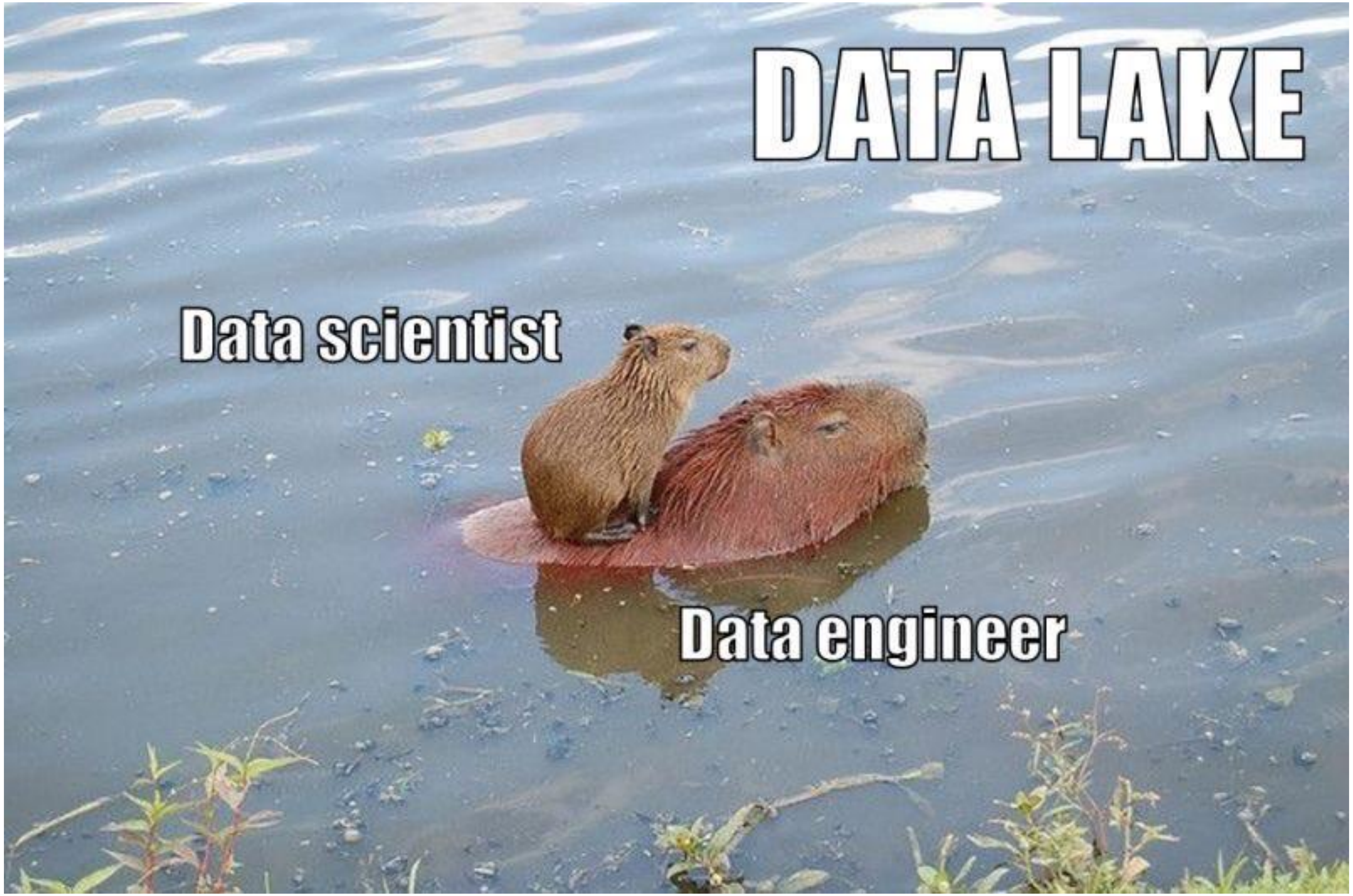
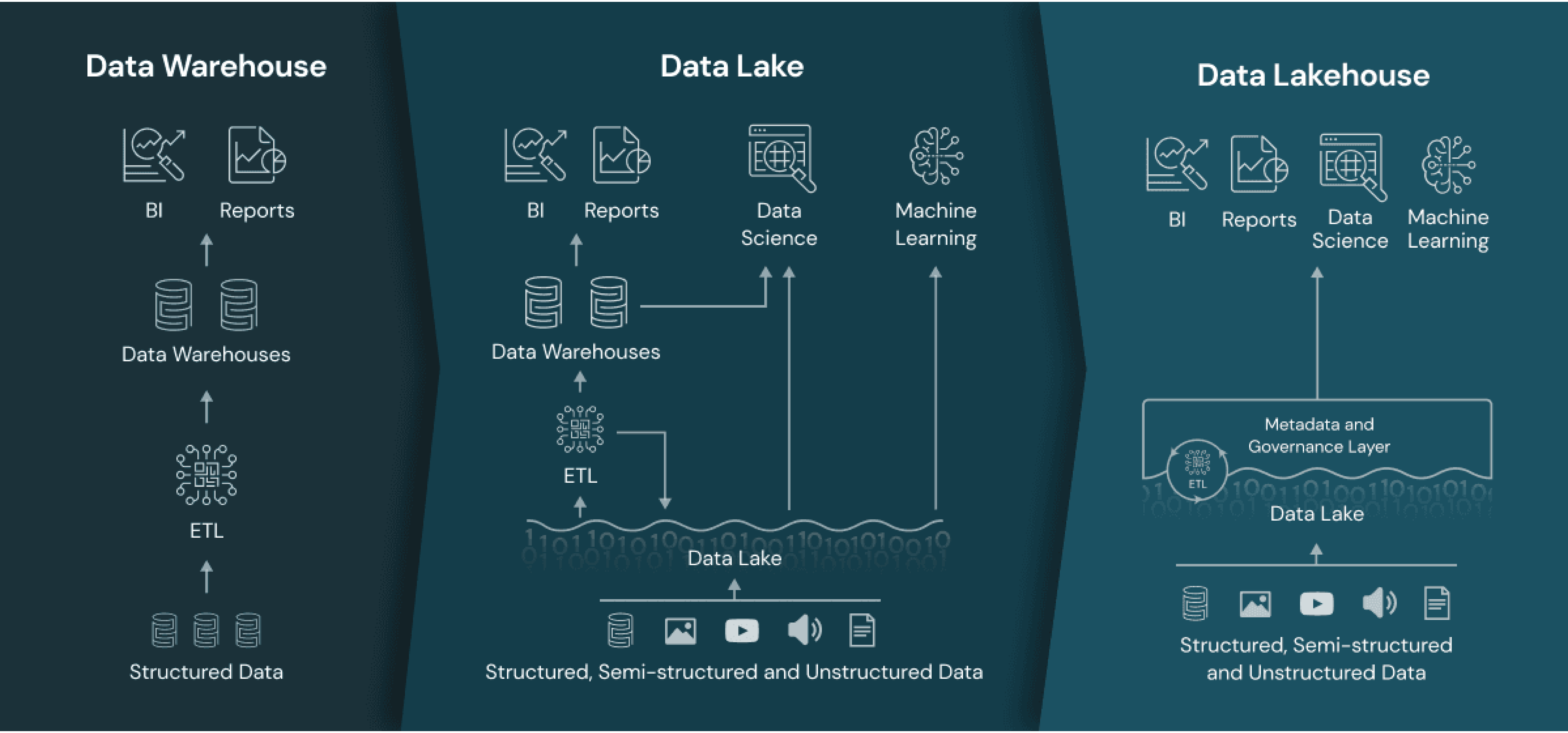


Image source: https://www.google.com/url?sa=i&url=https%3A%2F%2Ftwitter.com%2Frokar9%2Fstatus%2F1452339921629302784&psig=AOvVaw2IUSKtgUWxeaplk56f7CoK&ust=1668004535620000&source=images&cd=vfe&ved=0CA4QjhxqFwoTCJDHwbjnnvsCFQAAAAAdAAAAABAM

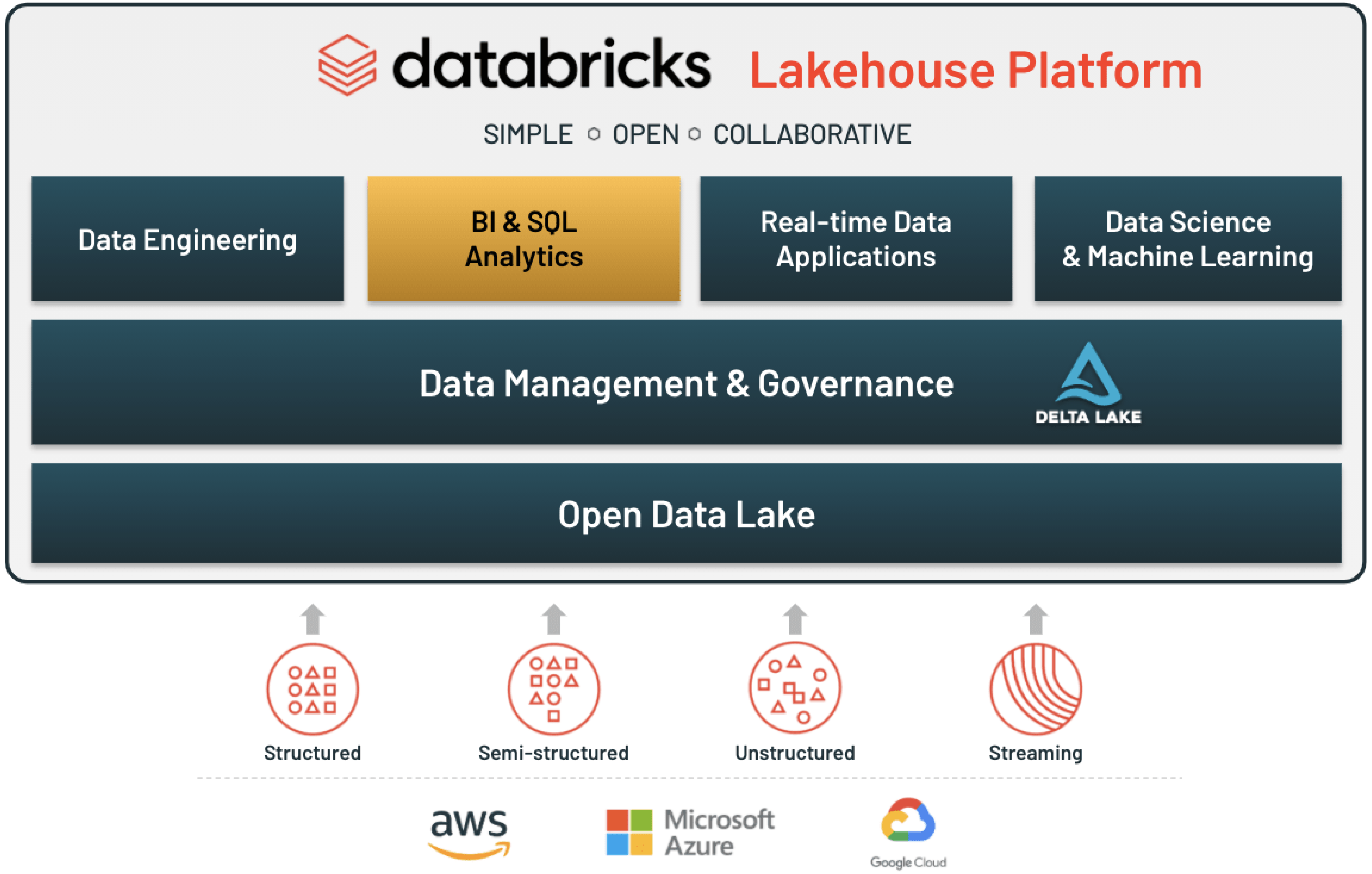
DATA LAKE & DATA WAREHOUSE

DATA LAKEHOUSE

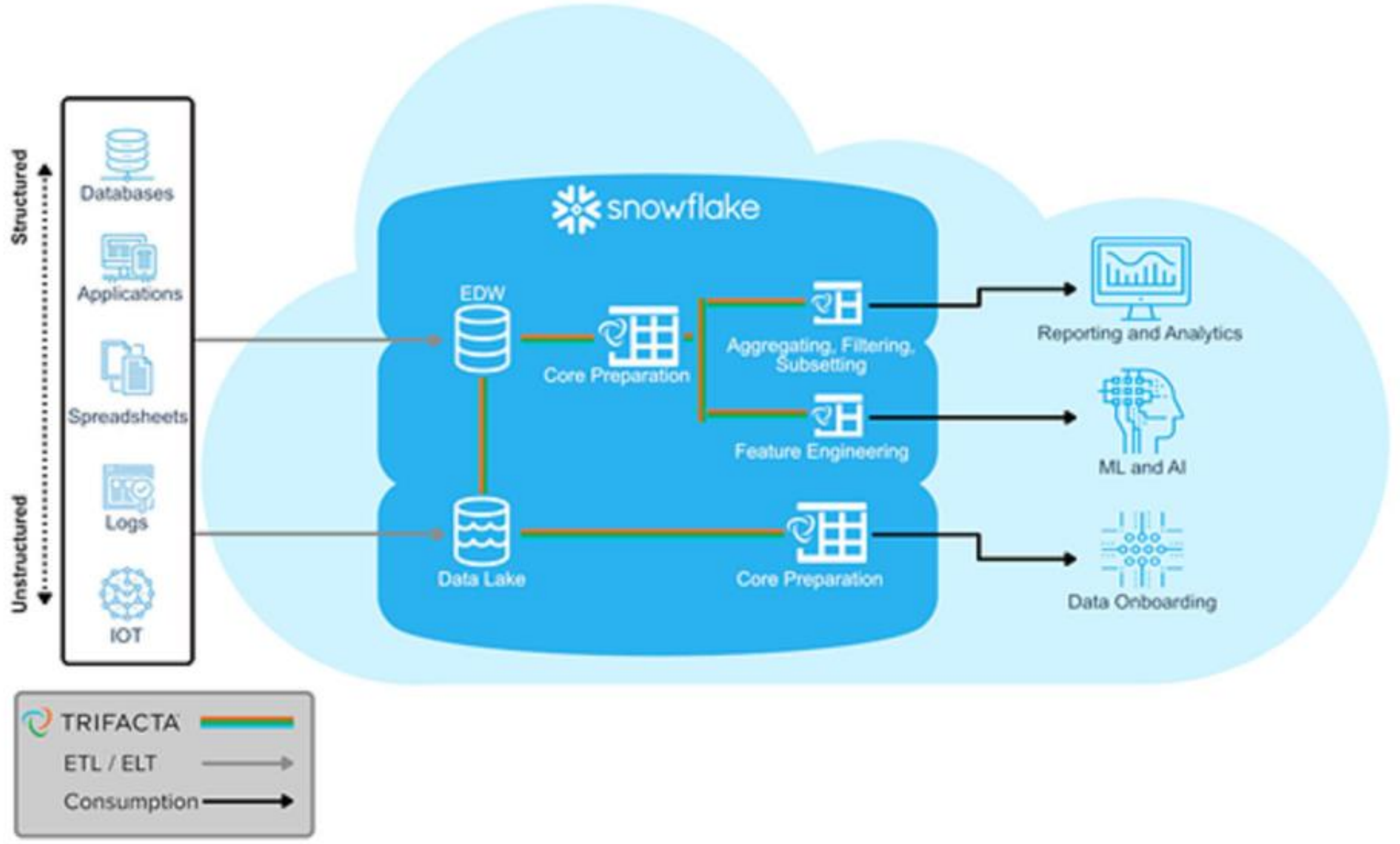


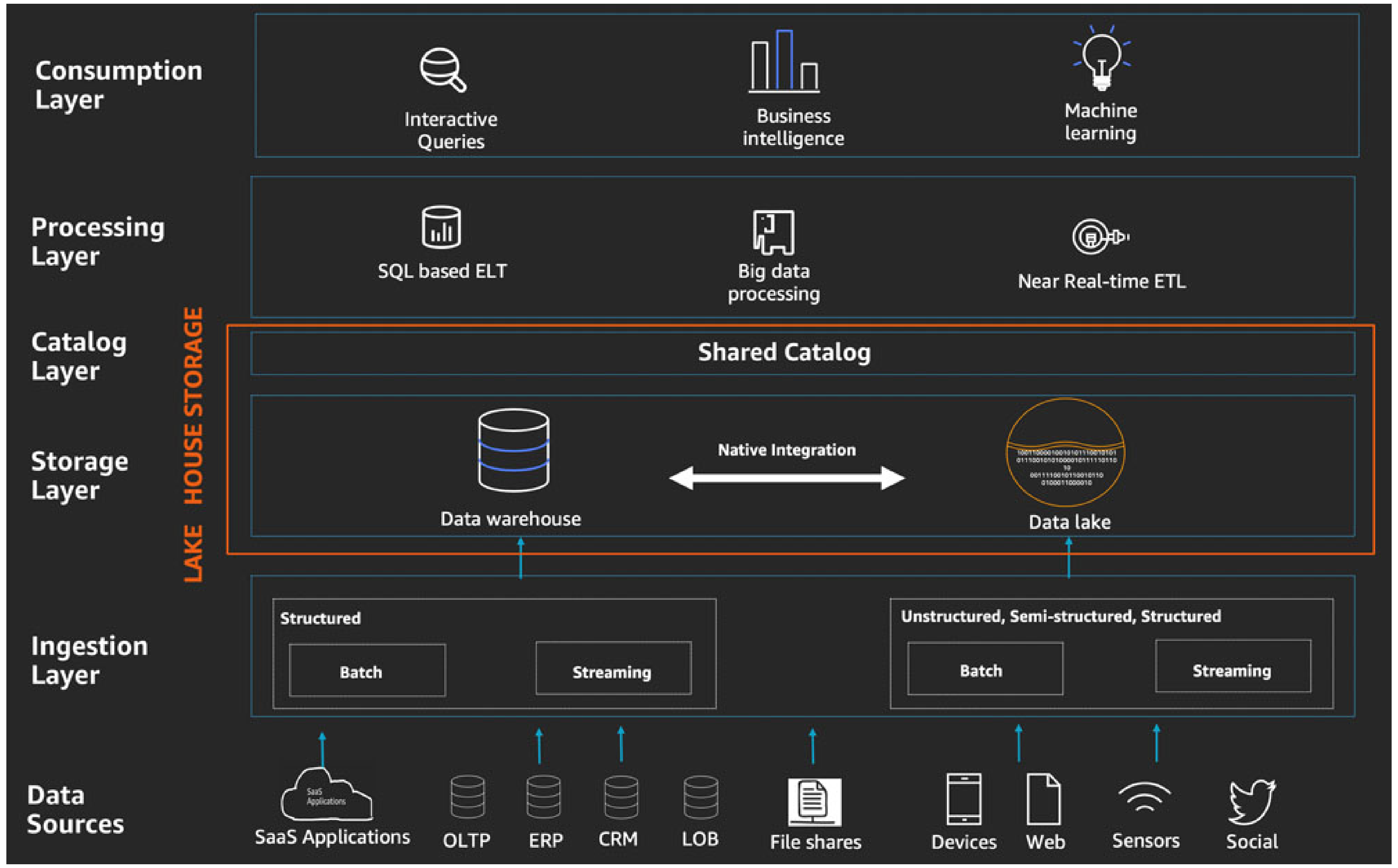
Data lakehouse is seen as a combination of data warehousing workloads & data lake economics

[Running Analytics on the Data Lake - The Databricks Blog]



databricks Lakehouse Platform
SIMPLE ○ OPEN ○ COLLABORATIVE
Data Engineering
BI & SQL Analytics
Real-time Data Applications
Data Science & Machine Learning
Data Management & Governance
DELTA LAKE
Open Data Lake
Structured
Semi-structured
Unstructured
Streaming
aws
Microsoft Azure
Google Cloud





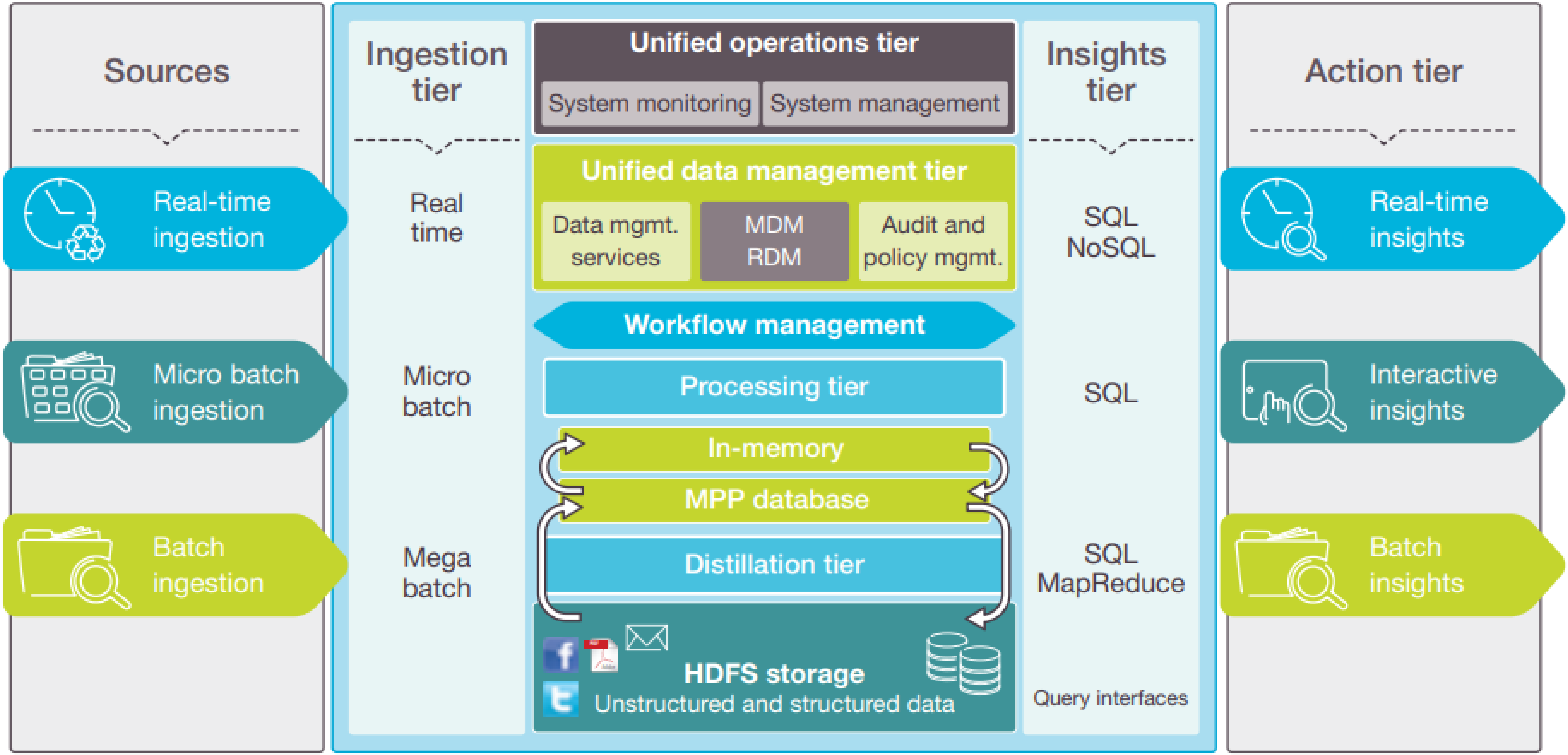
Running Analytics on the Data Lake - The Databricks Blog, Build a Lake House Architecture on AWS | AWS Big Data Blog (amazon.com), The Data Lakehouse, the Data Warehouse and a Modern Data platform architecture - Microsoft Community Hub

DATA LAKE FOR BUSINESS INTELLIGENCE

BUSINESS DATA LAKE

The Technology of the Business Data Lake

https://www.capgemini.com/wp-content/uploads/2017/07/pivotal_data_lake_vs_traditional_bi_20140805.pdf



Sources

Real-time ingestion

Micro batch ingestion

Batch ingestion

Ingestion tier

Real time

Micro batch

Mega batch

Unified operations tier

System monitoring

System management

Unified data management tier

Data mgmt. services

MDM RDM

Audit and policy mgmt.

Workflow management

Processing tier

In-memory

MPP database

Distillation tier

HDFS storage

Unstructured and structured data

Insights tier

SQL NoSQL

SQL

SQL MapReduce

Query interfaces

Action tier

Real-time insights

Interactive insights

Batch insights

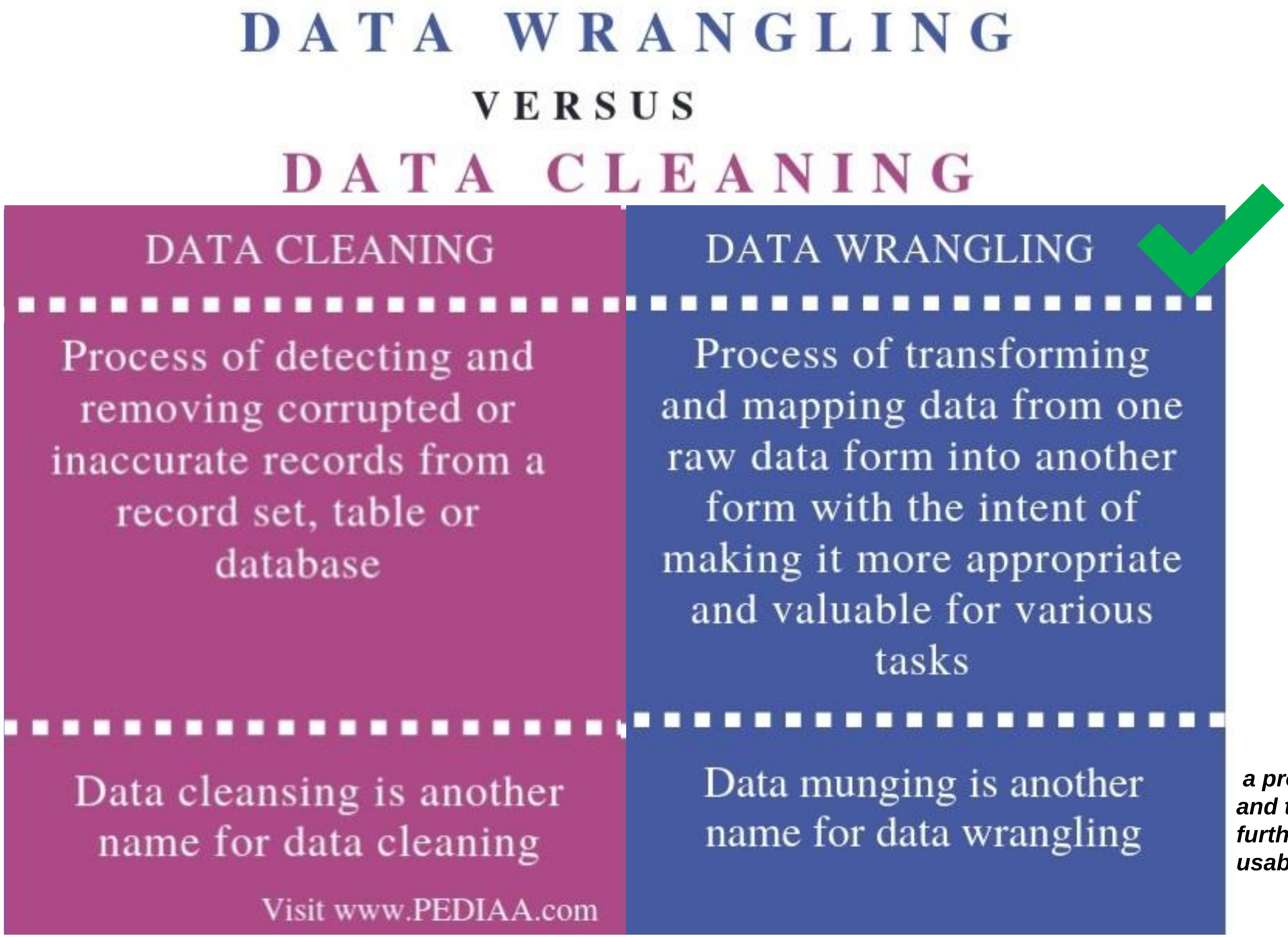
https://www.capgemini.com/wp-content/uploads/2017/07/pivotal_data_lake_vs_traditional_bi_20140805.pdf

Or how to avoid GIGO*?

***"garbage in, garbage out"**

Image source: The abstracted future of data engineering | by Justin Gage | Datalogue | Medium

DATA CLEANING or DATA WRANGLING?



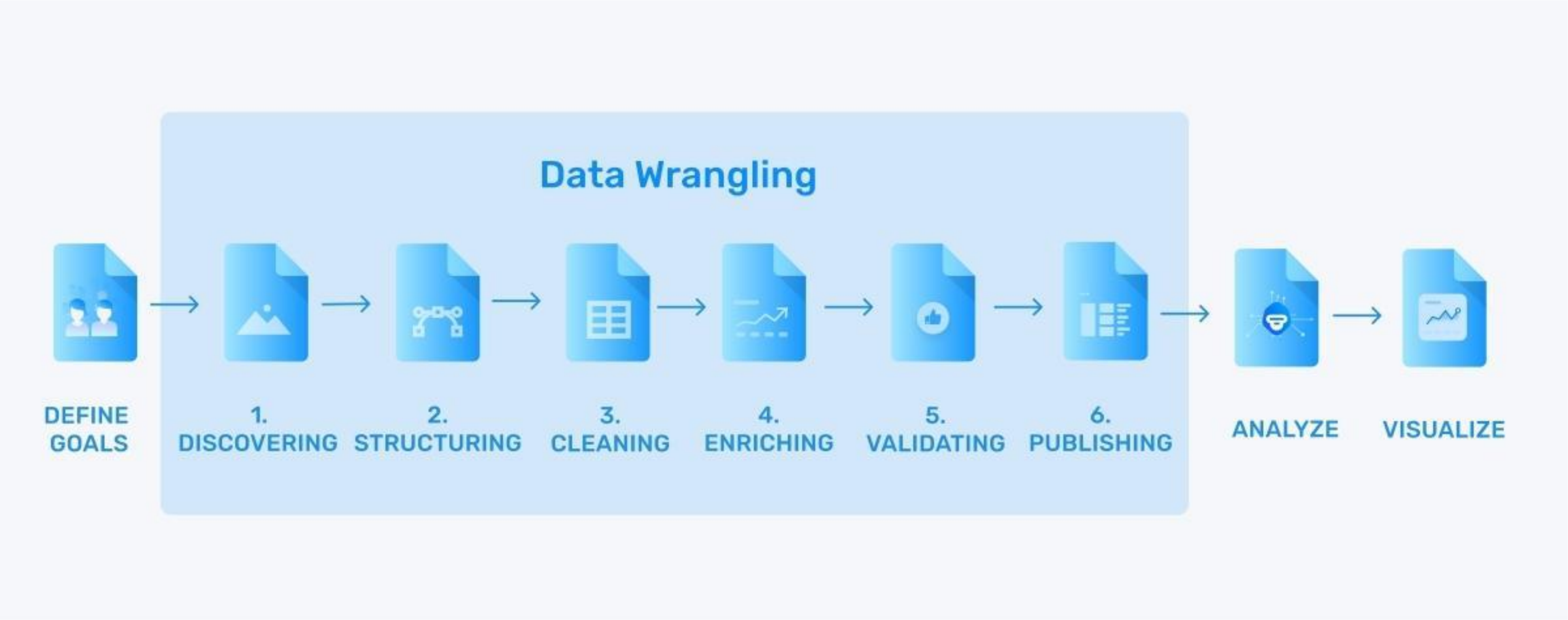
DATA WRANGLING VERSUS DATA CLEANING

DATA CLEANING	DATA WRANGLING ✓
Process of detecting and removing corrupted or inaccurate records from a record set, table or database	Process of transforming and mapping data from one raw data form into another form with the intent of making it more appropriate and valuable for various tasks
Data cleansing is another name for data cleaning	Data munging is another name for data wrangling

Visit www.PEDIAA.com

a process of iterative data exploration and transformation that enables their further analysis by making them (1) usable, (2) credible and (3) useful

https://pediaa.com/what-is-the-difference-between-data-wrangling-and-data-cleaning/



- ➢ **The nature of data lake allows to store a variety of data within the memory**

BUT

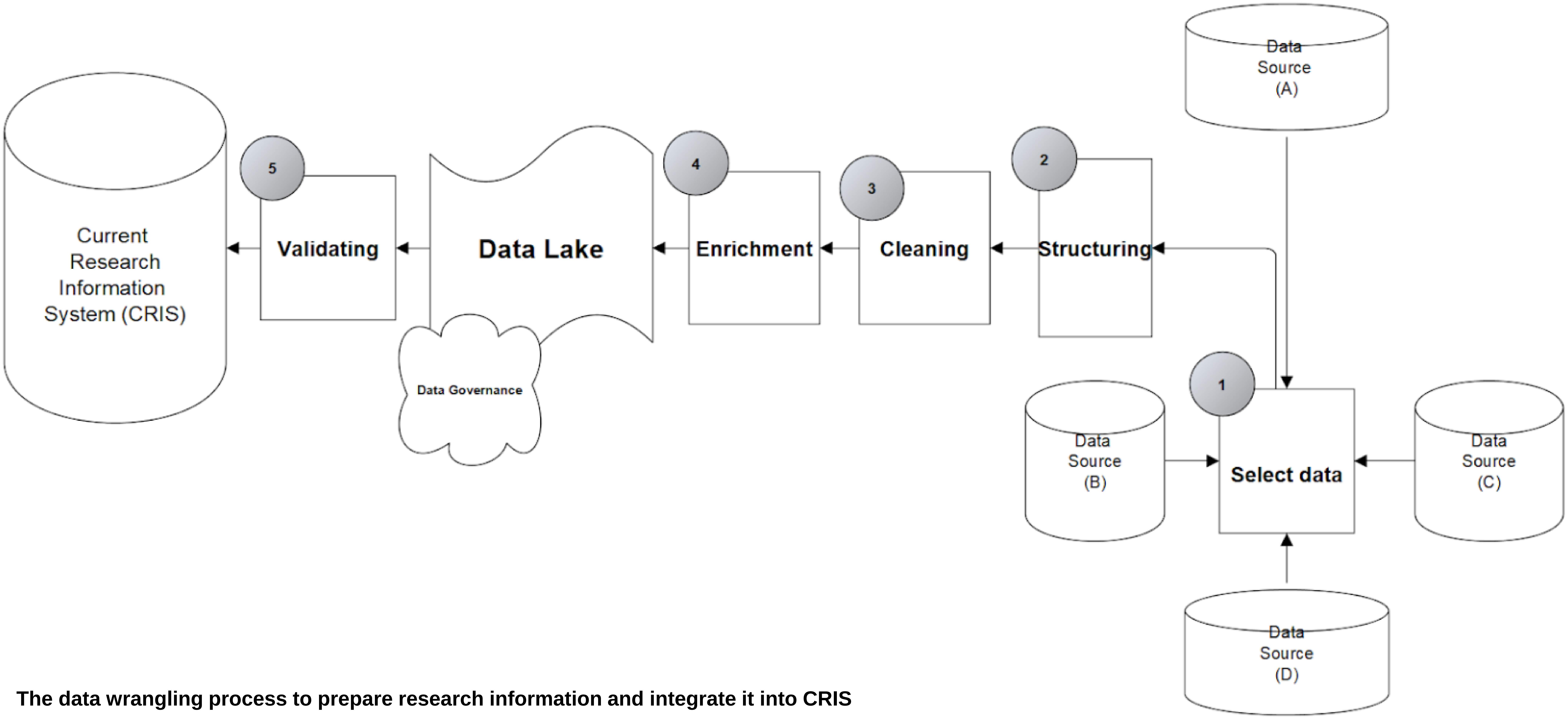
- ➢ **there is a need to clean up dirty data and enrich them in a pre-processing process, where data wrangling is found to be suitable for these purposes.**

- ➢ **The goal is to convert complex data types and data formats into structured data without programming efforts ➔ users should be able to prepare and transform their data without the need of using the ETL tools or familiarity and use of programming languages, where these transformations should be automatically suggested after reading the data based on machine learning algorithms that greatly speeds up this process.**

Image source: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.ecloudvalley.com%2Fwhat-is-datalake-and-datawarehouse%2F&psig=AOvVaw1vGuz42Qfu2-0J_bpmhpbJ&ust=1651908882490000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCNiaquD6yvcCFQAAAAAdAAAAABAO
Source: https://monkeylearn.com/blog/data-wrangling/, https://www.altair.com/what-is-data-wrangling/

DATA LAKE + DATA WRANGLING
=
DATA QUALITY IN IS

[an asset, not a silver bullet]



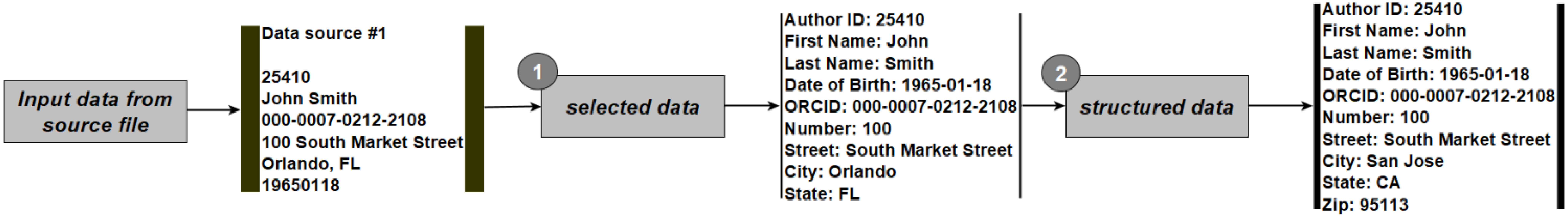
The data wrangling process to prepare research information and integrate it into CRIS

Depending on the IS and the <u>desired or required target quality*</u>, individual steps should be carried out several times ➔ data wrangling is a continuous process that repeats itself repeatedly at regular intervals.

Step	Description
Select data	**The required data records are identified in different data sources. When selecting data, a record is evaluated by its value □ if there is added value, the availability and terms of use of the data and subsequent data from this data source are checked**
Structure	**In most cases, there is little or no structure in the data □ change the structure of the data for easier accessibility.**
Clean	**Almost every dataset contains some outliers that can skew the analysis results □ the data are extensively cleaned for better analysis *(processing of null values, removing duplicates and special characters, and standardization of the formatting to improve data consistency)***
Enrich	**The data needs to be enriched - an inventory of the data set and a strategy for improving it by adding additional data should be carried out. The data set is enriched with various metadata:** ✓ **Schematic metadata provide basic information about the processing and ingestion of data □ the data wrangler analyzes / parses data records according to an existing schema.** ✓ **Conversation metadata are exchanged between accessing instances with the idea to document information obtained during the processing or analysis of these data for subsequent users.** **The recognized peculiarities/ features of a data set can be saved.**
Data lake	**The physical transfer of data in the data lake. Although data are prepared using metadata, the record is not pre-processed.** **The goal is to avoid a data swamp □ estimate the value of the data and decide on their lifespan depending on the data quality and its interconnectedness with the rest of the DB.** **Analyzes are not performed directly in the data lake, but only on the relevant data. To be able to use the data, the requester needs the appropriate access rights □ Data Wrangler performs data extraction, however, general viewing and exploration of the data should be possible directly in the data lake.**
Data governance	**The contents of the data lake, technologies and hardware used are subject to change □ an audit is required to take care of the care and maintenance of the data lake. The main principles / guidelines and measures that regulates data maintenance coordinating all processes in the data lake and responsibilities are defined**
Validate	**the data are checked one more time before they are integrated into the target CRIS to identify problems with the data quality and consistency of the data, or to confirm that the transformation has been successful.** **Verify that the values of the attribute are correct and conform to the syntactic and distribution constraints, thus ensuring high data quality AND document every change so that older versions can be restored, or history of changes can be viewed. If new data are generated during data analysis in CRIS, it can be re-included in Data Lake**** *****New data go through the data wrangling process, starting with the step 2 of data validating and structuring the data.***

At the end of this process, research information can be used by analytical applications and protected from unauthorized access by access control

USE-CASE



Input data from source file

Data source #1

25410
John Smith
000-0007-0212-2108
100 South Market Street
Orlando, FL
19650118

1 *selected data*

Author ID: 25410
First Name: John
Last Name: Smith
Date of Birth: 1965-01-18
ORCID: 000-0007-0212-2108
Number: 100
Street: South Market Street
City: Orlando
State: FL

2 *structured data*

Author ID: 25410
First Name: John
Last Name: Smith
Date of Birth: 1965-01-18
ORCID: 000-0007-0212-2108
Number: 100
Street: South Market Street
City: San Jose
State: CA
Zip: 95113

✓ **Data formatting**

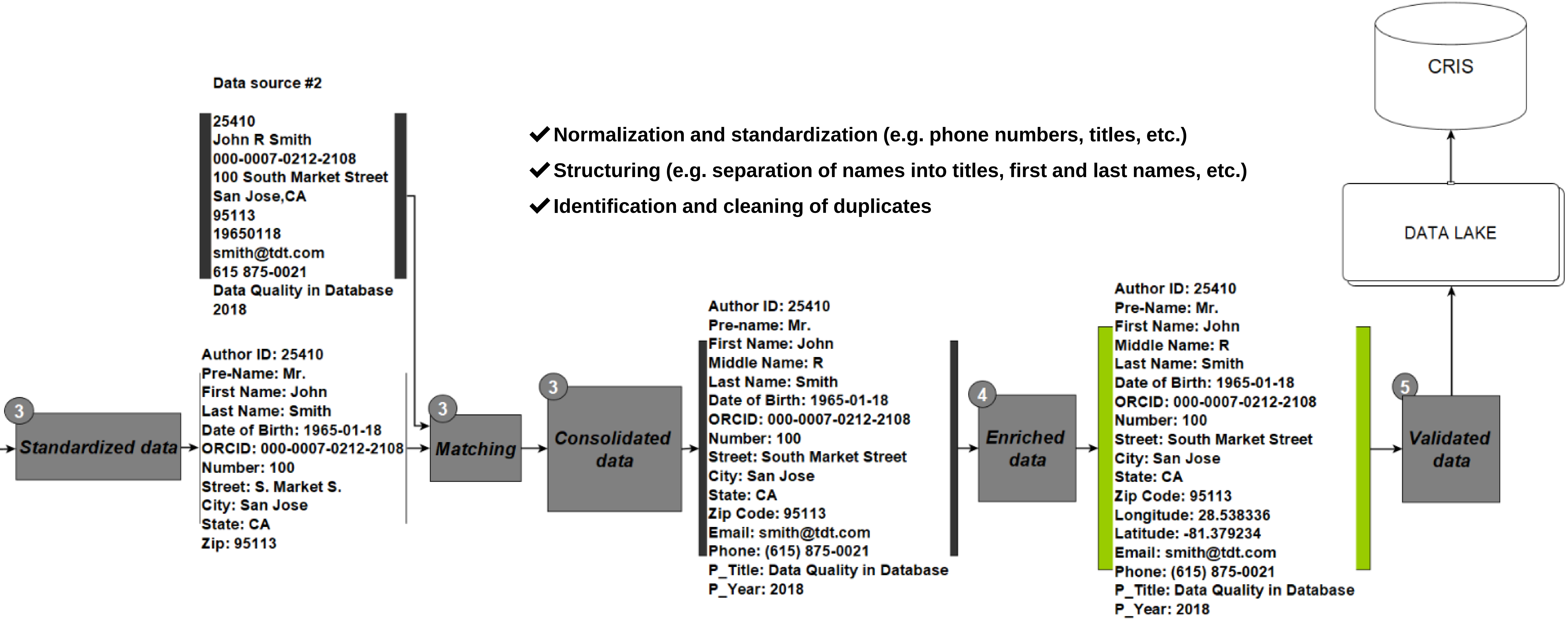
✓ **Correction of incorrect data (e.g. address data)**

USE-CASE

✓ Normalization and standardization (e.g. phone numbers, titles, etc.)

✓ Structuring (e.g. separation of names into titles, first and last names, etc.)

✓ Identification and cleaning of duplicates



Data source #2

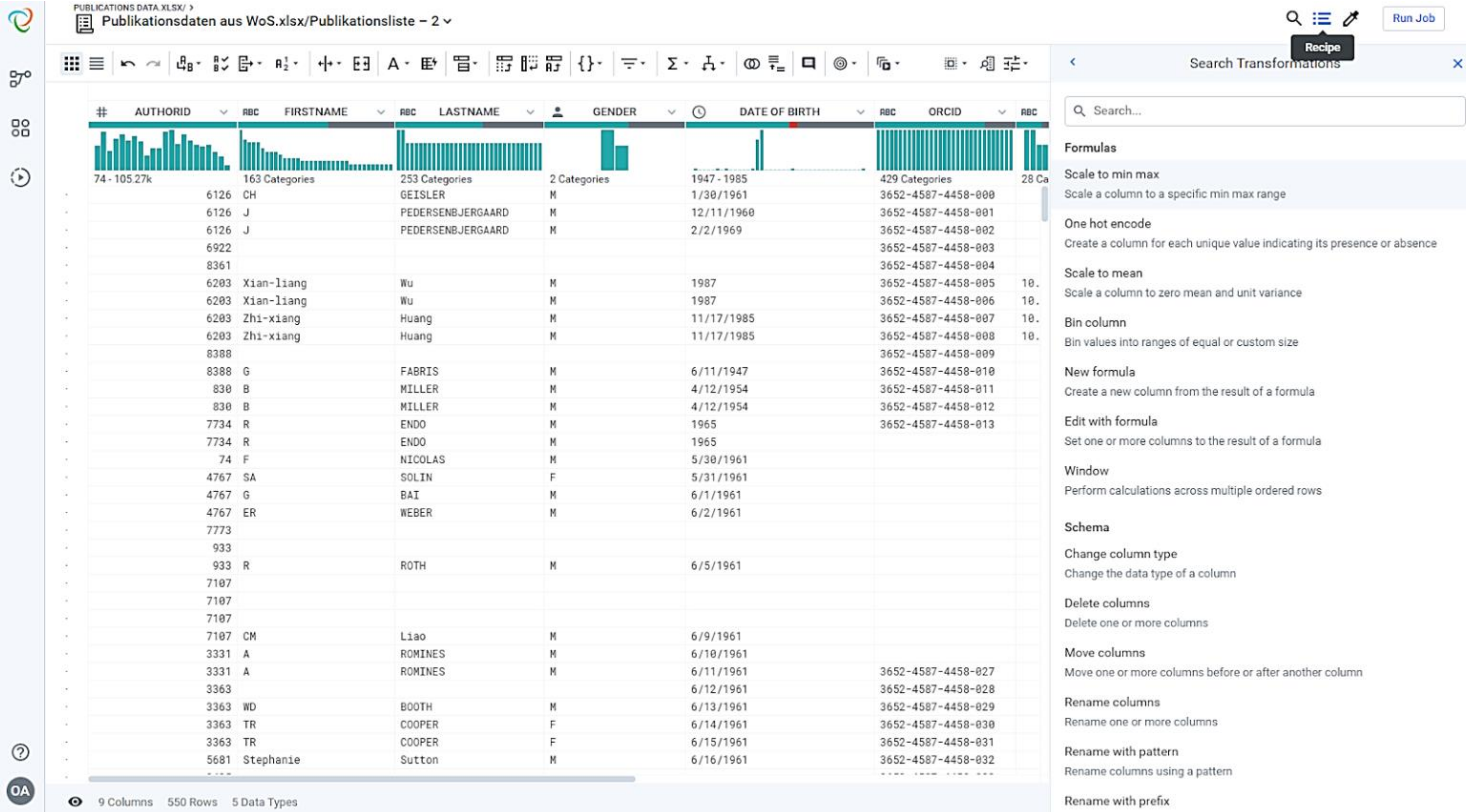
25410
John R Smith
000-0007-0212-2108
100 South Market Street
San Jose,CA
95113
19650118
smith@tdt.com
615 875-0021
Data Quality in Database
2018

Author ID: 25410
Pre-Name: Mr.
First Name: John
Last Name: Smith
Date of Birth: 1965-01-18
ORCID: 000-0007-0212-2108
Number: 100
Street: S. Market S.
City: San Jose
State: CA
Zip: 95113

Author ID: 25410
Pre-name: Mr.
First Name: John
Middle Name: R
Last Name: Smith
Date of Birth: 1965-01-18
ORCID: 000-0007-0212-2108
Number: 100
Street: South Market Street
City: San Jose
State: CA
Zip Code: 95113
Email: smith@tdt.com
Phone: (615) 875-0021
P_Title: Data Quality in Database
P_Year: 2018

Author ID: 25410
Pre-Name: Mr.
First Name: John
Middle Name: R
Last Name: Smith
Date of Birth: 1965-01-18
ORCID: 000-0007-0212-2108
Number: 100
Street: South Market Street
City: San Jose
State: CA
Zip Code: 95113
Longitude: 28.538336
Latitude: -81.379234
Email: smith@tdt.com
Phone: (615) 875-0021
P_Title: Data Quality in Database
P_Year: 2018

USE-CASE: TRIFACTA FOR DATA WRANGLING



Publikationsdaten aus WoS.xlsx/Publikationsliste – 2

AUTHORID	FIRSTNAME	LASTNAME	GENDER	DATE OF BIRTH	ORCID	
74 - 105.27k	163 Categories	253 Categories	2 Categories	1947 - 1985	429 Categories	28 Ca
6126	CH	GEISLER	M	1/30/1961	3652-4587-4458-000	
6126	J	PEDERSENBJERGAARD	M	12/11/1960	3652-4587-4458-001	
6126	J	PEDERSENBJERGAARD	M	2/2/1969	3652-4587-4458-002	
6922					3652-4587-4458-003	
8361					3652-4587-4458-004	
6203	Xian-liang	Wu	M	1987	3652-4587-4458-005	10.
6203	Xian-liang	Wu	M	1987	3652-4587-4458-006	10.
6203	Zhi-xiang	Huang	M	11/17/1985	3652-4587-4458-007	10.
6203	Zhi-xiang	Huang	M	11/17/1985	3652-4587-4458-008	10.
8388					3652-4587-4458-009	
8388	G	FABRIS	M	6/11/1947	3652-4587-4458-010	
830	B	MILLER	M	4/12/1954	3652-4587-4458-011	
830	B	MILLER	M	4/12/1954	3652-4587-4458-012	
7734	R	ENDO	M	1965	3652-4587-4458-013	
7734	R	ENDO	M	1965		
74	F	NICOLAS	M	5/30/1961		
4767	SA	SOLIN	F	5/31/1961		
4767	G	BAI	M	6/1/1961		
4767	ER	WEBER	M	6/2/1961		
7773						
933						
933	R	ROTH	M	6/5/1961		
7107						
7107						
7107						
7107	CM	Liao	M	6/9/1961		
3331	A	ROMINES	M	6/10/1961		
3331	A	ROMINES	M	6/11/1961	3652-4587-4458-027	
3363				6/12/1961	3652-4587-4458-028	
3363	WD	BOOTH	M	6/13/1961	3652-4587-4458-029	
3363	TR	COOPER	F	6/14/1961	3652-4587-4458-030	
3363	TR	COOPER	F	6/15/1961	3652-4587-4458-031	
5681	Stephanie	Sutton	M	6/16/1961	3652-4587-4458-032	

9 Columns 550 Rows 5 Data Types

Search...

Formulas

Scale to min max
Scale a column to a specific min max range

One hot encode
Create a column for each unique value indicating its presence or absence

Scale to mean
Scale a column to zero mean and unit variance

Bin column
Bin values into ranges of equal or custom size

New formula
Create a new column from the result of a formula

Edit with formula
Set one or more columns to the result of a formula

Window
Perform calculations across multiple ordered rows

Schema

Change column type
Change the data type of a column

Delete columns
Delete one or more columns

Move columns
Move one or more columns before or after another column

Rename columns
Rename one or more columns

Rename with pattern
Rename columns using a pattern

Rename with prefix

CONCLUSIONS

- ✓ As the volume of research information and data sources increases, the prerequisite for data to be complete, findable, comprehensively accessible, interoperable, reusable (compliant with FAIR principles), but also securely stored, structured, and networked in order to be useful remain critical but at the same time become more difficult to fulfill ➔ data wrangling can be seen a valuable asset in ensuring this.

- ✓ The goal is to counteract the growing number of data silos that isolate data from different areas of the organization. Once successfully implemented, data can be retrieved, managed and made available and accessible to everyone within the entity.

- ✓ A data lake and data wrangling can be implemented to improve and simplify IT infrastructure and architecture, governance and compliance. They provide valuable support for predictive analytics and self-service analysis by making it easier and faster to access large amount of data from multiple sources.

- ✓ The proper organization of the data lake makes it easier to find the data the user needs. Managing the data that have already been pre-processed results in an increased efficiency and cost saving, as preparing data for their further use is the most resource-consuming part of data analysis.
- ✓ By providing pre-processed data, users with limited or no experience in data preparation (low level of data literacy) can be supported and analyzes can be carried out faster and more accurately.

THANK YOU FOR ATTENTION!

QUESTIONS?

For more information, see ***ResearchGate****,* ***anastasijanikiforova.com***

For questions or any queries, contact me via ***Nikiforova.Anastasija@gmail.com****,*