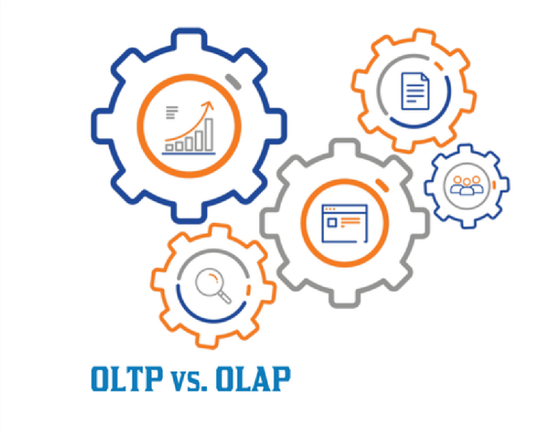
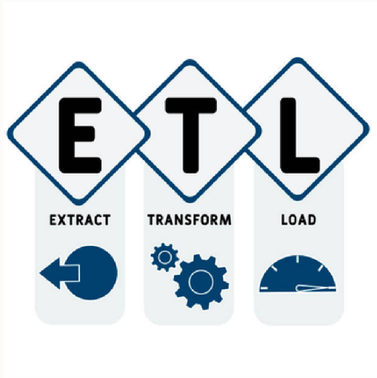
databricks



DATA STORAGES, AND BASIC ETL PROCESS



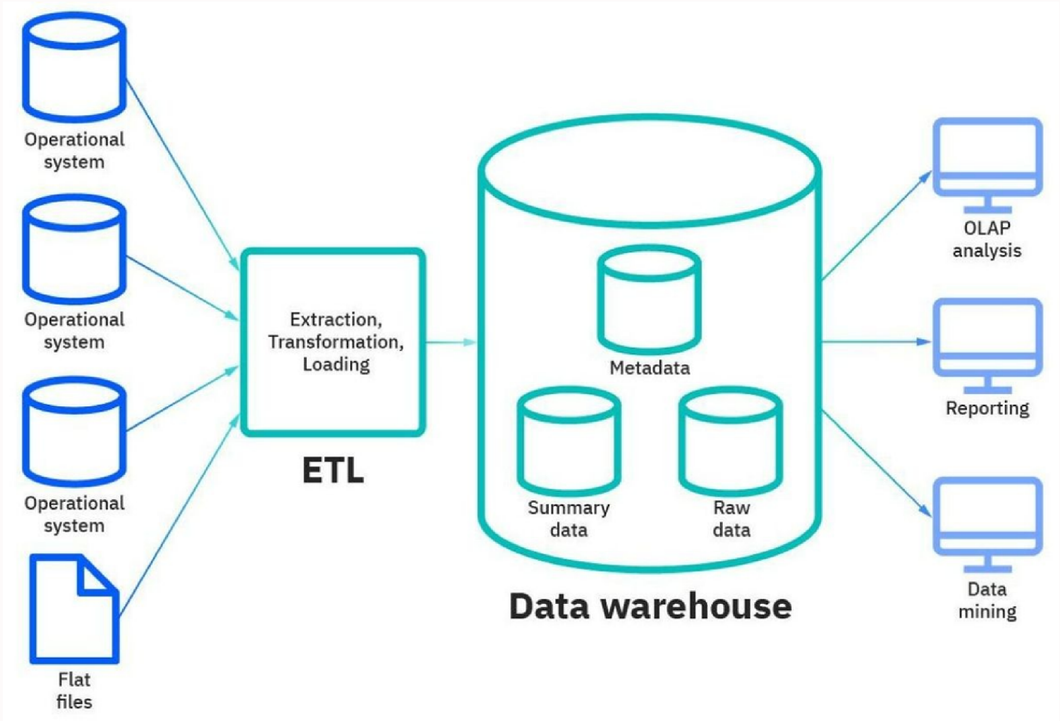
Mentor/Intern: Nguyen Ngoc Thien / Tran Quoc Hai

Le Trung Viet / Tran Thi Anh Thu

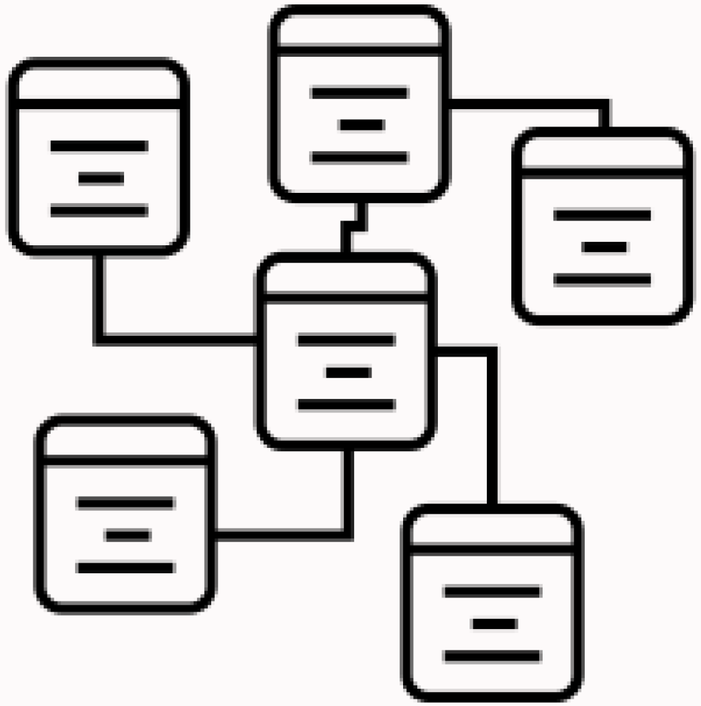
Phan Thien Huu

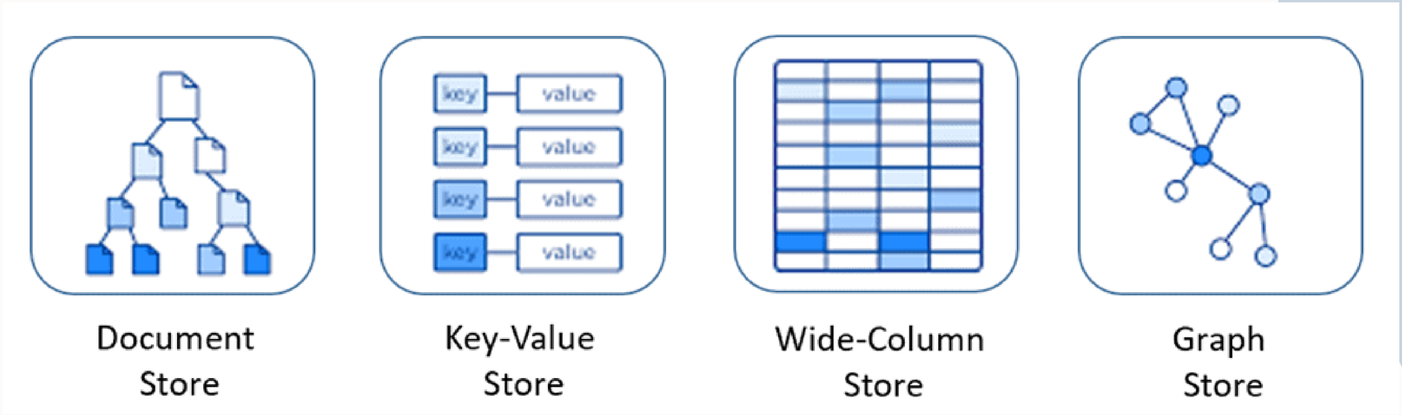
TMA

TABLE OF CONTENTS



DATABASE





Relational Database

Non-Relational Database

DATABASE

Relational Database

Customer

ID	Email	Address
1	joe@litware.com	1 Main St.
2	samir@northwind.com	123 Elm Pl.

Product

ID	Name	Price
123	Hammer	2.99
162	Screwdriver	3.49
201	Wrench	4.25

Order

OrderNo	OrderDate	Customer
1000	1/1/2022	1
1001	1/1/2022	2

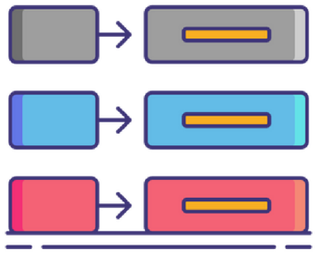
LineItem

OrderNo	ItemNo	ProductID	Quantity
1000	1	123	1
1000	2	201	2
1001	1	123	2

DATABASE

Non-Relational Database

1. Key - Value



Products	
Key	**Value**
123	"Hammer ($2.99)"
162	"Screwdriver ($3.49)"
201	"Wrench ($4.25)"

DATABASE

Non-Relational Database

2. Document

Customers

Key	Document
1	{ "name": "Joe Jones", "email": "joe@litware.com" }
2	{ "name": "Samir Nadoy", "email": "Samir@northwind.com" }

DATABASE

Non-Relational Database

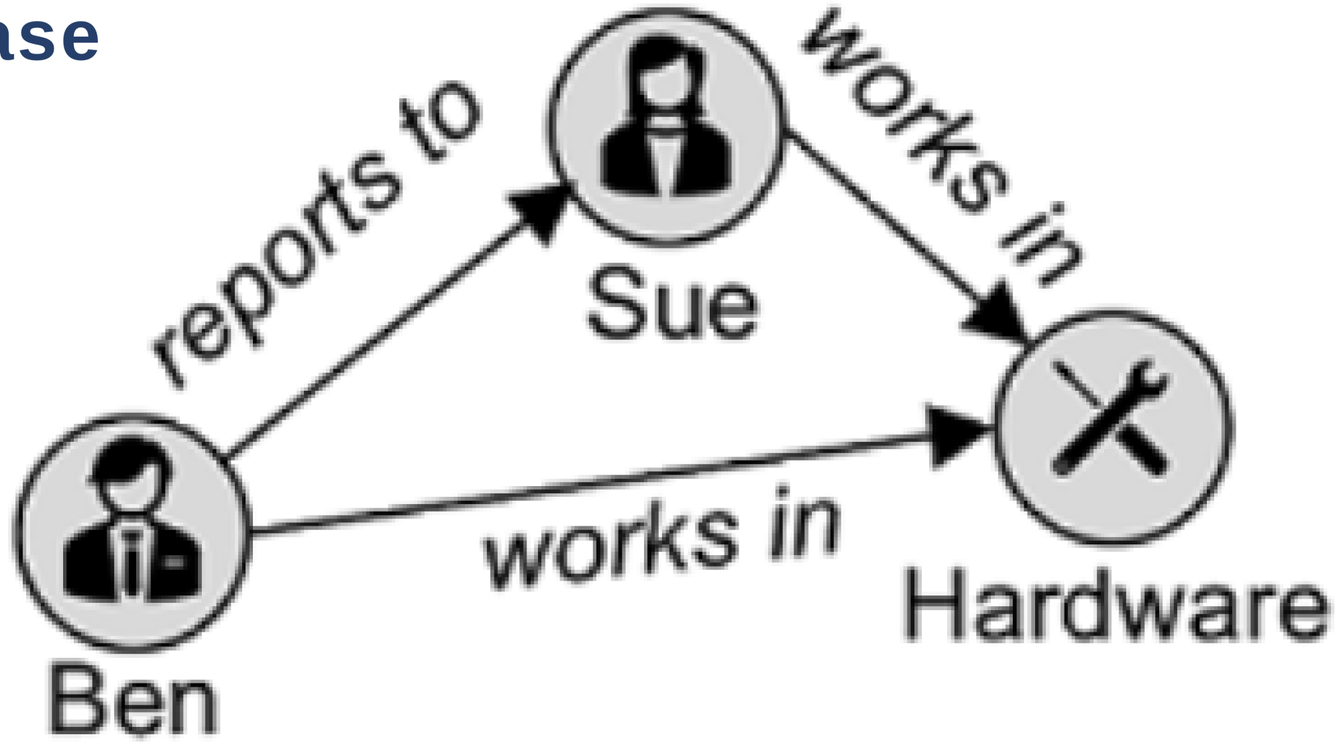
3. Column family

Orders				
Key	Customer		Product	
	Name	Address	Name	Price
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

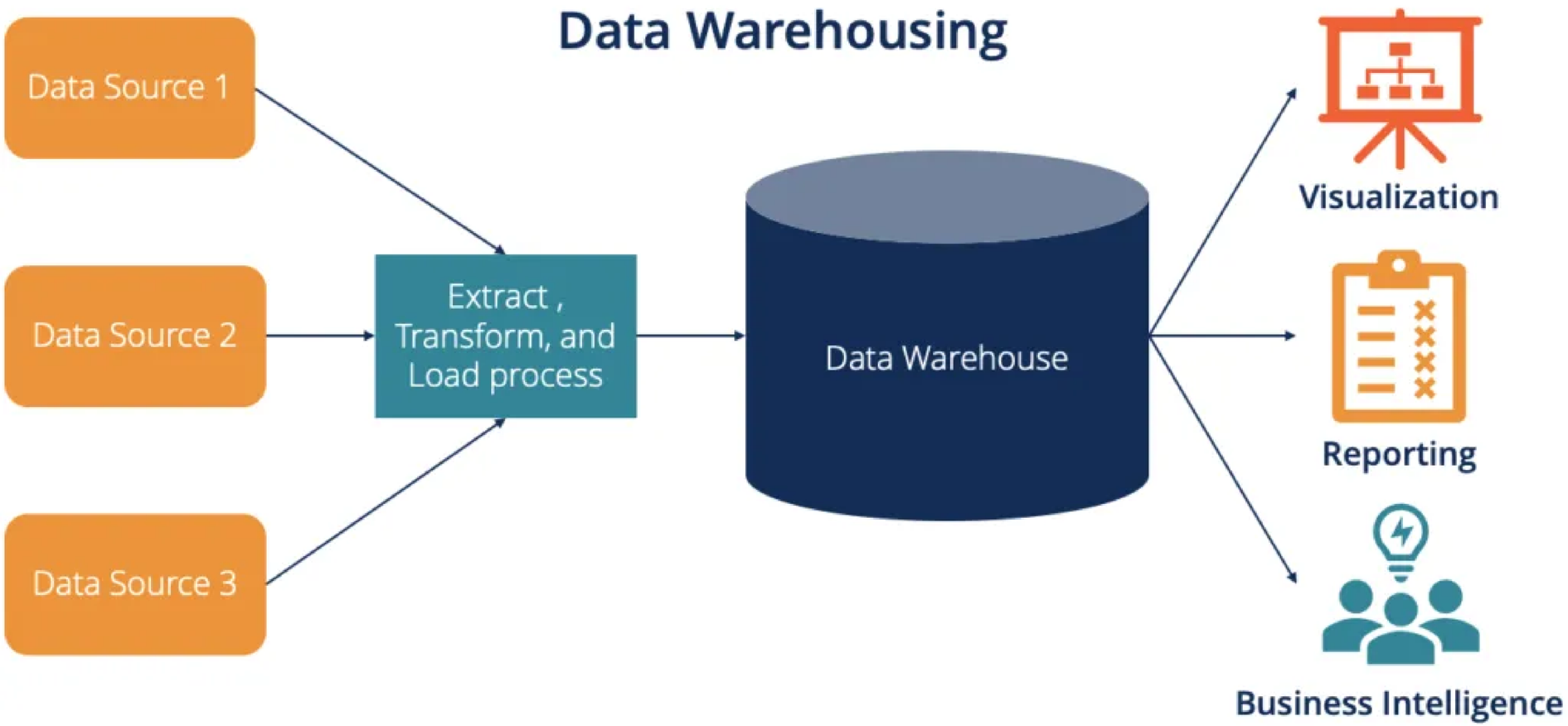
DATABASE

Non-Relational Database

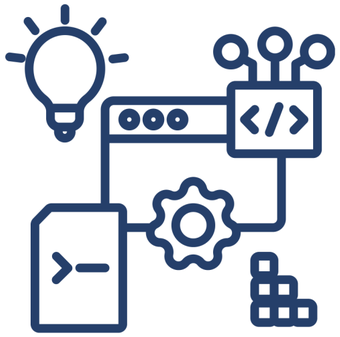
4. Graph



DATA WAREHOUSE



DATA WAREHOUSE



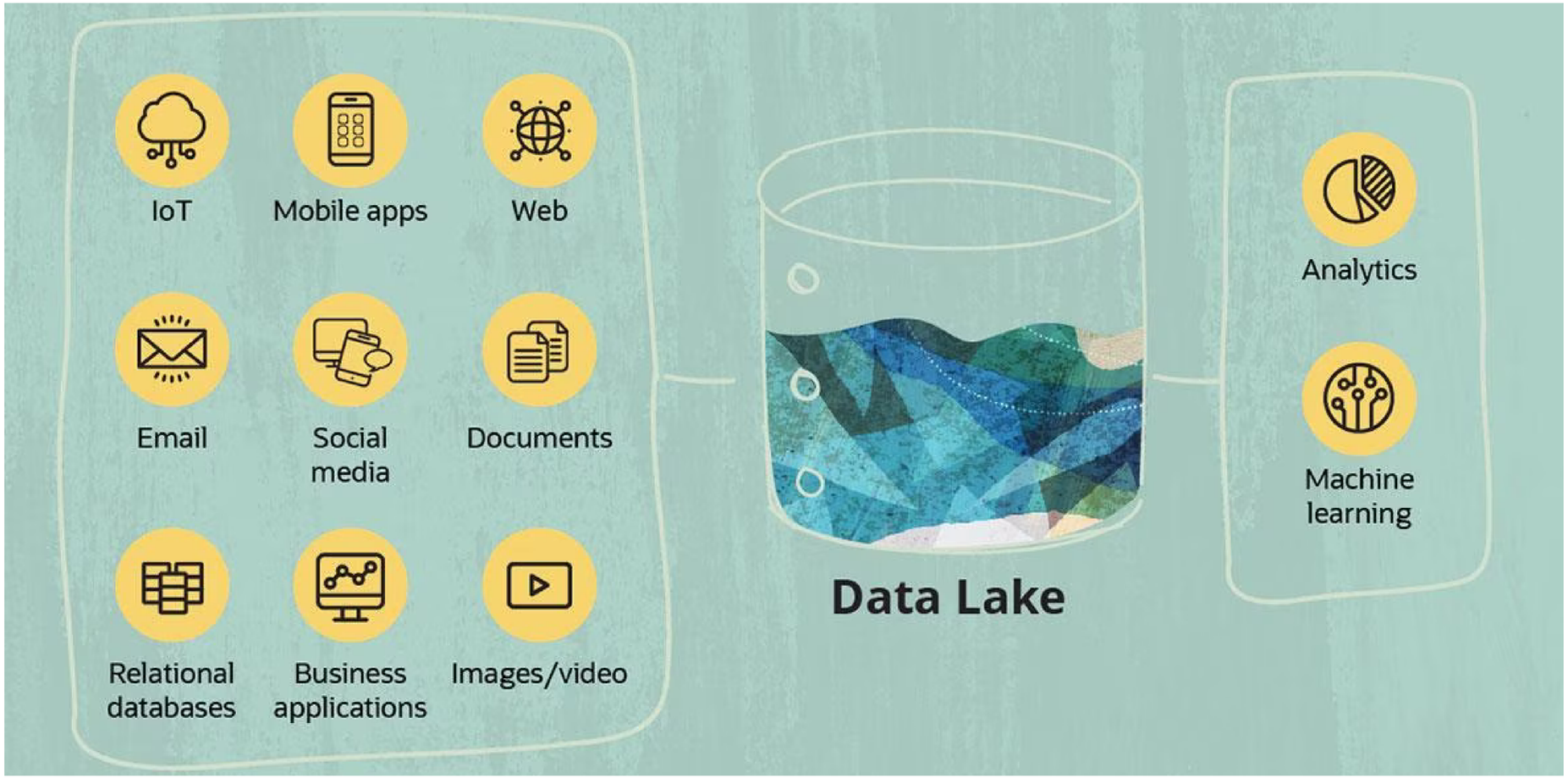
1
Integrated

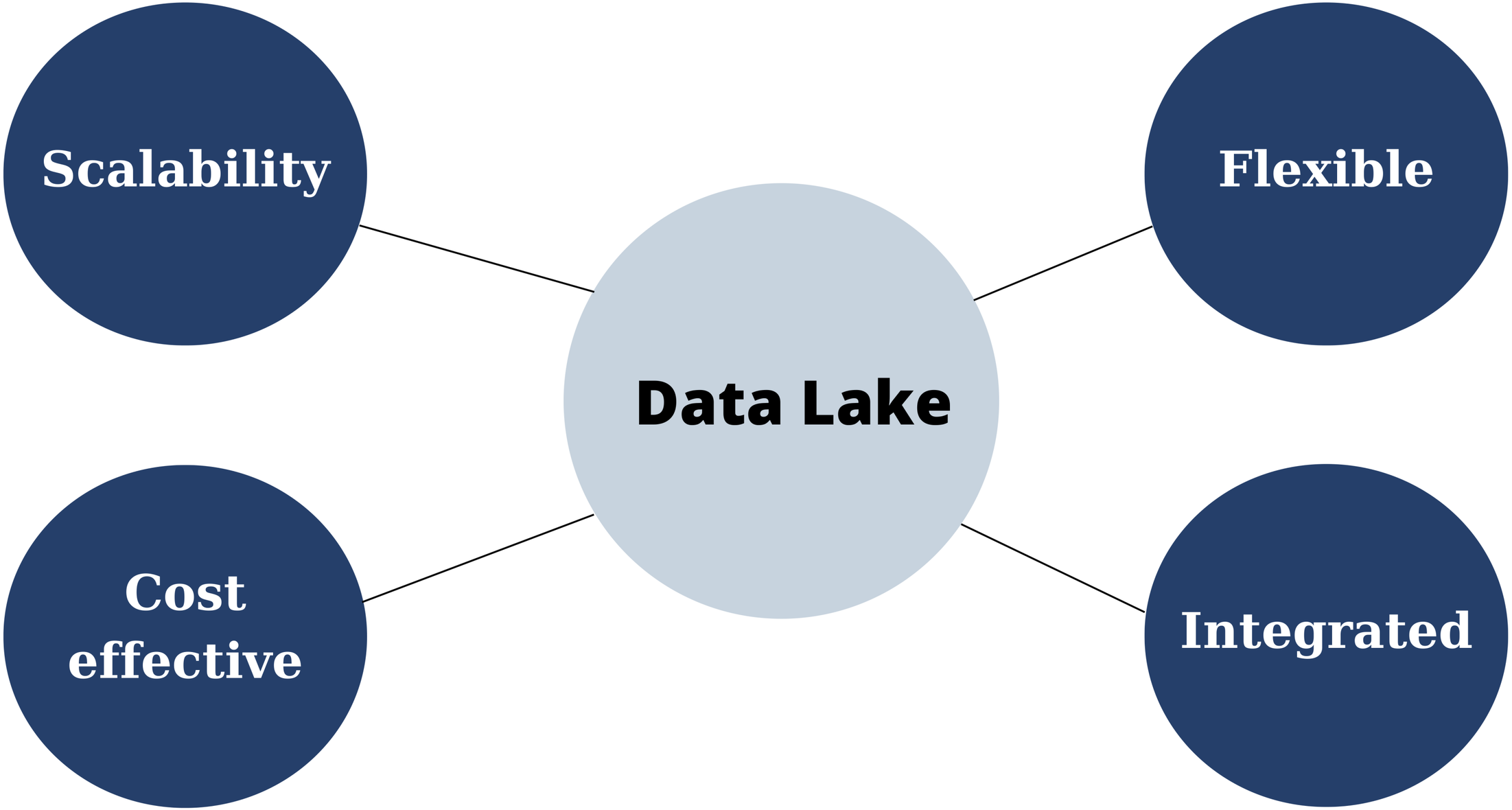
2
Subject-oriented

3
Time variant

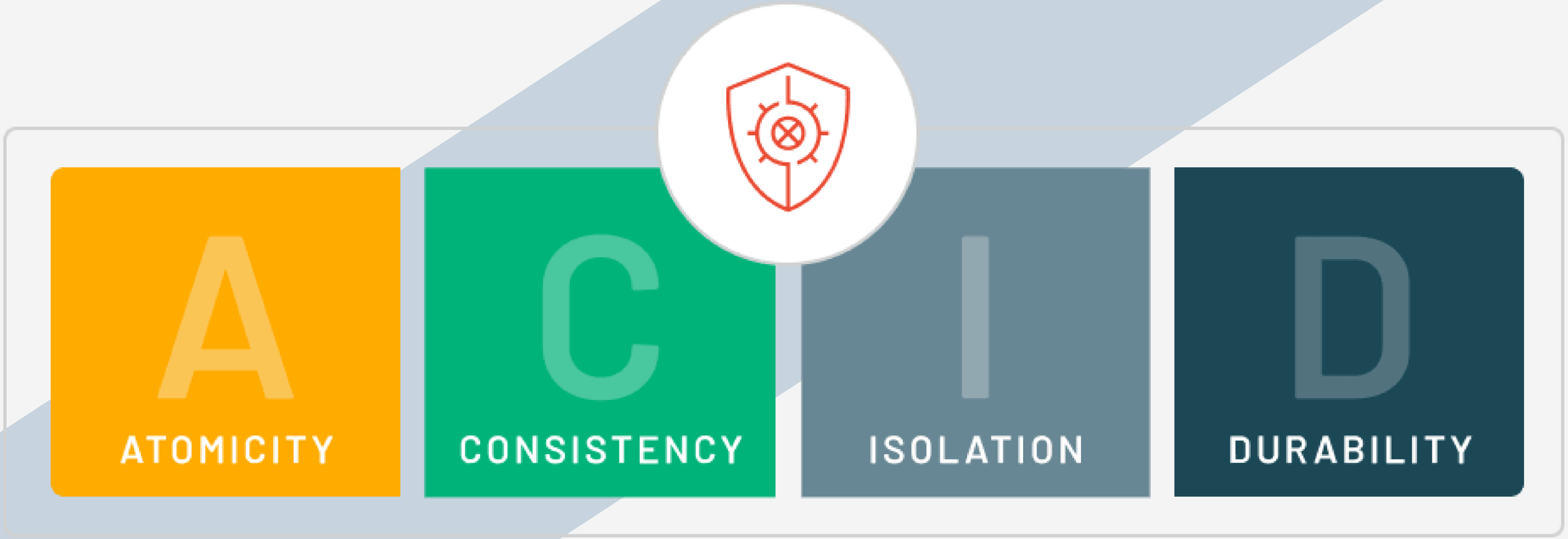
4
Non-volatile

DATA LAKE

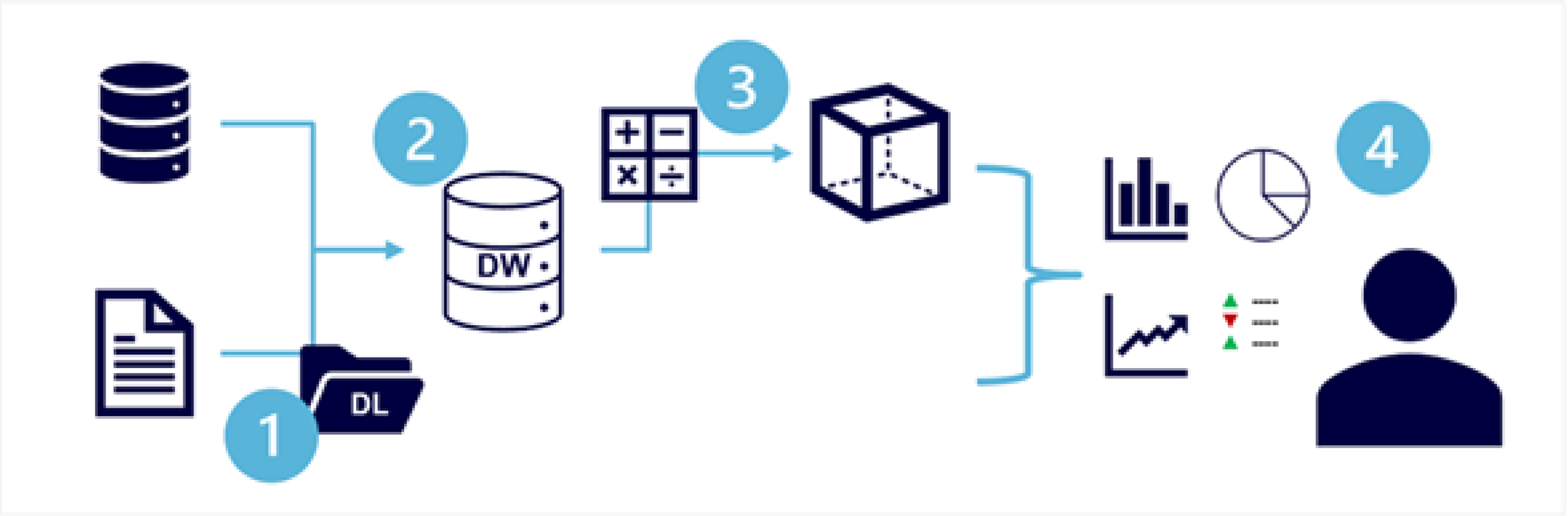

IoT
Mobile apps
Web
Email
Social media
Documents
Relational databases
Business applications
Images/video
Data Lake
Analytics
Machine learning


Scalability
Flexible
Data Lake
Cost effective
Integrated

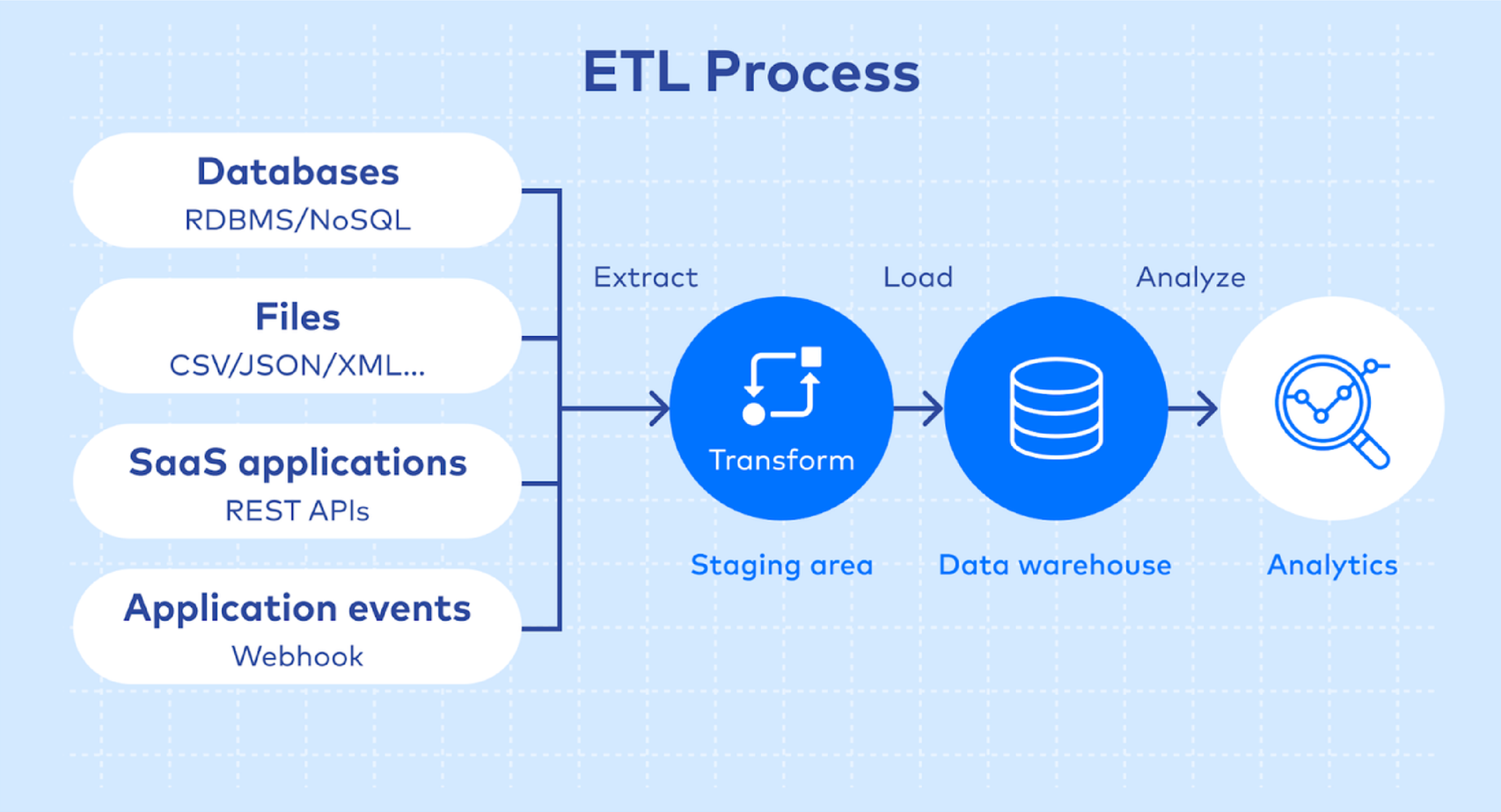
OLTP



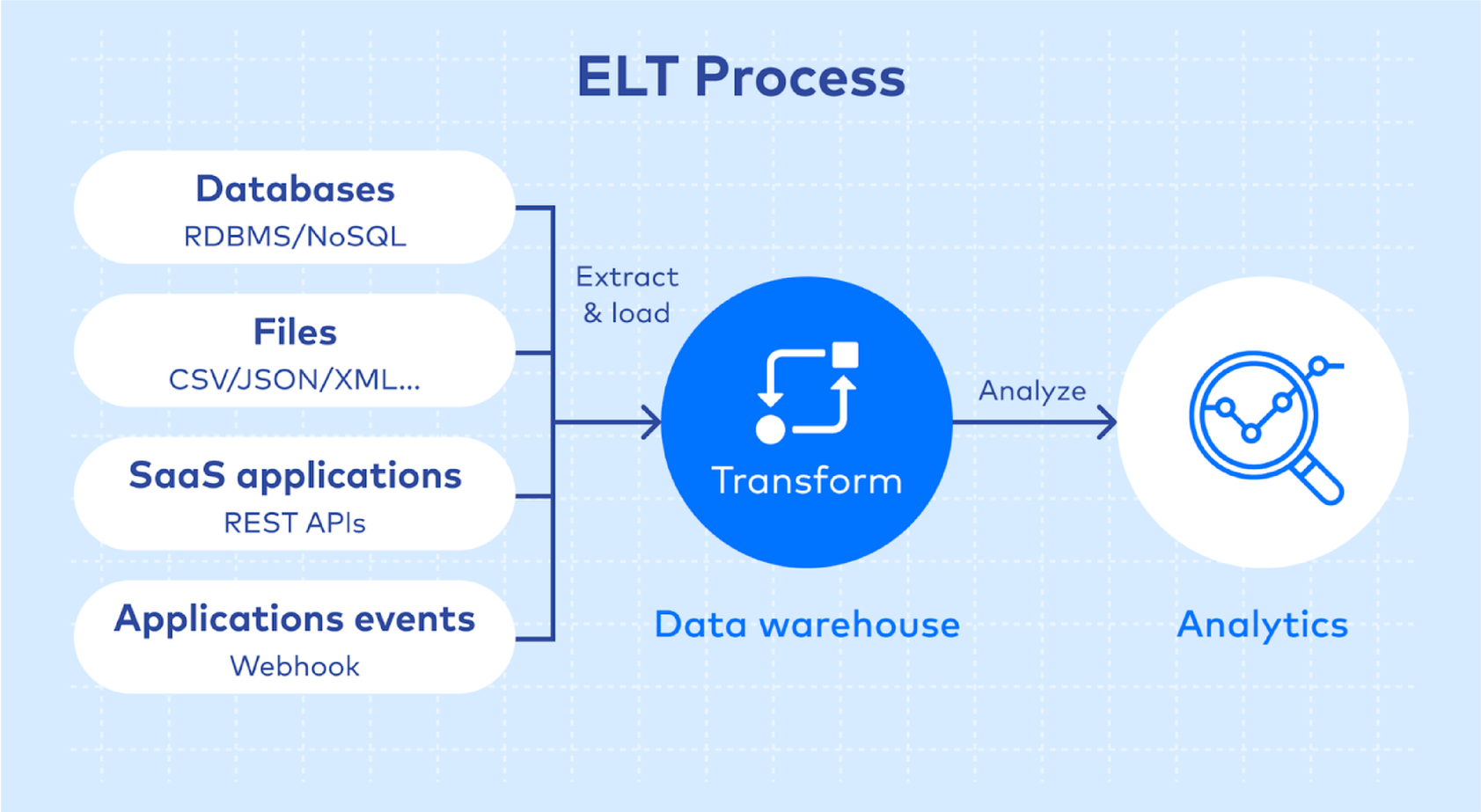
OLAP



ETL PROCESS


ETL Process
Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS applications
REST APIs
Application events
Webhook
Extract
Load
Analyze
Transform
Staging area
Data warehouse
Analytics

ELT PROCESS


ELT Process
Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS applications
REST APIs
Applications events
Webhook
Extract
& load
Transform
Data warehouse
Analyze
Analytics

Brief Comparison

	ETL	ELT
Privacy	Yes	No
Load speed	Slow	Fast
Powerful target system	No	Yes
Capital for hosting required	Yes	No
Data Lake support	No	Yes

DATA

Tokyo 2020 Olympic Summer Games

Athletes

Coaches

Medals

Athletes

name	short_nan	gender	birth_date	birth_plac	birth_cou	country	country_c	discipline	discipline_	residence_	residence_	height_m/	url
AALERUD	AALERUD	Female	12/4/1994	VESTBY	Norway	Norway	NOR	Cycling Ro	CRD				../../../en/
ABAD Nes	ABAD N	Male	3/29/1993	ALCOI	Spain	Spain	ESP	Artistic Gy	GAR	MADRID	Spain	1.65/5'4''	../../../en/
ABAGNALI	ABAGNALI	Male	1/11/1995	GRAGNAN	Italy	Italy	ITA	Rowing	ROW	SABAUDIA	Italy	1.98/6'5''	../../../en/
ABALDE A	ABALDE A	Male	12/15/1995	FERROL	Spain	Spain	ESP	Basketball	BKB			2.00/6'6''	../../../en/

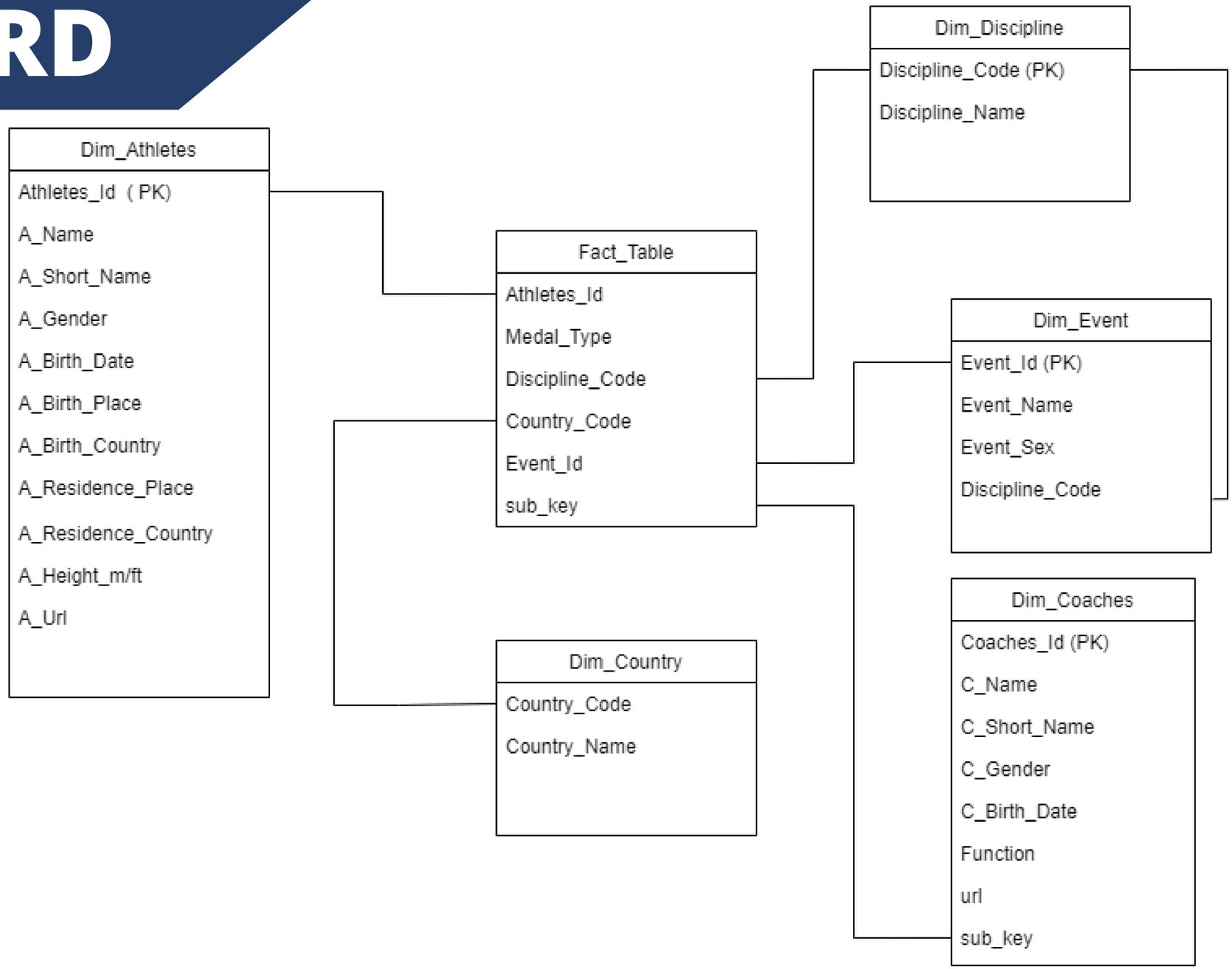
Medals

	medal_typ	medal_code	medal_date	athlete_short_name	athlete_name	athlete_sex	athlete_lir	country_code	discipline_code	event	country	discipline
0	Gold Med	1	00:00.0	KIM JD	KIM Je Deok		../../../en/	KOR	ARC	Mixed Tea	Republic c	Archery
1	Gold Med	1	00:00.0	AN S	AN San	X	../../../en/	KOR	ARC	Mixed Tea	Republic c	Archery
2	Silver Me	2	00:00.0	SCHLOESSER G	SCHLOESSER G	X	../../../en/	NED	ARC	Mixed Tea	Netherlan	Archery

Coaches

name	short_name	gender	birth_date	country_code	discipline	function	event	url
ABDELMA	ABDELMAGID W	Male	8/2/1982	EGY	Football	Head Coach		../../../en/
ABE Junya	ABE J	Male	7/25/1990	JPN	Volleyball	Head Coach		../../../en/
ABE Katsu	ABE K	Male	9/23/1979	JPN	Basketball	Coach		../../../en/

ERD


Dim_Discipline
Discipline_Code (PK)
Discipline_Name
Dim_Athletes
Athletes_Id (PK)
A_Name
A_Short_Name
A_Gender
A_Birth_Date
A_Birth_Place
A_Birth_Country
A_Residence_Place
A_Residence_Country
A_Height_m/ft
A_Url
Fact_Table
Athletes_Id
Medal_Type
Discipline_Code
Country_Code
Event_Id
sub_key
Dim_Event
Event_Id (PK)
Event_Name
Event_Sex
Discipline_Code
Dim_Country
Country_Code
Country_Name
Dim_Coaches
Coaches_Id (PK)
C_Name
C_Short_Name
C_Gender
C_Birth_Date
Function
url
sub_key

```python
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
spark = SparkSession.builder.getOrCreate()

df = spark.read.csv('/FileStore/tables/medals2.csv', header=True, inferSchema=True)
display(df)
```

df: pyspark.sql.dataframe.DataFrame = [_c0: integer, medal_type: string ... 5 more fields]

DEMO

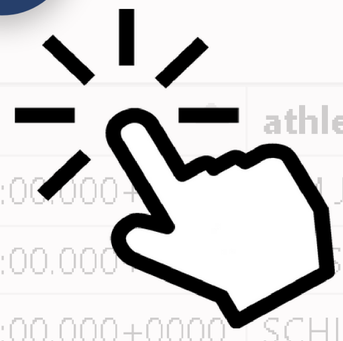
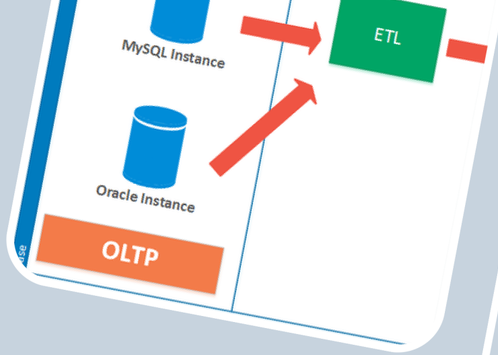


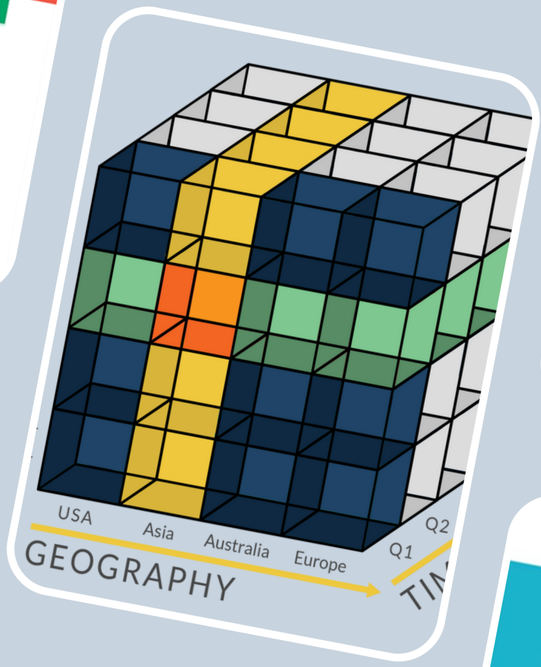
Table +

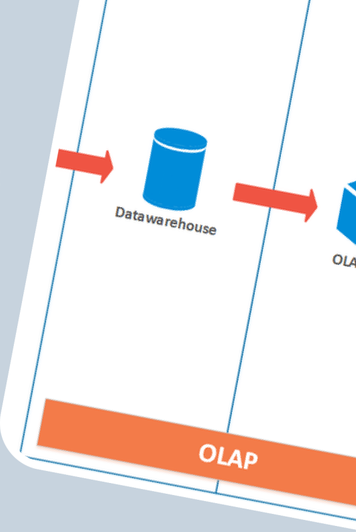
	_c0	medal_type	medal_code	medal_date	athlete_short_name	athlete_name
1	0	Gold Medal	1	2021-07-24T00:00:00.000+0000	KIM JD	KIM Je Deok
2	1	Gold Medal	1	2021-07-24T00:00:00.000+0000	AN S	AN San
3	2	Silver Medal	2	2021-07-24T00:00:00.000+0000	SCHLOESSER G	SCHLOESSER Gabriela
4	3	Silver Medal	2	2021-07-24T00:00:00.000+0000	WIJLER S	WIJLER Steve
5	4	Bronze Medal	3	2021-07-24T00:00:00.000+0000	ALVAREZ L	ALVAREZ Luis
6	5	Bronze Medal	3	2021-07-24T00:00:00.000+0000	VALENCIA A	VALENCIA Alejandra
7	6	Gold Medal	1	2021-07-24T00:00:00.000+0000	CARAPAZ R	CARAPAZ Richard

2,401 rows | 2.26 seconds runtime

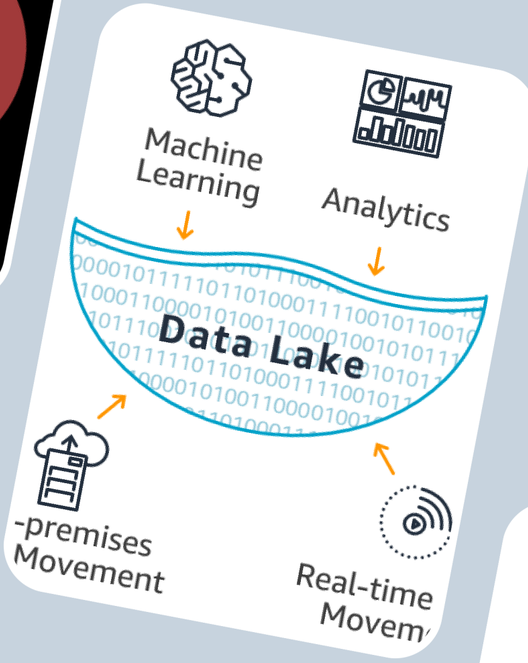
THANK YOU


MySQL Instance
Oracle Instance
ETL
OLTP


USA
Asia
Australia
Europe
Q1
Q2
GEOGRAPHY


Datawarehouse
OLAP


CAP Theorem
Consistency
CA
CP
Availability
AP
Partition Tolerance
Cassandra
CouchDB
Dynamo
Voldermort


Machine Learning
Analytics
Data Lake
-premises Movement
Real-time Movem


Data Warehouse