DATA ENGINEERING COMMUNITY

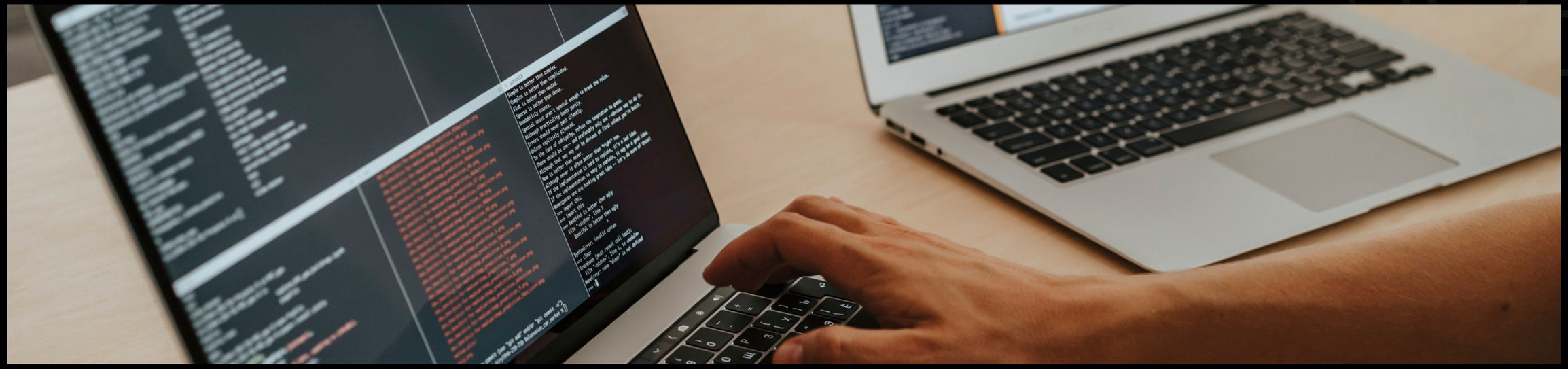
DATA PIPELINE DESIGN AND BEST PRACTICES

BY NAJEEB SULAIMAN

DATA ENGINEERING COMMUNITY

ABOUT ME

I am a data engineer with years of experince working across industries and organizations. I currently work as a data engineer at JLR UK .



WHAT IS A DATA PIPELINE?

A data pipeline refers to the integrated set of processes, tools, and infrastructure set up to automate the flow of data from its various sources to its end destinations.

DATA ENGINEERING COMMUNITY

WHY IS DATA PIPELINE IMPORTANT?

Volume

The amount of data qualifying as big data

Velocity

The speed at which the data is created and how fast it moves

Veriety

The diversity that exists in the types of data

Veracity

The data's quality and accuracy

Value

The value the data provides.

COMPONENTS OF DATA PIPELINE

01

Data sources

These are the original repositories where the data resides before it is processed

02

Data ingestion

The initial step of capturing raw data and introducing it into the data pipeline

03

Data processing

Once data is ingested, it often needs to be transformed to be useful

04

Data storage

After processing, data is stored in a system suitable for its intended use

05

Data consumption

This refers to the end use of the processed data

06

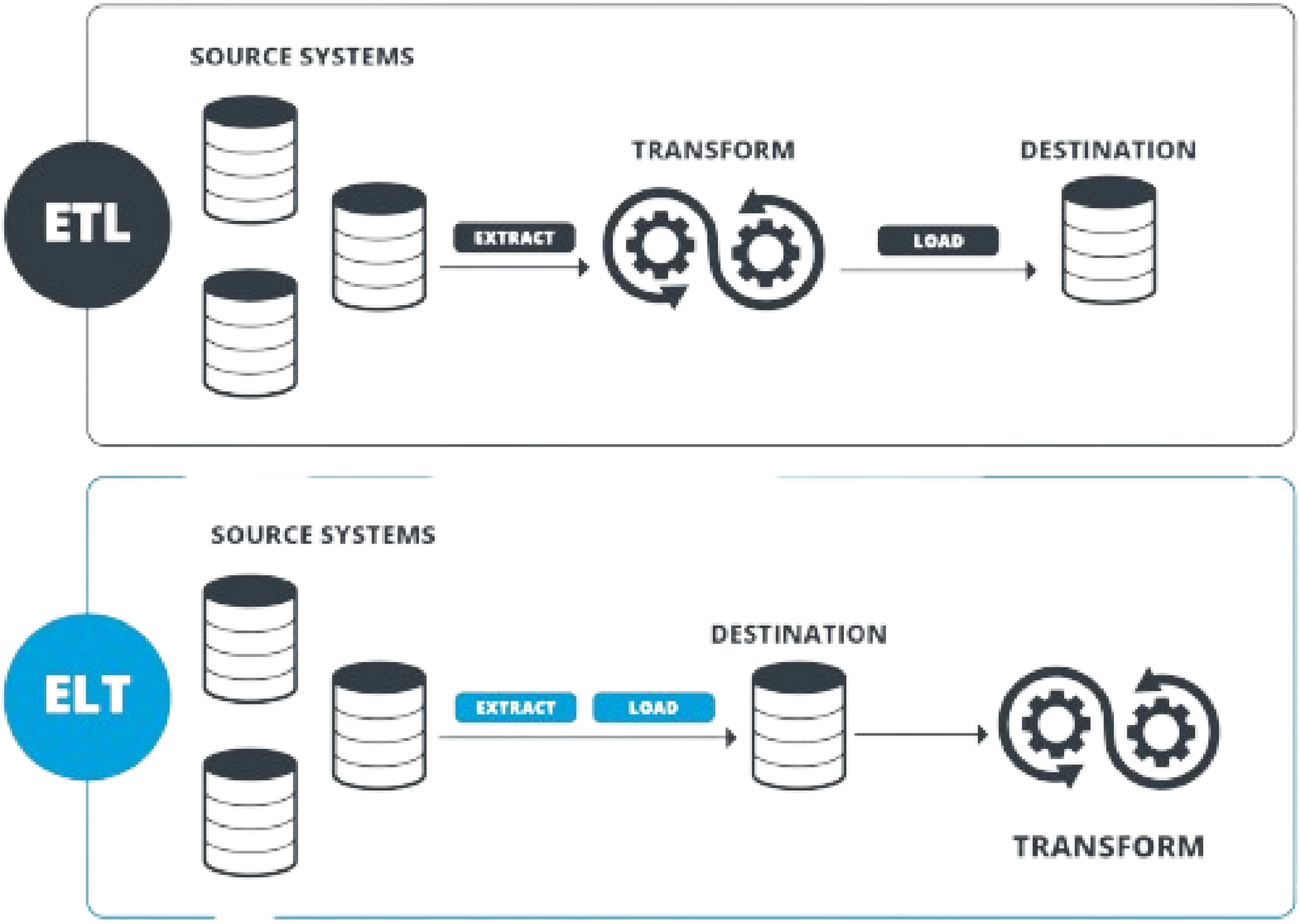
Orchestration

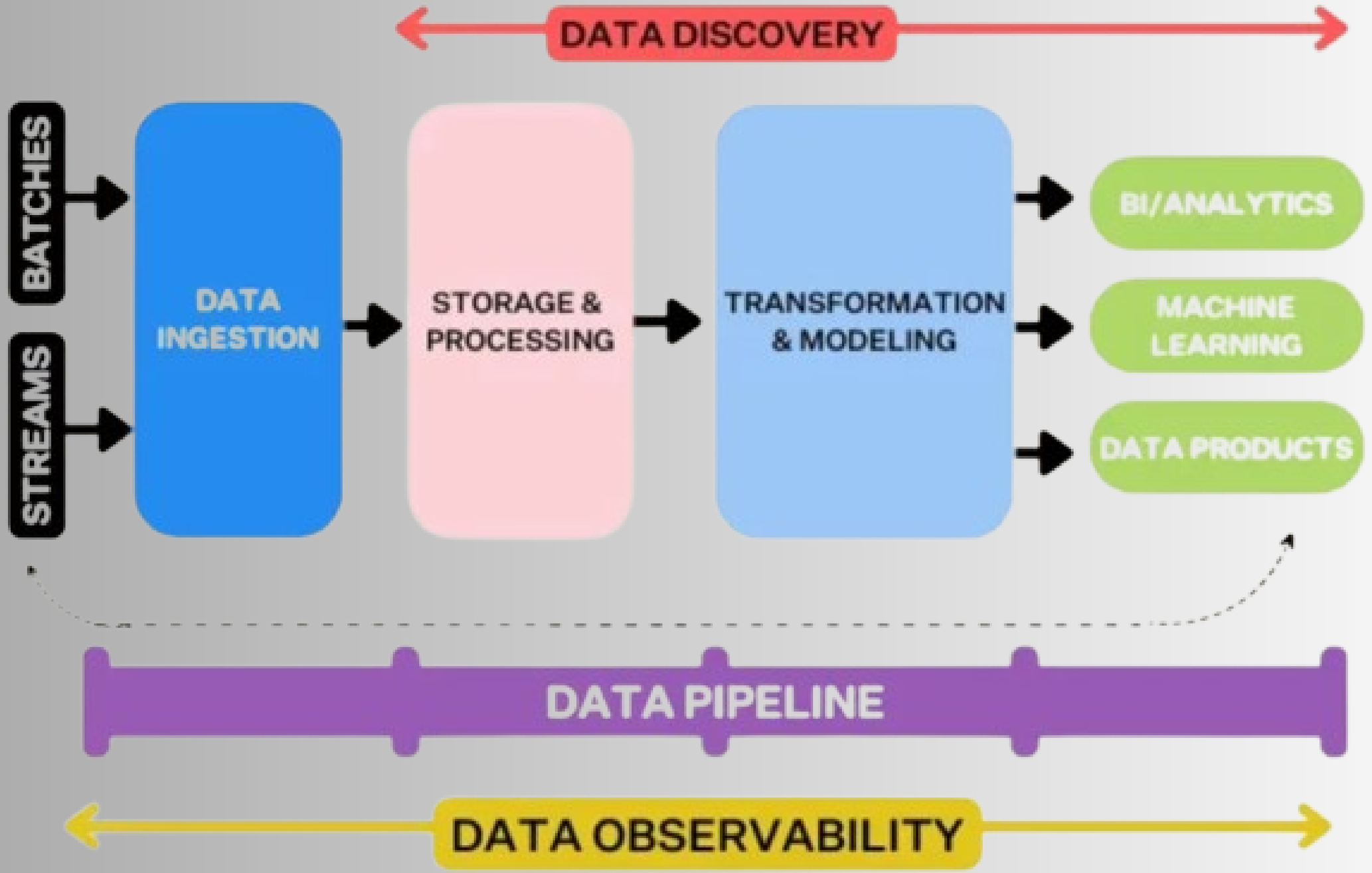
As data moves through various stages, there needs to be coordination of tasks, error handling, retries, and logging

07

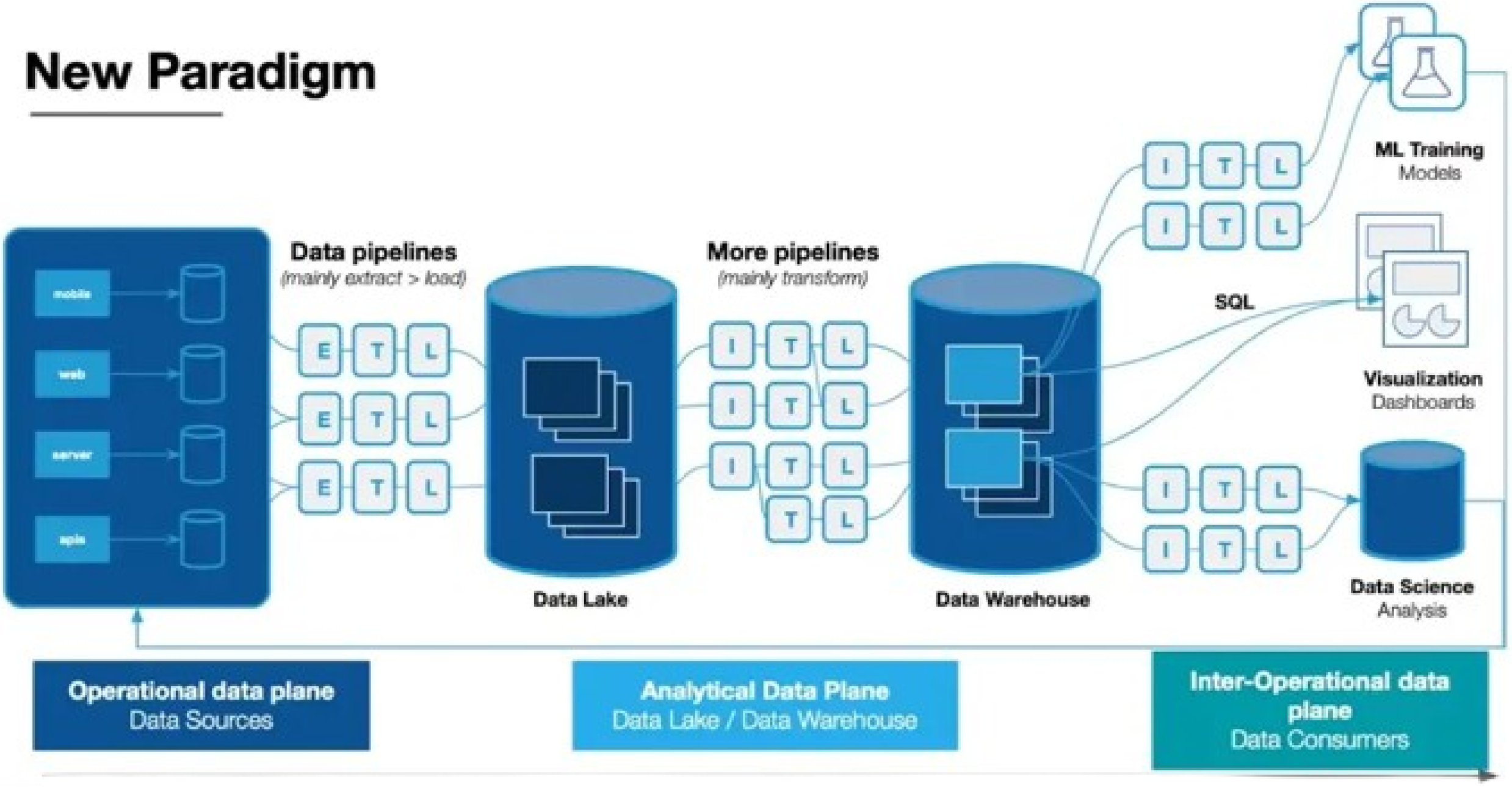
Data managmt and gov

This encompasses practices, processes, and tools used to ensure data accuracy, quality, privacy, and security


SOURCE SYSTEMS
ETL
TRANSFORM
DESTINATION
EXTRACT
LOAD
SOURCE SYSTEMS
ELT
DESTINATION
EXTRACT
LOAD
TRANSFORM


DATA DISCOVERY
BATCHES
STREAMS
DATA INGESTION
STORAGE & PROCESSING
TRANSFORMATION & MODELING
BI/ANALYTICS
MACHINE LEARNING
DATA PRODUCTS
DATA PIPELINE
DATA OBSERVABILITY

New Paradigm



DATA ENGINEERING COMMUNITY

DATA PIPELINE BEST PRACTICES

1. Have good domain knowledge
2. Define Clear Objectives
3. Clearly define and understand data sources
4. Ensure data quality at entry
5. Prioritize scalability

DATA ENGINEERING COMMUNITY

DATA PIPELINE BEST PRACTICES

6 Deterministic, idempotent and immutable

7 Use version control for development and collaboration

8 Choose the right processing paradigm

9 DRY

10 Prioritize incremental data ingestion

DATA ENGINEERING COMMUNITY

DATA PIPELINE BEST PRACTICES

11 KISS - Pipeline should be optimized against cost

12 Design modular and automated pipeline

13 Appropriate testing, exception handling and logging

14 Prioritize data security and compliance

15 Implement robust documentation

Do you have any questions?

i am all ears! I hope you learned something new.

THANK YOU

DATA ENGINEERING COMMUNITY