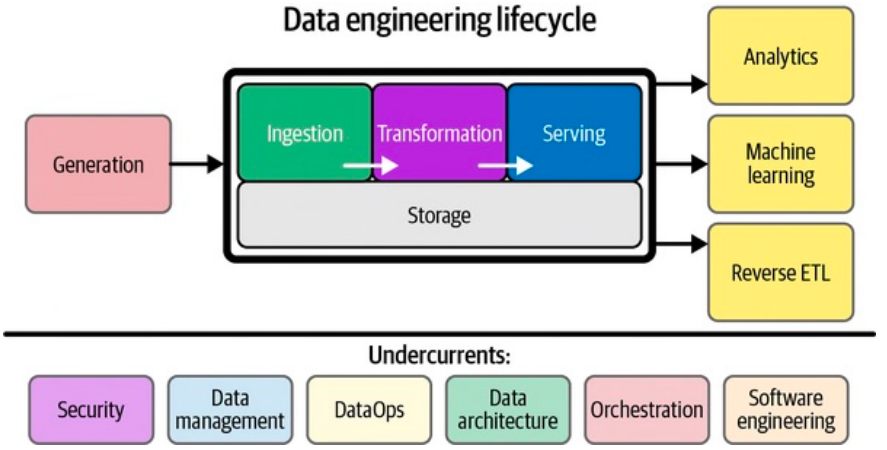
Data Engineering Life Cycle

Generation:
- Web Sites
- Sensors
- Forms/Feedback
- Analytics



Serving: Data Warehouse

Ingestion: Fetching data from multiple sources

Storage: Storing all the data that you fetched from multiple sources

Transformation: Based on business logic or other requirements, doing a transformation

What is Data Warehouse

A data warehouse is a centralized repository that stores data from various sources, such as transactional systems, applications, and external sources.

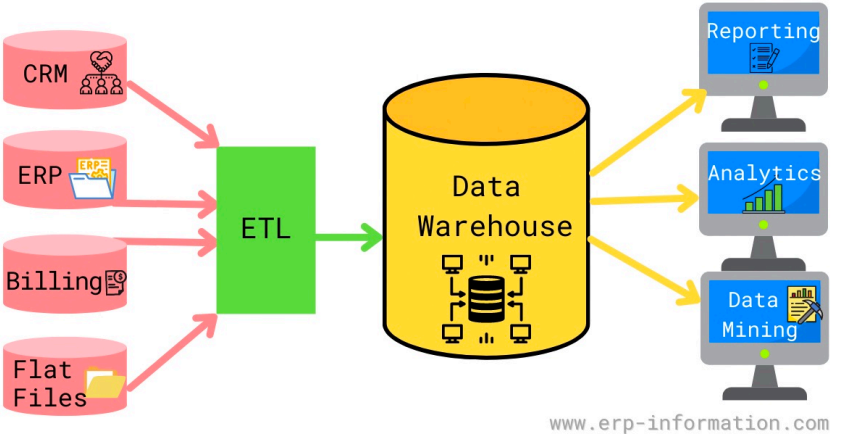
It is designed to support business intelligence activities, such as reporting, analysis, and data mining. Unlike transactional databases, which are designed for day-to-day operations, data warehouses are designed for complex queries and analysis.

Characteristics of Data Warehouse

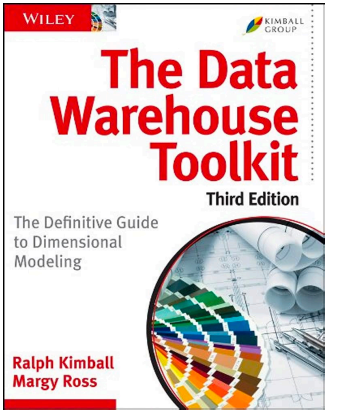
1. Subject-Oriented
2. Integrated
3. Time-Variant
4. Non-Volatile

Benefits of Data Warehouse

1. Improved Decision Making
2. Increased Efficiency
3. Enhanced Data Quality
4. Competitive Advantage



Kimball's Approach to Data Warehousing

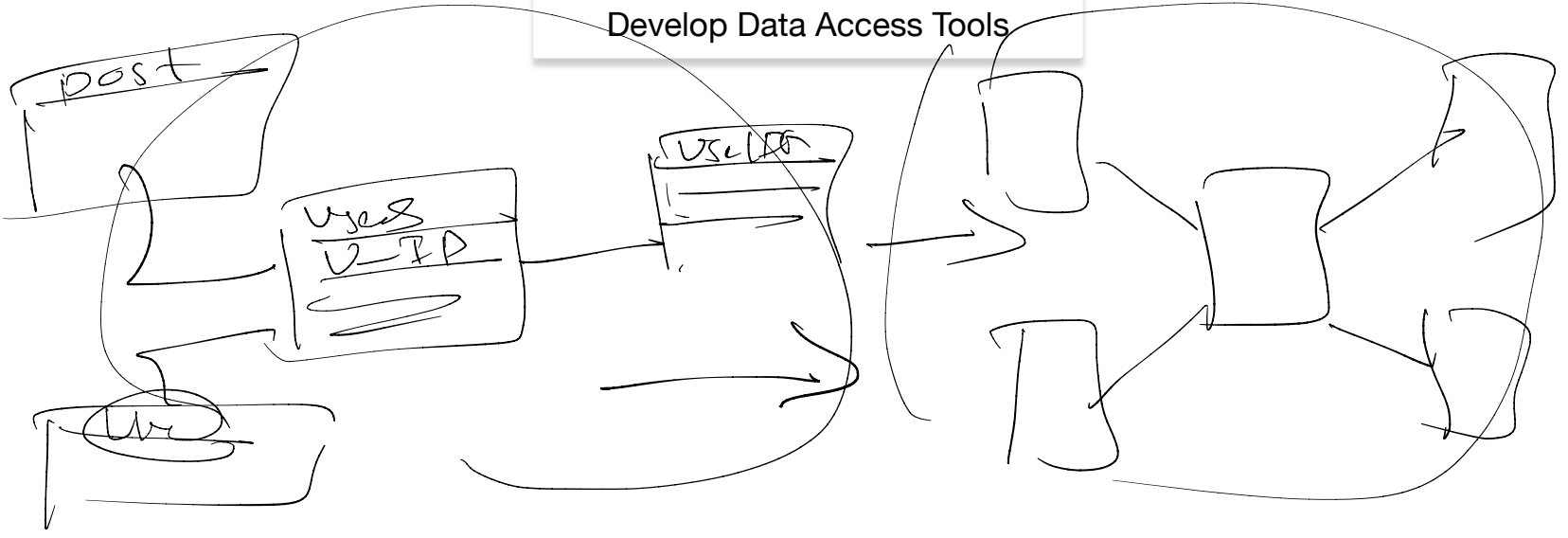


Identify the Business Requirements

Design the Dimensional Model

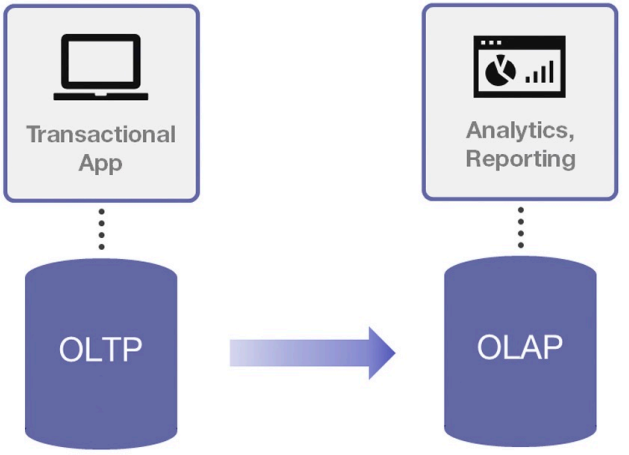
Implement the ETL Processes

Develop Data Access Tools

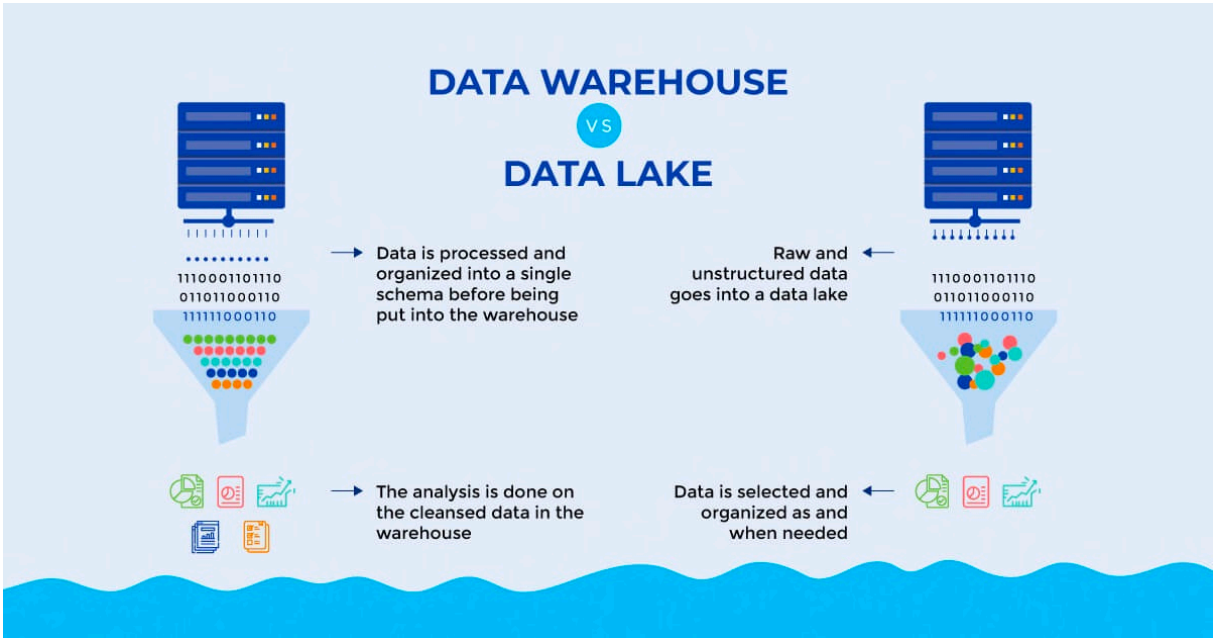


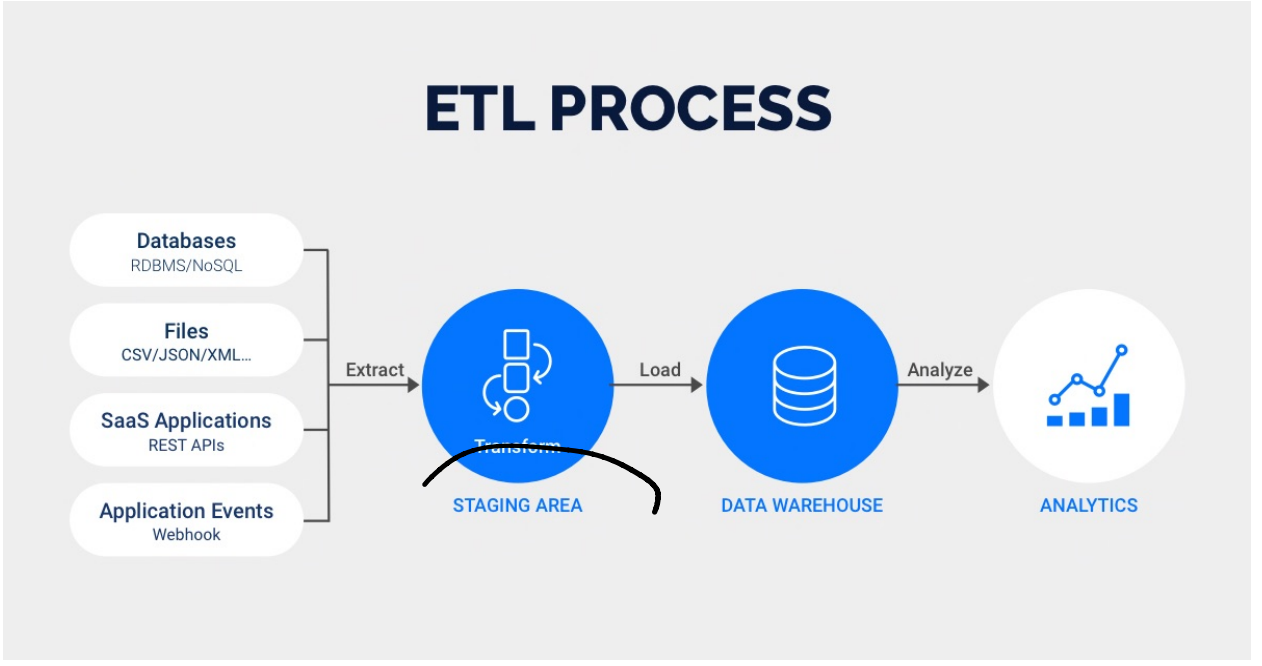
OLTP vs OLAP

- High volume of transactions
- Fast processing
- Normalized data
- Many tables
- *"Who bought X?"*

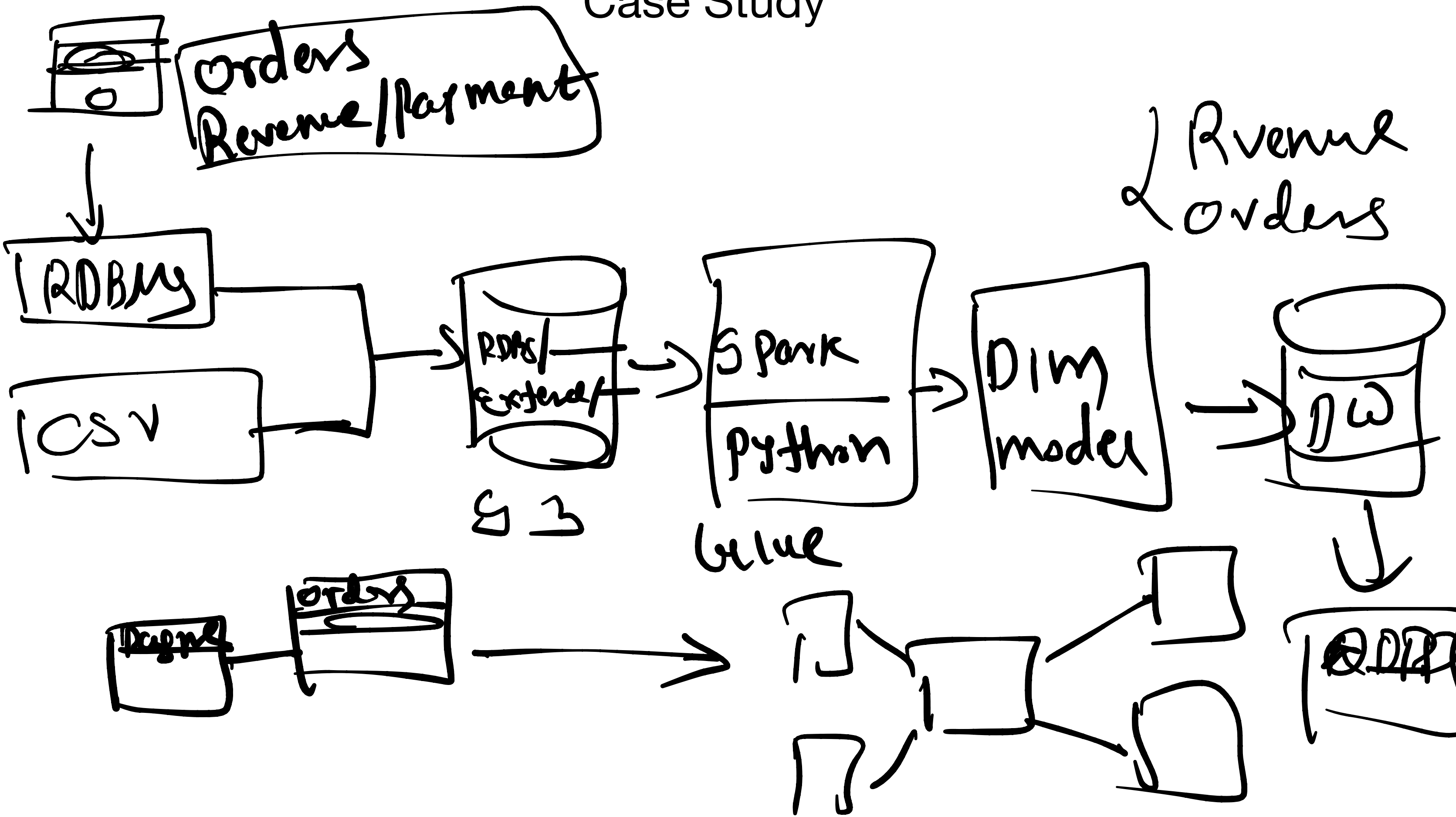


- High volume of data
- Slow queries
- Denormalized data
- Fewer tables
- *"How many people bought X?"*

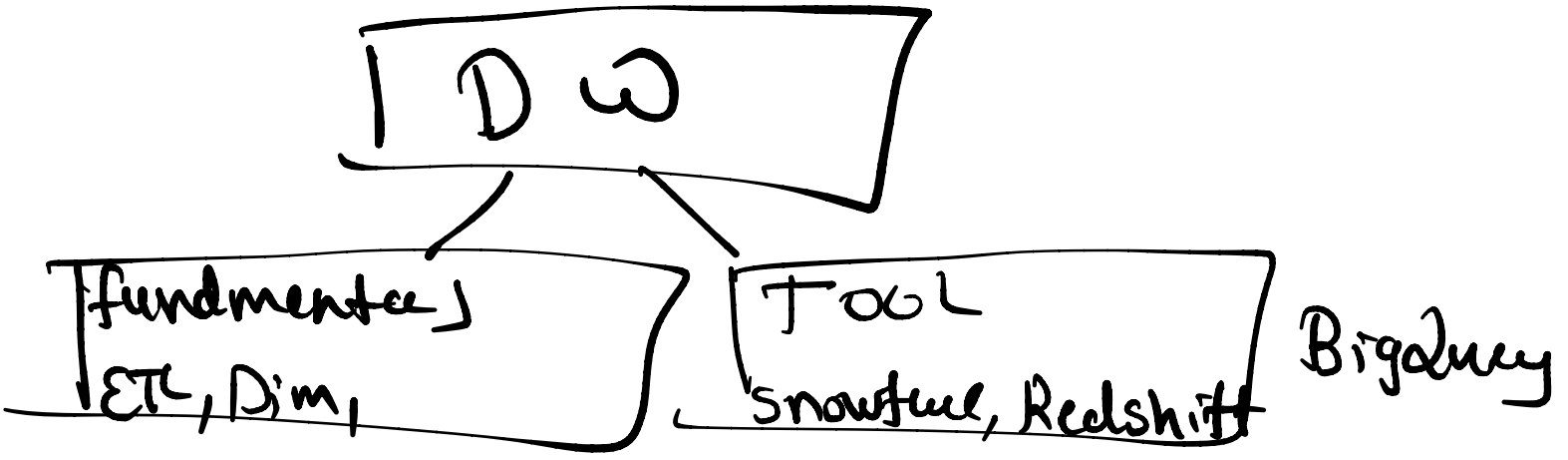

DATA WAREHOUSE
VS
DATA LAKE
1110001101110
0110110000110
1111110000110
Data is processed and organized into a single schema before being put into the warehouse
Raw and unstructured data goes into a data lake
1110001101110
0110110000110
1111110000110
The analysis is done on the cleansed data in the warehouse
Data is selected and organized as and when needed


ETL PROCESS
Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS Applications
REST APIs
Application Events
Webhook
Extract
Transform
Load
Analyze
STAGING AREA
DATA WAREHOUSE
ANALYTICS

Case Study



Orders
Revenue/Payment
Rvenue
orders
RDBMS
CSV
RDMS/
external
S3
Spark
Python
Glue
Dim
model
DW
orders


DW
Fundmentals
ETL, Dim,
Tool
Snowflake, Redshift
Bigquery


Data Sources
OLTP Databases
Enterprise Applications
Third-Party
Web/Log Data
IoT
ETL, Streaming
Data Warehouse
Data Lake
Data Engineering
Data Exchange
Data Applications
Data Science
snowflake
Google Cloud
aws
Azure
Data Consumers
Data Monetization
Operational Reporting
Ad Hoc Analysis
Real-time Analytics