INFORMATION TECHNOLOGY UNIVERSITY

DATA SCIENCE LAB

Data Warehouse

A Business Perspective

CS 537- Big Data Analytics

Dr. Faisal Kamiran

Operational vs Analytical Business Processes

Operational Processes

Make it work!

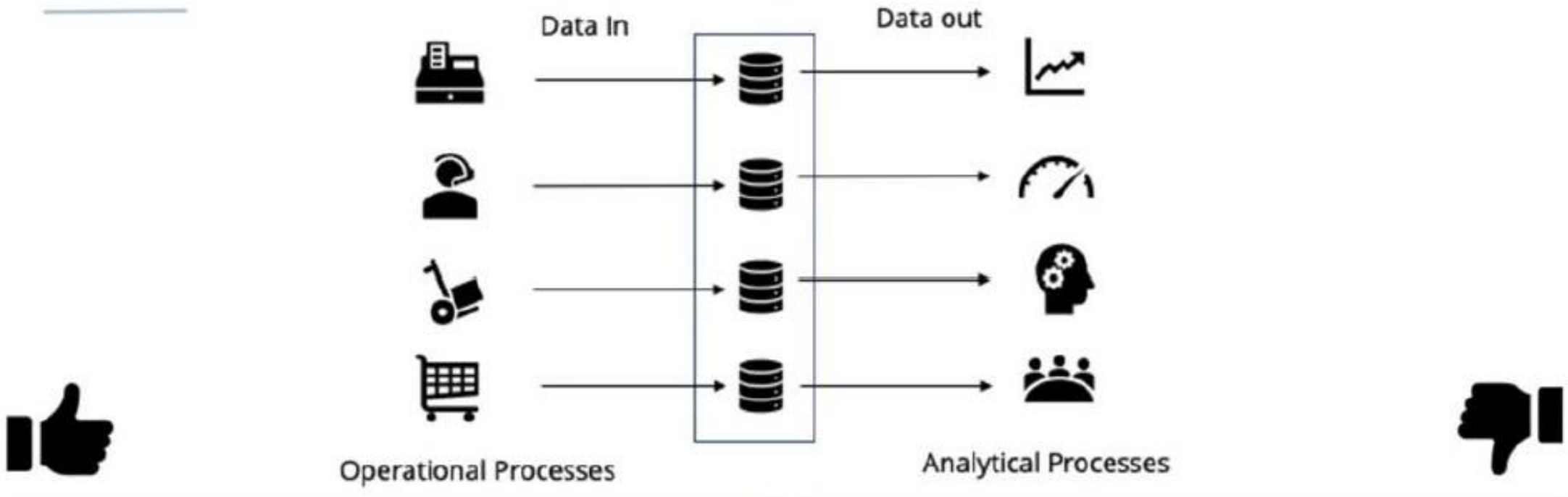
- Find goods & make orders (for customers)
- Stock and find goods (for inventory staff)
- Pick up & deliver goods (for delivery staff)

Analytical Processes

What is going on?

- Assess the performance of sales staff (for HR)
- See the effect of different sales channels (for marketing)
- Monitor sales growth (for management)

Same data source for operational & analytical processes?



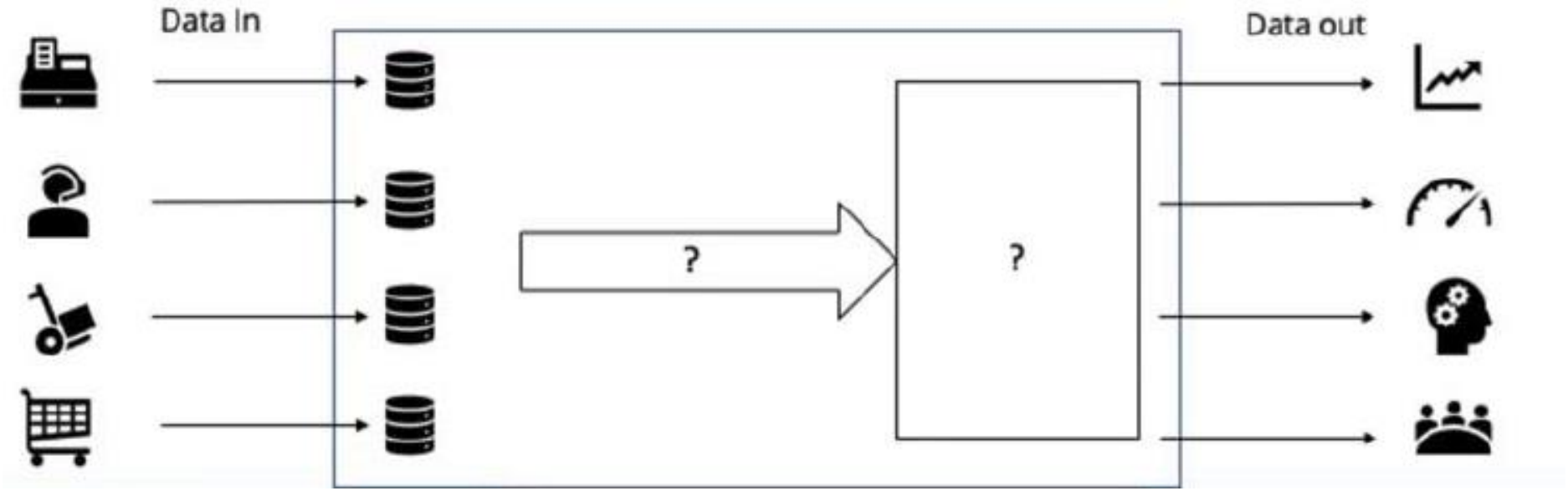
Operational Databases

- Excellent for operations
- No redundancy, high integrity

Operational Databases

- Too slow for analytics, too many joins
- Too hard to understand

Solution: Create two processing modes
Create a system for them to co-exist



OLTP
Online **transactional** processing

OLAP
Online **analytical** processing

Data Warehouse is a system (including processes, technologies & data representations) that enables us to support analytical processes

What is a Data Warehouse?

Tech Perspective: DWH Definition 1

A data warehouse is a copy of transaction data specifically structured for query and analysis.

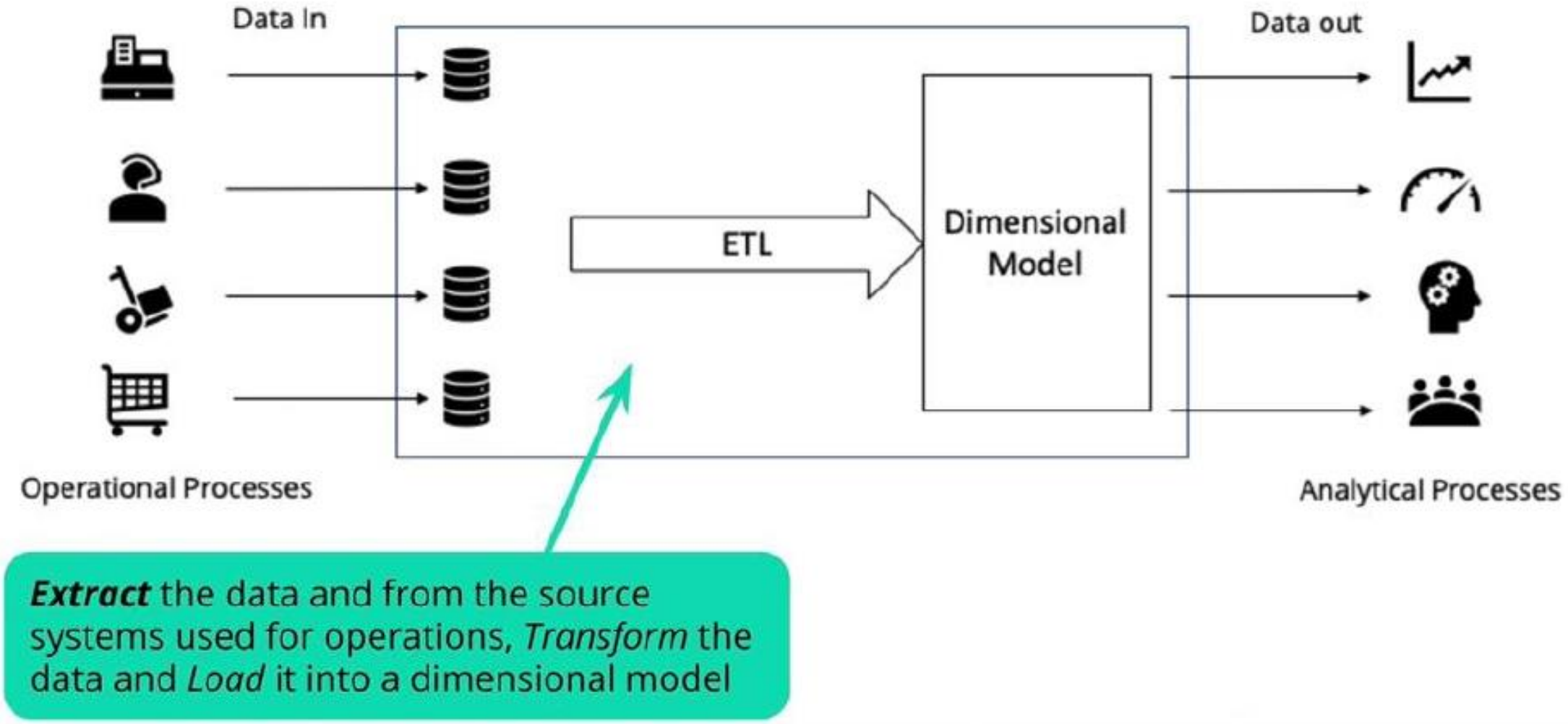
REF: KIMBALL

Tech Perspective: DWH Definition 2

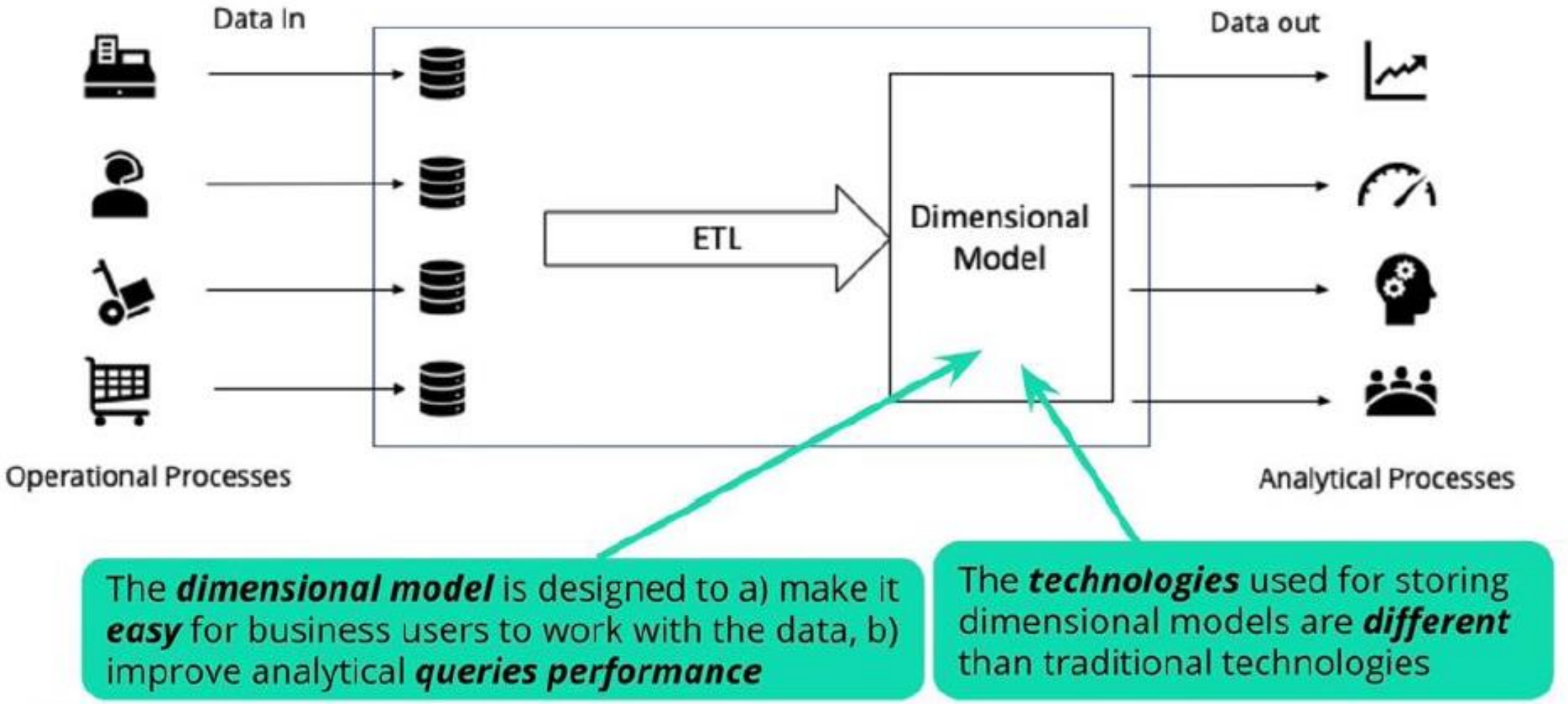
A data warehouse is a system that **retrieves** and **consolidates** data **periodically** from the source systems into a **dimensional or normalized** data store. It usually **keeps years of history** and is **queried for business intelligence** or other **analytical activities.** It is typically **updated in batches,** not every time a transaction happens in the source system.

REF: INMON

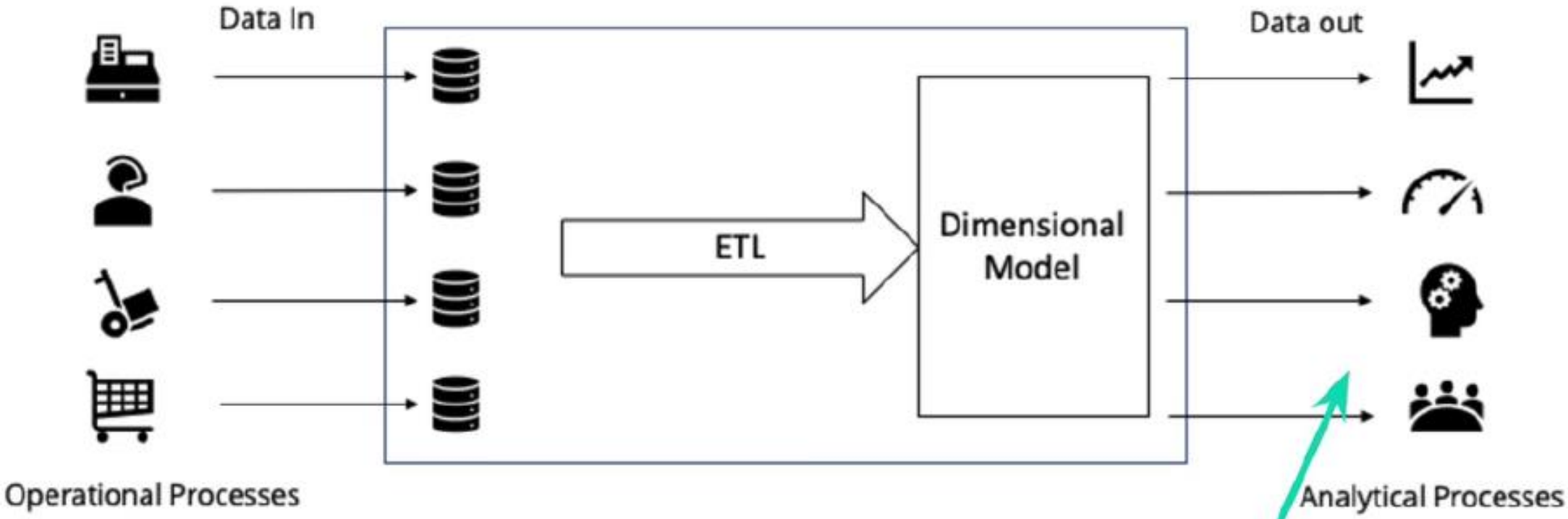
DWH: Tech Perspective



DWH: Tech Perspective



DWH: Tech Perspective



Data Warehouse Goals

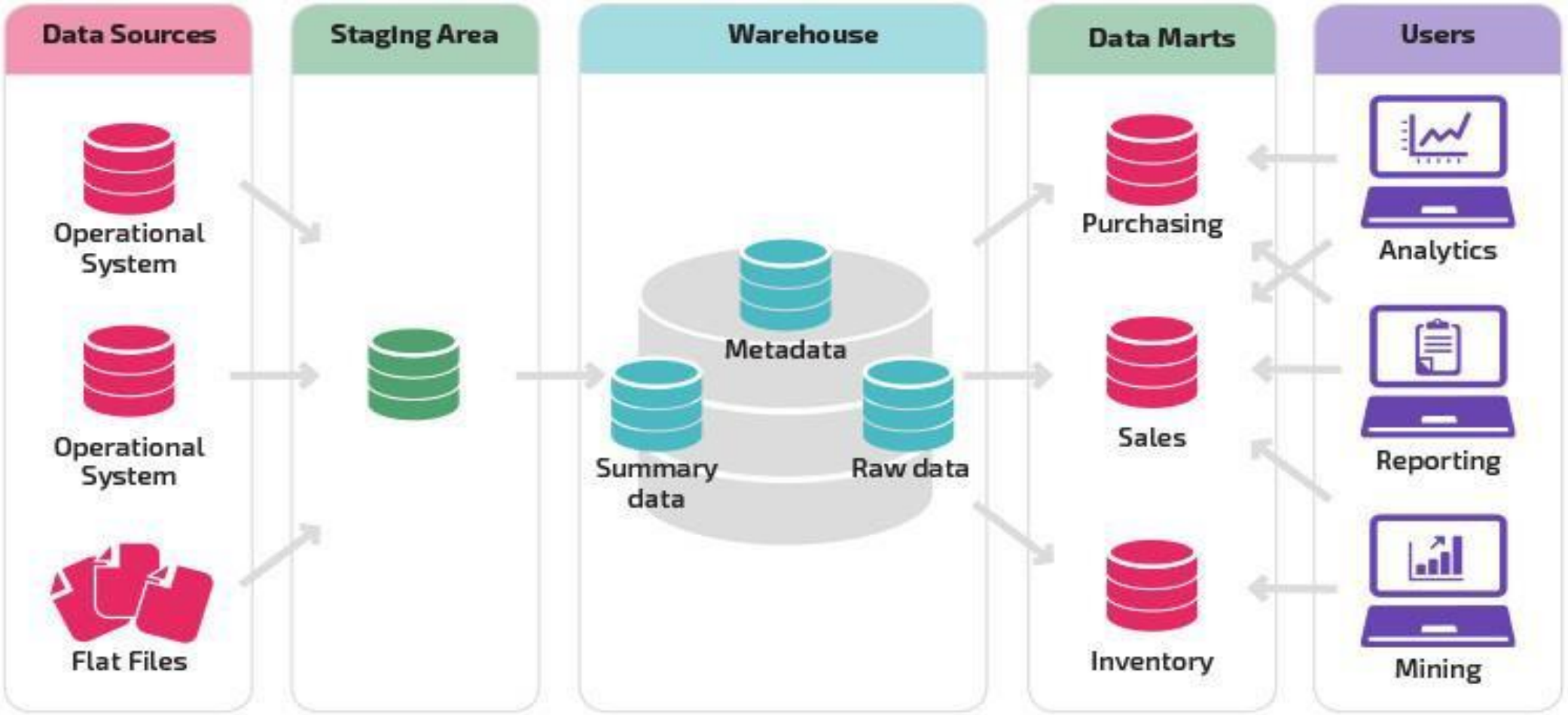
- Simple to understand
- Enterprise level information Consolidation
- Adaptive and resilient to change
- Handles new questions well
- Secure
- Improved business decision making

Data Marts VS Data Warehouse

- **Data Warehouse** is a large centralized repository of data that contains information from many sources within an organization. The collated data is used to guide business decisions through analysis, reporting, and data mining tools.

- **Data Mart** is a subset of a data warehouse oriented to a specific business line. Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization, for example, the sales department.

Data Marts VS Data Warehouse

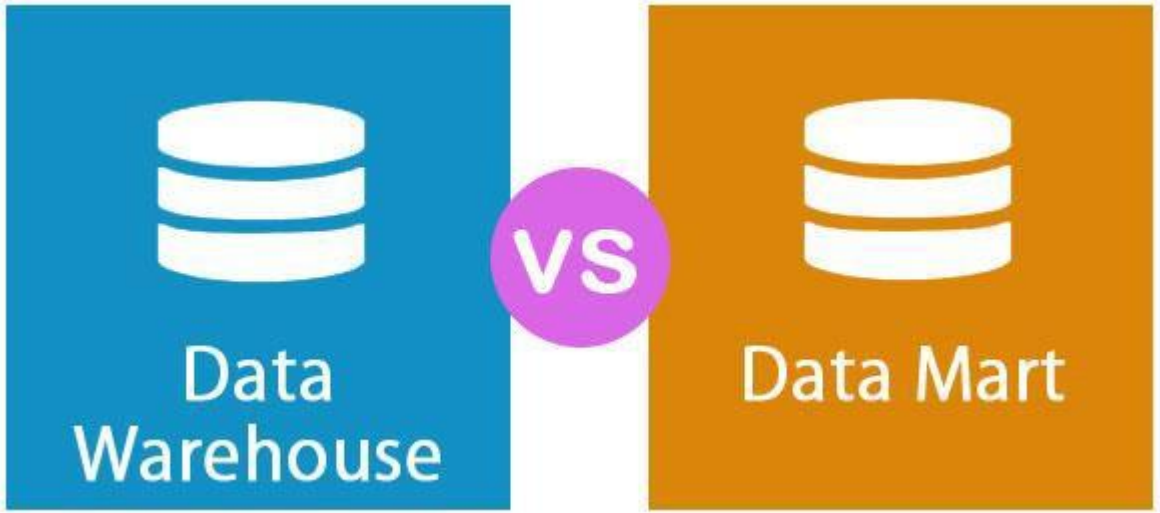


https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Inmon vs. Kimball

(Data Warehouse Structures)

- Two data warehouse pioneers, Bill Inmon and Ralph Kimball differ in their views on how data warehouses should be designed from the organization's perspective.



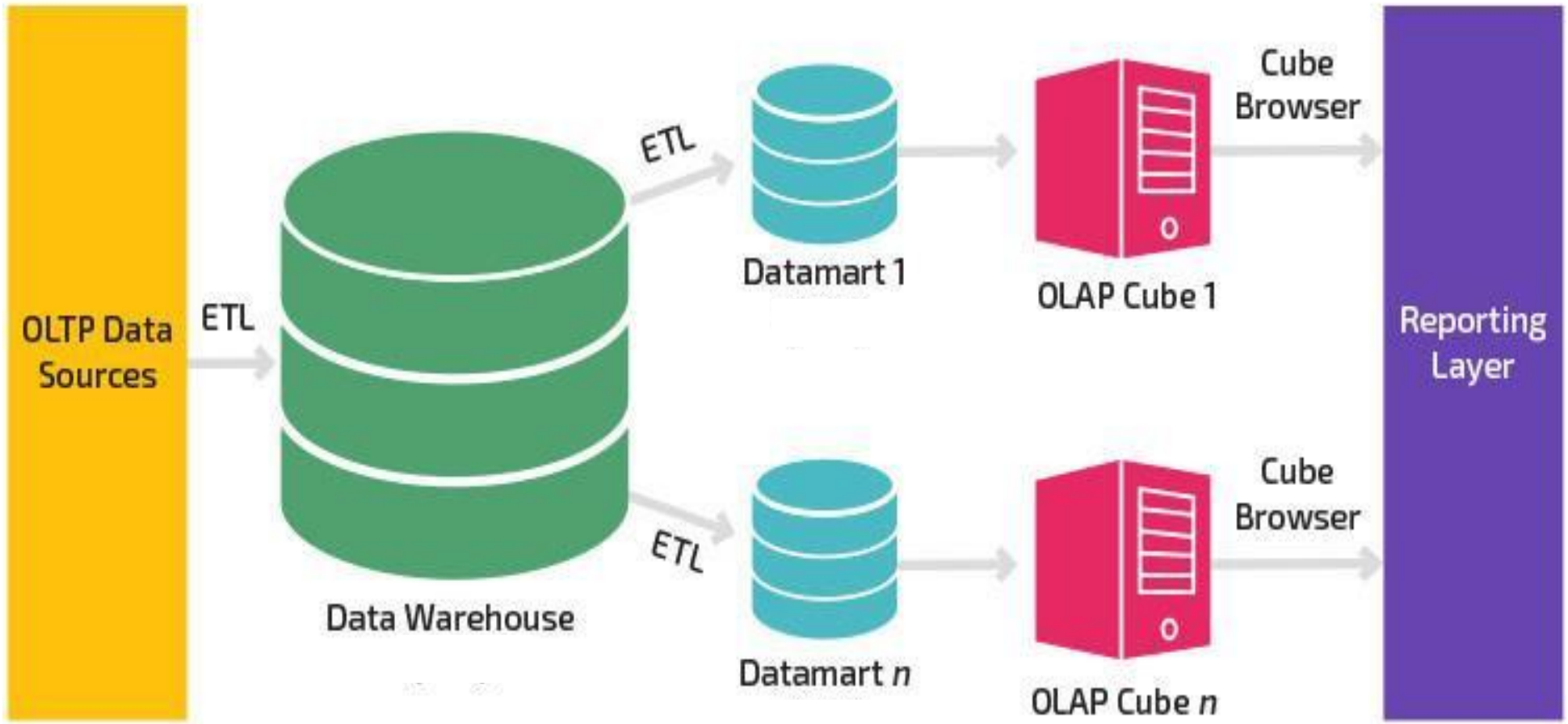
https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Bill Inmon's approach

- Favors a top-down design in which the data warehouse is the centralized data repository and the most important component of an organization's data systems.
- The Inmon approach first builds the centralized corporate data model, and the data warehouse is seen as the physical representation of this model.
- Dimensional data marts related to specific business lines can be created from the data warehouse when they are needed.

https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Inmon Model

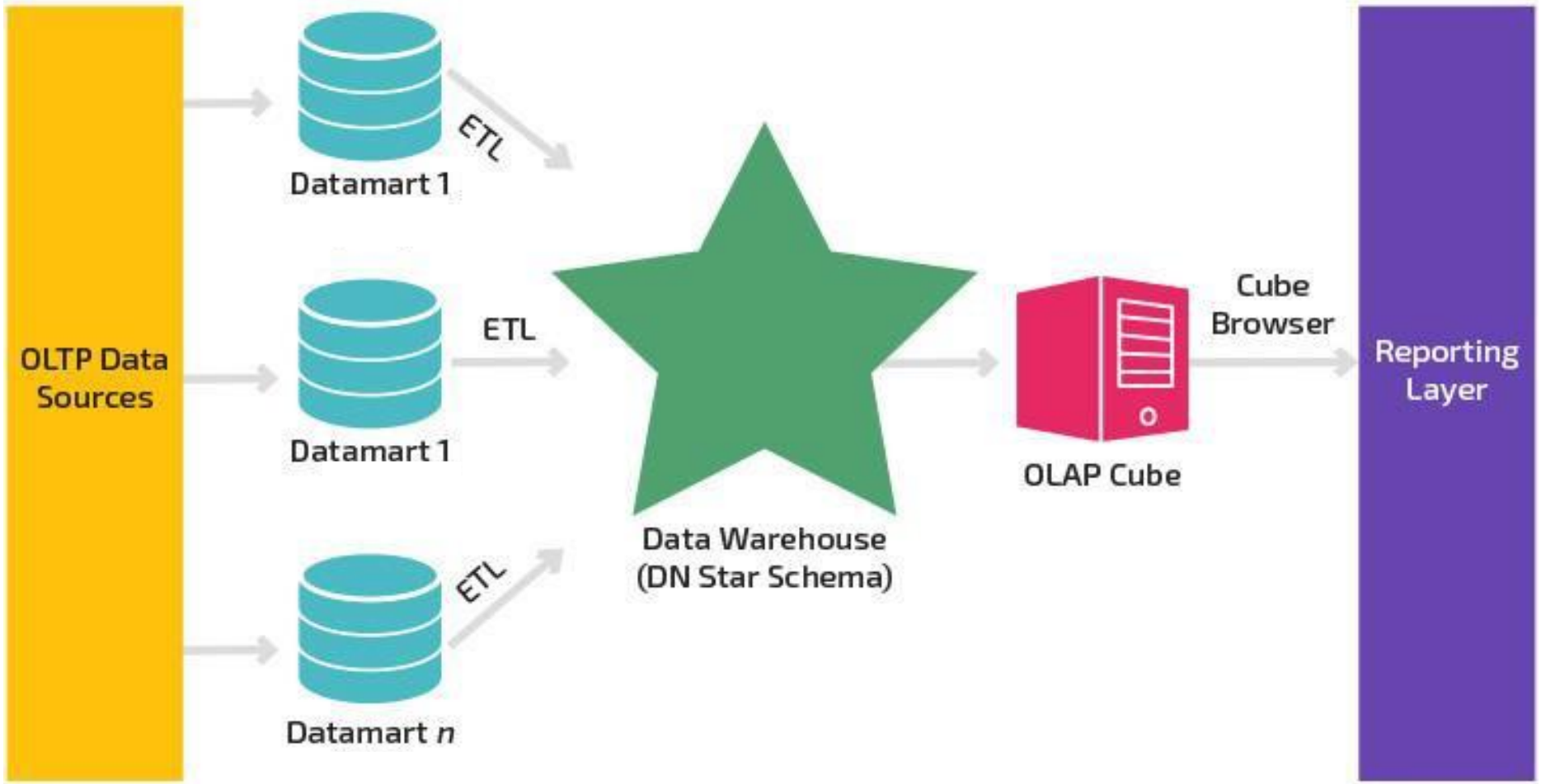

OLTP Data Sources
ETL
Data Warehouse
ETL
Datamart 1
OLAP Cube 1
Cube Browser
ETL
Datamart n
OLAP Cube n
Cube Browser
Reporting Layer

Ralph Kimball's approach

- **Ralph Kimball's** data warehouse design starts with the most important business processes. In this approach, an organization creates data marts that aggregate relevant data around subject-specific areas.
- The data warehouse is the combination of the organization’s individual data marts.
- Data warehouse is the conglomerate of a number of data marts. This is in contrast to Inmon's approach, which creates data marts based on information in the warehouse.

https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Kimball Model



Data Marts – Use Cases

- Marketing analysis and reporting favor a data mart approach because these activities are typically performed in a specialized business unit, and do not require enterprise-wide data.
- A financial analyst can use a finance data mart to carry out financial reporting.

https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Data Warehouse – Use Cases

- A company considering an expansion needs to incorporate data from a variety of data sources across the organization to come to an informed decision. This requires a data warehouse that aggregates data from sales, marketing, store management, customer loyalty, supply chains, etc.

- Many factors drive profitability at an insurance company. An insurance company reporting on its profits needs a centralized data warehouse to combine information from its claims department, sales, customer demographics, investments, and other areas.

https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Fact and Dimension Tables

Fact and Dimension Tables

- Work together to create an organized data model
- While fact and dimension are not created differently in the DDL, they are conceptual and extremely important for organization.

Fact Tables

Fact tables consist of the measurements, metrics of facts of a business process.

- Fact tables are made up of facts (events that have actually happened).
- Fact tables can be aggregations of data and aren’t meant to be updated at place.
- Fact tables normally have integers or numbers.
- Fact tables also typically have quantitative data. The quantity sold, the price per item, total price, and so on.

Dimension

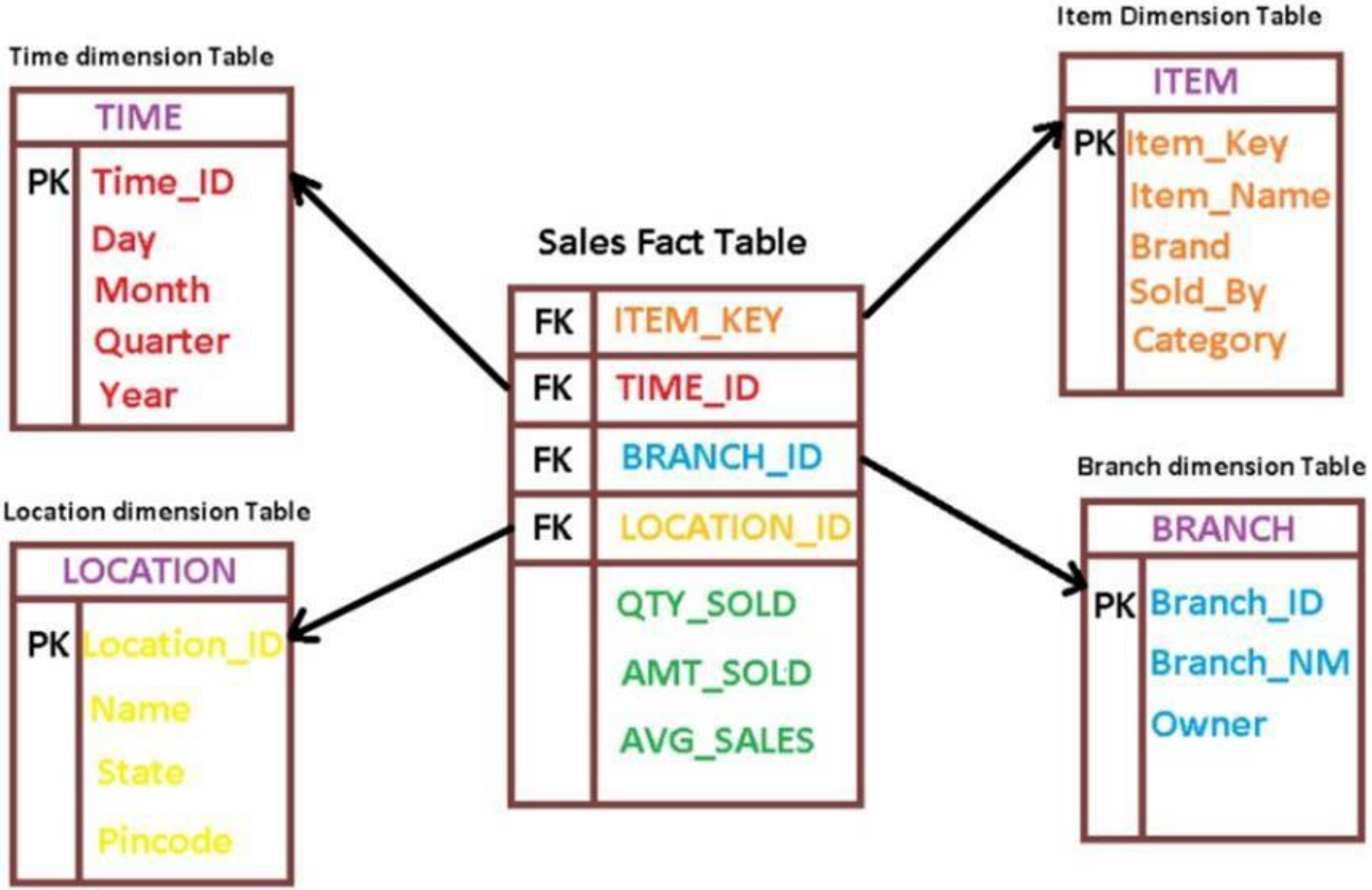
A structure that categorizes facts and measures in order to enable users to answer business questions. Dimensions are people, products, place and time.

- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- Dimension tables are de-normalized tables.
- The Dimension Attributes are the various columns in a dimension table
- Dimensions offers descriptive characteristics of the facts with the help of their attributes

Fact or Dimension Dilemma

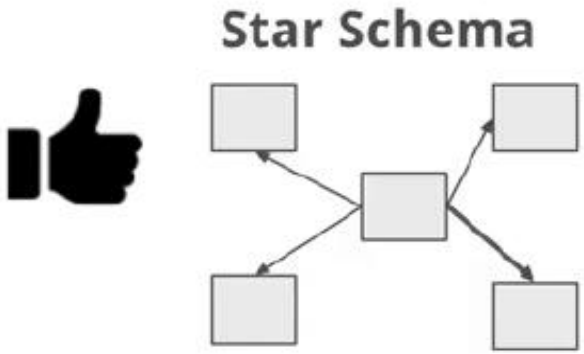
- For facts, If you're unsure if a column is a fact or dimension, the simplest rule is that a fact is usually: **Numeric & Additive**
- Example facts:
- A comment on an article represents an *event* but we can not easily make a statistic out of its content per se (Not a good fact)
- Invoice number is numeric but adding it does not make sense (Not a good fact)
- Total amount of an invoice could be added to compute total sales (A good fact)

- Example dimensions:
  - Date & time are always a dimension
  - Physical locations and their attributes are good candidates dimensions
  - Human Roles like customers and staff always good candidates for dimensions
  - Goods sold always good candidates for dimensions

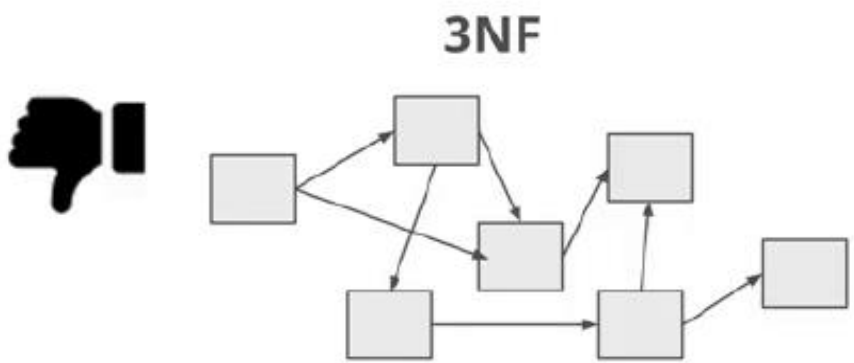

Time dimension Table
TIME
PK
Time_ID
Day
Month
Quarter
Year
Item Dimension Table
ITEM
PK
Item_Key
Item_Name
Brand
Sold_By
Category
Sales Fact Table
FK
ITEM_KEY
FK
TIME_ID
FK
BRANCH_ID
FK
LOCATION_ID
QTY_SOLD
AMT_SOLD
AVG_SALES
Location dimension Table
LOCATION
PK
Location_ID
Name
State
Pincode
Branch dimension Table
BRANCH
PK
Branch_ID
Branch_NM
Owner

Dimensional Modelling Goals

- Easy to understand
- Fast analytical query performance



Joins with dimensions only
Good for OLAP not OLTP



Lots of expensive joins
Hard to explain to business users

Implementing Different Schemas

Two of the most popular (because of their simplicity) data mart schemas for data warehouses are:

- Star Schema
- Snowflake Schema

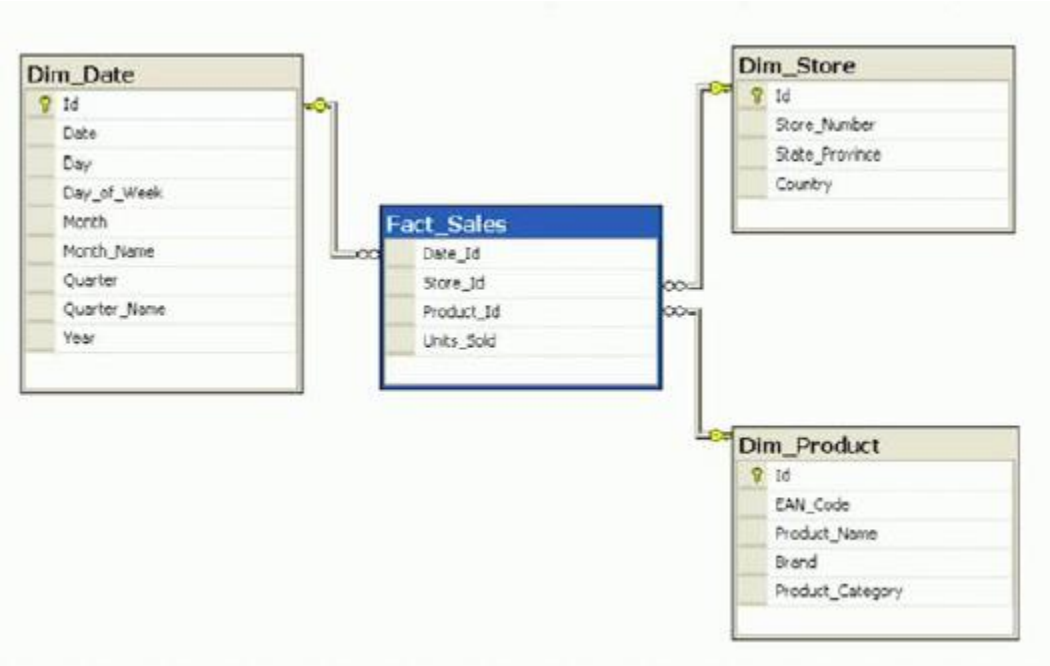
Star Schema

Star Schema is the simplest style of data mart schema.

The star schema consists of one fact table referencing any number of dimension tables.

Why "star" schema?

- Gets its name from the physical model resembling a star shape
- A fact table is at its center
- Dimension table surrounds the fact table representing the star's points.



Music Store DataBase with Star Schema

Customer

Customer ID	Name	Rewards
1	Amanda	Y
2	Toby	N

Customer Transactions Fact Table

Customer ID	Store Id	Spent
1	1	20.50
2	1	35.21

Items Purchased

Customer ID	Item number	Item Name
1	1	Rubber Soul
2	3	Let It Be

Store

Store Id	State
1	CA
2	WA

Star Schema: Characteristics

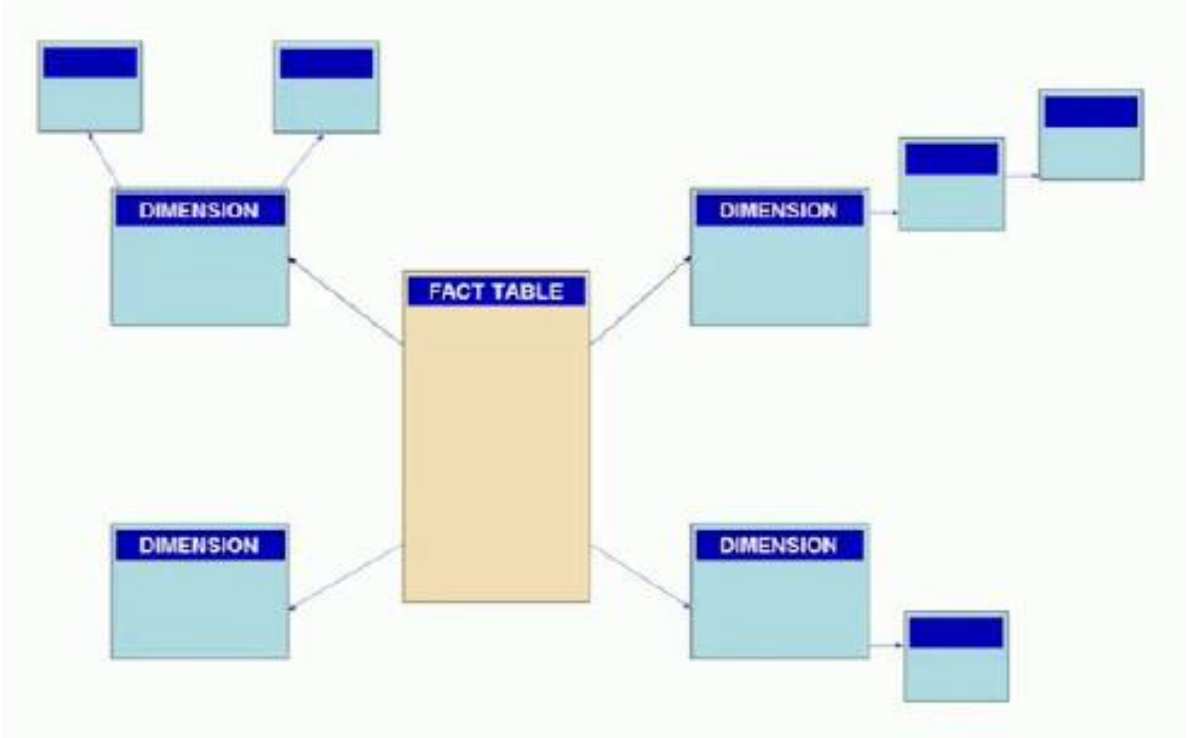
- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized.
- The schema is widely supported by BI Tools

Snowflake Schema

Logical arrangement of tables in a multidimensional database represented by centralized fact tables which are connected to multiple dimensions.

Why "snowflake" schema?

"A complex snowflake shape emerges when the dimensions of a snowflake schema are elaborated, having multiple levels of relationships, child tables having multiple parents."



Snowflake Schema: Characteristics

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Snowflake vs Star

1. Star schema dimension tables are not normalized, snowflake schemas dimension tables are normalized.
2. Snowflake schemas will use less space to store dimension tables but are more complex.
3. Star schemas will only join the fact table with the dimension tables, leading to simpler, faster SQL queries.
4. Snowflake schemas have no redundant data, so they're easier to maintain.