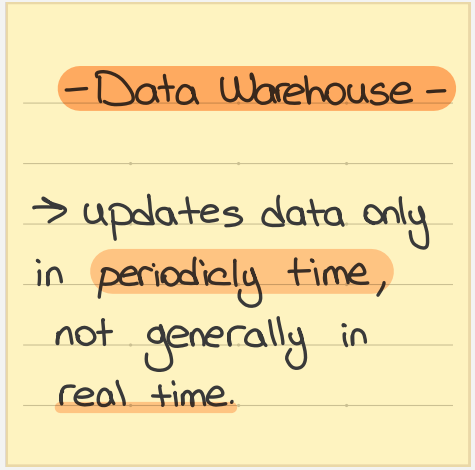
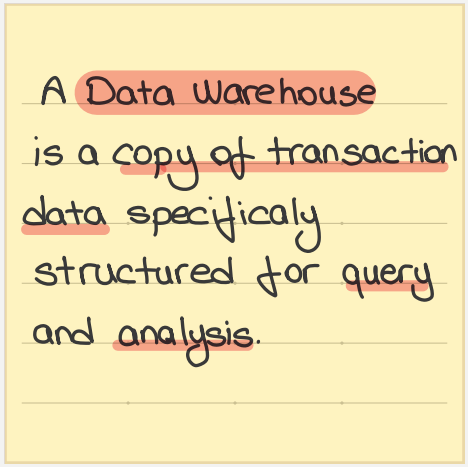
BD II

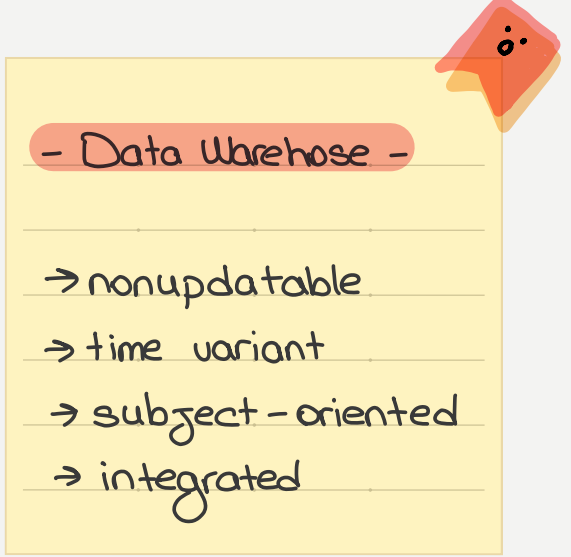
MOTIVATION

- "Modern organization is drowning in data but starving for information".
- Operational processing (transaction processing) captures, stores and manipulates data to support daily operations.
- Information processing is the analysis of data or other forms of information to **support decision making.**
- Data warehouse can consolidate and integrate information from many internal and external sources and arrange it in a meaningful format for **making business decisions**.





DEFINITION



- **Data Warehouse**: (W.H. Immon)
- A subject-oriented, integrated, time-variant, nonupdatable collection of data used in support of management decision-making processes
  - **Subject-oriented:** e.g. customers, patients, students, products
  - **Integrated:** Consistent naming conventions, formats, encoding structures; from multiple data sources
  - **Time-variant:** Can study trends and changes
  - **Non updatable:** Read-only, periodically refreshed

characteristics

- **Data Warehousing**: The process of constructing and using a data warehouse

DATA WAREHOUSE – SUBJECT-ORIENTED

↳ around major subjects

- Organized around major subjects (ana konular), such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view (özlü görünüm) around particular subject issues by excluding data (→ verileri hariç tutmak) that are not useful in the decision support process

– Subject-Oriented–

→ organized arround major subjects.

→ Focusing the modelling, analysis of data to help decision makers.

– in decision support process;

→ provide a simple and concise view by excluding data.

Not in daily operations or transaction processing.

↳ hareket işleme

→A data warehouse integrates data from multiple sources

DATA WAREHOUSE – INTEGRATED

multiple sources

→birden fazla, heterojen veri kaynağı elde edilerek oluşur.

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
- When data is moved to the warehouse, it is converted.

→veriler depoya (DW) taşındığında dönüştürülür.

– integrated –

- integrating multiple data sources.
- data cleaning
- data integration
- naming conventions
- encoding structures

entegre
→birbiriyle bağlanma yoluyla tüm duruma gelen, tümleşik

DATA WAREHOUSE – TIME VARIANT

historical data

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years) → tarihsel bir perspektiften bilgi sağlar (ör. son 5-10 yıl)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly (açıkça) or implicitly (dolaylı)
  - But the key of operational data may or may not contain "time element".

– time variant –

→ longer

→ contains time

→ key of operational data may or may not contain "time element"

DATA WAREHOUSE – NON UPDATABLE

↗ Çalışma ortamında dönüştürülen fiziksel olarak ayrı bir veri deposu.

A physically separate store of data transformed from the operational environment.

Operational update of data does not occur in the data warehouse environment.

Does not require transaction processing, recovery, and concurrency control mechanisms.

Requires only two operations in data accessing:

Initial loading of data and access of data.

↳ verilerin ilk yüklenmesi ↳ verilere erişim

– non updatable –

→ a physically seperate store

→ does'n accur in DW

* initial loading

* access data

NEED FOR DATA WAREHOUSING

Integrated, company-wide view of high-quality information (from disparate databases)

Separation of operational and informational systems and data (for improved performance)

güncel → | genel →

Characteristic	Operational Systems	Informational Systems
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time(snapshots) and predictions
Primary Users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

- Operational System-
(OLTP) → (OnLine Transaction Process)
→ optimize for quick inserts and updates
→ looks up of single rows or a small collection rows

- Information System-
→ organized system that affects the interplay between people, processes and technology organization

NEED TO SEPARATE OPERATIONAL AND INFORMATION SYSTEMS

- **Three primary factors:**

  → farklı işletim sistemlerine dağılmış verileri merkezleştirir

- A data warehouse centralizes data that are scattered throughout disparate operational systems and makes them available for DS.
- A well-designed data warehouse adds value to data by improving their quality and consistency.
- A separate data warehouse eliminates much of the contention for resources that results when information applications are mixed with operational processing.

→ A Data Warehouse is organized around important subject areas.

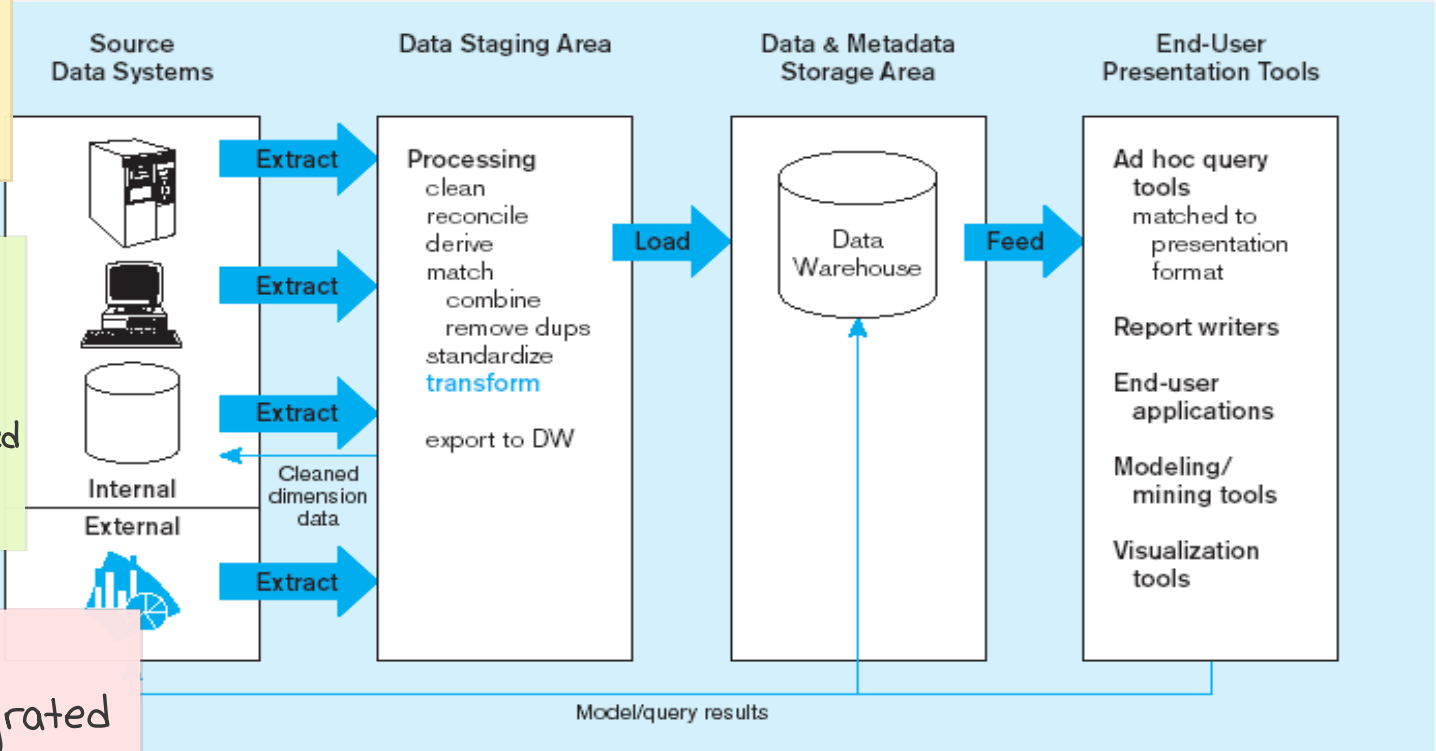
DATA WAREHOUSE ARCHITECTURES

1. Generic Two-Level Architecture
2. Independent Data Mart
3. Dependent Data Mart and Operational Data Store
4. Logical Data Mart and Active Warehouse
5. Three-Layer architecture

→ no intermediate application between client and data base layer.

GENERIC TWO-LEVEL ARCHITECTURE

→ it is a client-server applications.
→ there is a direct communication between client and data source server (data layer)



Periodic extraction → data is not completely current in warehouse

- Advantages -

- easy to maintain
- modification of the stored data is easy

- Disadvantages -

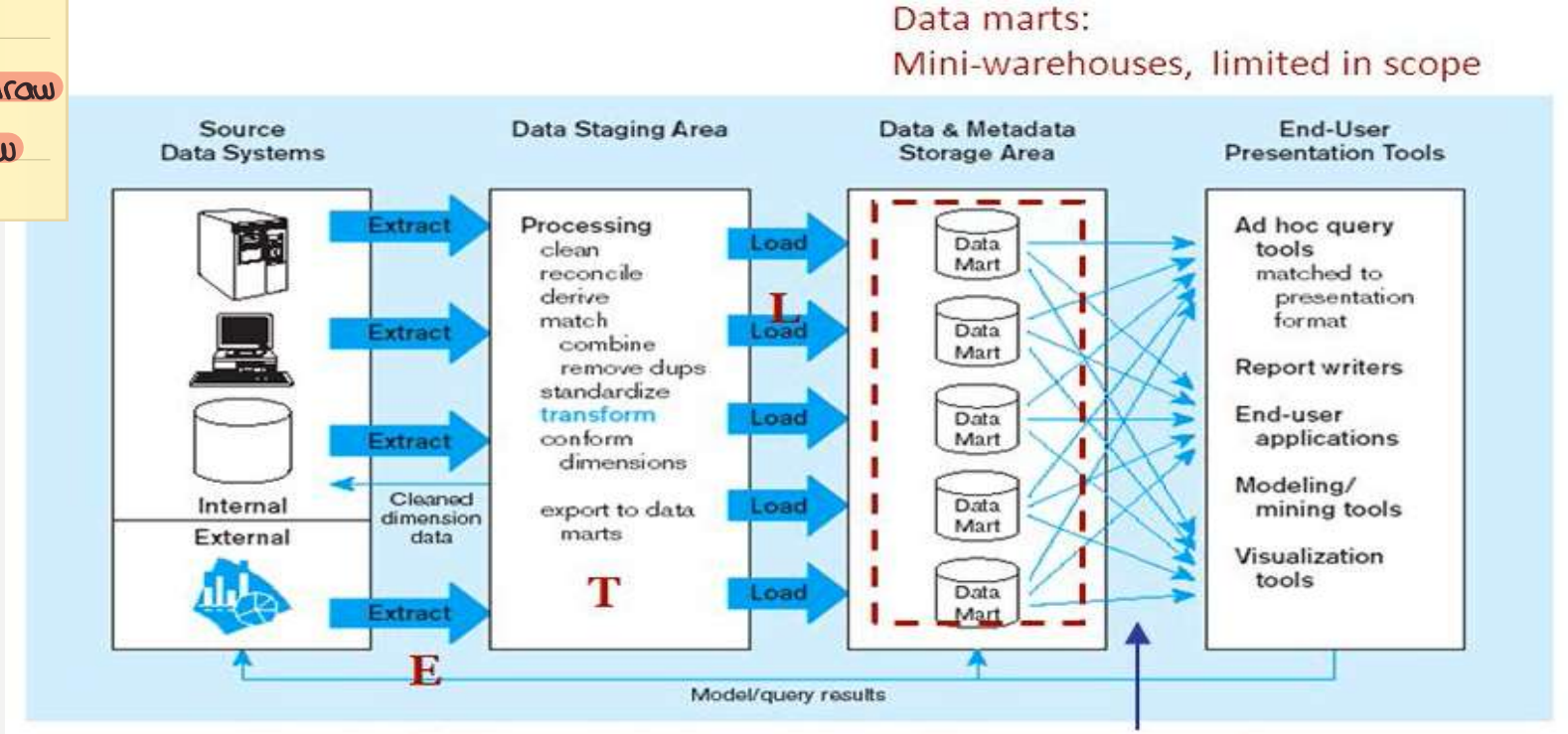
- performance will be degrated with increase user traffic.
- cost - ineffective

– Data Mart –

→ it is a simple form of a DW that is focused on a single subject.

→ Data mart usually draw data from only a few resources

INDEPENDENT DATA MART



Separate ETL for each ***independent*** data mart

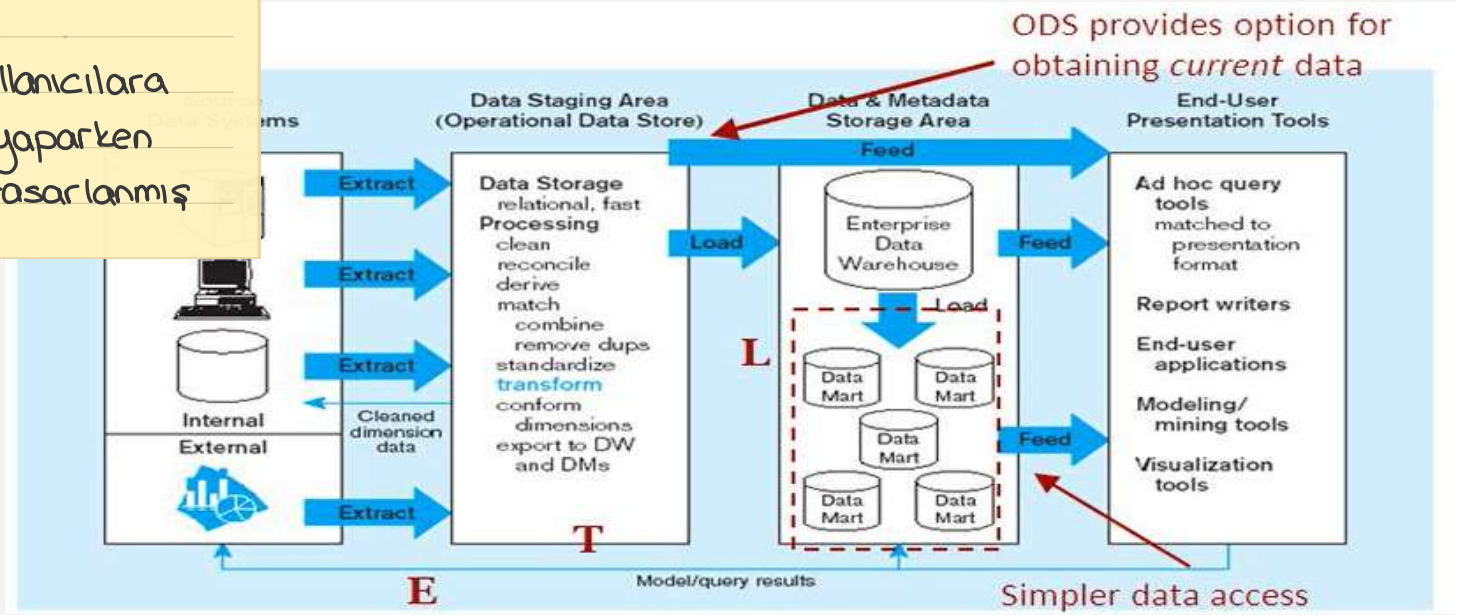
Data access complexity due to ***multiple*** *data marts*

• Independent data mart: a data mart filled with data extracted from the operational environment without benefits of a data warehouse. → Bir veri ambarının faydaları olmadan operasyonel ortamdan çıkarılan verilerle doldurulmuş bir data martı

DEPENDENT DATA MART WITH OPERATIONAL DATA STORE

→yalnızca kurumsal veri ambarından ve onun verilerinden doldurulan bir data martı.
→ODS: operasyonel kullanıcılara karar destek işlemleri yaparken hizmet vermek üzere tasarlanmış

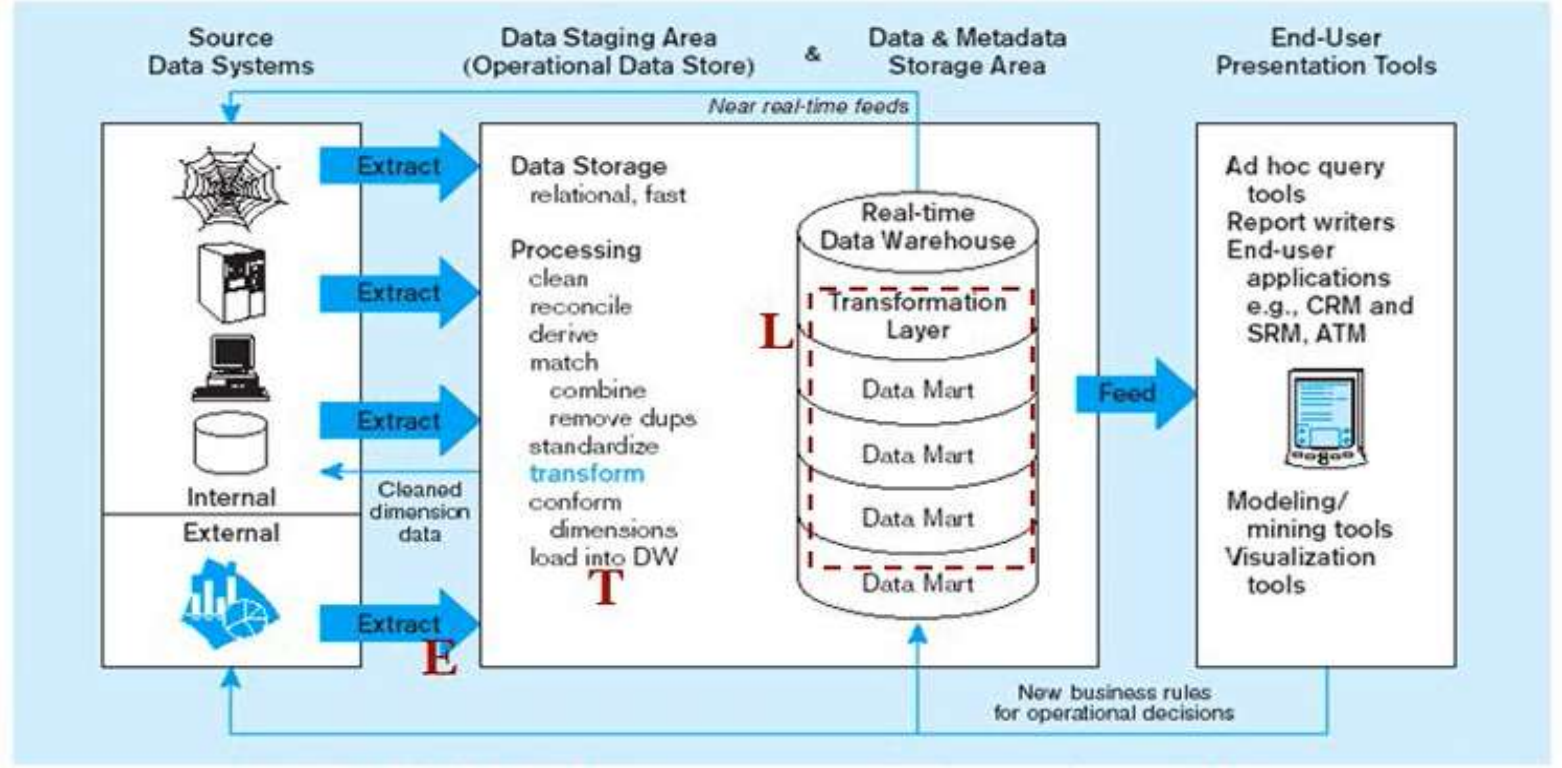
- entegre
- konu odaklı
- güncellenebilir
- güncel değerli
- işletme açısından ayrıntılı bir veri tabanı



- **Dependent data mart:** A data mart filled exclusively from the enterprise data warehouse and its reconciled data. (→uzlaştırılan veriler)
- **Operational data store (ODS):** An integrated, subject-oriented, updatable, current-valued, enterprise-wise, detailed database designed to serve operational users as they do decision support processing.

LOGICAL DATA MART AND ACTIVE DATA WAREHOUSE

(Hybrid DW)



@Active

ACTIVE DATA WAREHOUSE

• **Active Data Warehouse:** An enterprise data warehouse that accepts near-real-time feeds of transactional data from the systems of record, analyzes warehouse data, and in near-real-time relays business rules to the data warehouse and systems of record so that immediate actions can be taken in response to business events.

↳ Kayıt sistemlerinden gerçek zamanlıya yakın işlem verisi beslemelerini kabul eden, depo verilerini analiz eden, ve gerçek zamana yakın iş kurallarını veri ambarına ve kayıt sistemlerine aktaran bir kurumsal veri ambarı; iş olaylarına yanıt olarak acil önlemler alınabilir.

→near real-time updates
→at least one data mart
→data than can extracted from numerous internal and external sources

DATA WAREHOUSE VS. DATA MART

- DW -

Scope
→ centralize

Data
→ historical, detailed

Subjects
→ multiple subjects

Source
→ many internal and external sources

Other characteristic
→ Data-oriented
→ Large

- DM -

Scope
→ decentralized

Data
→ some history

Subjects
→ one central subjects

Sources
→ few internal and external

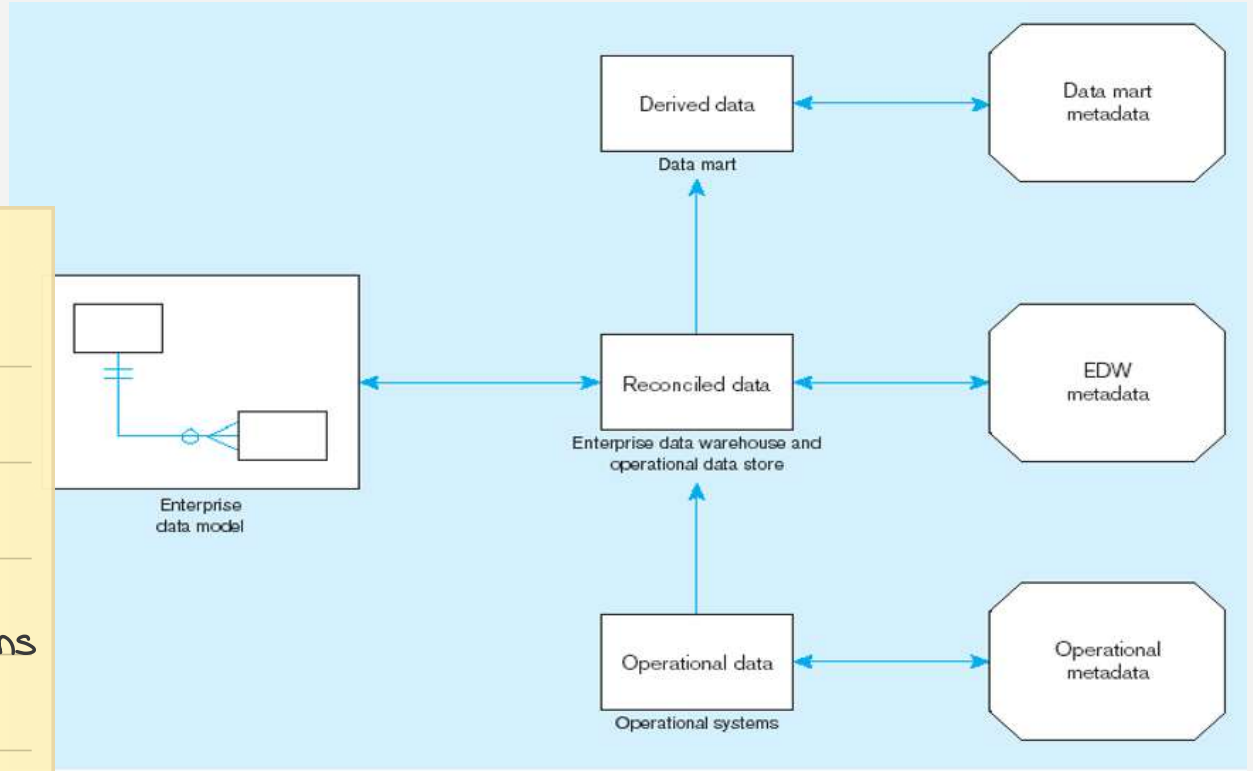
Other characteristics
→ project-oriented
→ start small and becomes larger

Data Warehouse	Data Mart
Scope	**Scope** ↗ Decision Support System
- Application independent	- Specific DSS application
- Centralized, possibly enterprise-wide	- Decentralized by user are
- Planned	- Organic, possibly not planned
Data	**Data**
- Historical, detailed and summarized	- Some history, detailed and summarized
- Lightly denormalized	- Highly denormalized
Subjects	**Subjects**
- Multiple subjects	- One central subject of concern to users
Sources	**Sources**
- Many internal and external sources	- Few internal and external sources
Other Characteristics	**Other Characteristics**
- Flexible	- Restrictive
- Data-oriented	- Project-oriented
- Long life	- Short file
- Large	- Start small, becomes lager
- Single complex structure	- Multi, semi-complex structures, together complex

THREE-LAYER ARCHITECTURE

client server application

→güncellenirken, programcılara yardım sağlar. Biri başka bir parte güncellerken, diğerleri bundan etkilenmez.



3 tiers

1- Presentation Tier:

front end-user interface

2- Application Tier: contains

functional business logic

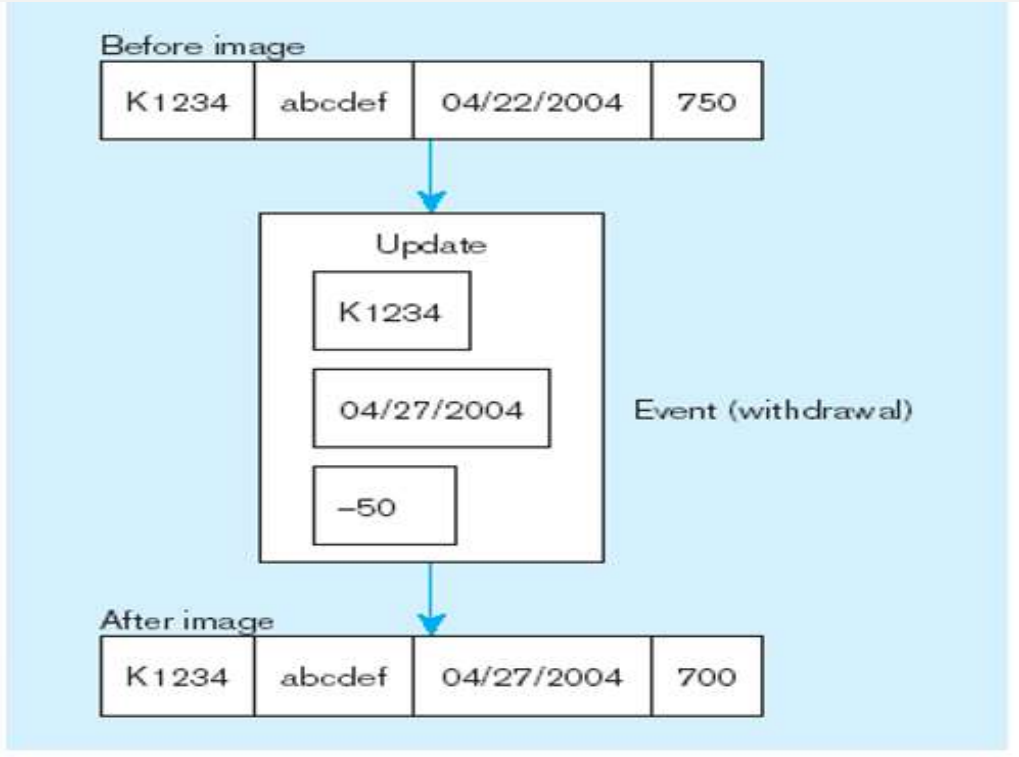
3- Data Tier: comprises

of the database system and access layer

THREE-LAYER ARCHITECTURE RECONCILED AND DERIVED DATA

- **Reconciled data:** detailed, current data intended to be the single, authoritative source for all decision support.
- **Derived data:** Data that have been selected, formatted, and aggregated for end-user decision support application.
- **Metadata:** technical and business data that describe the properties or characteristics of other data.

DATA CHARACTERISTICS - STATUS VS. EVENT DATA



Depoya eklendikten sonra veriler asla fiziksel olarak değiştirilmez veya silinmez.

DATA CHARACTERISTICS – TRANSIENT VS. PERIODIC DATA

Transient

Table X (10/05)

Key	A	B
001	a	b
002	c	d
003	e	f
004	g	h

Table X (10/06)

Key	A	B
001	a	b
002	r	d
003	e	f
004	y	h
005	m	n

Table X (10/07)

Key	A	B
001	a	b
002	r	d
003	e	t
005	m	n

Periodic

Table X (10/05)

Key	Date	A	B	Action
001	10/03	a	b	C
002	10/03	c	d	C
003	10/03	e	f	C
004	10/03	g	h	C

Table X (10/06)

Key	Date	A	B	Action
001	10/05	a	b	C
002	10/05	c	d	C
002	10/06	r	d	U
003	10/05	e	f	C
004	10/05	g	h	C
004	10/06	y	h	U
005	10/06	m	n	C

Table X (10/07)

Key	Date	A	B	Action
001	10/05	a	b	C
002	10/05	c	d	C
002	10/06	r	d	U
003	10/05	e	f	C
003	10/07	e	t	U
004	10/05	g	h	C
004	10/06	y	h	U
004	10/07	y	h	D
005	10/06	m	n	C

- Data are never physically altered or deleted once they have been added to the store

OTHER DATA WAREHOUSE CHANGE

- New descriptive attributes → yeni tanımlayıcı özellikler
- New business activity attributes → yeni iş etkinliği özellikleri
- New classes of descriptive attributes → yeni tanımlayıcı nitelik sınıfları
- Descriptive attributes become more refined → açıklayıcı özellikler daha rafine hale gelir
- Descriptive data are related to one another → açıklayıcı veriler birbirleriyle ilişkilidir
- New source of data → yeni veri kaynağı

DATA RECONCILIATION

- leads to inaccurate insight and issues with customer service
- reconcilation of data is also important for enterprise - control integration.

- Typical operational data is:
  - Transient – not historical
  - Not normalized (perhaps due to denormalization for performance)
  - Restricted in scope – not comprehensive
  - Sometimes poor quality – inconsistencies and errors
- After ETL, data should be:
  - Detailed – not summarized yet
  - Historical – periodic
  - Normalized – 3rd normal form or higher
  - Comprehensive – enterprise-wide perspective
  - Quality controlled – accurate with full integrity

- Use of Data Reconciliation helps you for extracting accurate and reliable information about the state of industry process from raw measurement data.

- helps you to produces a single consistent set of data representing the most likely process operation

THE ETL PROCESS

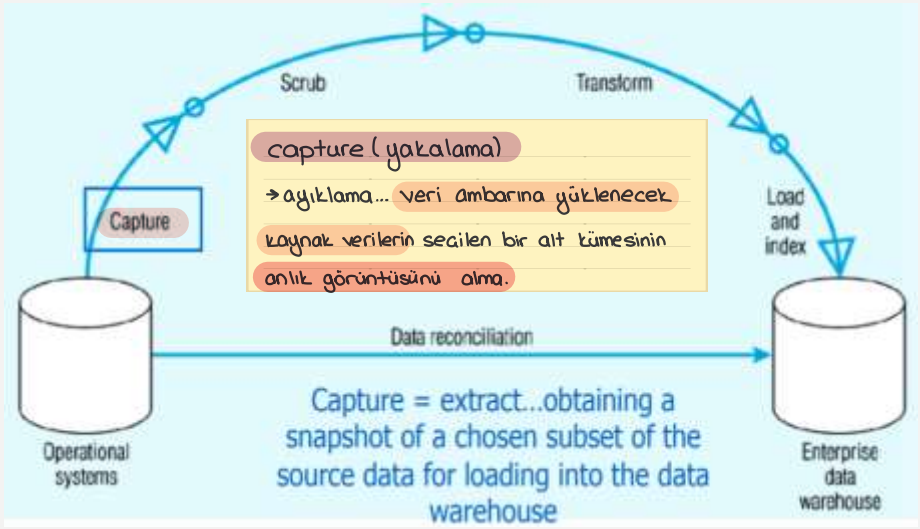
- Capture
- Scrub or data cleansing
- Transform
- Load and Index

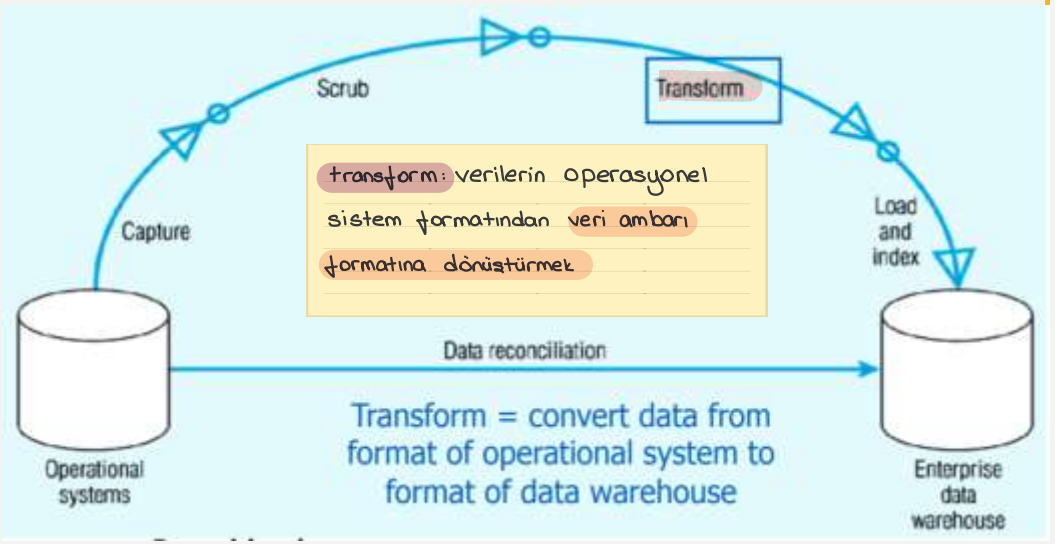
Extracts the data from different source systems, then transform the data and loads to DW system.

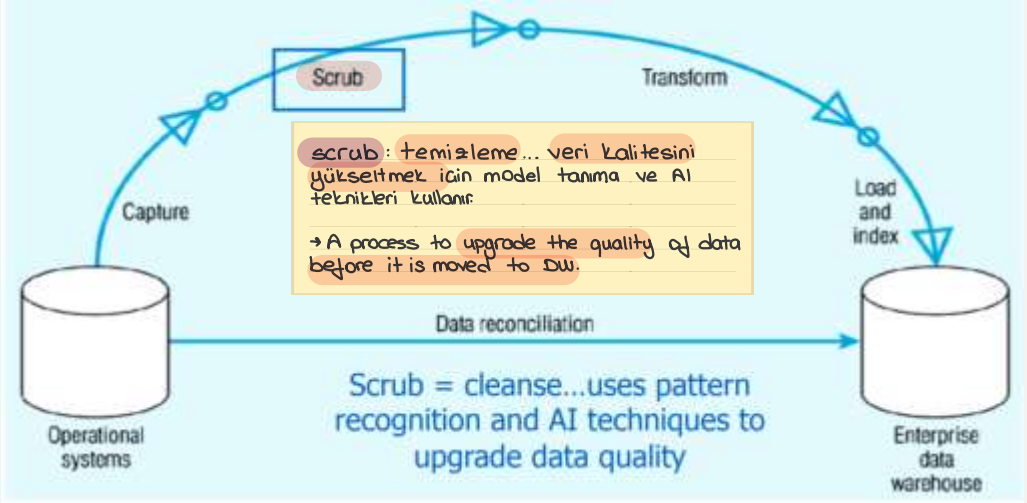
Operational System
→ a system that is used to run the business in real time and based on current data

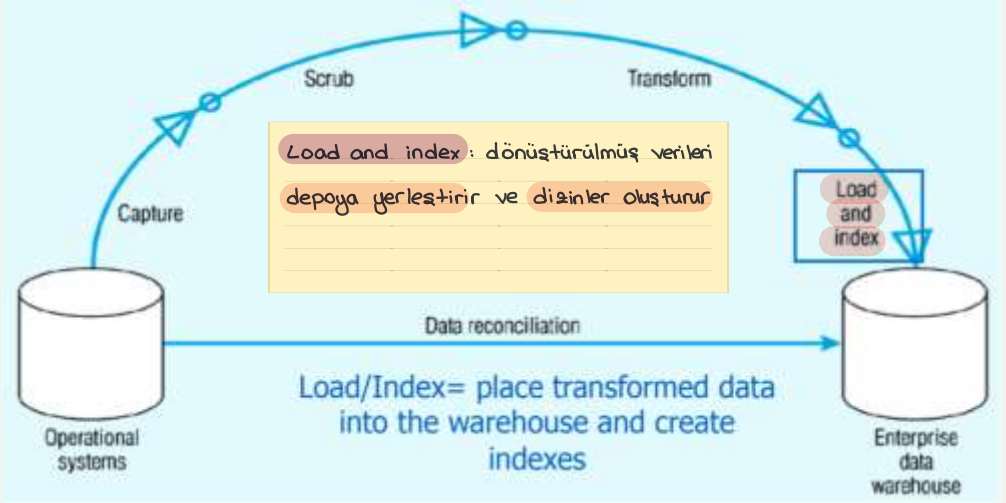
ETL = Extract, transform, and load

STEPS IN DATA RECONCILIATION









DATA TRANSFORMATION

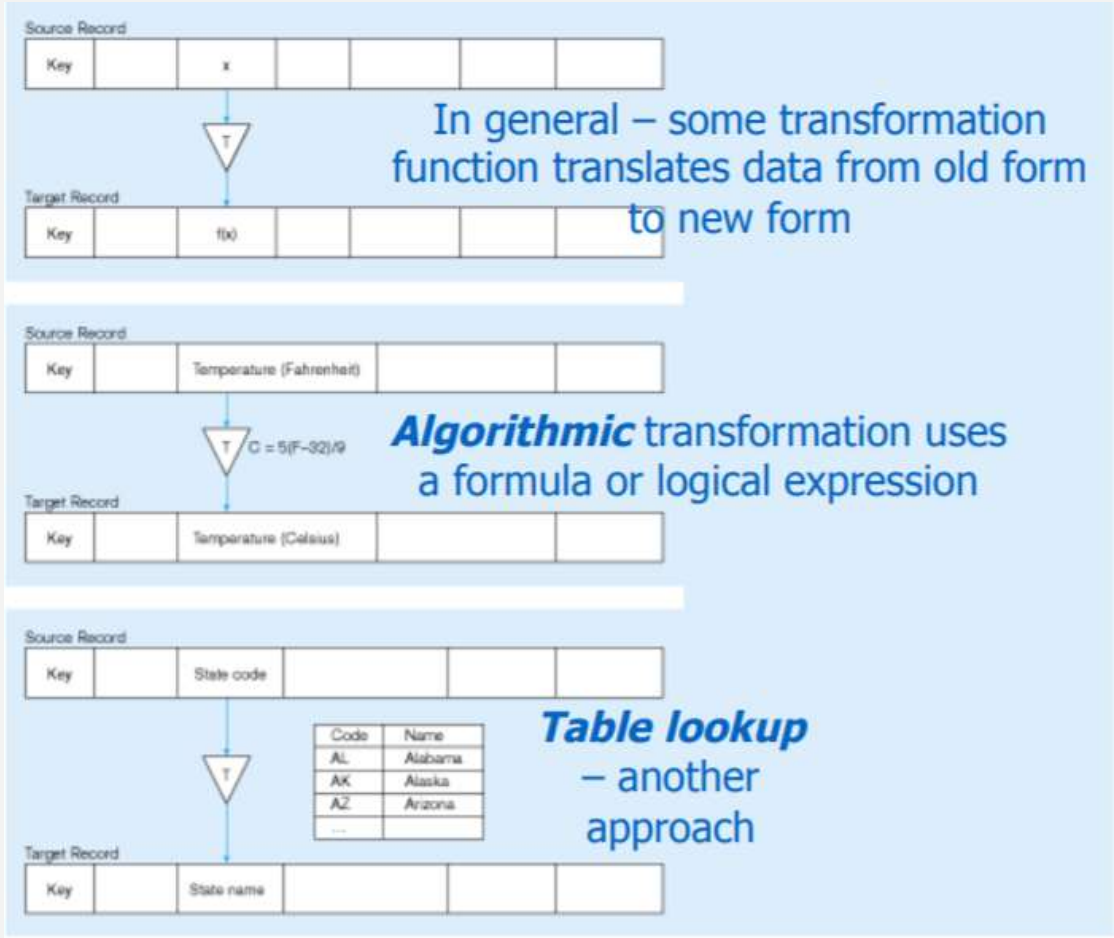
Veri dönüşümü, verileri kaynak işletim sistemlerinin biçimden kurumsal veri ambarı biçimine dönüştüren veri uzlaşmasının (data reconcilation) bileşenidir.

- Data transformation is the component of data reconcilation that converts data from the format of the source operational systems to the format of enterprise data warehouse.
- Data transformation consists of a variety of different functions:
  - record-level functions,
  - field-level functions and
  - more complex transformation.

RECORD-LEVEL FUNCTION & FIELD-LEVEL FUNCTION

- **Record-level functions**
  - Selection: data partitioning → seçim: veri bölümleme
  - Joining: data combining → birleştirme : veri birleştirme
  - Normalization → normalleştirme
  - Aggregation: data summarization → toplama: veri özetleme
- **Field-level functions**
  - Single-field transformation: from one field to one field → bir alandan bir alana ○→○
  - Multi-field transformation: from many fields to one, or one field to many → birçok alandan bir alana veya bir alandan birçok alana

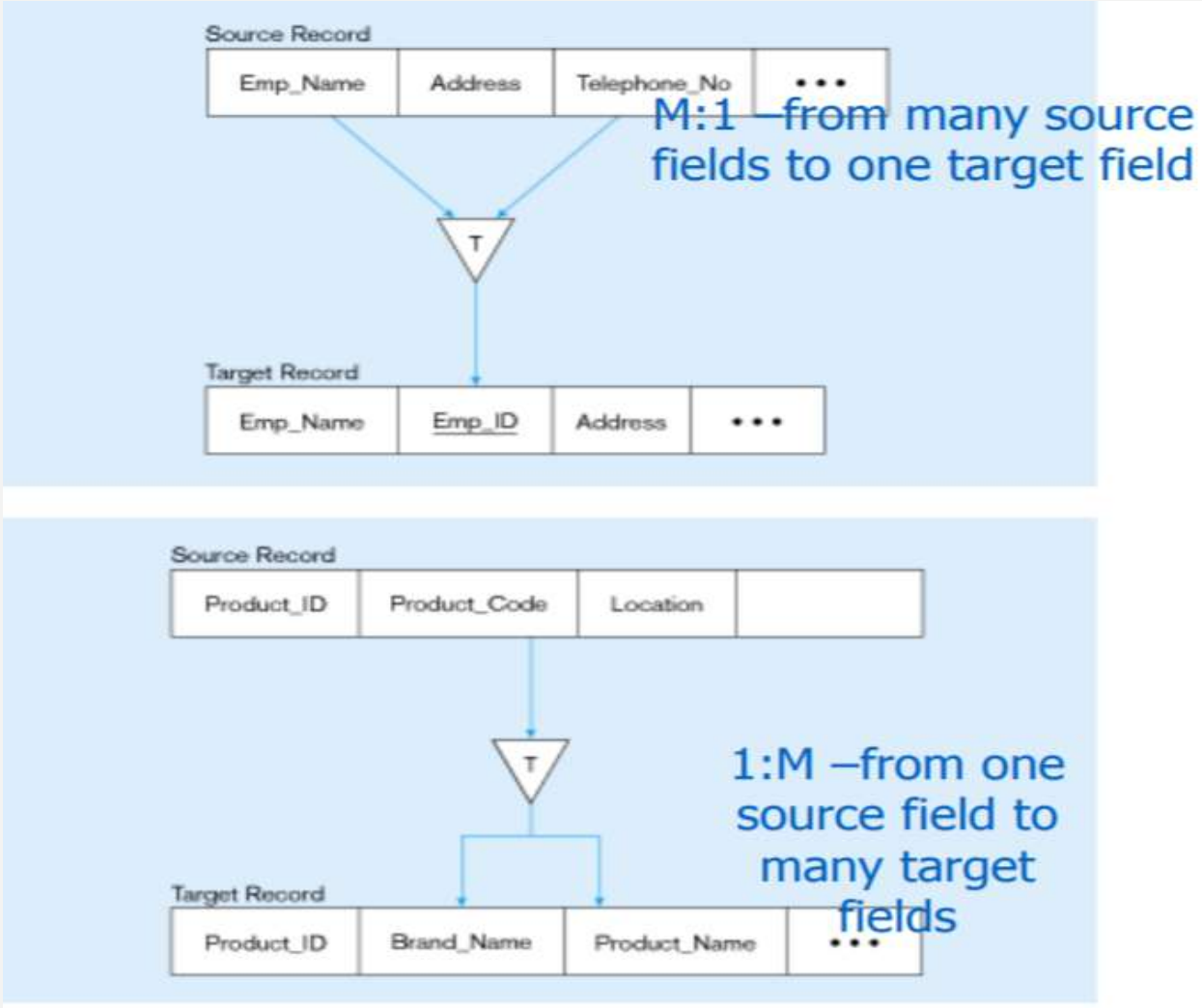
SINGLE-FIELD TRANSFORMATION



→ genel olarak – bazı dönüştürme işlemleri verileri eskiden yeni forma çevirir.

→ algoritmik dönüşüm bir formül veya mantıksal ifade kullanır

MULTIFIELD TRANSFORMATION



DERIVED DATA

objectives

→ decision support app.
→ fast response to queries
→ Ad-hoc query
→ Data mining

- Objectives
  - Ease of use for decision support applications
  - Fast response to predefined user queries
  - Customized data for particular target audiences
  - Ad-hoc query support → geçici sorgu desteği
  - Data mining capabilities
- Characteristics
  - Detailed (mostly periodic) data
  - Aggregate (for summary)
  - Distributed (to departmental servers)

characteristics

→ detailed
→ aggregate
→ distributed

A derived data element is a data element derived from other data element using a mathematical, logical or other type of transformation.

e.g, arithmetic formula, composition, aggregation

Most common data model = **star schema** (also called "dimensional model")

Fundamental schema among the data mart schema and its simplest.
→ used to develop or build a data WH and dimensional data marts.
→ includes one or more fact tables

THE STAR SCHEMA

- **Star schema:** is a simple database design in which dimensional (describing how data are commonly aggregated) are separated from fact or event data.
- A star schema consists of two types of tables: **fact tables** and **dimension table.**

A Fact Table is a primary table in a dimensional model.
– contains –
→ measurements / facts
→ foreign key to dimensional tables

– FACT TABLE –
contains
→ meauserement, metrics and facts about business process

location
→ center of star or snowflake schema

defined
→ by their grain, atomic level

⇒ helps to store report labels
⇒ does not contain hierarchy

– DIMENSION TABLE –
contains
→ descriptive attributes

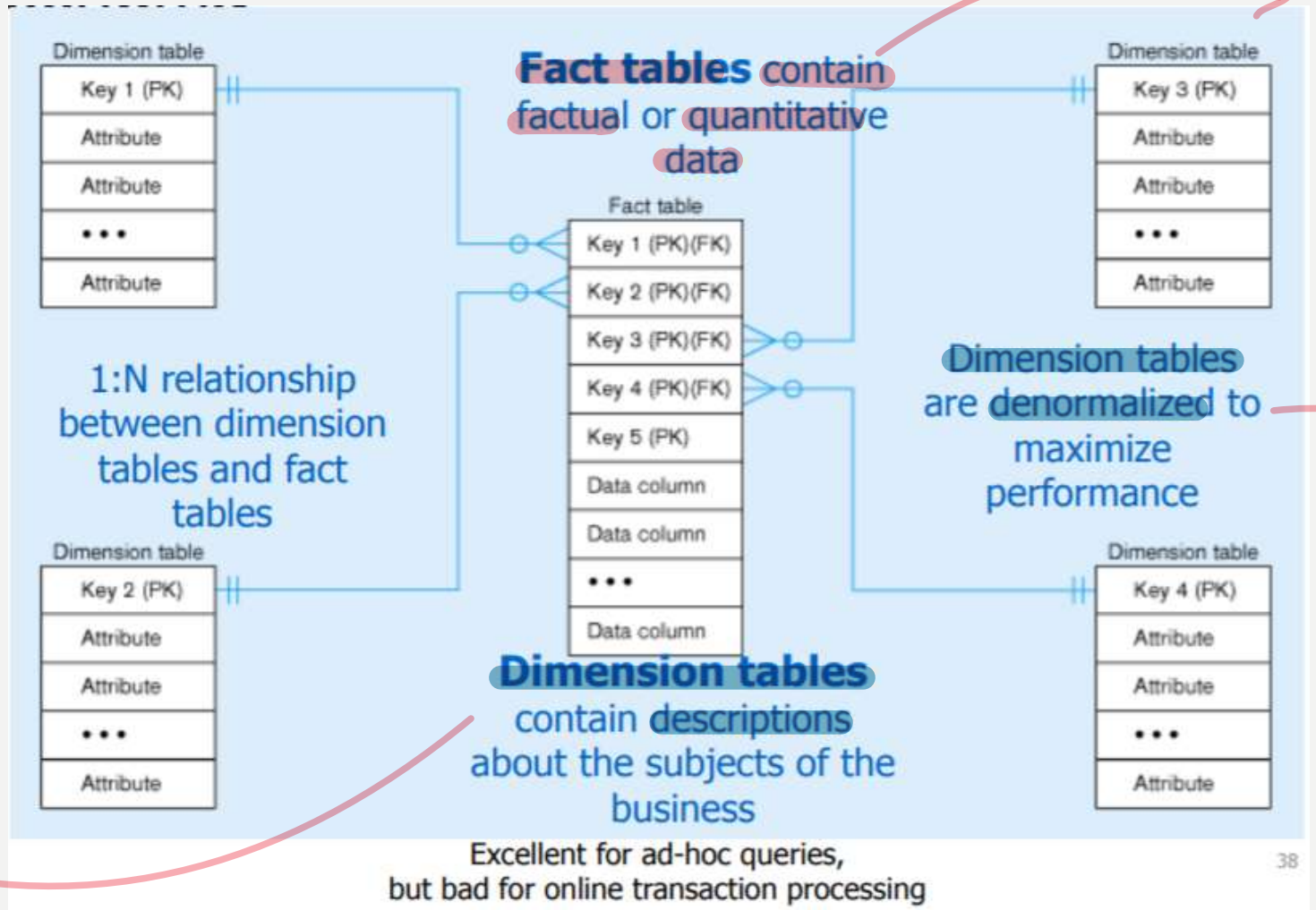
location
→ edges of star or snowflake schema

defined
→ wordy, descriptive, complete, quality assured

⇒ contains detailed data
⇒ contains hierarchy

– Dimension Table –
- contains dimensions of fact
- joined to fact table via foreign key.
- de-normalize tables
- can contain one or more hierarchial relationship.

COMPONENTS OF A STAR SCHEMA

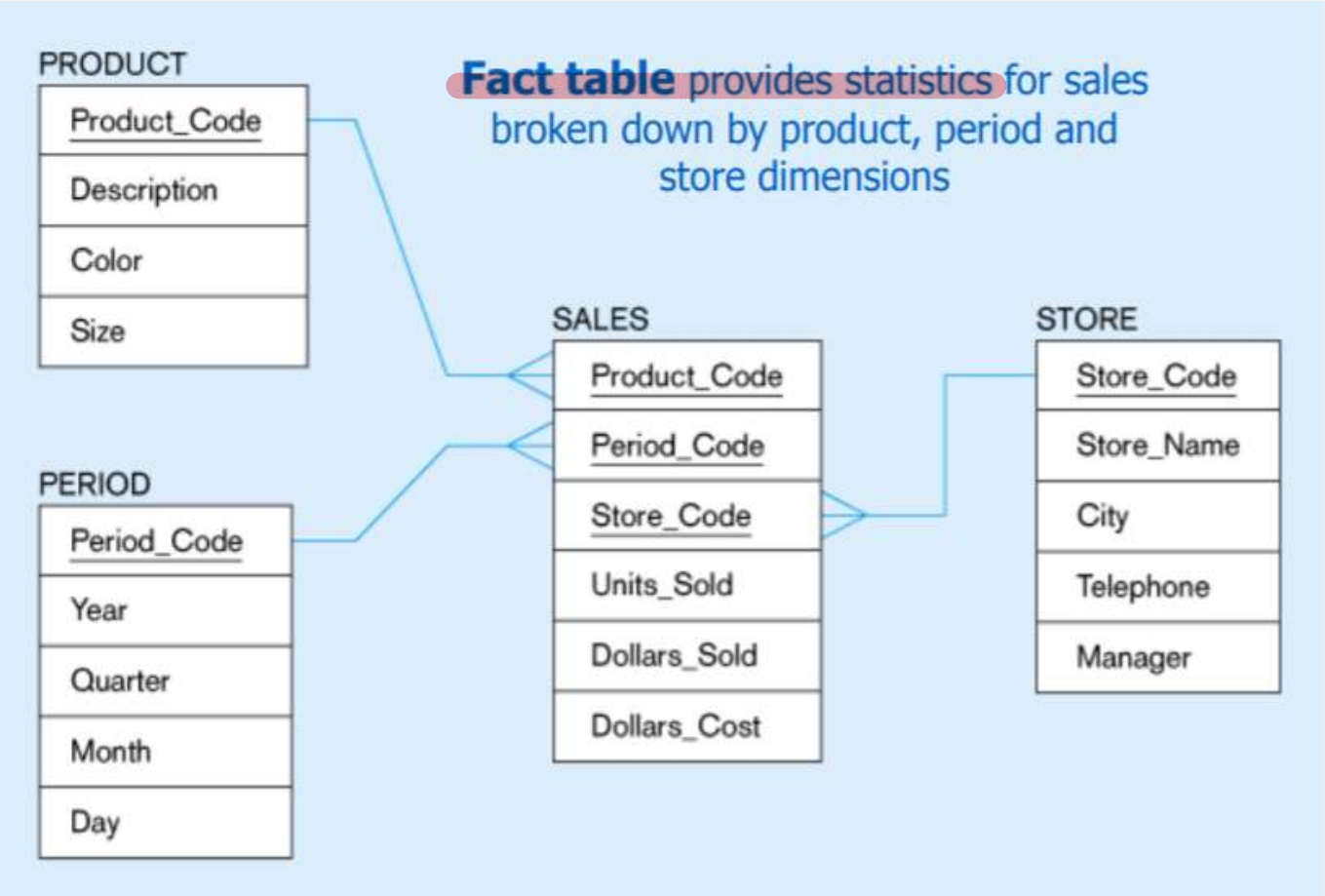


Fact Tables olgusal veya nicel veriler içerir

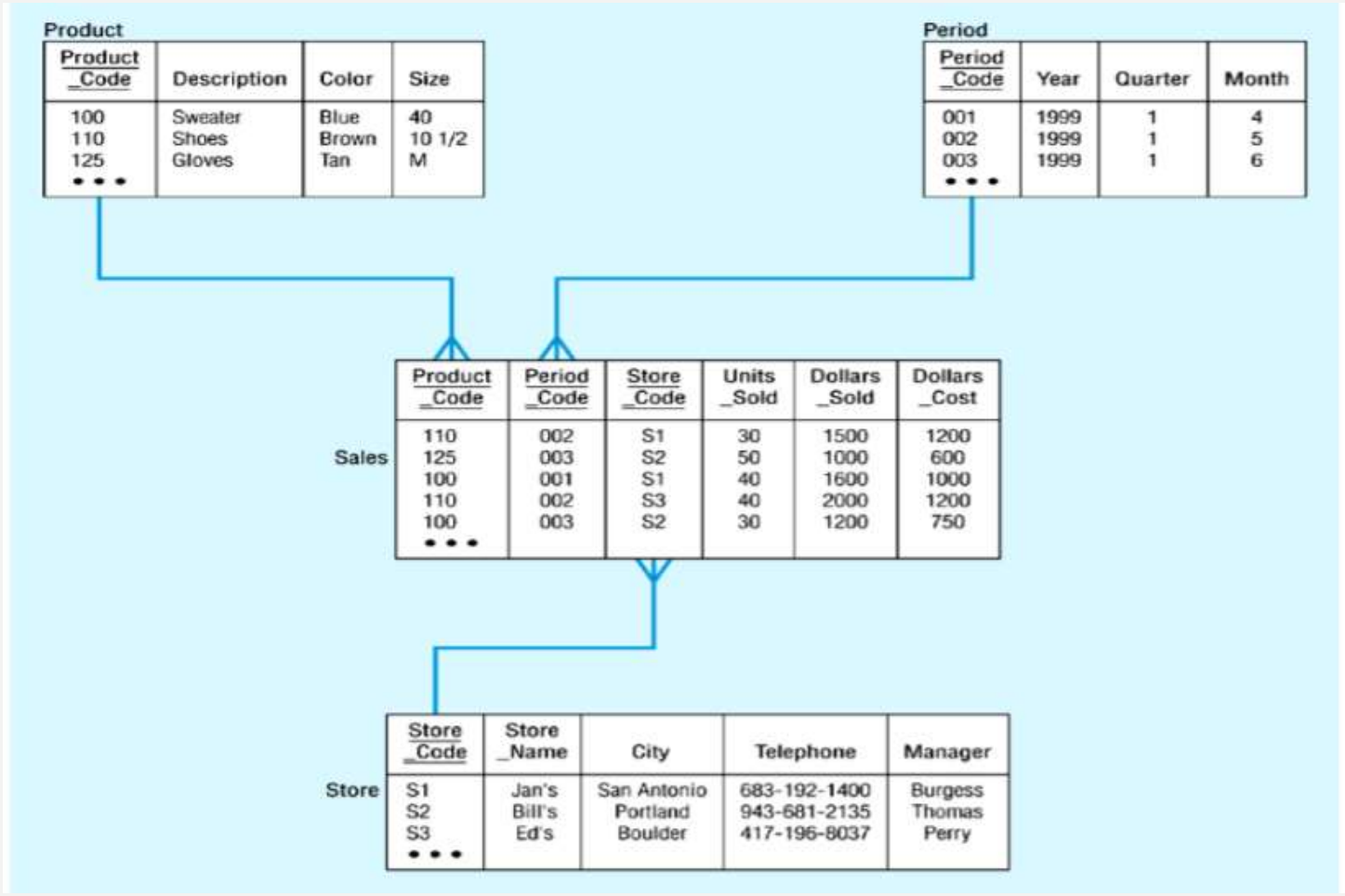
Dimension Table performansı en üst düzeye çıkarmak için denormalize edilir.

Dimension Tables işin konuları hakkında açıklamalar içerir.

STAR SCHEMA EXAMPLE



STAR SCHEMA WITH SAMPLE DATA

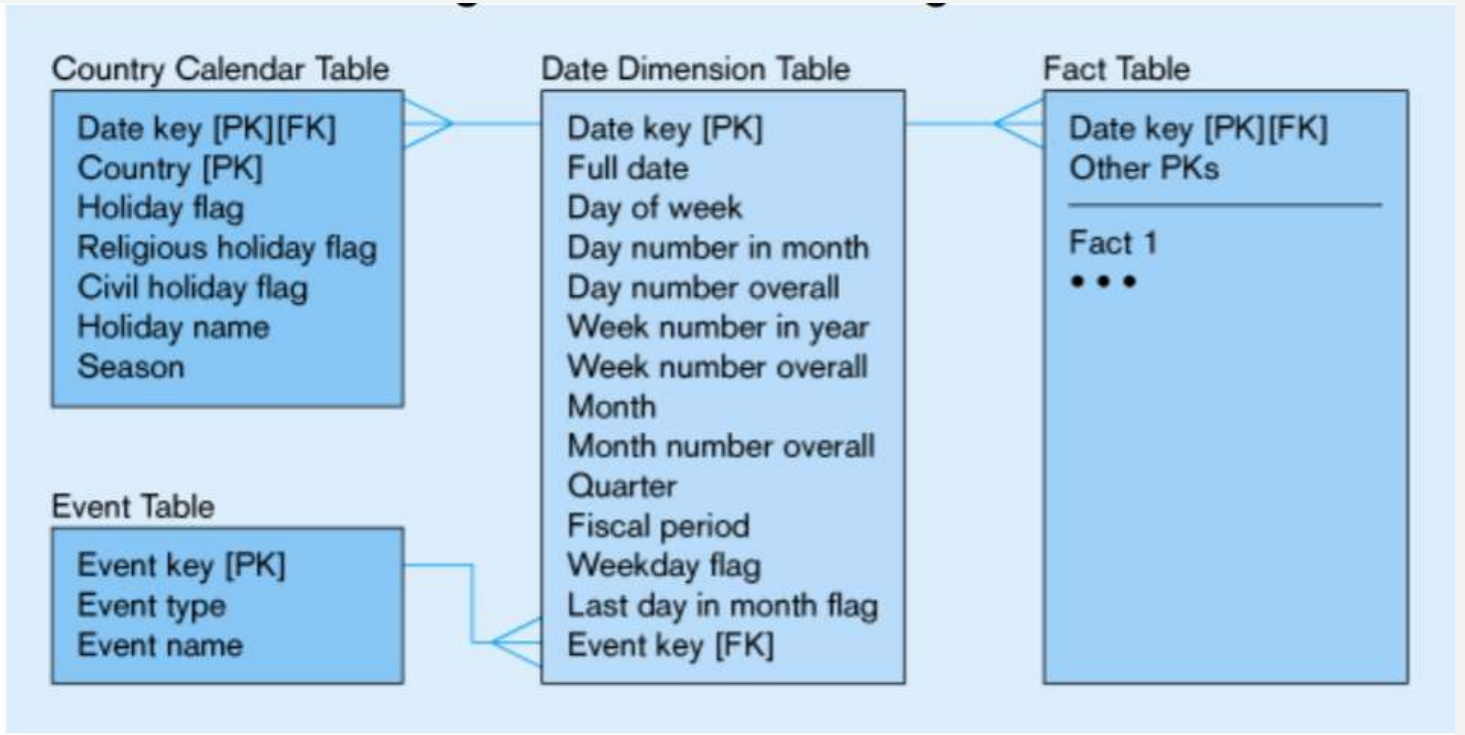


ISSUES REGARDING STAR SCHEMA

- Dimension table keys must be **surrogate** (non-intelligent and nonbusiness related), because:
  - Keys may change over time
  - Length/format consistency
- Granularity of Fact Table – what level of detail do you want?
  - Transactional grain – finest level
  - Aggregated grain – more summarized
  - Finer grains: better **market basket analysis** capability
  - Finer grain: more dimension tables, more rows in fact table
- Duration of the database
  - Ex: 13 months or 5 quarters
  - Some businesses need for a longer durations.
- Size of the fact table
  - Estimate the number of possible values for each dimension associated with the fact table.
  - Multiply the values obtained in the first step after making any necessary adjustments.

The granulatiy is the lowest level of information stored in fact table.

MODELING DATES

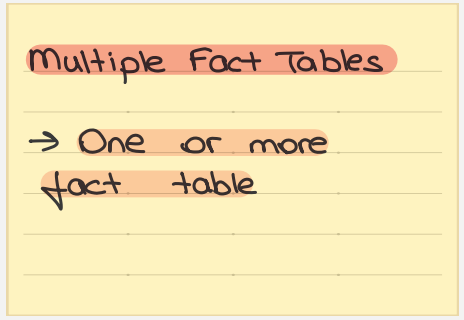


Fact tables contain time-period data
➔ Date dimensions are important

VARIATIONS OF THE STAR SCHEMA

↓ Types

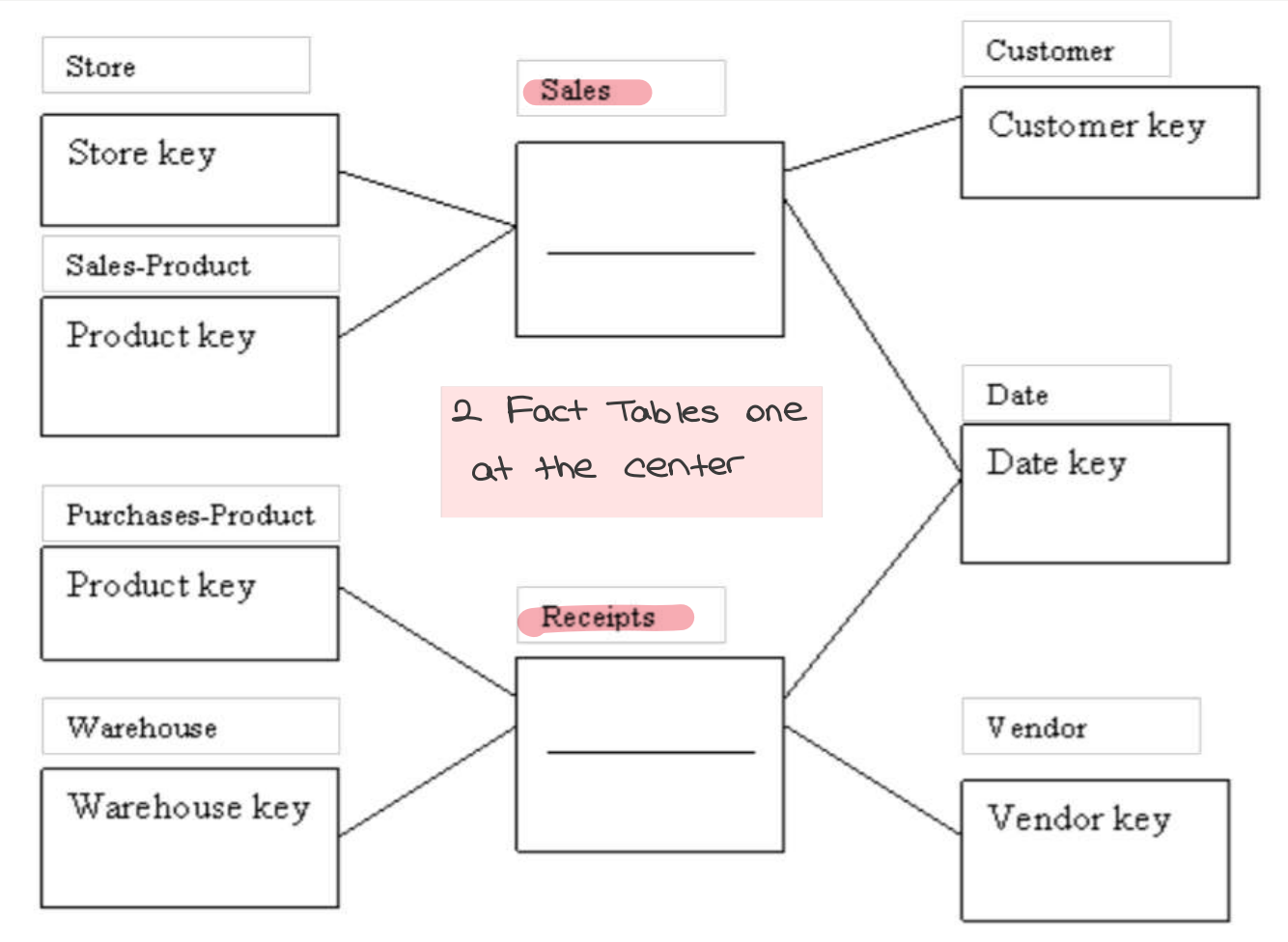
1. Multiple fact tables
2. Factless fact tables
3. Normalizing Dimension Tables
4. Snowflake schema



MULTIPLE FACT TABLES

- More than one fact table in a given star schema.
- Ex: There are 2 fact tables, one at the center of each star:
  - Sales – facts about the sale of a product to a customer in a store on a date.
  - Receipts - facts about the receipt of a product from a vendor to a warehouse on a date.
  - Two separate product dimension tables have been created.
  - One date dimension table is used.

MULTIPLE FACT TABLES



→ to record events

→ coverage informations

FACTLESS FACT TABLES

- There are applications in which fact tables do not have nonkey data but that do have foreign keys for the associated dimensions. → Fact tablolarının anahtar olmayan verilere sahip olmadığı, ancak ilişkili boyutlar için foreign keylere sahip olduğu uygulamalar vardır.
- The two situations:
  - To track events → olayları izlemek için
  - To inventory the set of possible occurrences(called coverage)
    ↳ olası olayların envanterini çıkarmak için (kapsama denir)

Factless Fact Table

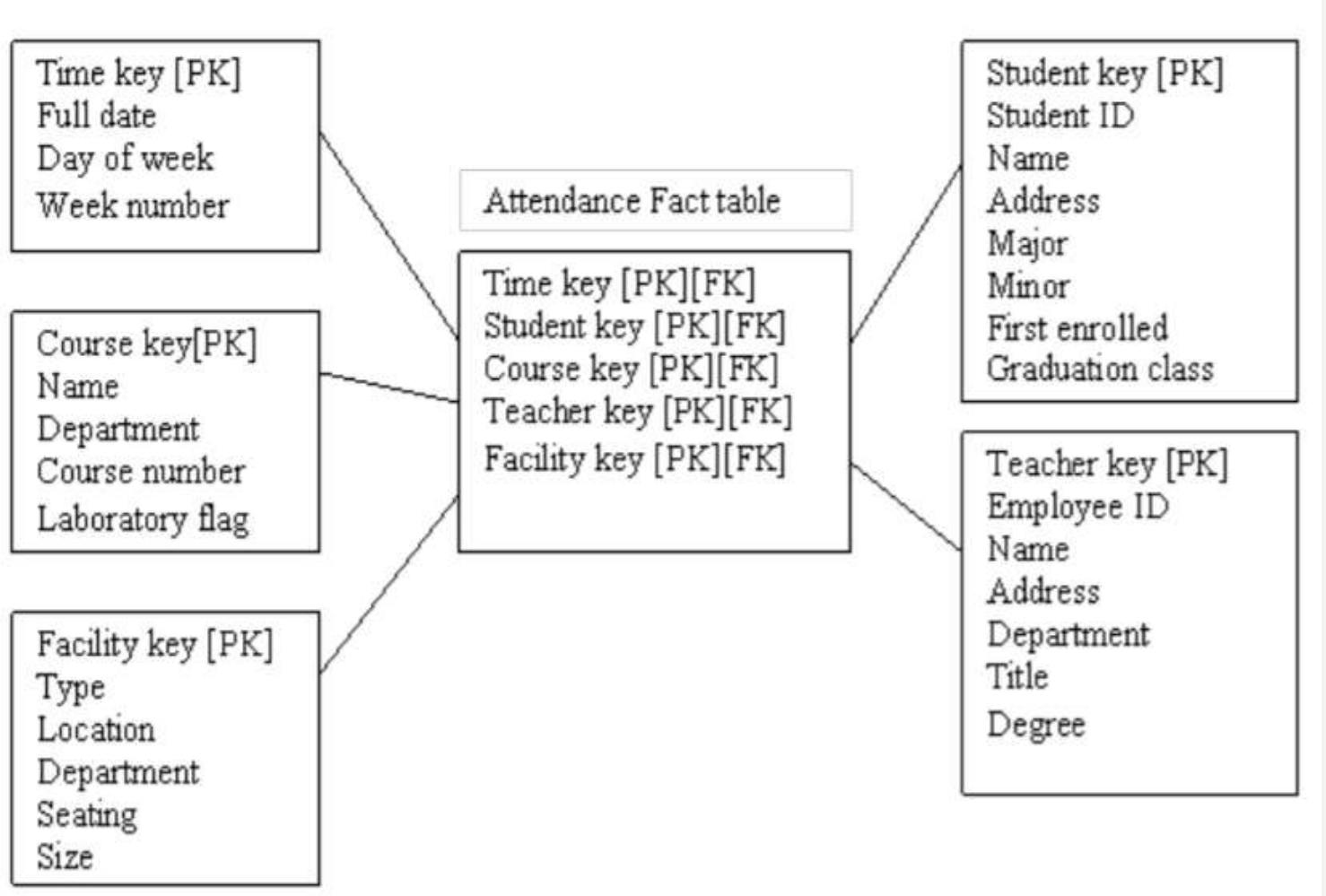
→ it is a table that does not contain fact.

→ contain only dimensional keys

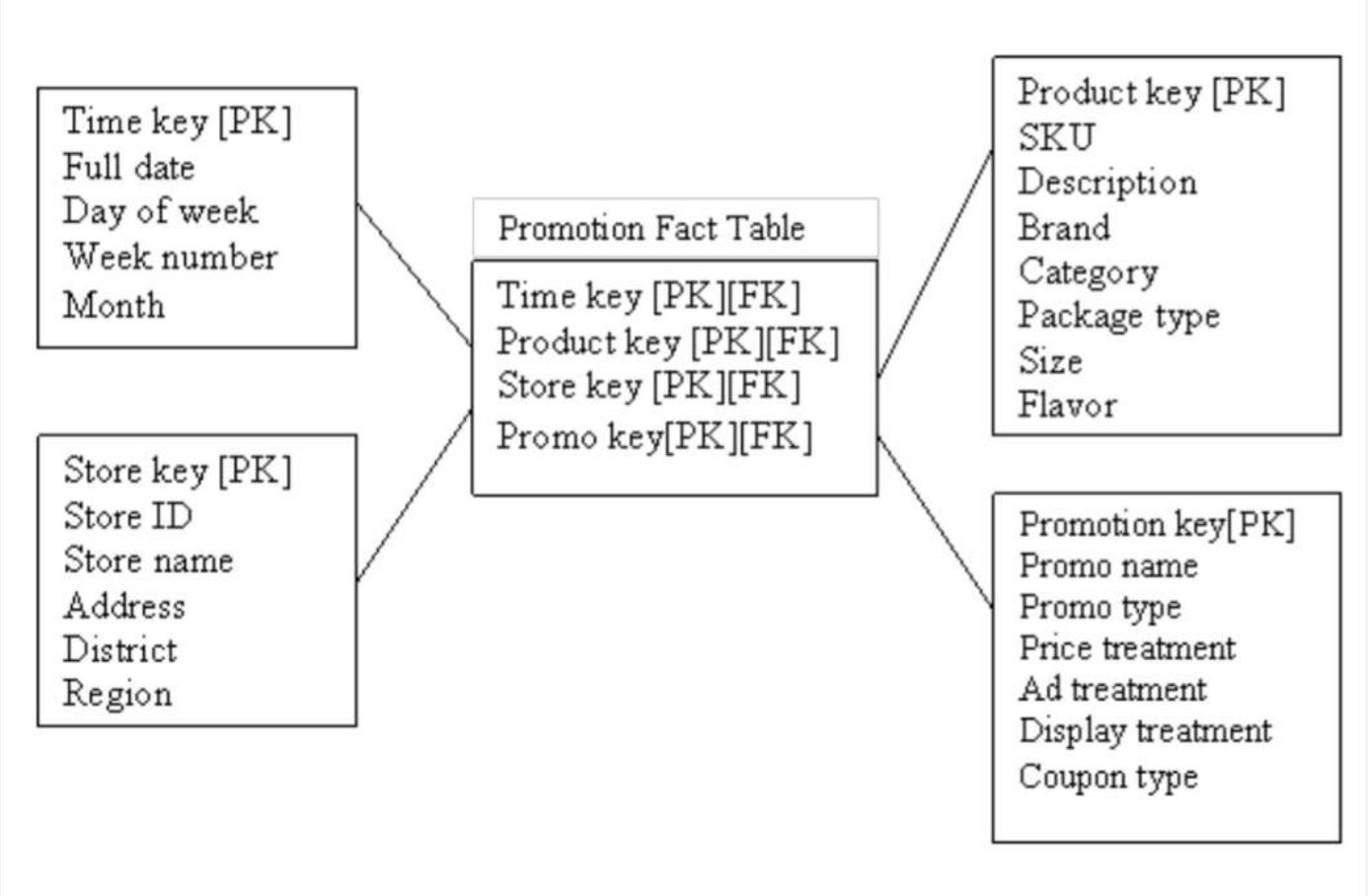
it captures events that happen only at information level but not included in the calculations level

→ captures m:m between dimensions, but contains no numeric or textual facts.

FACTLESS FACT TABLE SHOWING OCCURRENCE OF AN EVENT



FACTLESS FACT TABLE SHOWING COVERAGE

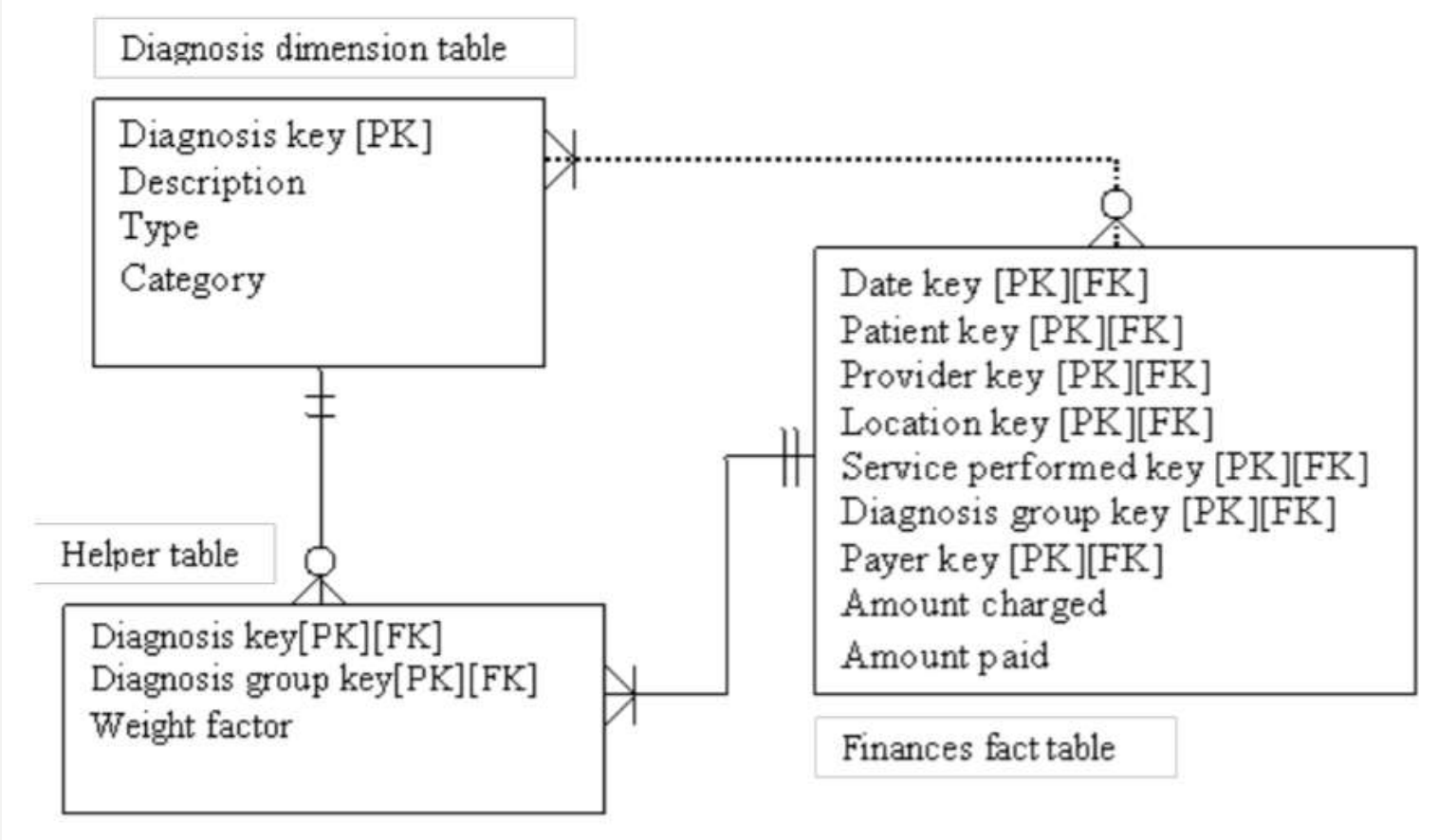


NORMALIZING DIMENSION TABLES

↳ snowflake schema

- Dimension tables may not be normalized. Most data warehouse experts find this acceptable.
- In some situations in which it makes sense to further normalize dimension tables.
- Multivalued dimensions:
  - Ex: Hospital charge/payment for a patient on a date is associated with one or more diagnosis.
  - N:M relationship between the Diagnosis and Finances fact table.
  - Solution: create an associative entity (helper table) between Diagnosis and Finances.

MULTIVALUED DIMENSION



SNOWFLAKE SCHEMA

- Snowflake schema is an expanded version of a star schema in which dimension tables are normalized into several related tables.
- **Advantages**
  - Small saving in storage space
  - Normalized structures are easier to update and maintain
- **Disadvantages**
  - Schema less intuitive →sema data az segjsel
  - Ability to browse through the content difficult
  - Degraded query performance because of additional joins.

– Advantage –

→ small saving in storage

→ normalized structure easy to update

– Disadvantages –

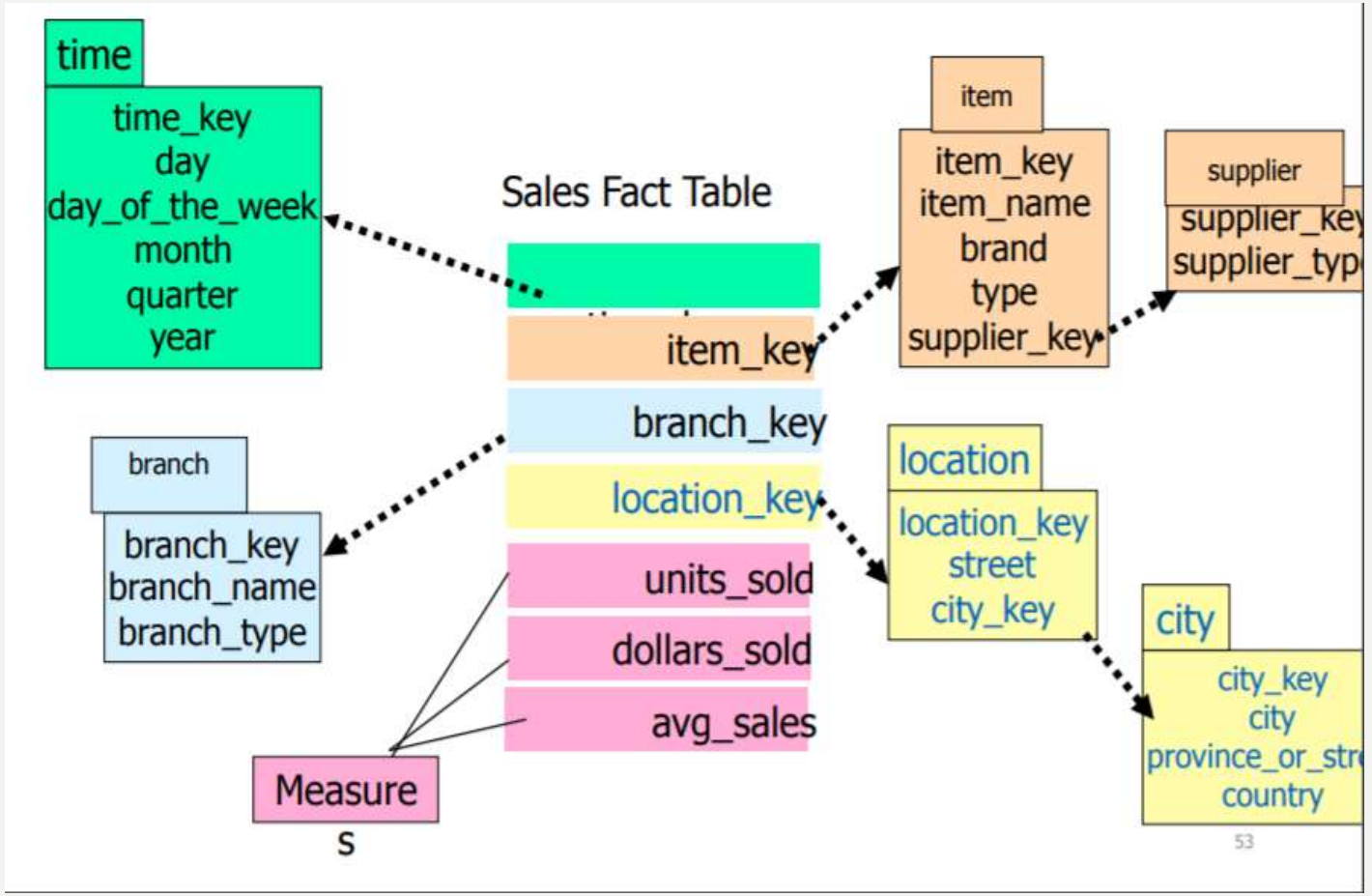
→ less intuitive

→ browse difficult

→ degraded query performance

↳ because of joins

EXAMPLE OF SNOWFLAKE SCHEMA



THE USER INTERFACE

- A variety of tools are available to query and analyze data stored in data warehouses.

1. Querying tools
2. On-line Analytical processing (OLAP, MOLAP, ROLAP) tools
3. Data Mining tools
4. Data Visualization tools

ROLE OF METADATA (DATA CATALOG)

- Identify subjects of the data mart
- Identify dimensions and facts
- Indicate how data is derived from enterprise data warehouses, including derivation rules
- Indicate how data is derived from operational data store, including derivation rules
- Identify available reports and predefined queries
- Identify data analysis techniques (e.g. drill-down)
- Identify responsible people

QUERYING TOOLS

- SQL is not an analytical language
- SQL-99 includes some data warehousing extensions
- SQL-99 still is not a full-featured data warehouse querying and analysis tool.
- Different DBMS vendors will implement some or all of the SQL-99 OLAP extension commands and possibly others.

ON-LINE ANALYTICAL PROCESSING (OLAP)

making analysis faster

- OLAP is the use of a set of graphical tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques

With OLAP data can pre-calculated and pre-aggragated.

- **Relational OLAP (ROLAP)**
  - OLAP tools that view the database as a traditional relational database in either a star schema or other normalized or denormalized set of tables.
- **Multidimensional OLAP (MOLAP)**
  - OLAP tools that load data into an intermediate structure, usually a three or higher dimensional array. (Cube structure)

OLAP database are devided into one or more cubes. Cubes are design in such a way that creating and viewing reports become easy.

Analysist frequently need to group, aggregate and join data. These operations in relational DB are resource intensive.

- ROLAP -

→ works with data that exist in a relational database.
→ Fact and dimension tables are stored as relational tables.
→ Allows multidimensional analysis of data it is the fastest growing OLAP.

- MOLAP -

→ uses array-based multidimensional storage engines to display multidimensional views of data.

=> They use OLAP Cube

Advantages of ROLAP

→ high data efficiency
→ scability

Disadvantages of ROLAP

→ demand for higher source
→ aggregately data limitations
→ slow query performance

MOLAP	ROLAP
Information retrieval is fast.	İnformaton retrieval is comparatively slow.
Uses sparse array to store data-sets. (an array of data in which many elements have a value of ZERO.)	Uses relational table.
MOLAP is best suited for inexperienced users, since it is very easy to use.	ROLAP is best suited for experienced users.
Maintains a separate database for data cubes.	İt may not require space other than available in the Data Warehouse.
DBMS facility is weak.	DBMS facility is strong.

cube: allows data to be modeled and viewed in multiple dimensions

FROM TABLES TO DATA CUBES

- A data warehouse is based on a **multidimensional data model** which views data in the form of a **data cube**
- A **data cube**, such as *sales*, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as *item* (*item_name, brand, type*), *or time (day, week, month, quarter, year)*
  - Fact table contains measures (such as *dollars_sold*) and keys to each of the related dimension tables

MOLAP OPERATIONS

- Roll up (drill-up): summarize data
  - by climbing up hierarchy or by dimension reduction
- Drill down (roll down): reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- Slice and dice:
  - project and select

-MOLAP-

- Drill-down
- Slice and dice
- Roll up

-Advantages-

→excelente performance since pre-aggregated
→optimal for slice and dice operations

-Disadvantages-

→difficult to change dimension without re-aggregation
→main disadvantage is that it is not scalable. It can handle only limited amount of data at the same time.

Drill-down: fragmented into smaller parts. Opposite of roll-up process, it can be done via
→moving down hierarchy
→increasing a dimension

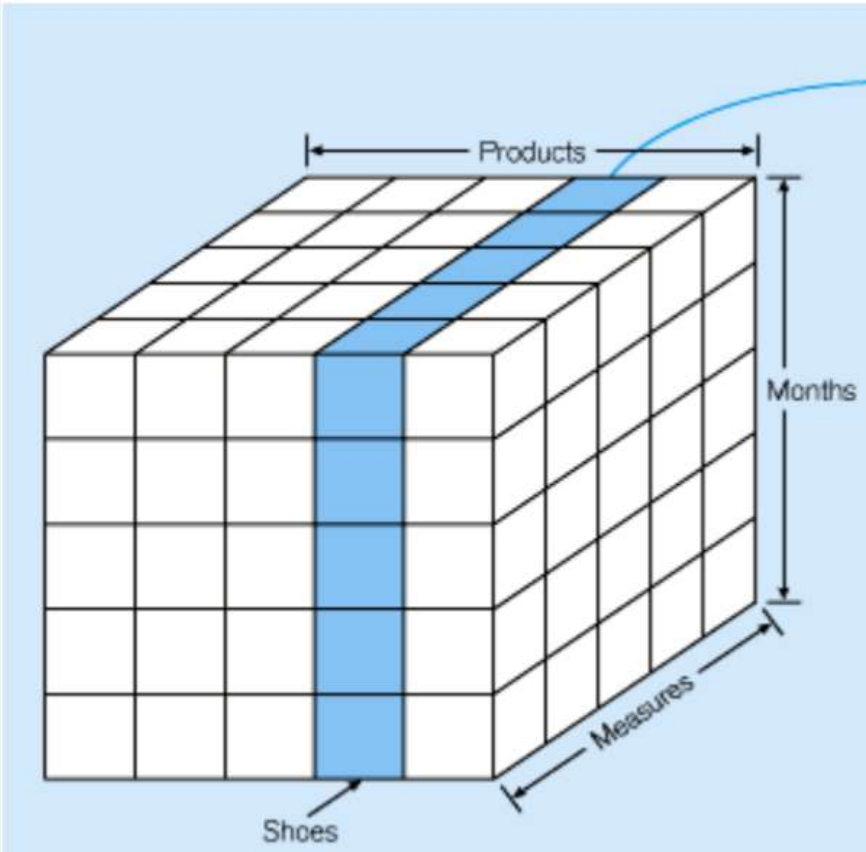
Roll-up: can be performed as 2 ways
① reducing dimensions
② climbing up concept hierarchy
↳concept hierarchy is a system of grouping things based on their order or level.

ex: hierarchy moves up
city → country
in roll-up process at least one or more dimension need to be removed.

Slice: one dimension is selected and a new sub-cube is created

Dice: select 2 or more dimensions

SLICING A DATA CUBE



	Measure		
	Units	Revenue	Cost
January	250	1564	1020
February	200	1275	875
March	350	1800	1275
April	400	1935	1500
May	485	2000	1560

Product: Shoes

EXAMPLE OF DRILL-DOWN

Summary report

Brand	Package size	Sales
SofTowel	2-pack	$75
SofTowel	3-pack	$100
SofTowel	6-pack	$50

Drill-down with color added

Brand	Package size	Color	Sales
SofTowel	2-pack	White	$30
SofTowel	2-pack	Yellow	$25
SofTowel	2-pack	Pink	$20
SofTowel	3-pack	White	$50
SofTowel	3-pack	Green	$25
SofTowel	3-pack	Yellow	$25
SofTowel	6-pack	White	$30
SofTowel	6-pack	Yellow	$20

to explain some observed event or condition

DATA MINING

- Data mining is knowledge discovery using a blend of statistical, AI, and computer graphics techniques
- Goals:
  - Explain observed events or conditions
  - Confirm hypotheses
  - Explore data for new or unexpected relationships
- Techniques
  - Case-based reasoning
  - Rule discovery
  - Signal processing
  - Neural nets
  - Fractals

DATA VISUALIZATION

- •Data visualization is the representation of data in graphical/multimedia formats for human analysis