Data Mining: Data Warehouse

Presented by

Dr. Siddique Ibrahim S P

SCOPE

VIT-AP University


py
PMG


Whirlpool


HUDSONS
HUDSONS

August 21, 2024


gettyimages
NOAH SEELAM

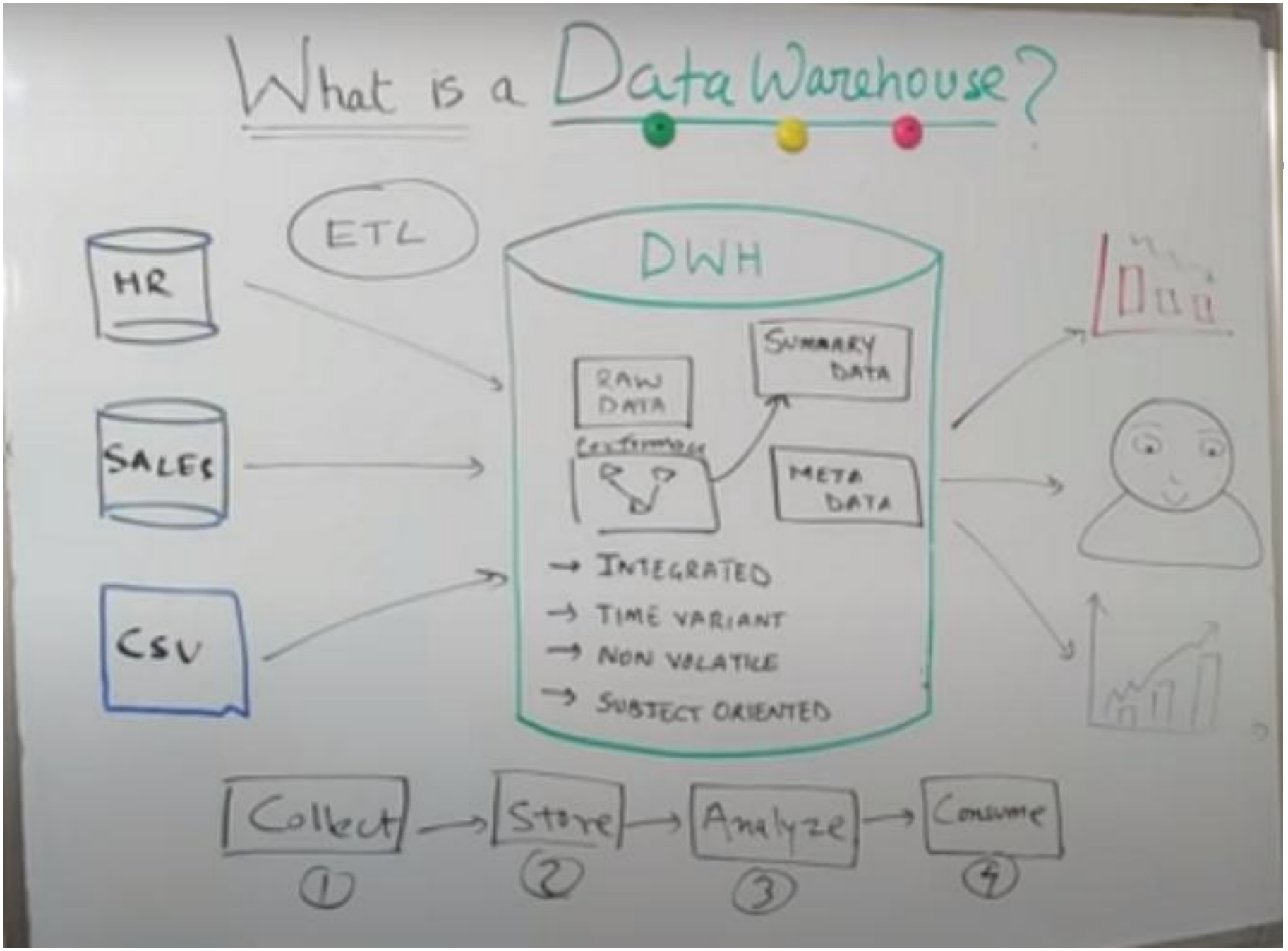
August 21, 2024

- Data warehouse is centralize area to store organization/enterprise data where you can do further analysis.

- Databases, transaction, flat files

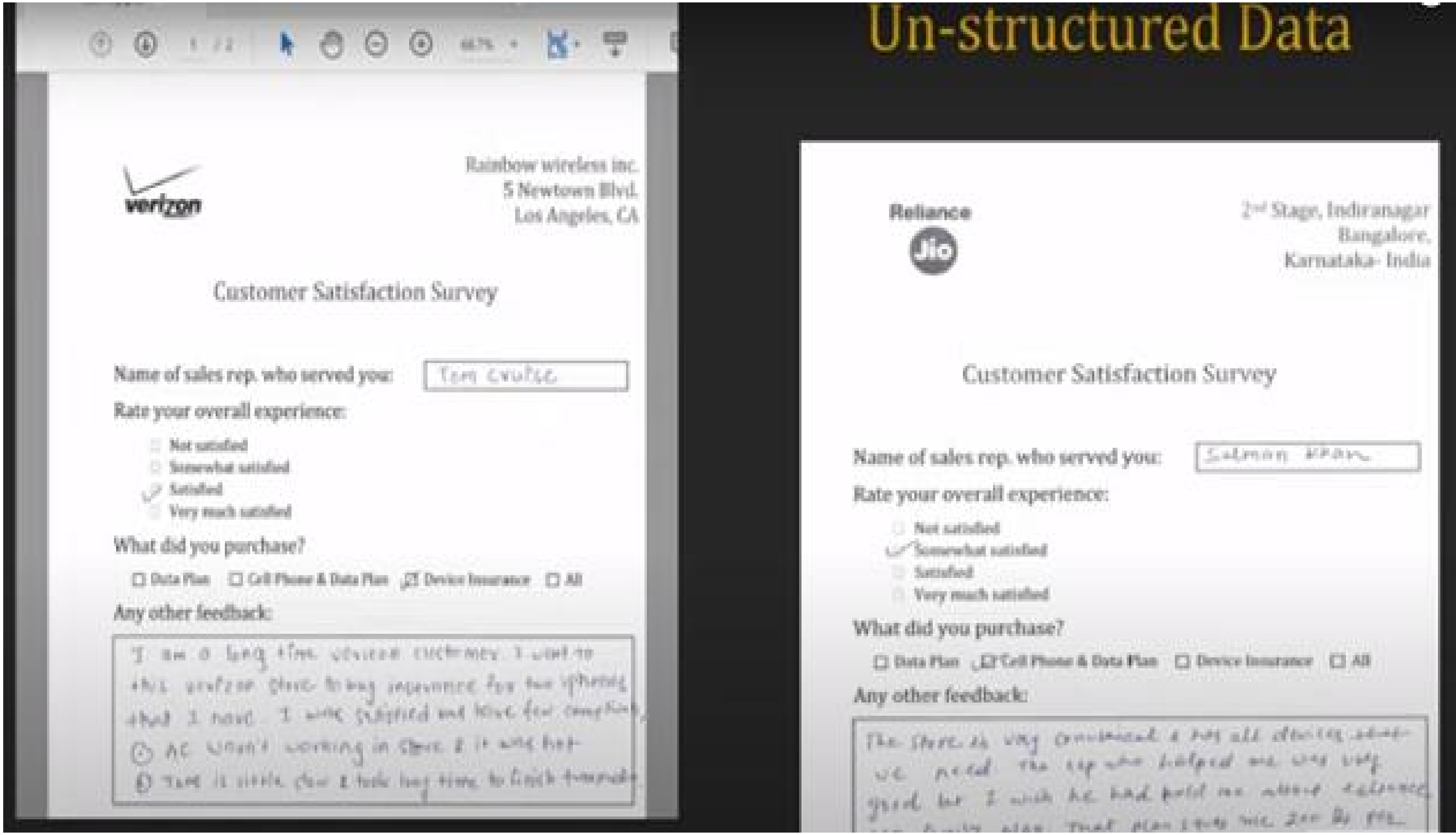
- How many employee will exceed the sales target within a year of joining?

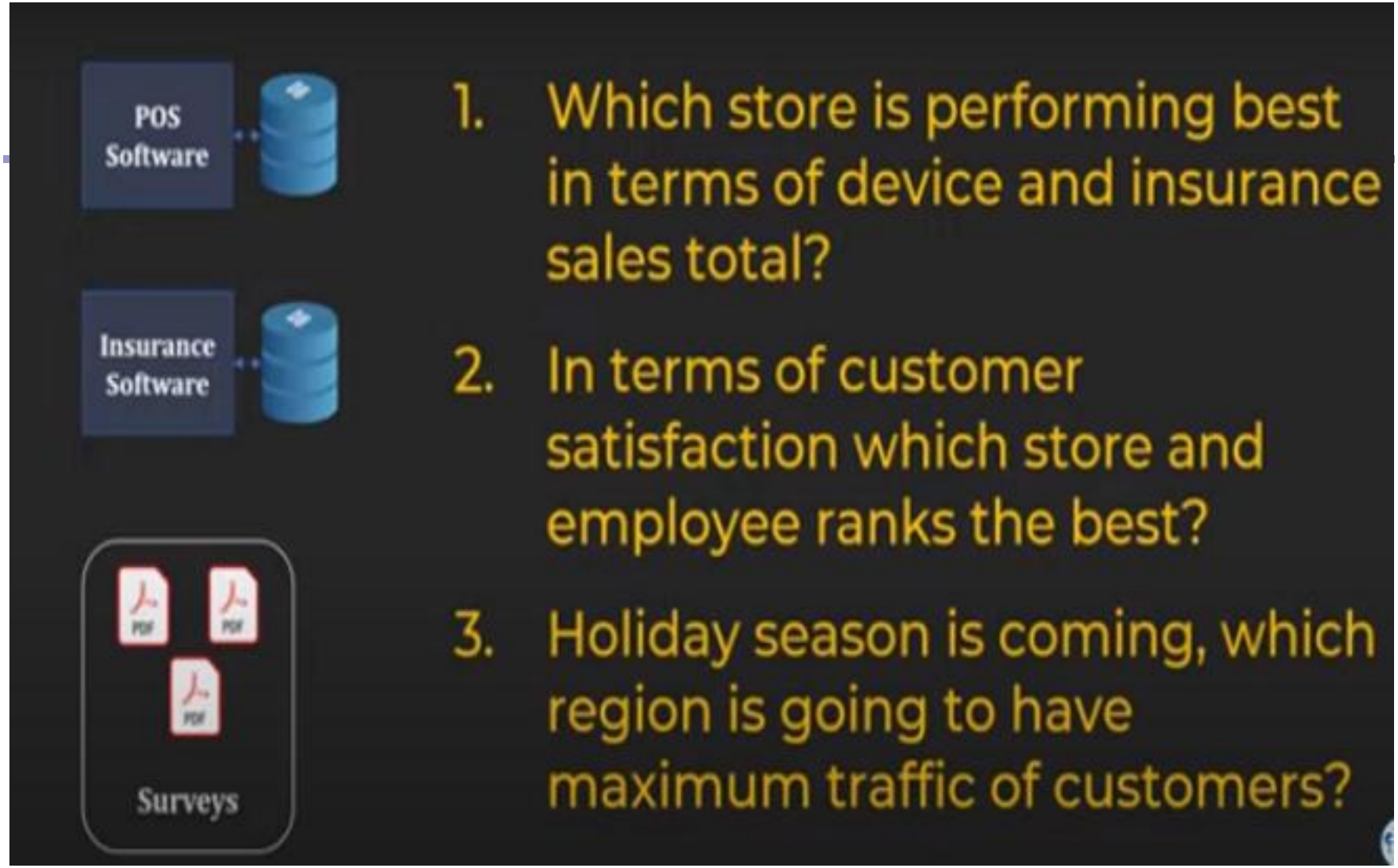
- How many employee’s are eligible for on site work?



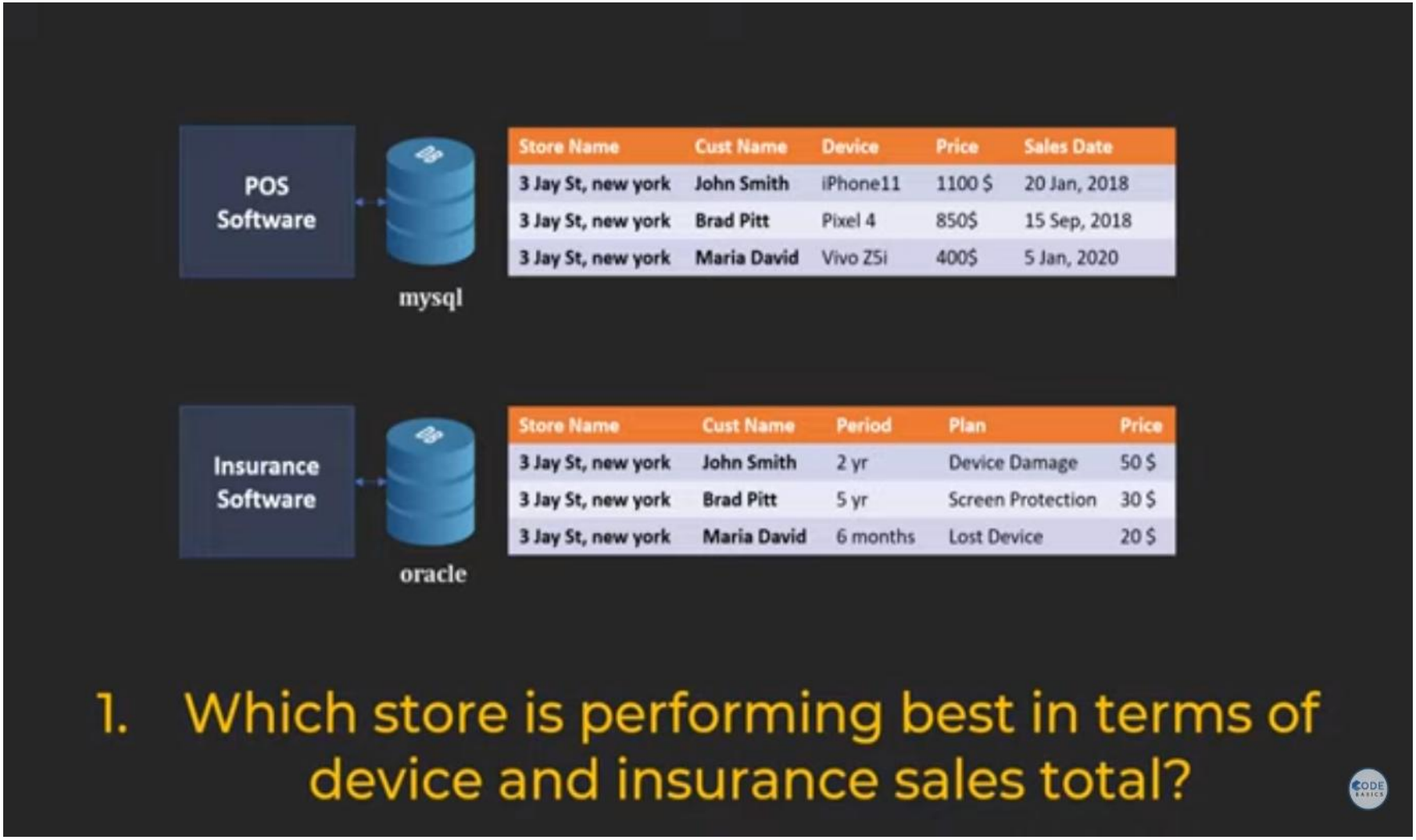
August 21, 2024

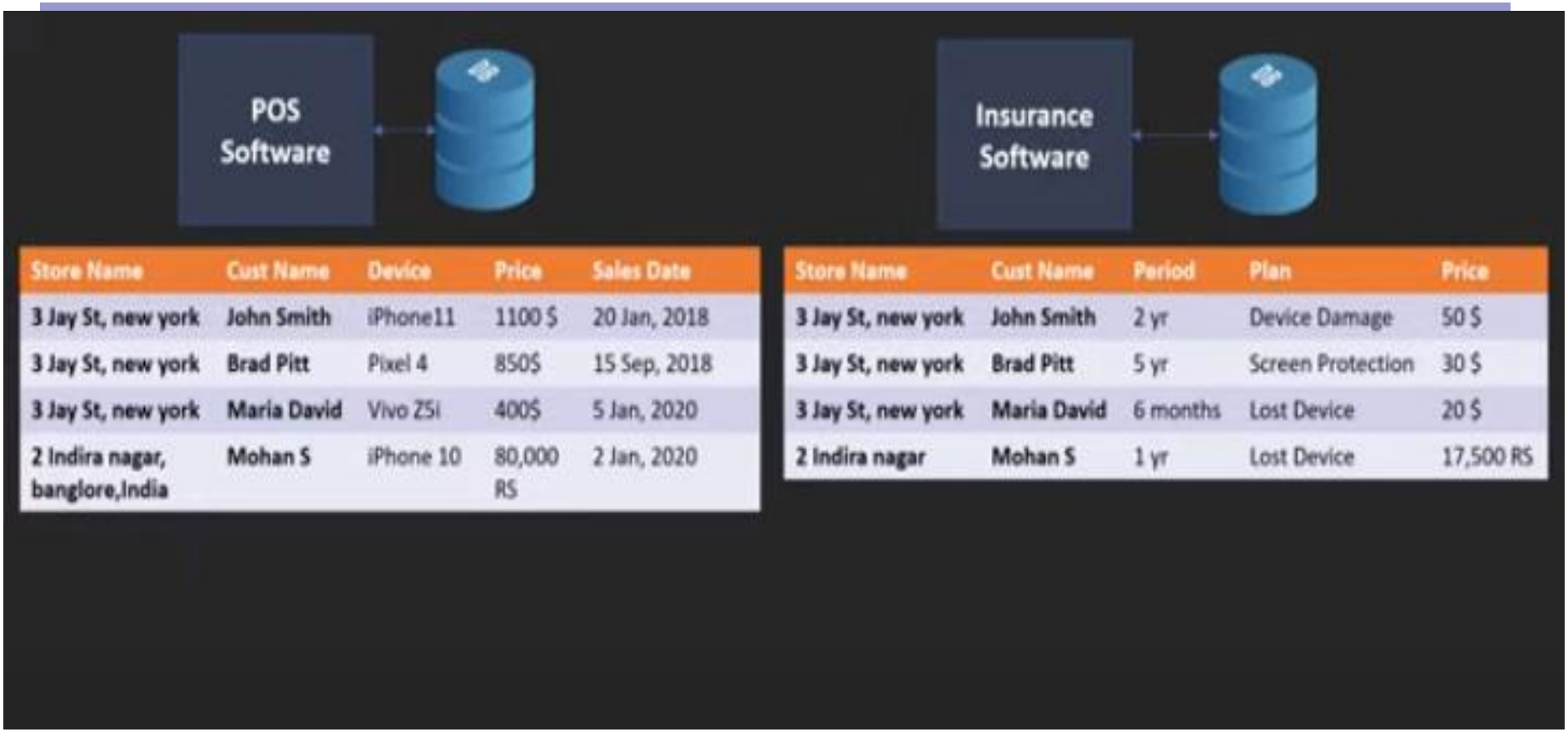

POS Software
mysql
Store Name | Cust Name | Device | Price | Sales Date
3 Jay St, new york | John Smith | iPhone11 | 1100 $ | 20 Jan, 2018
3 Jay St, new york | Brad Pitt | Pixel 4 | 850$ | 15 Sep, 2018
3 Jay St, new york | Maria David | Vivo Z5i | 400$ | 5 Jan, 2020
Insurance Software
oracle
Store Name | Cust Name | Period | Plan | Price
3 Jay St, new york | John Smith | 2 yr | Device Damage | 50 $
3 Jay St, new york | Brad Pitt | 5 yr | Screen Protection | 30 $
3 Jay St, new york | Maria David | 6 months | Lost Device | 20 $
Structured Data


Un-structured Data
Rainbow wireless inc.
5 Newtown Blvd.
Los Angeles, CA
Customer Satisfaction Survey
Name of sales rep. who served you:
Tom Cruise
Rate your overall experience:
Not satisfied
Somewhat satisfied
Satisfied
Very much satisfied
What did you purchase?
Data Plan
Cell Phone & Data Plan
Device Insurance
All
Any other feedback:
Reliance
Jio
2nd Stage, Indiranagar
Bangalore,
Karnataka- India
Customer Satisfaction Survey
Name of sales rep. who served you:
Salman Khan
Rate your overall experience:
Not satisfied
Somewhat satisfied
Satisfied
Very much satisfied
What did you purchase?
Data Plan
Cell Phone & Data Plan
Device Insurance
All
Any other feedback:


POS Software
Insurance Software
Surveys
1. Which store is performing best in terms of device and insurance sales total?
2. In terms of customer satisfaction which store and employee ranks the best?
3. Holiday season is coming, which region is going to have maximum traffic of customers?

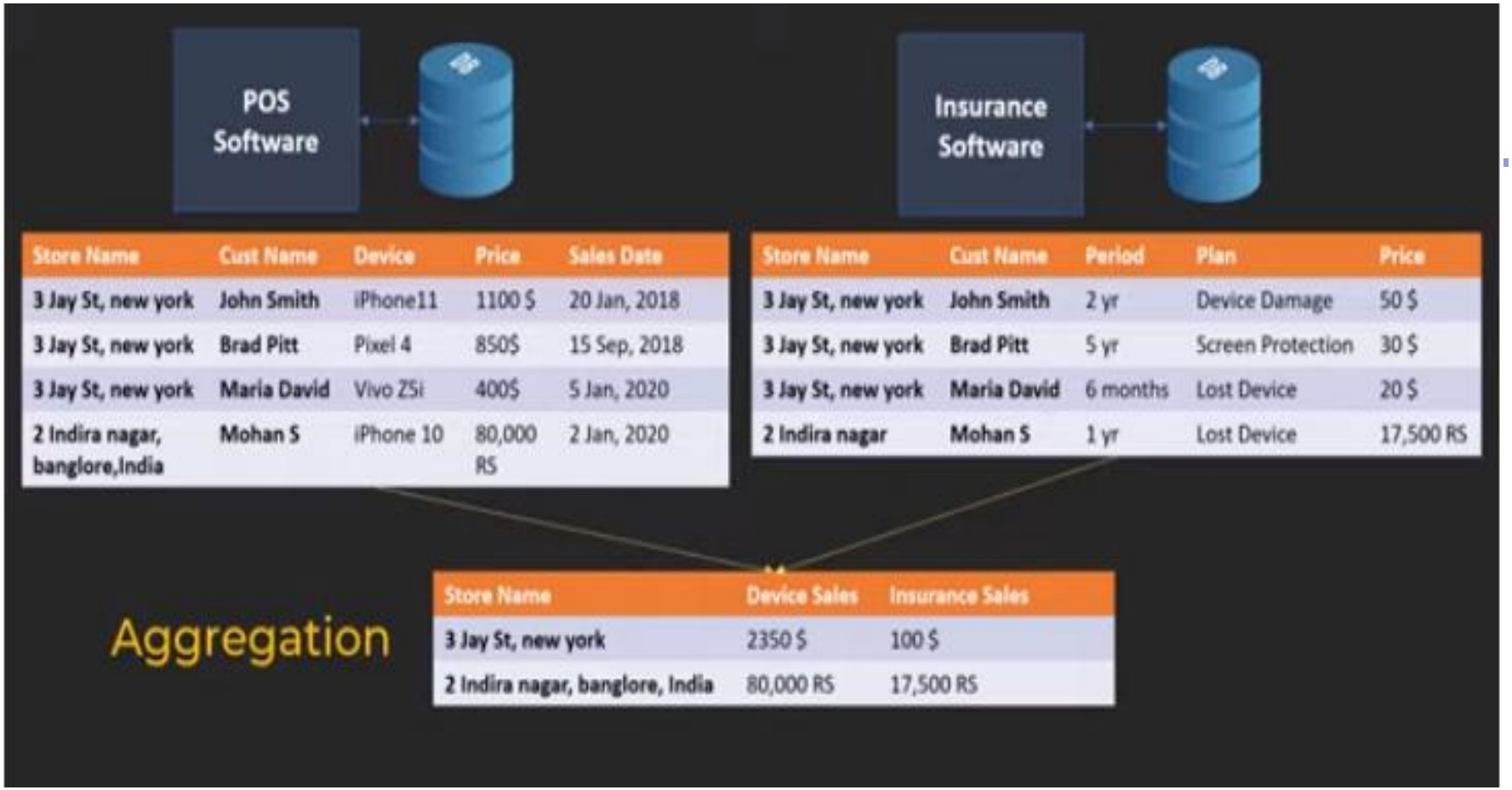
To see insight



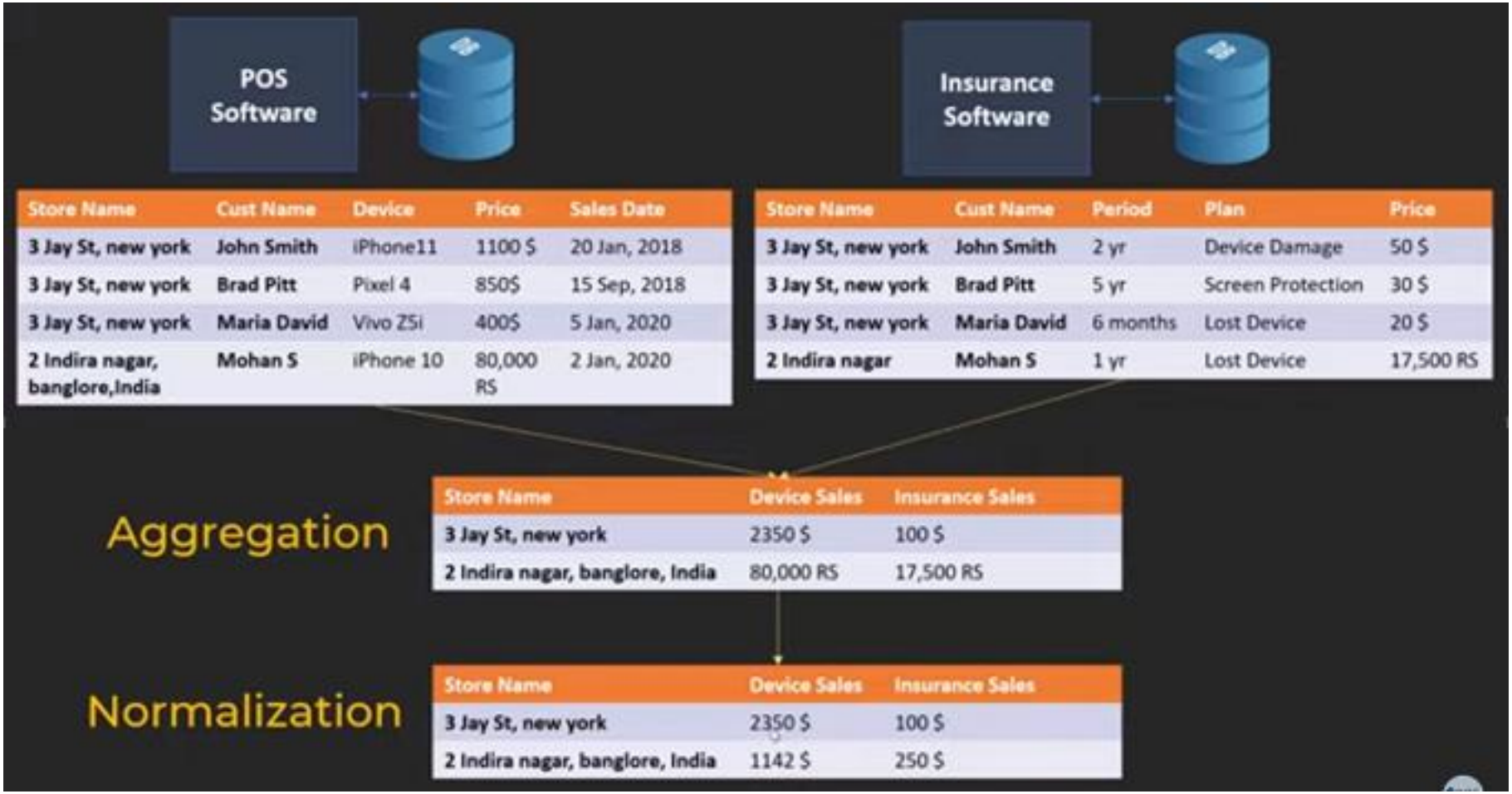


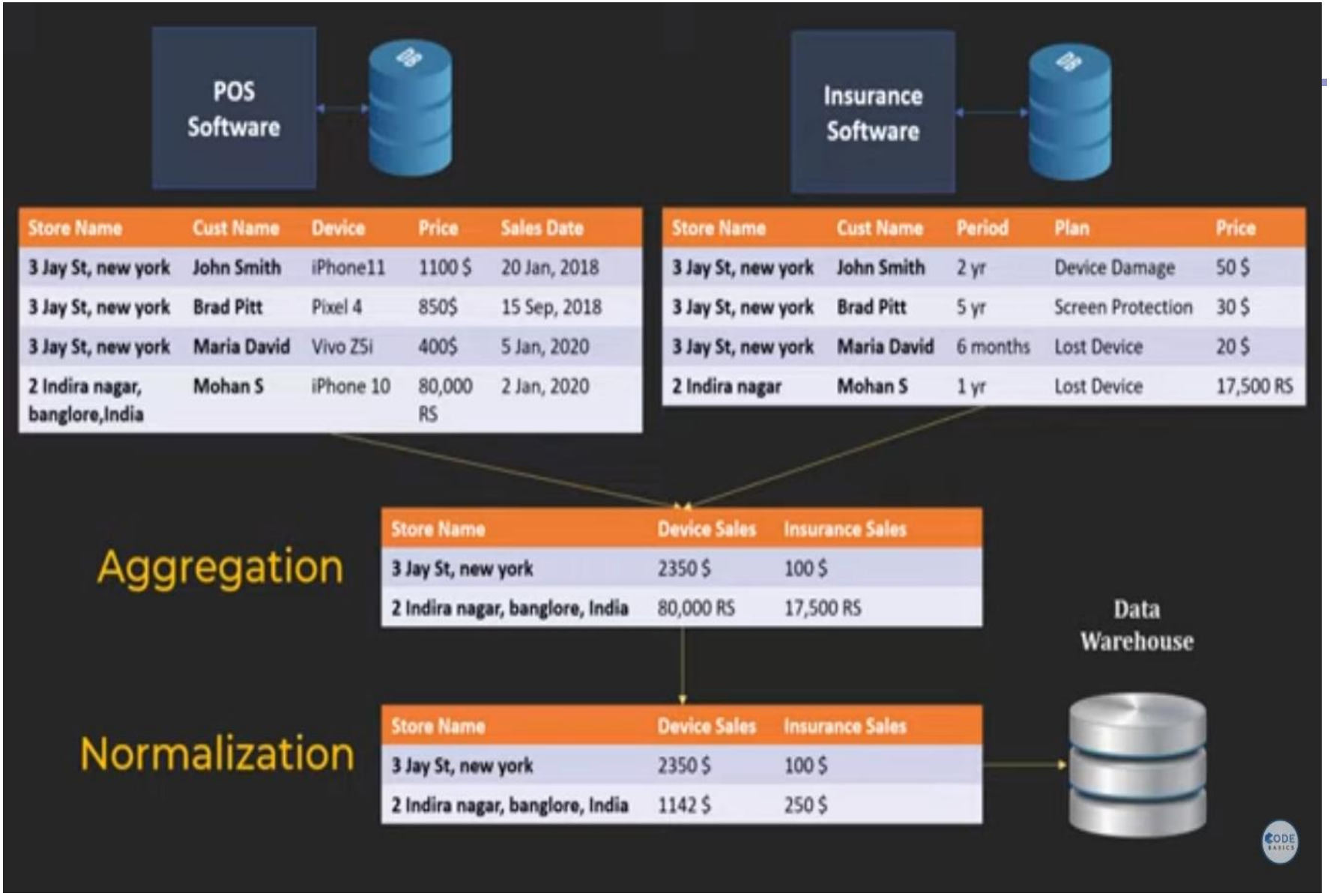
Store Name	Cust Name	Device	Price	Sales Date
3 Jay St, new york	John Smith	iPhone11	1100 $	20 Jan, 2018
3 Jay St, new york	Brad Pitt	Pixel 4	850$	15 Sep, 2018
3 Jay St, new york	Maria David	Vivo Z5i	400$	5 Jan, 2020
2 Indira nagar, banglore,India	Mohan S	iPhone 10	80,000 RS	2 Jan, 2020

Store Name	Cust Name	Period	Plan	Price
3 Jay St, new york	John Smith	2 yr	Device Damage	50 $
3 Jay St, new york	Brad Pitt	5 yr	Screen Protection	30 $
3 Jay St, new york	Maria David	6 months	Lost Device	20 $
2 Indira nagar	Mohan S	1 yr	Lost Device	17,500 RS

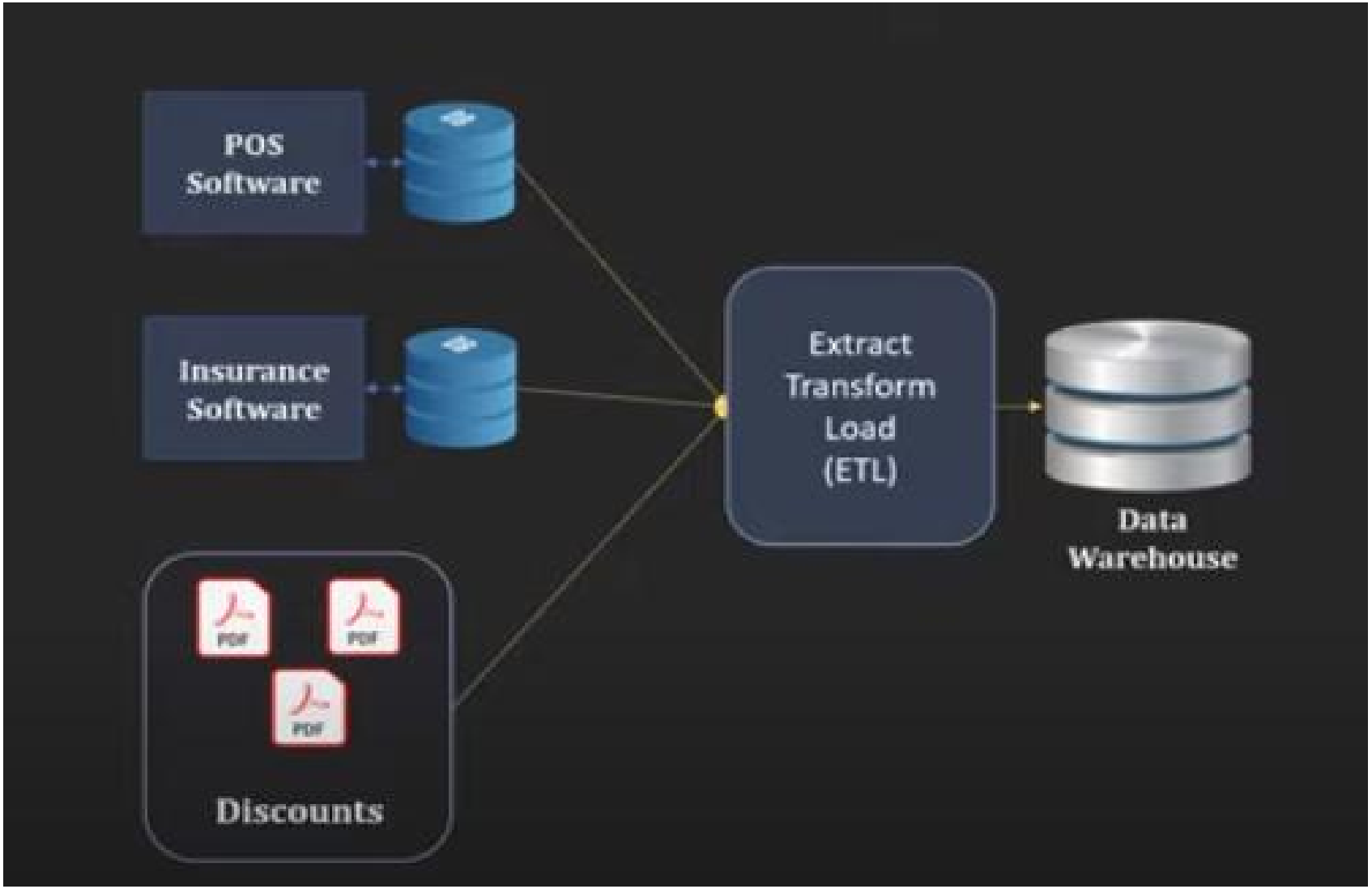


1. Which store is performing best in terms of device and insurance sales total?

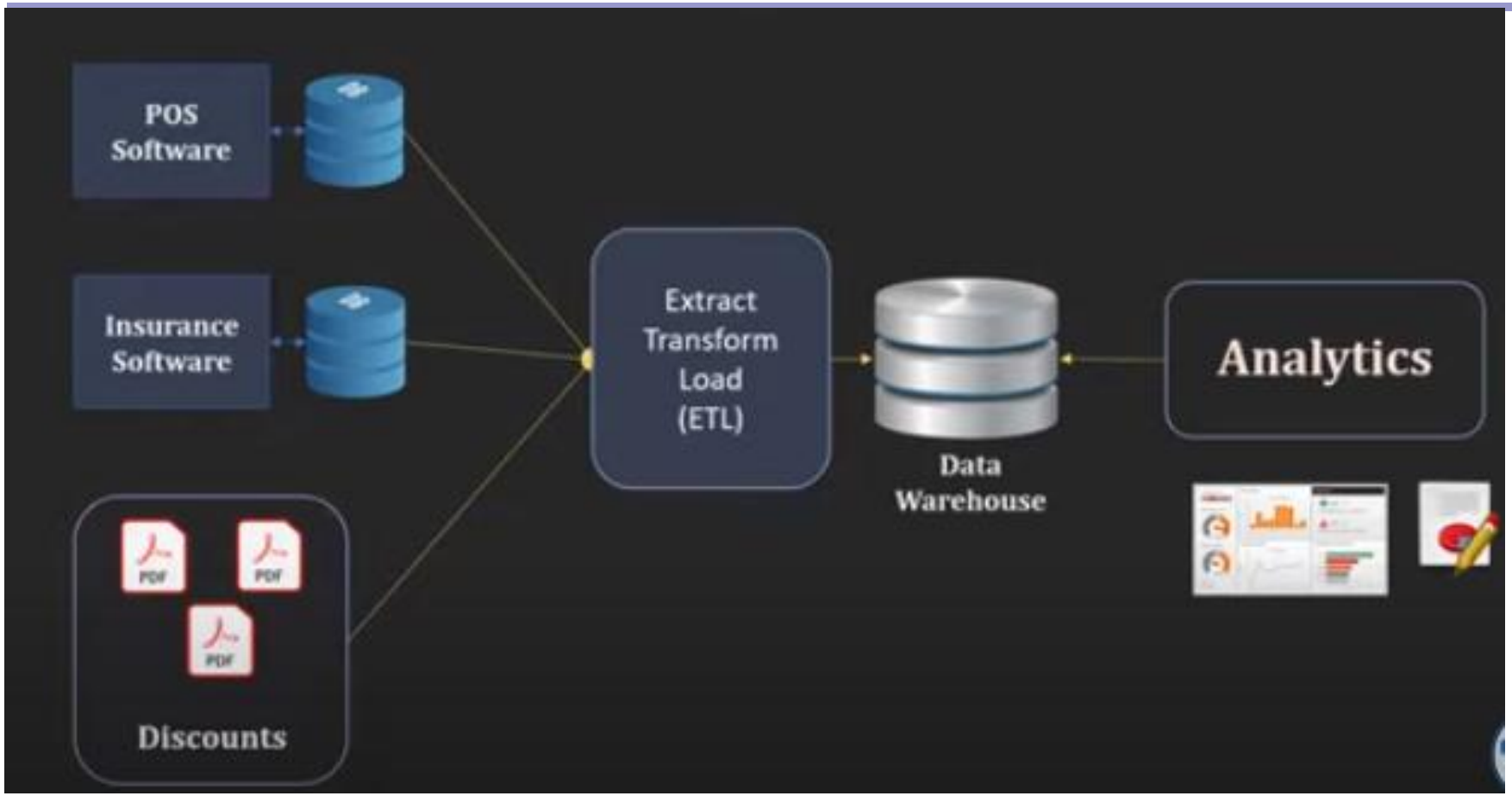

POS Software
Insurance Software
Store Name | Cust Name | Device | Price | Sales Date
3 Jay St, new york | John Smith | iPhone11 | 1100 $ | 20 Jan, 2018
3 Jay St, new york | Brad Pitt | Pixel 4 | 850$ | 15 Sep, 2018
3 Jay St, new york | Maria David | Vivo Z5i | 400$ | 5 Jan, 2020
2 Indira nagar, banglore,India | Mohan S | iPhone 10 | 80,000 RS | 2 Jan, 2020
Store Name | Cust Name | Period | Plan | Price
3 Jay St, new york | John Smith | 2 yr | Device Damage | 50 $
3 Jay St, new york | Brad Pitt | 5 yr | Screen Protection | 30 $
3 Jay St, new york | Maria David | 6 months | Lost Device | 20 $
2 Indira nagar | Mohan S | 1 yr | Lost Device | 17,500 RS
Aggregation
Store Name | Device Sales | Insurance Sales
3 Jay St, new york | 2350 $ | 100 $
2 Indira nagar, banglore, India | 80,000 RS | 17,500 RS
Normalization
Store Name | Device Sales | Insurance Sales
3 Jay St, new york | 2350 $ | 100 $
2 Indira nagar, banglore, India | 1142 $ | 250 $


POS Software
Insurance Software
Store Name | Cust Name | Device | Price | Sales Date
3 Jay St, new york | John Smith | iPhone11 | 1100 $ | 20 Jan, 2018
3 Jay St, new york | Brad Pitt | Pixel 4 | 850$ | 15 Sep, 2018
3 Jay St, new york | Maria David | Vivo Z5i | 400$ | 5 Jan, 2020
2 Indira nagar, banglore,India | Mohan S | iPhone 10 | 80,000 RS | 2 Jan, 2020
Store Name | Cust Name | Period | Plan | Price
3 Jay St, new york | John Smith | 2 yr | Device Damage | 50 $
3 Jay St, new york | Brad Pitt | 5 yr | Screen Protection | 30 $
3 Jay St, new york | Maria David | 6 months | Lost Device | 20 $
2 Indira nagar | Mohan S | 1 yr | Lost Device | 17,500 RS
Aggregation
Store Name | Device Sales | Insurance Sales
3 Jay St, new york | 2350 $ | 100 $
2 Indira nagar, banglore, India | 80,000 RS | 17,500 RS
Normalization
Store Name | Device Sales | Insurance Sales
3 Jay St, new york | 2350 $ | 100 $
2 Indira nagar, banglore, India | 1142 $ | 250 $
Data Warehouse

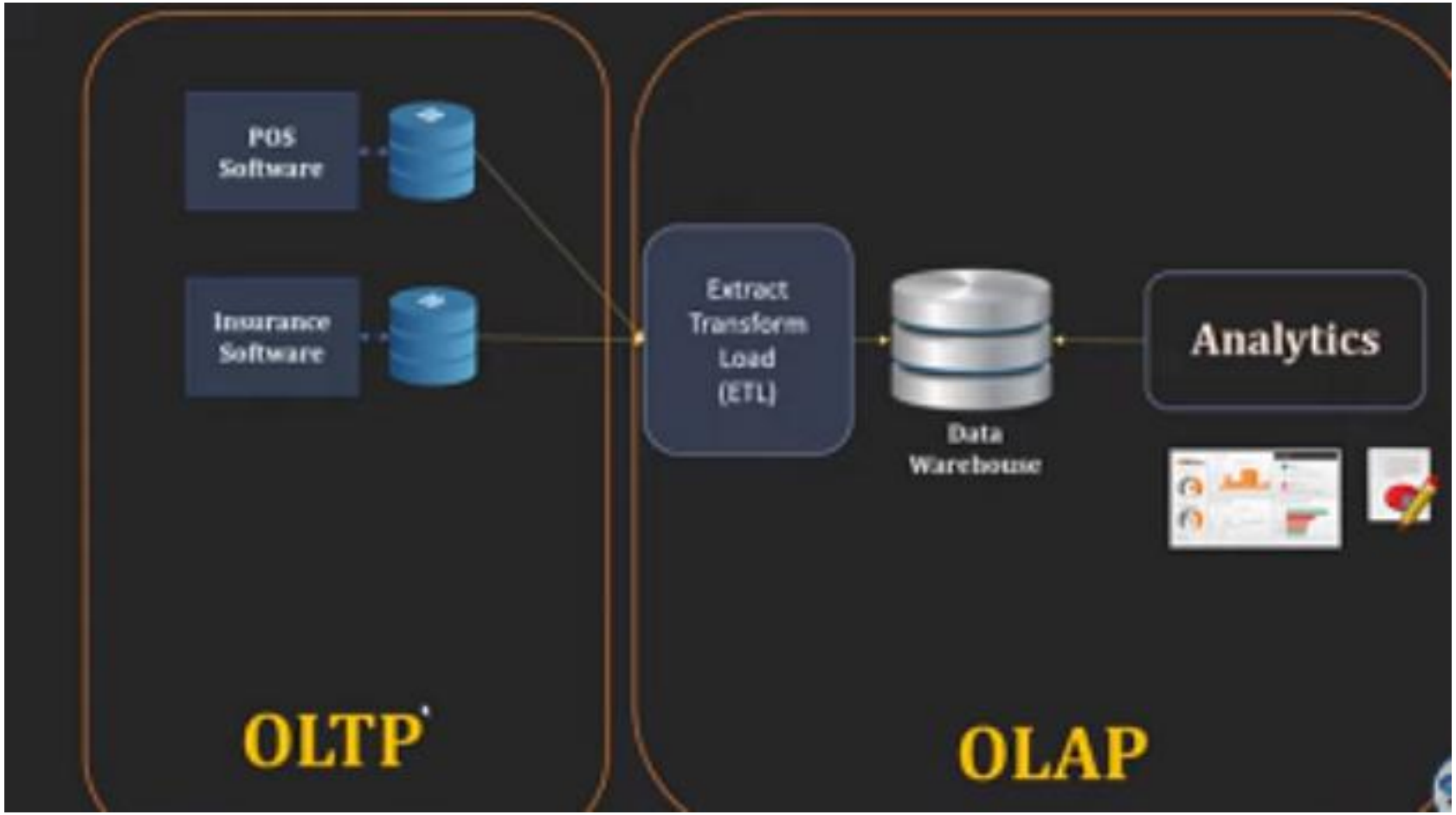
ETL


POS Software
Insurance Software
Discounts
Extract Transform Load (ETL)
Data Warehouse

Data Analysis



OLTP Vs OLAP


POS Software
Insurance Software
Extract Transform Load (ETL)
Data Warehouse
Analytics
OLTP
OLAP

ETL Tools

BEST ETL TOOLS

Enterprise Data Warehouses

What is a Data Warehouse?

- A data warehouse is a powerful tool that allows organizations to store, manage, and analyze large amounts of data.
- It is designed to support the decision-making process by providing a centralized location for all of an organization's data.

- **Characteristics of Data Wareshouse:**
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

Data Warehouse—Subject-Oriented

- Organized around major subjects, 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 around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- 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.

Data Warehouse—Time Variant

- 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)
- Every key structure in the data warehouse
  - Contains an element of **time, explicitly or implicitly**
  - But the key of operational data may or may not contain "time element"
  - This makes it possible to track trends and patterns over time.

Data Warehouse—Nonvolatile

- 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*

Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

What is OLAP?

- Online analytical processing (OLAP) is software technology you can use to analyze business data from different points of view.
- Organizations collect and store data from multiple data sources, such as websites, applications, smart meters, and internal systems.
- OLAP combines and groups this data into categories to provide actionable insights for strategic planning.

For example

- A retailer stores data about all the products it sells, such as color, size, cost, and location.
- The retailer also collects customer purchase data, such as the name of the items ordered and total sales value, in a different system.
- OLAP combines the datasets to answer questions such as
- which color products are more popular or
- how product placement impacts sales.

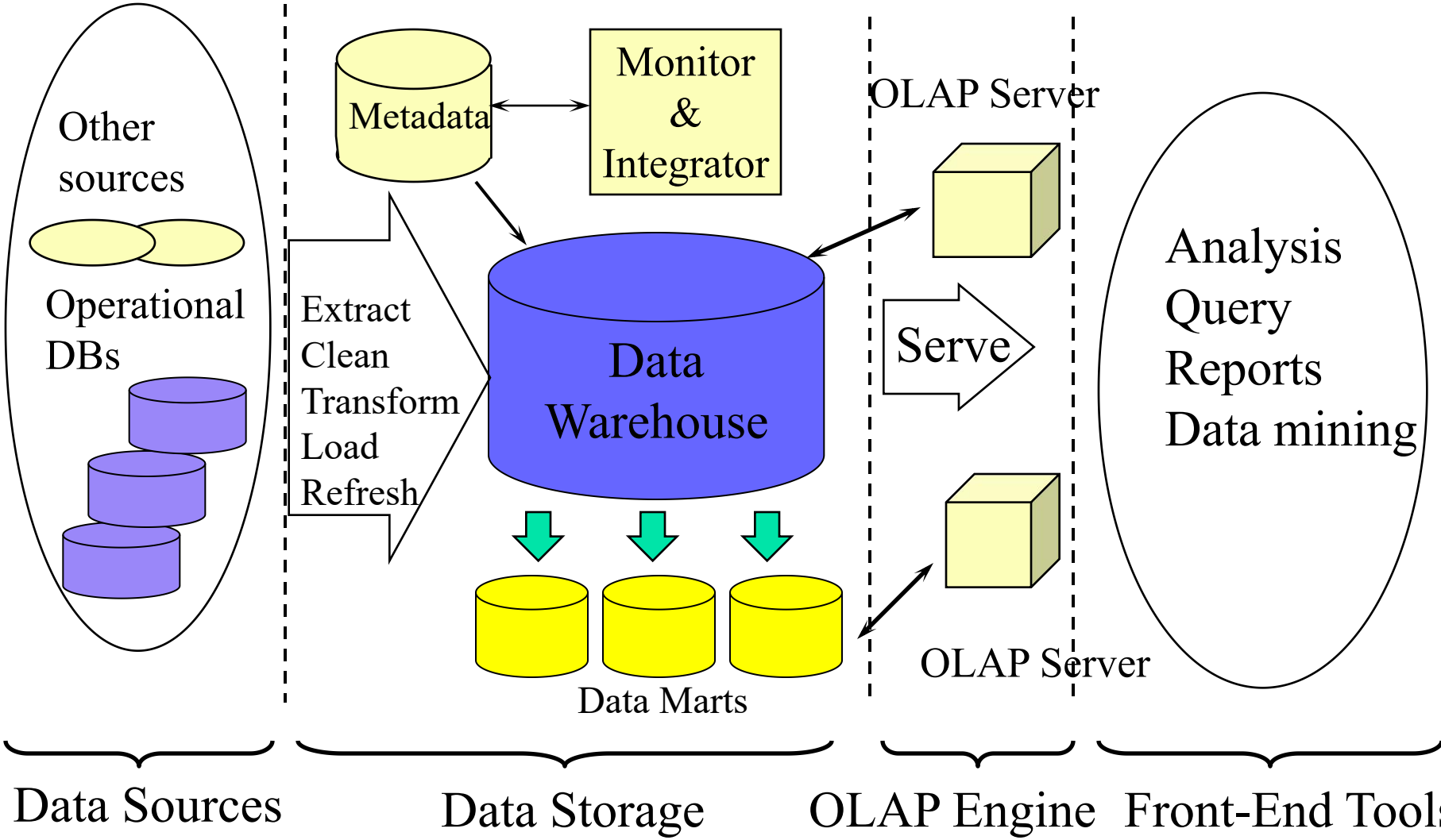
OLTP vs. OLAP

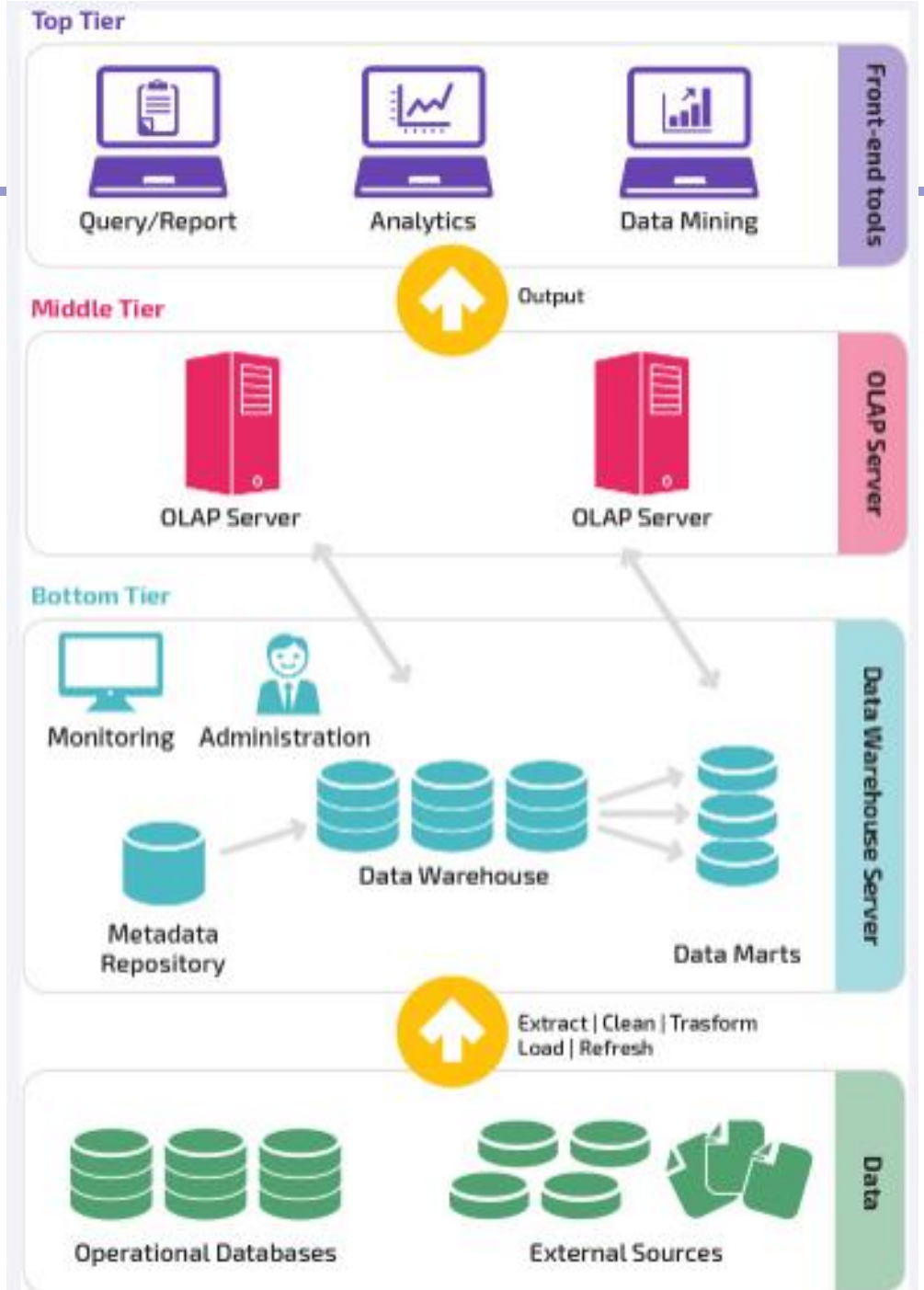
	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why a Separate Data Warehouse?

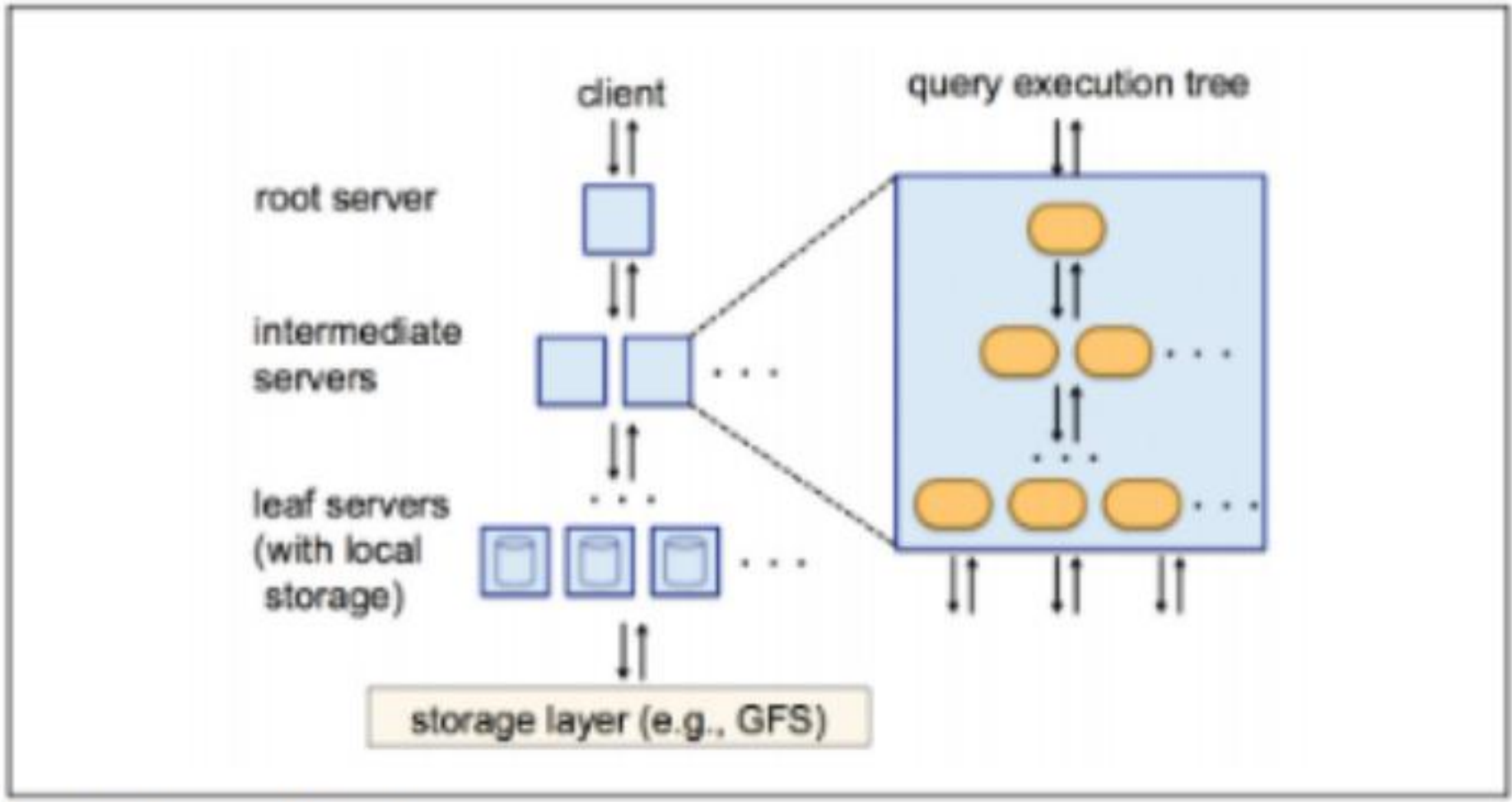
- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

Data Warehouse: A Multi-Tiered Architecture



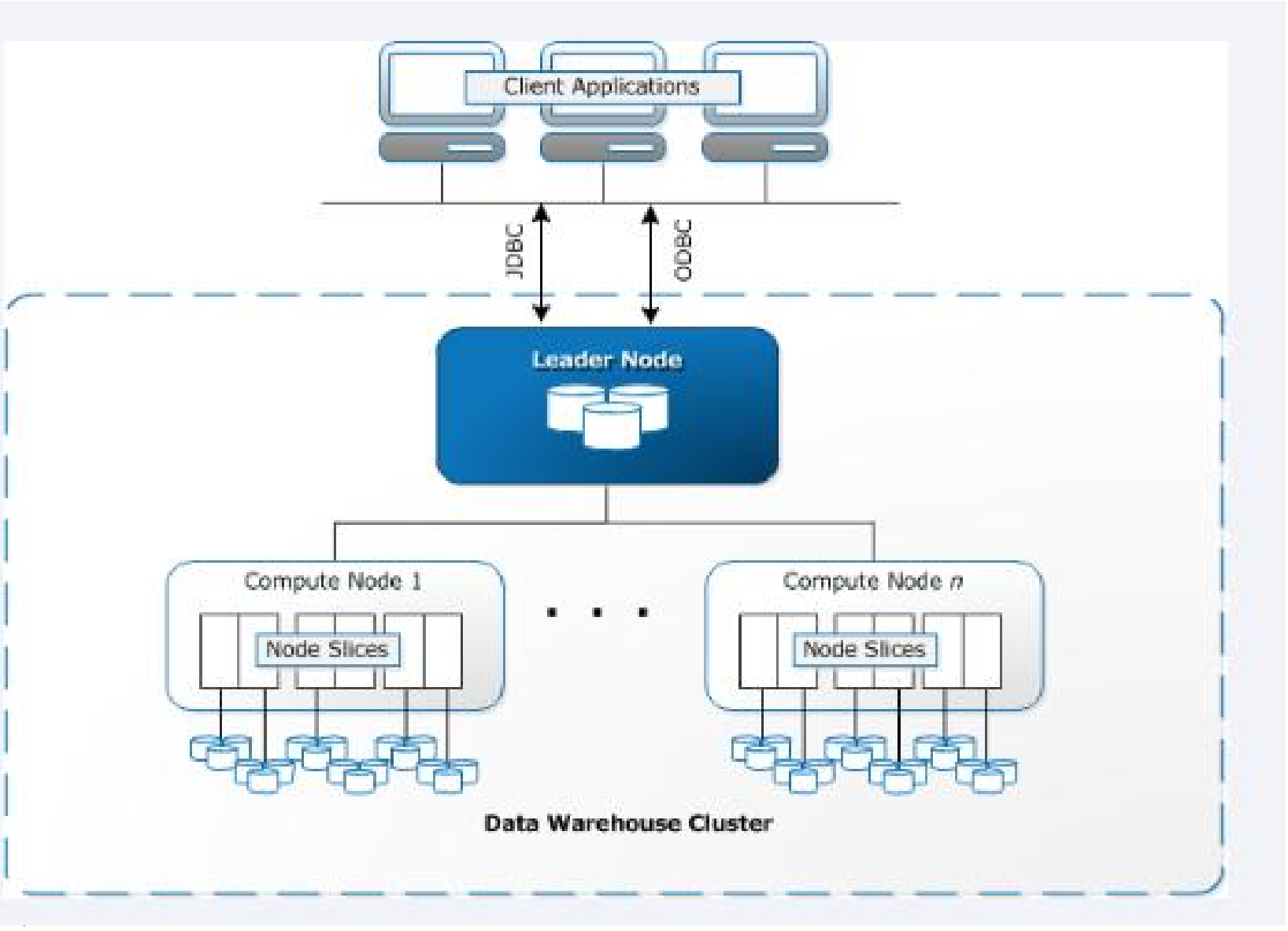

Top Tier
Query/Report
Analytics
Data Mining
Front-end tools
Output
Middle Tier
OLAP Server
OLAP Server
OLAP Server
Bottom Tier
Monitoring
Administration
Data Warehouse
Metadata Repository
Data Marts
Data Warehouse Server
Extract | Clean | Trasform
Load | Refresh
Operational Databases
External Sources
Data

Google's BigQuery



Tree architecture of Dremel

Amazon's Redshift



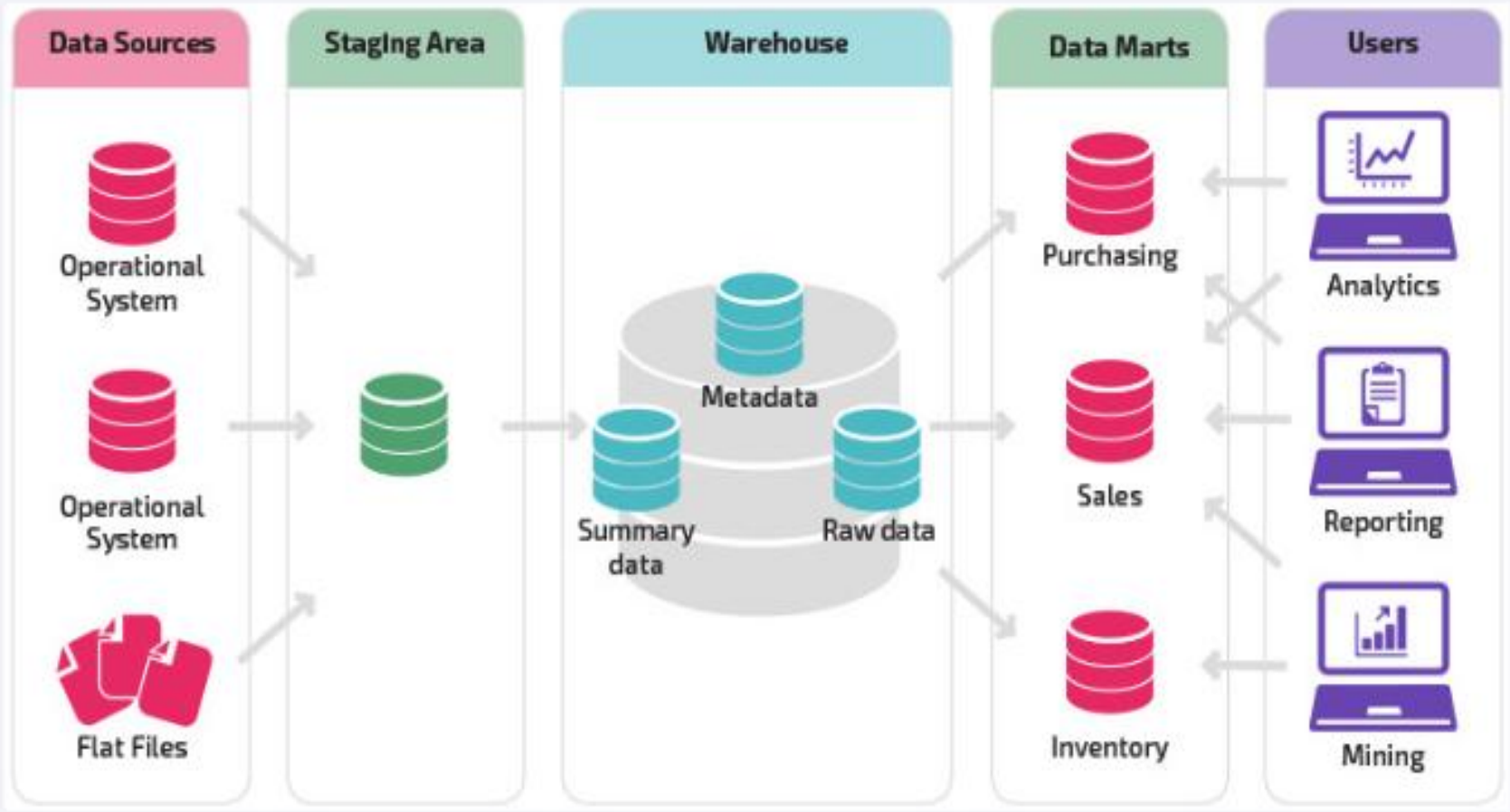
Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization. Corporate-wide data integration.
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

virtual warehouse

- With a virtual warehouse, employees can see all products currently in stock at a glance, as well as any products the company needs to make or order from other sources and, in turn, sell to customers. This means the company can keep and access a wider and deeper choice of products. Larger product inventory means businesses can satisfy customers who are looking for very specific products.

- For quick product delivery(Amazon)


Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
Operational System
Flat Files
Metadata
Summary data
Raw data
Purchasing
Sales
Inventory
Analytics
Reporting
Mining

Metadata Repository

- **Meta data** is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents.
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- Summary

From Tables and Spreadsheets to Data Cubes

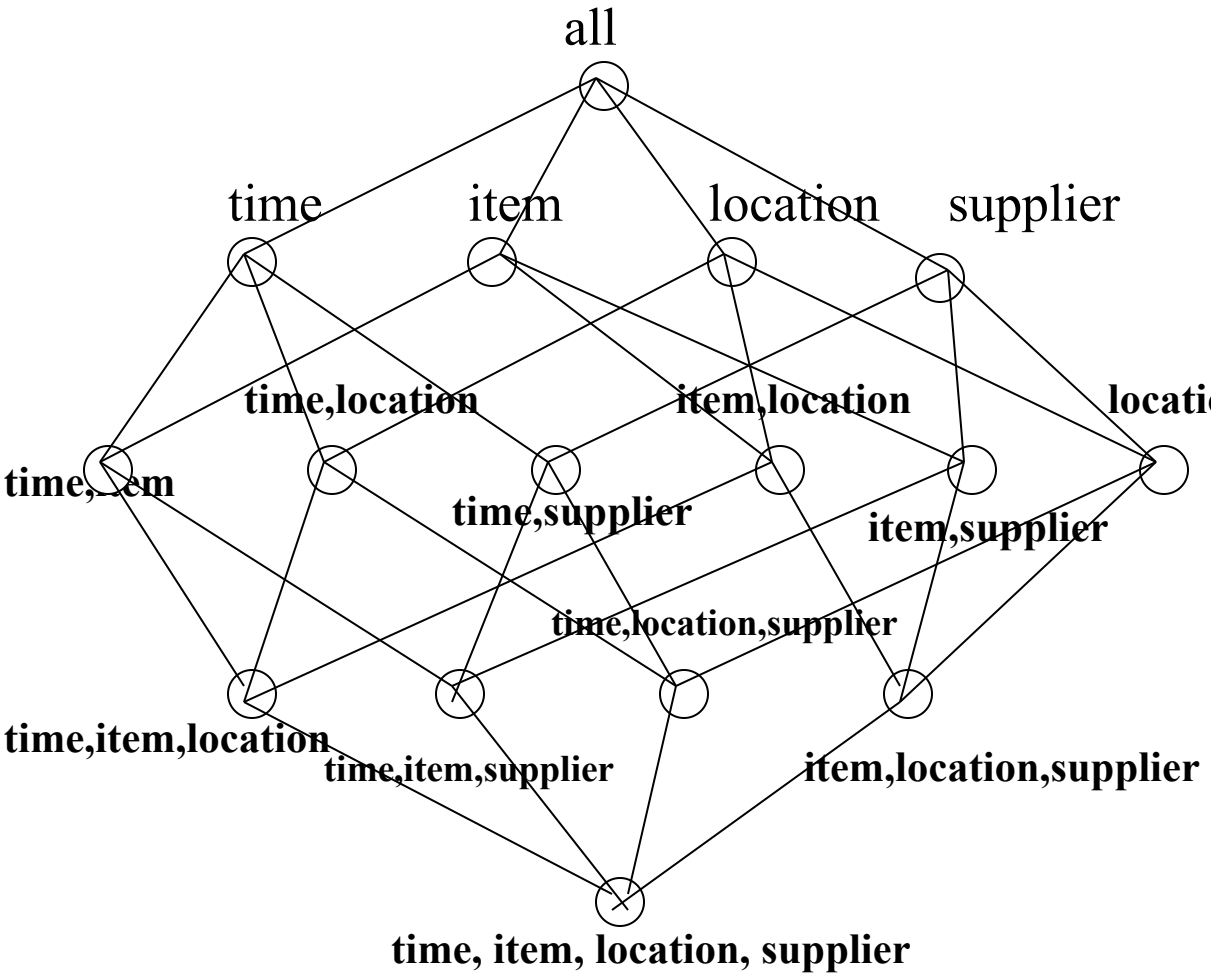
- A **data warehouse** is based on a multidimensional data model which views data in the form of a **data cube.**
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Example Data cube implementation

- suppose that you would like to create a data cube for Electronic store that contains the following: *item, city, year* and *sales in dollars*.

- Compute the sum of sales, grouping by "item and city"
- Compute the sum of sales, grouping by "item"
- Compute the sum of sales, grouping by "city"
- What is the total cubiod/group by can be computed?

Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

The "Compute Cube" Operator

- Cube definition and computation in DMQL

  define cube sales [item, city, year]: sum (sales_in_dollars)

  compute cube sales

- Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al.'96)

  SELECT item, city, year, SUM (amount)

  FROM SALES

  CUBE BY item, city, year

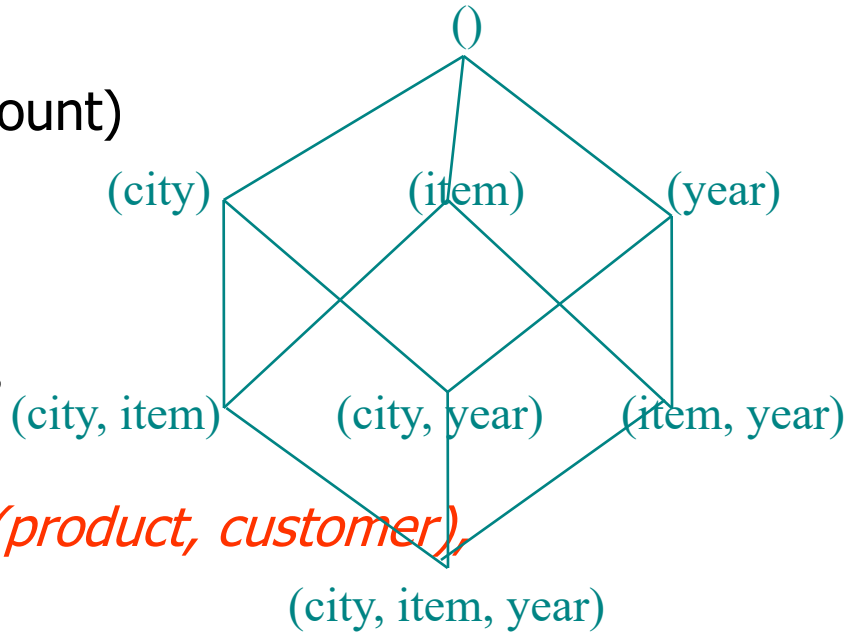
- Need compute the following Group-Bys

  (date, product, customer),

  (date,product),(date, customer), (product, customer),

  (date), (product), (customer)

  ()



Star schema vs snowflake schema

- The star schema and snowflake schema are two ways to **structure** a data warehouse.

- The star schema has a centralized data repository, stored in a fact table. The schema splits the fact table into a series of denormalized dimension tables. The fact table contains aggregated data to be used for reporting purposes while the dimension table describes the stored data.

- Denormalized designs are less complex because the data is grouped. The fact table uses only one link to join to each dimension table. The star schema's simpler design makes it much easier to write complex queries.

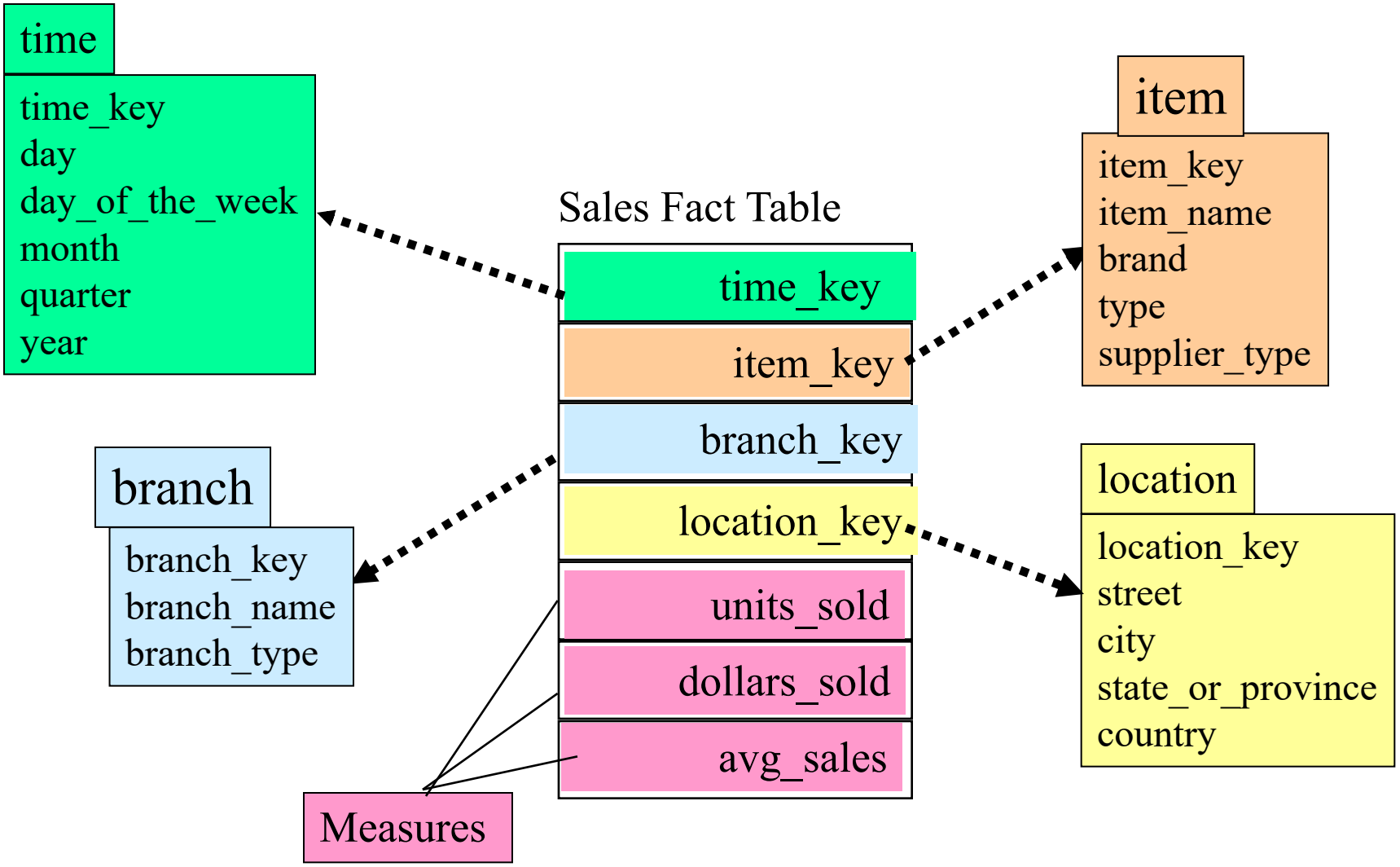
Conceptual Modeling of Data Warehouses

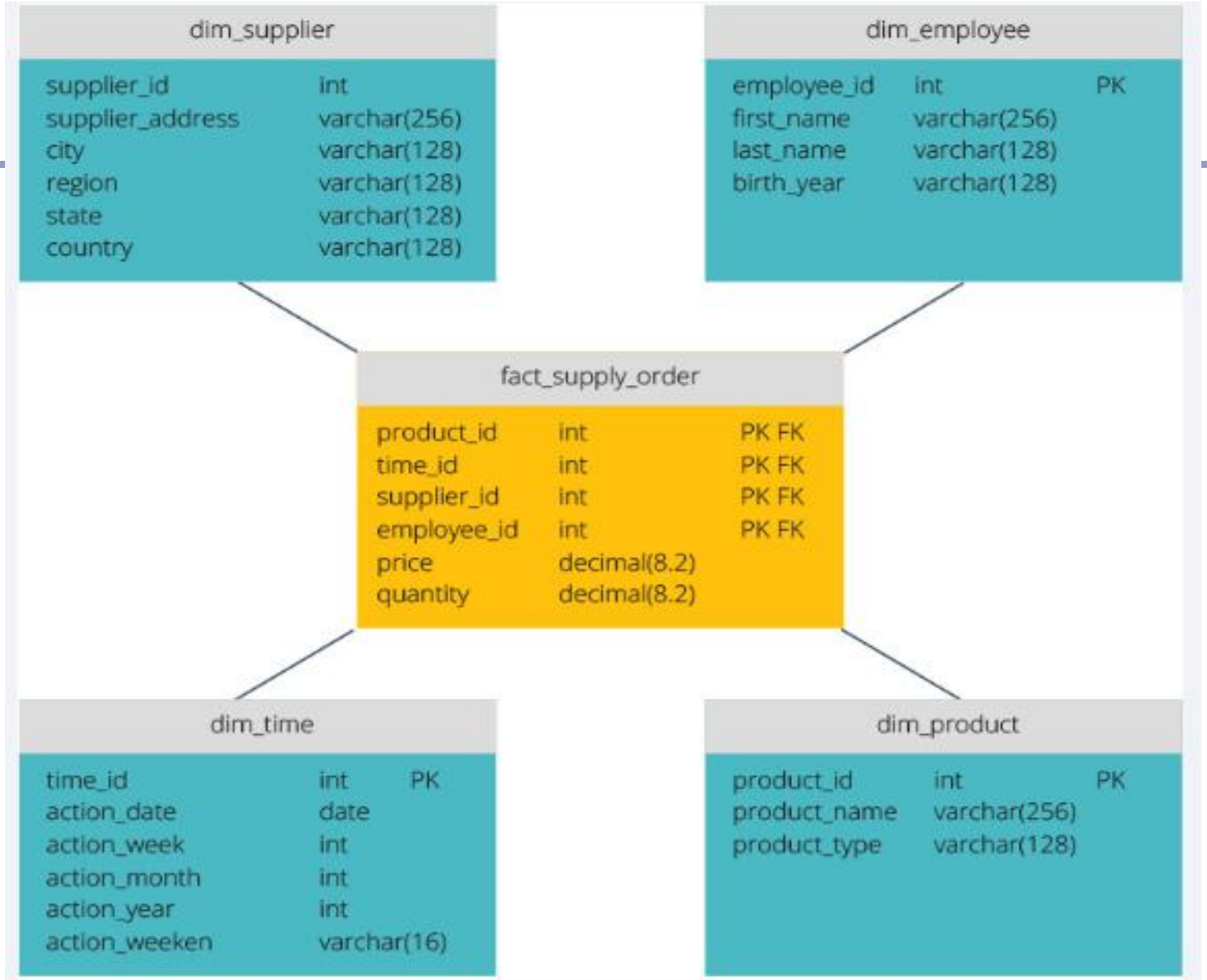
- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Steps in designing Star Schema:

- Identify a business process for analysis(like sales).
- Identify measures or facts (sales dollar).
- Identify dimensions for facts(product dimension, location dimension, time dimension, organization dimension).
- List the columns that describe each dimension.(region name, branch name, region name).
- Determine the lowest level of summary in a fact table(sales dollar).

Example of Star Schema



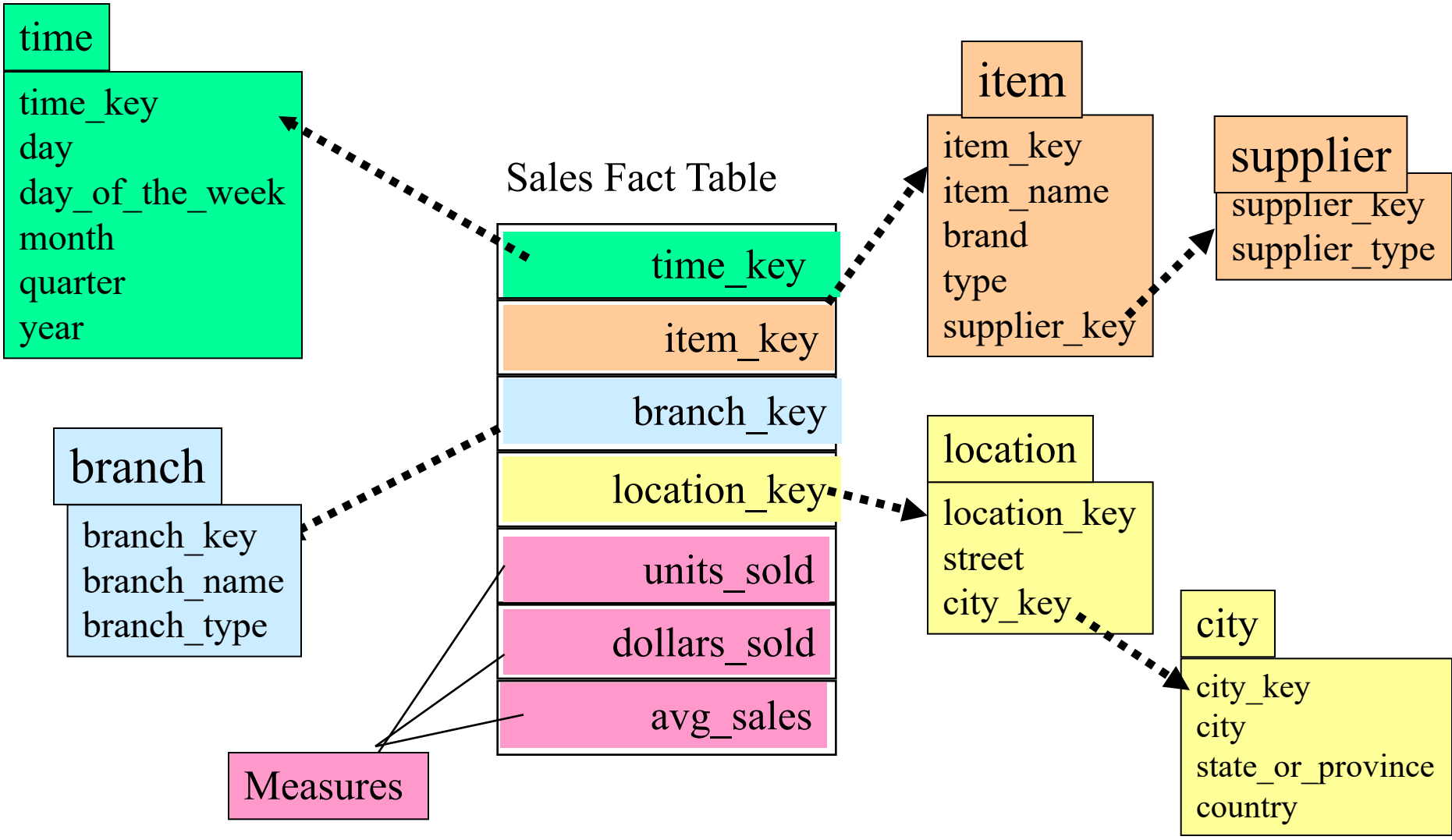

dim_supplier
supplier_id int
supplier_address varchar(256)
city varchar(128)
region varchar(128)
state varchar(128)
country varchar(128)
dim_employee
employee_id int PK
first_name varchar(256)
last_name varchar(128)
birth_year varchar(128)
fact_supply_order
product_id int PK FK
time_id int PK FK
supplier_id int PK FK
employee_id int PK FK
price decimal(8.2)
quantity decimal(8.2)
dim_time
time_id int PK
action_date date
action_week int
action_month int
action_year int
action_weeken varchar(16)
dim_product
product_id int PK
product_name varchar(256)
product_type varchar(128)

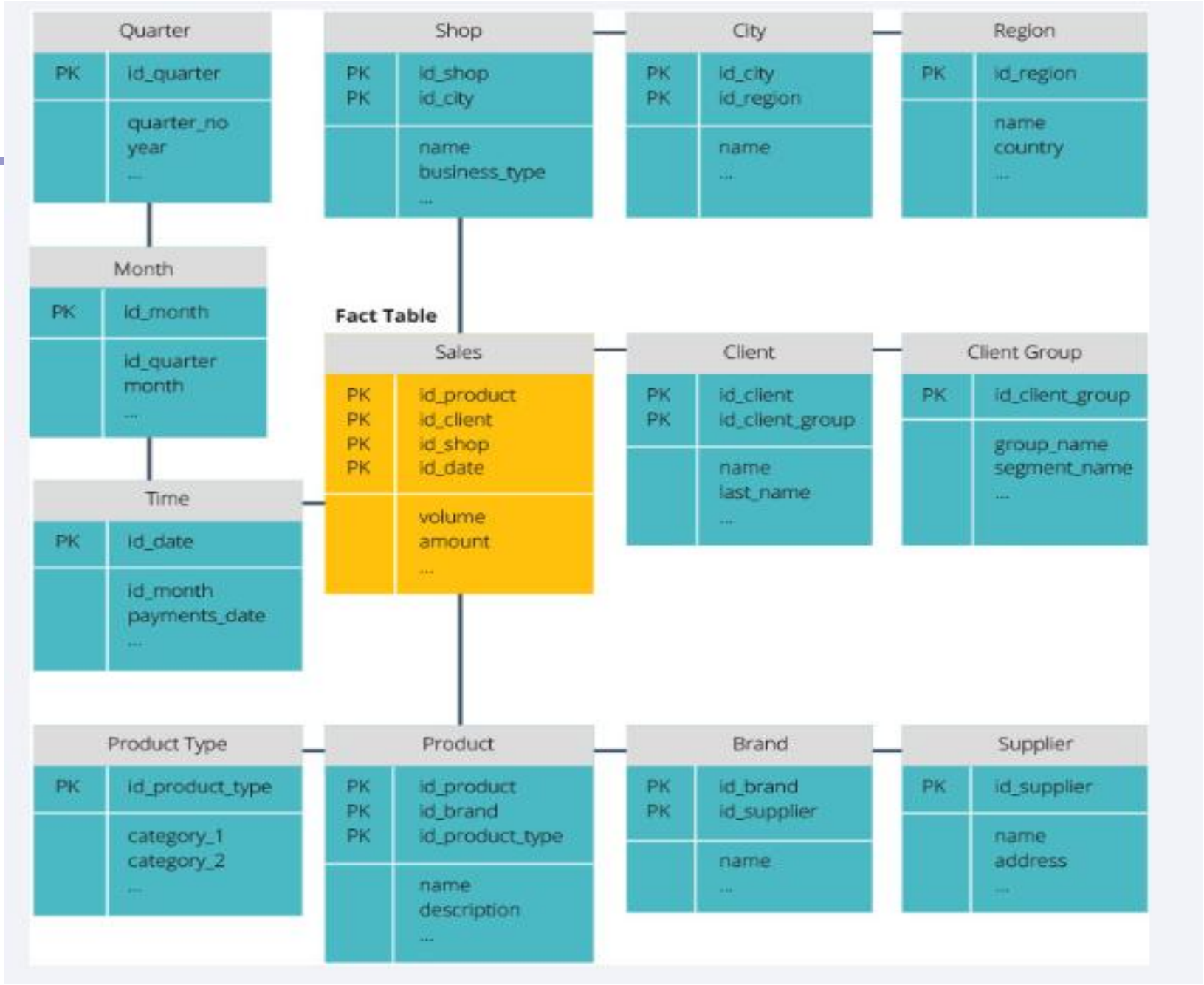
Snowflake Schema

- The snowflake schema is different because it normalizes the data. Normalization means efficiently organizing the data so that all data dependencies are defined, and each table contains minimal redundancies.
- Single dimension tables thus branch out into separate dimension tables.

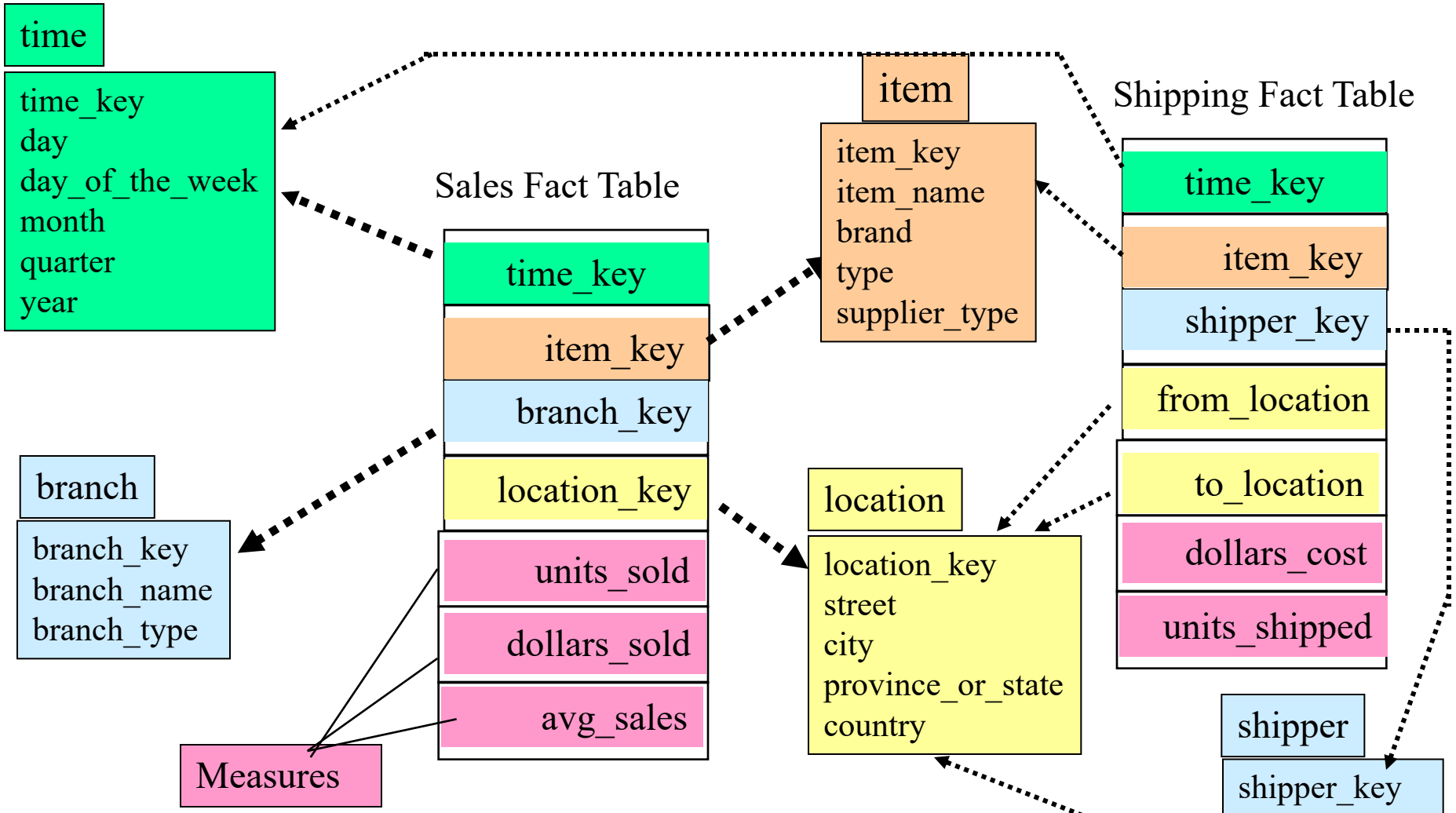
- The snowflake schema uses less disk space and better preserves data integrity.
- The main disadvantage is the complexity of queries required to access data—each query must dig deep to get to the relevant data because there are multiple joins.

Example of Snowflake Schema




Quarter
PK id_quarter
quarter_no
year
...
Shop
PK id_shop
PK id_city
name
business_type
...
City
PK id_city
PK id_region
name
...
Region
PK id_region
name
country
...
Month
PK id_month
id_quarter
month
...
Fact Table
Sales
PK id_product
PK id_client
PK id_shop
PK id_date
volume
amount
...
Client
PK id_client
PK id_client_group
name
last_name
...
Client Group
PK id_client_group
group_name
segment_name
...
Time
PK id_date
id_month
payments_date
...
Product Type
PK id_product_type
category_1
category_2
...
Product
PK id_product
PK id_brand
PK id_product_type
name
description
...
Brand
PK id_brand
PK id_supplier
name
...
Supplier
PK id_supplier
name
address
...

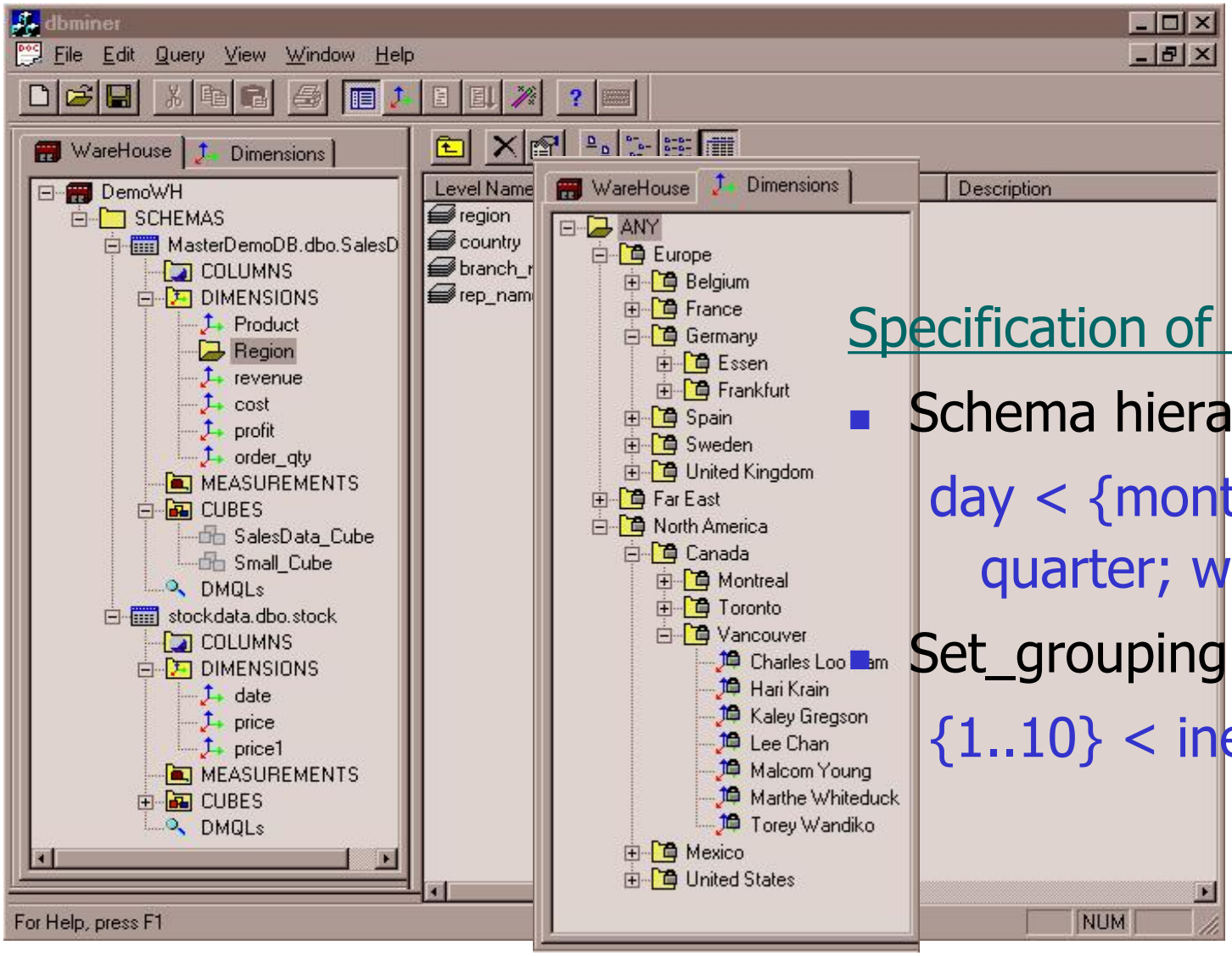
Example of Fact Constellation



Data Cube Measures: Three Categories

- Distributive: if the result derived by applying the function to *n* aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., count(), sum(), min(), max()
- Algebraic: if it can be computed by an algebraic function with *M* arguments (where *M* is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., avg(), min_N(), standard_deviation()
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., median(), mode(), rank()

View of Warehouses and Hierarchies



Specification of hierarchies

- Schema hierarchy

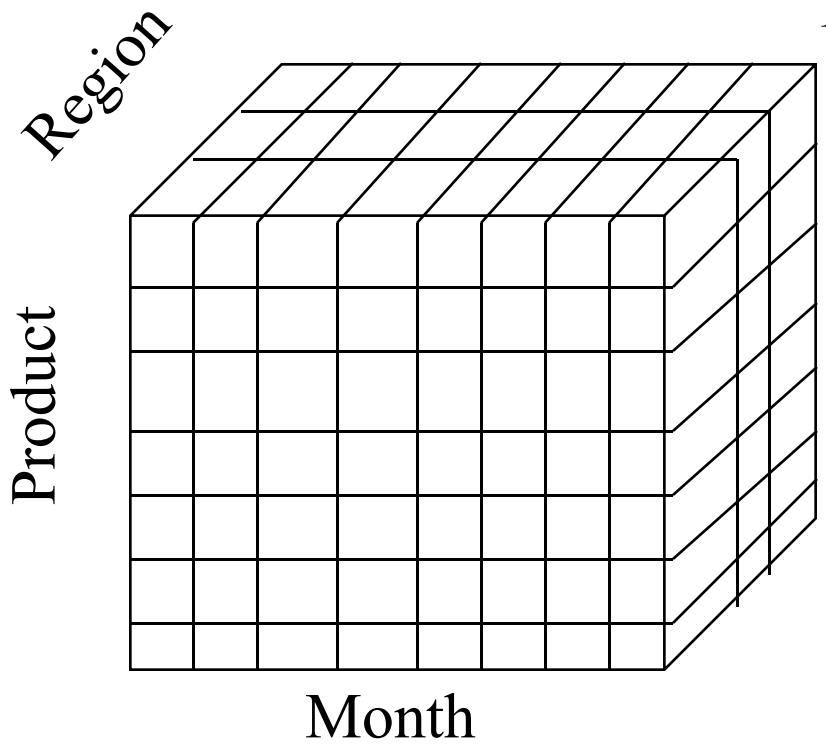
  day < {month < quarter; week} < year

Set_grouping hierarchy

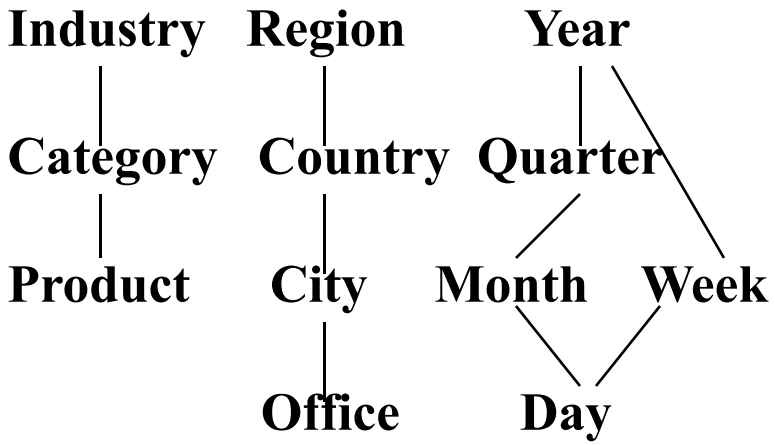
{1..10} < inexpensive

Multidimensional Data

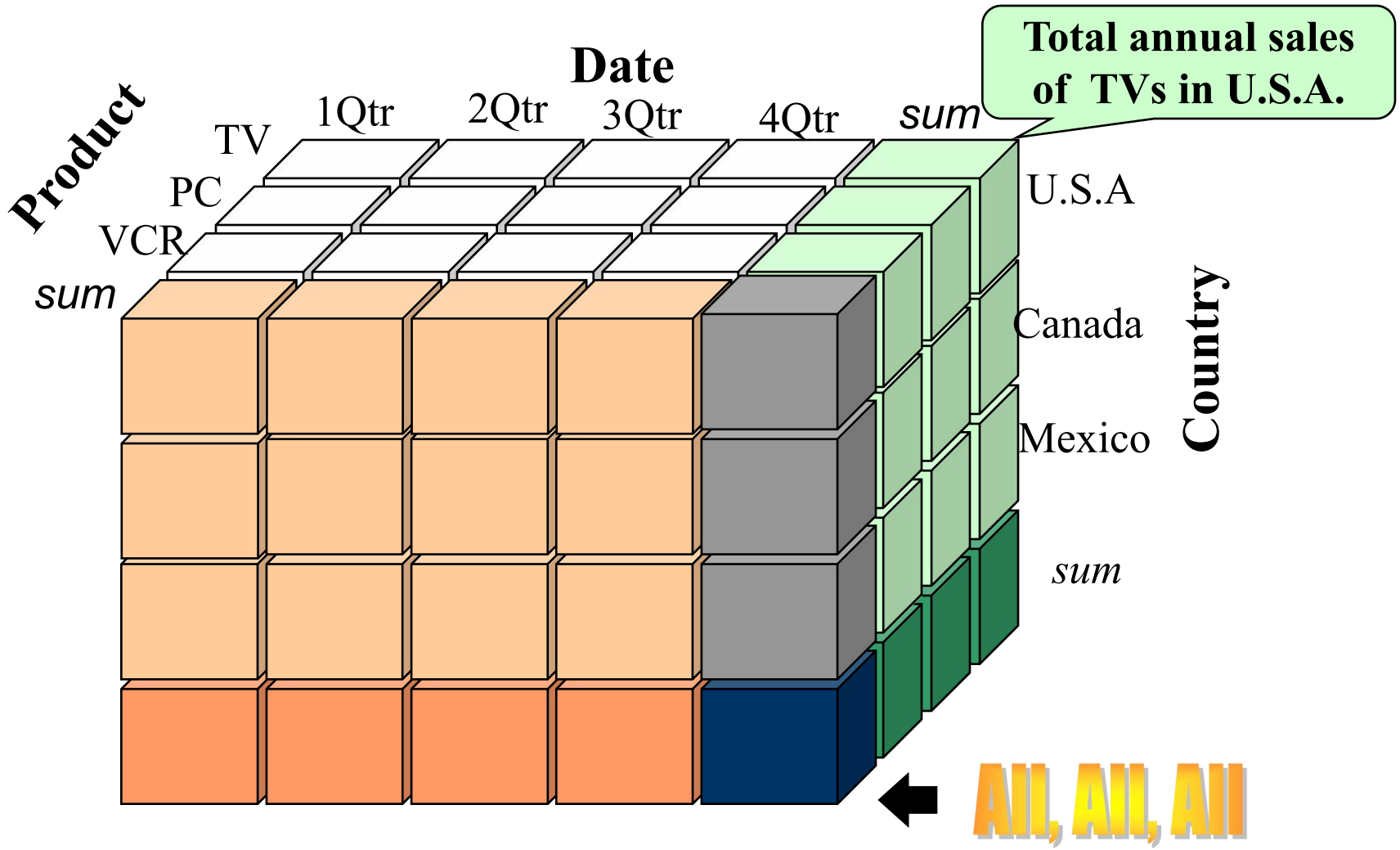
- Sales volume as a function of product, month, and region



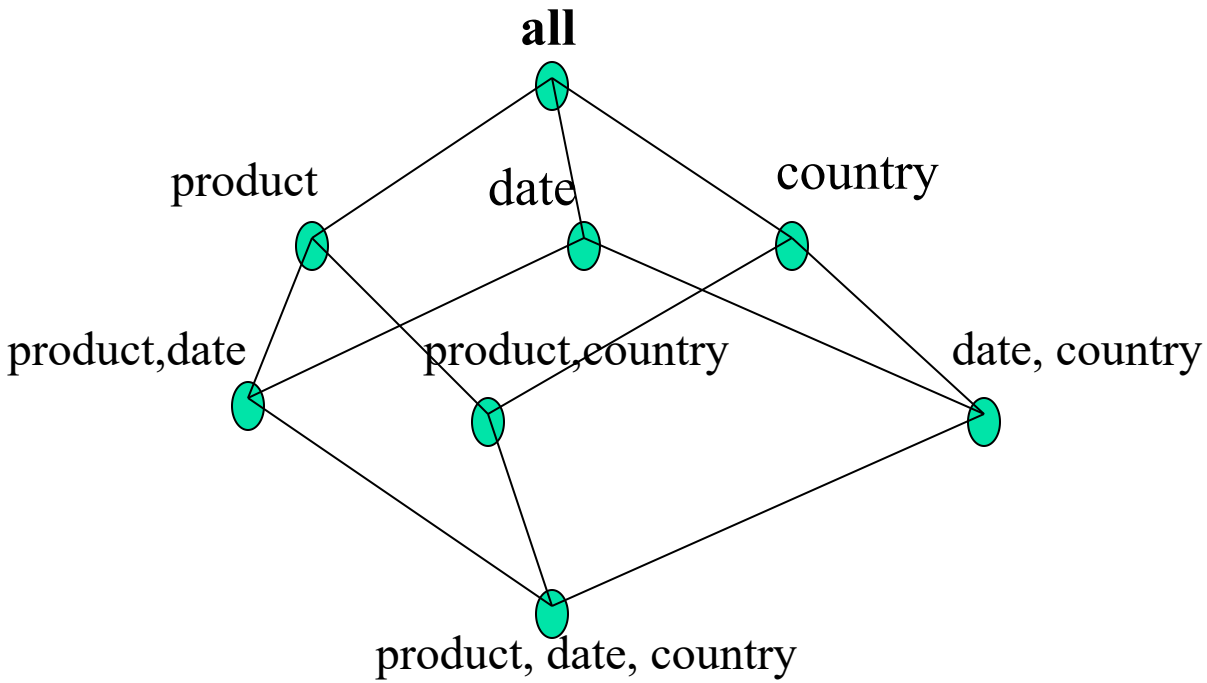
Dimensions: *Product, Location, Time*
Hierarchical summarization paths



A Sample Data Cube



Cuboids Corresponding to the Cube



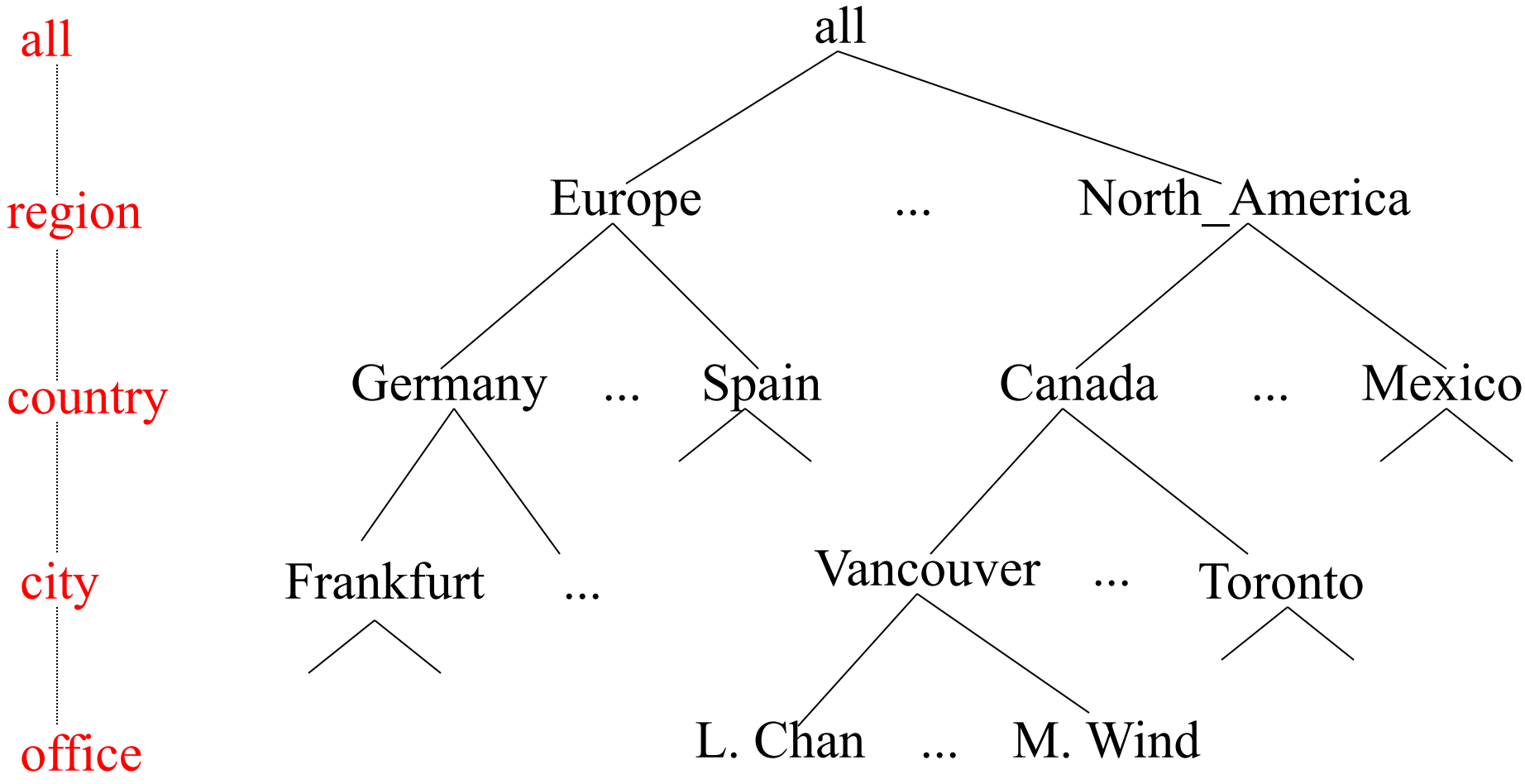
0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D (*base*) cuboid

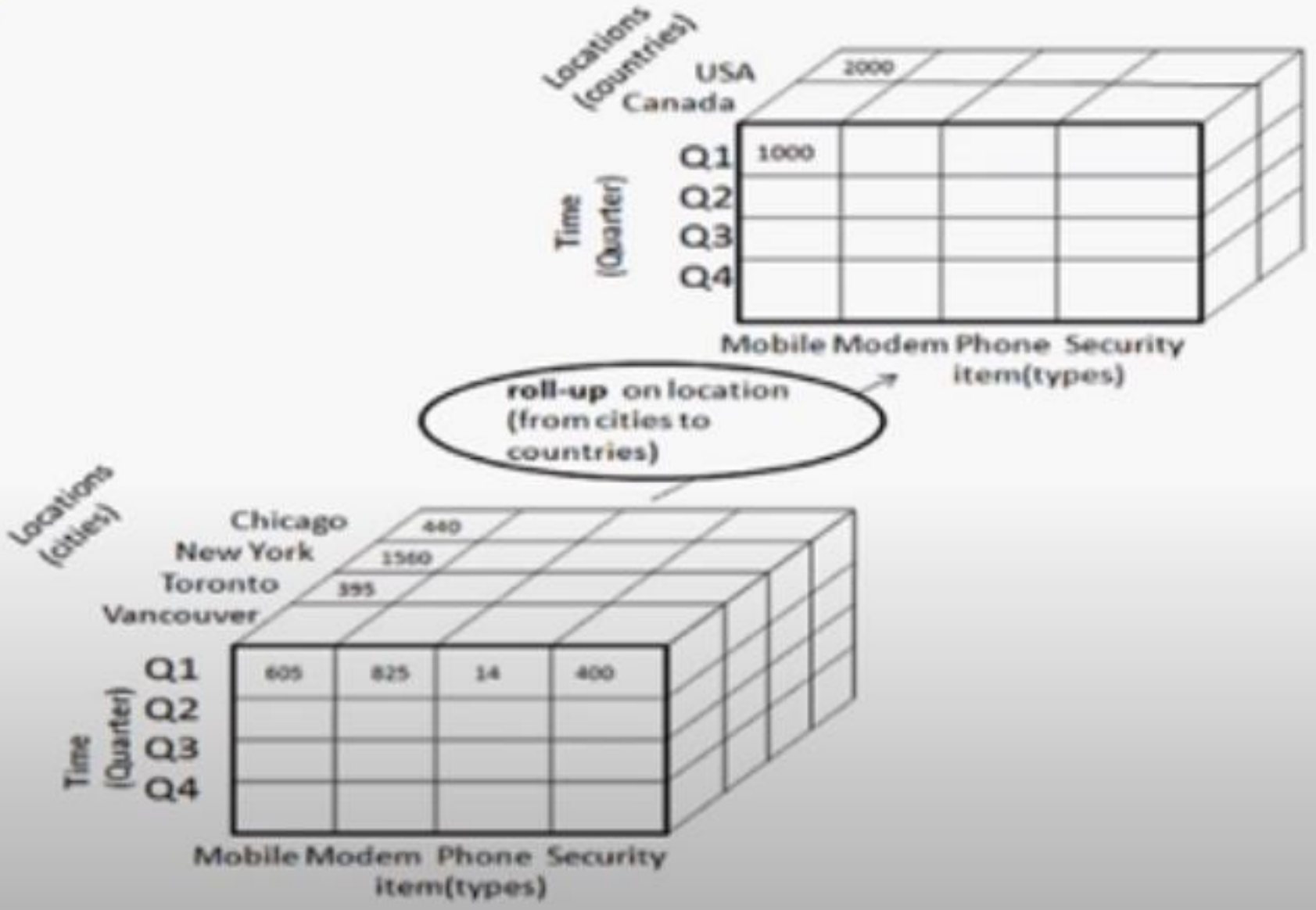
A Concept Hierarchy: Dimension (location)



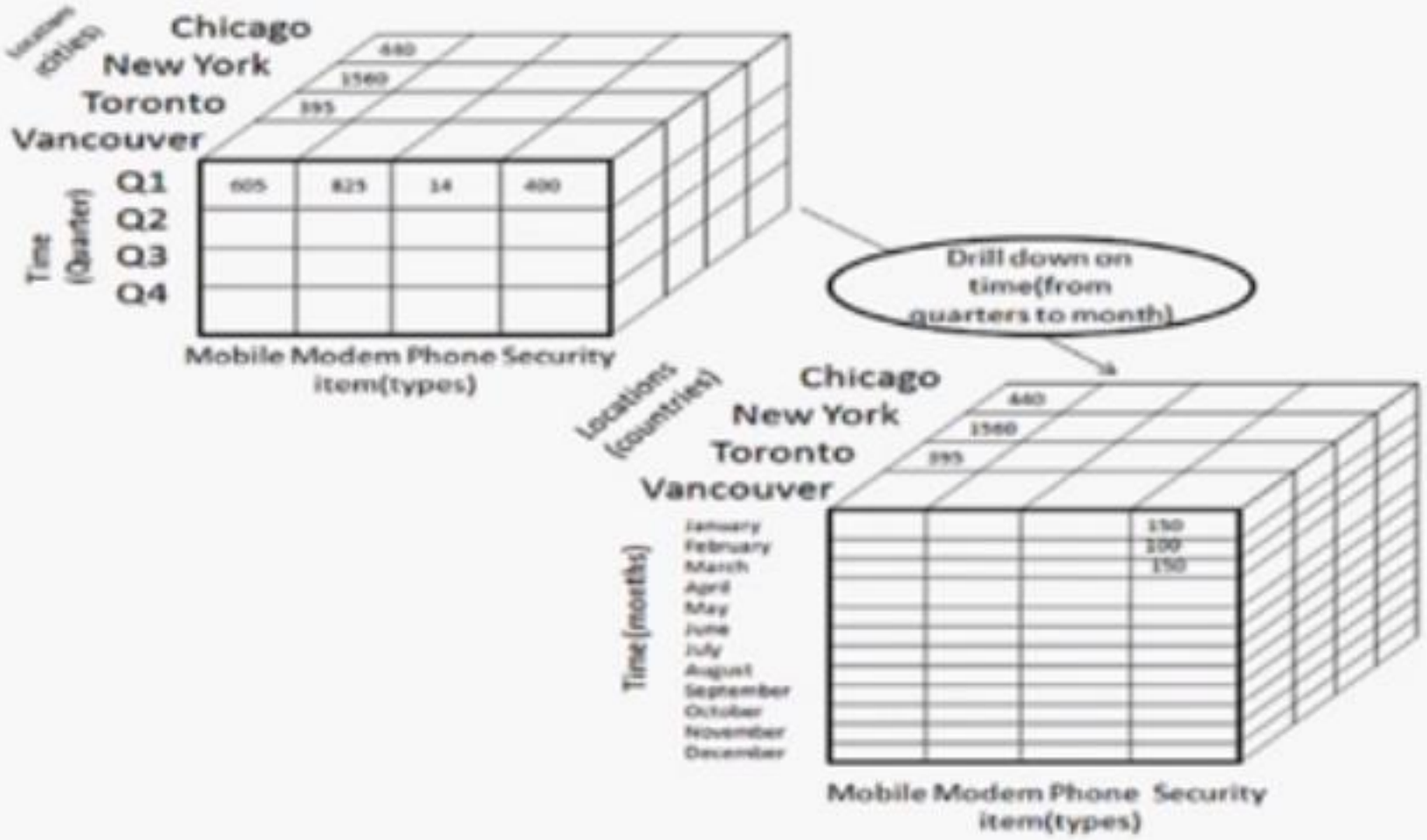
Typical OLAP Operations

- 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*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

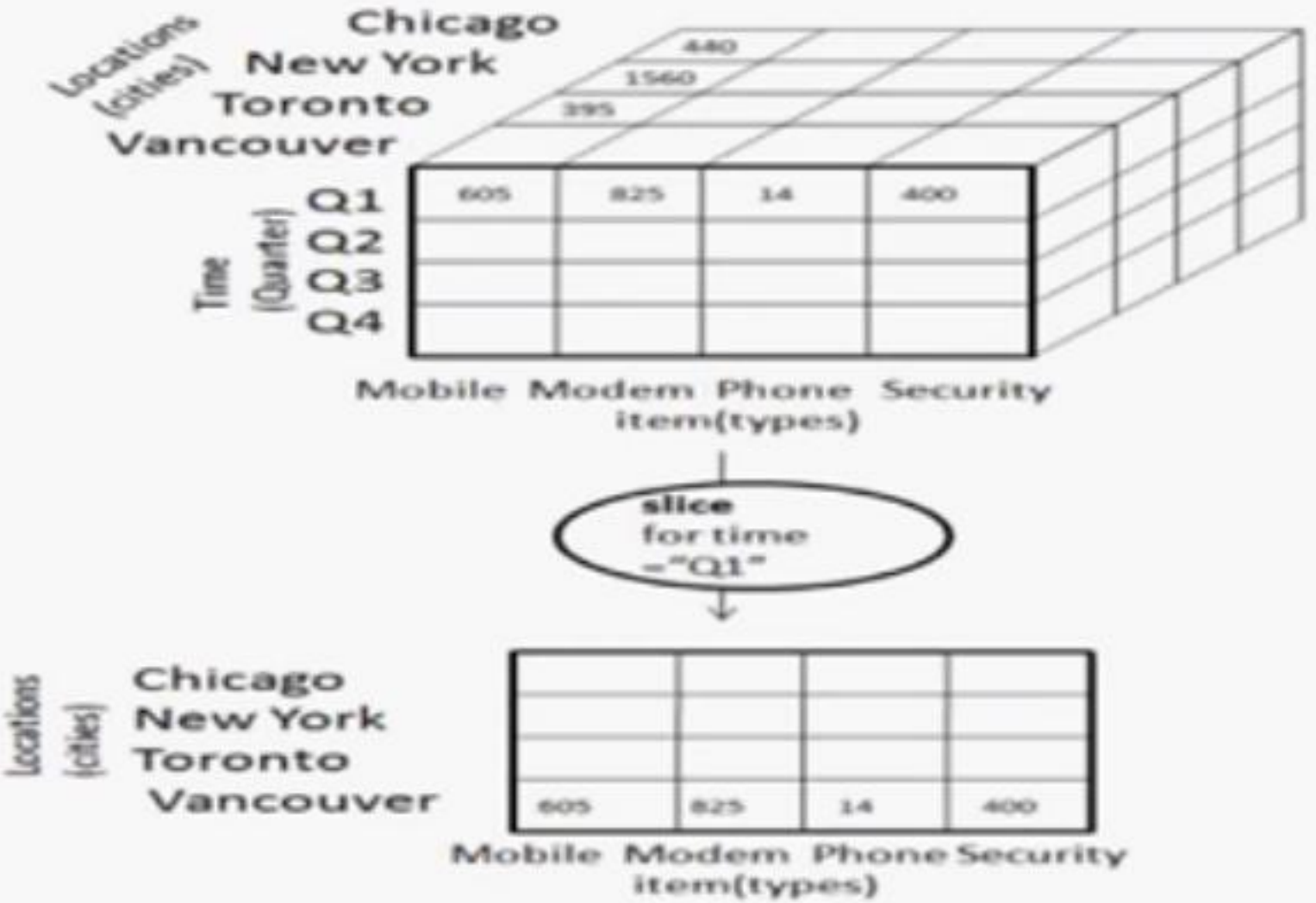
Roll-Up


Locations (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)

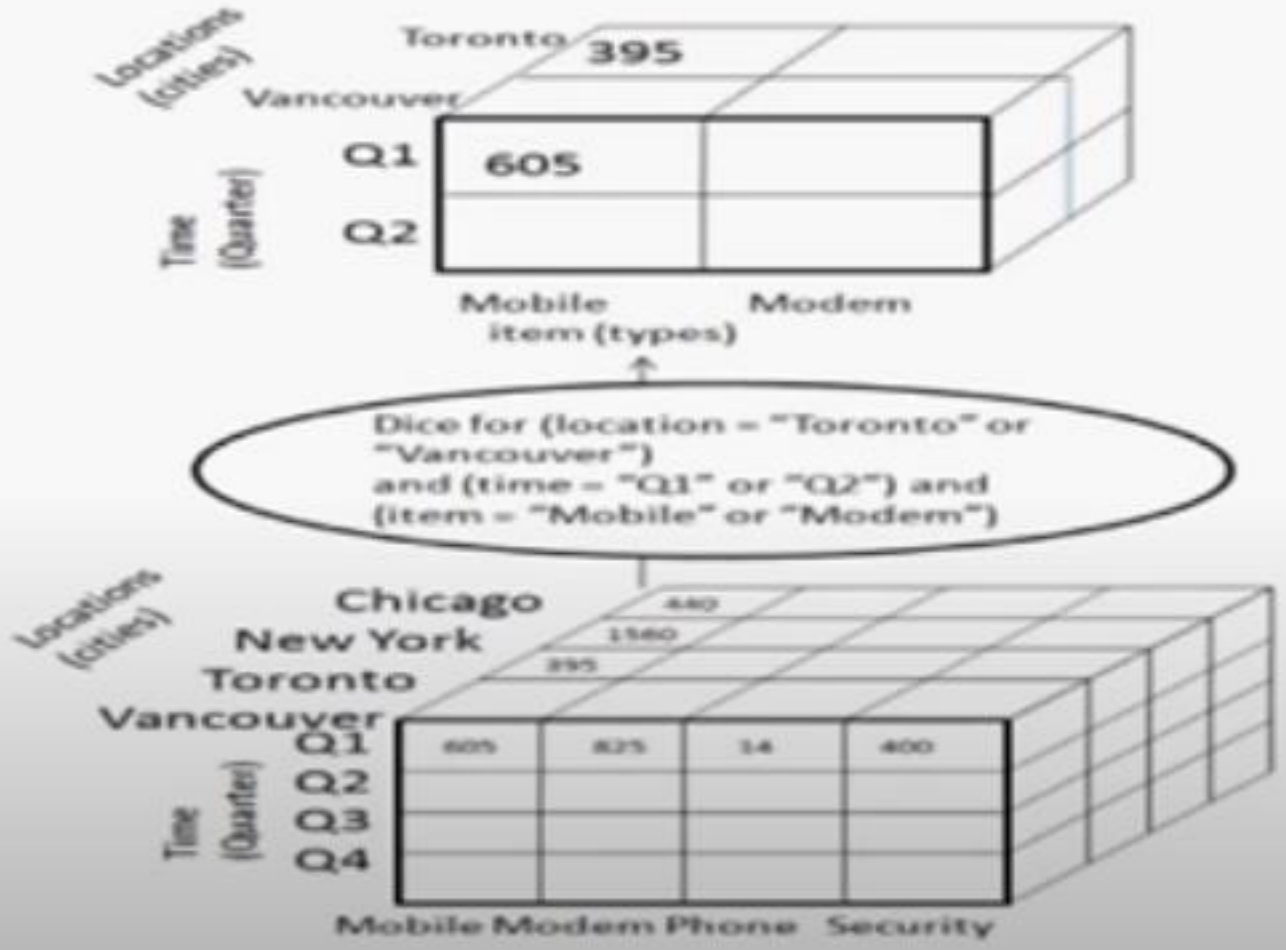
Drill-Down


Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)
Drill down on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
150
100
150
January
February
March
April
May
June
July
August
September
October
November
December
Time (months)
Mobile Modem Phone Security
item(types)

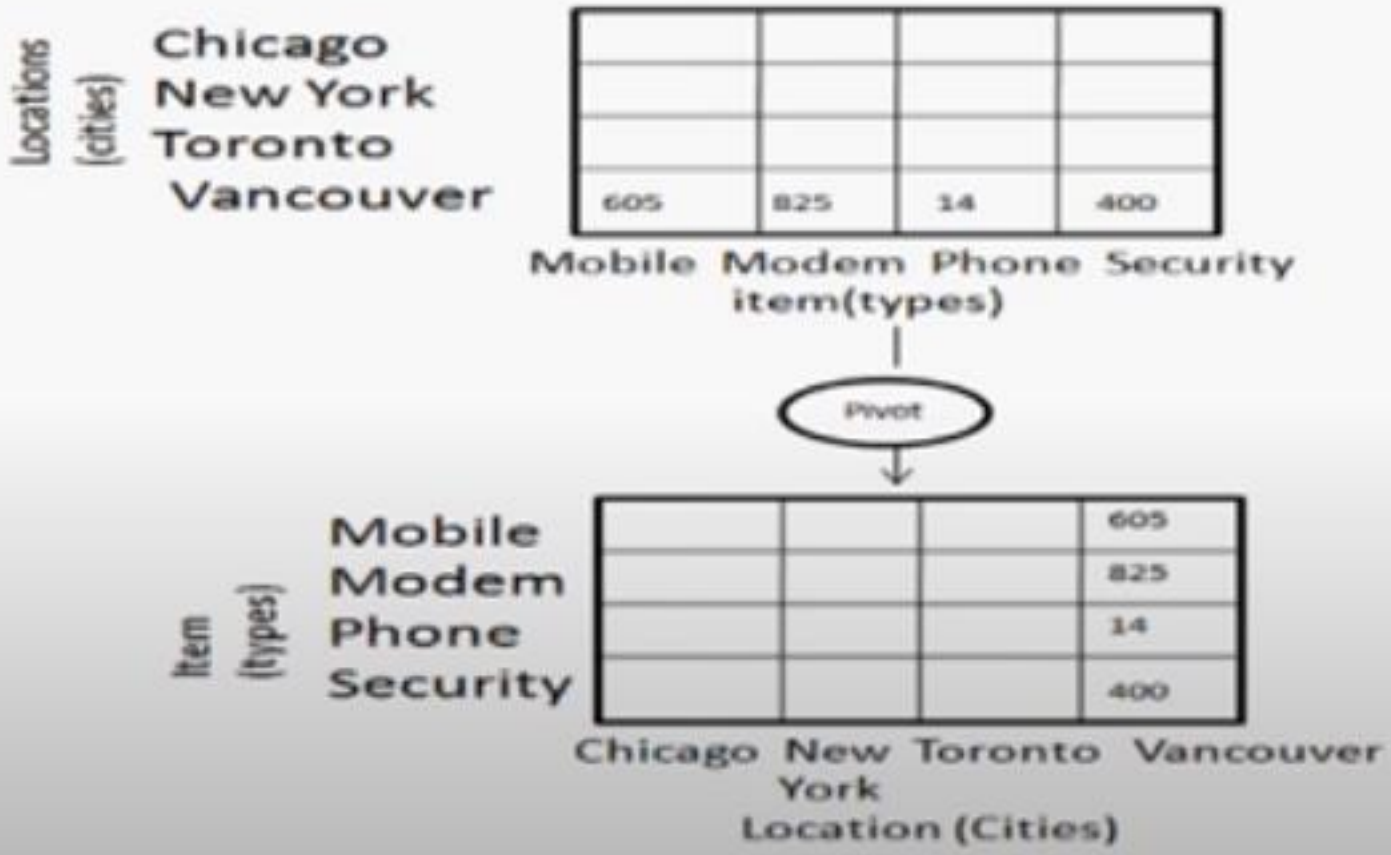
Slice


Locations (cities)
Chicago
New York
Toronto
Vancouver
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
440
1560
395
Mobile Modem Phone Security
item(types)
slice
for time
="Q1"
Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)

Dice


Locations (cities)
Toronto
395
Vancouver
Q1
605
Q2
Time (Quarter)
Mobile
Modem
item (types)
Dice for (location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "Mobile" or "Modem")
Locations (cities)
Chicago
New York
Toronto
Vancouver
480
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
Mobile
Modem
Phone
Security

Pivot


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)
Pivot
Item (types)
Mobile
Modem
Phone
Security
605
825
14
400
Chicago New York Toronto Vancouver
Location (Cities)

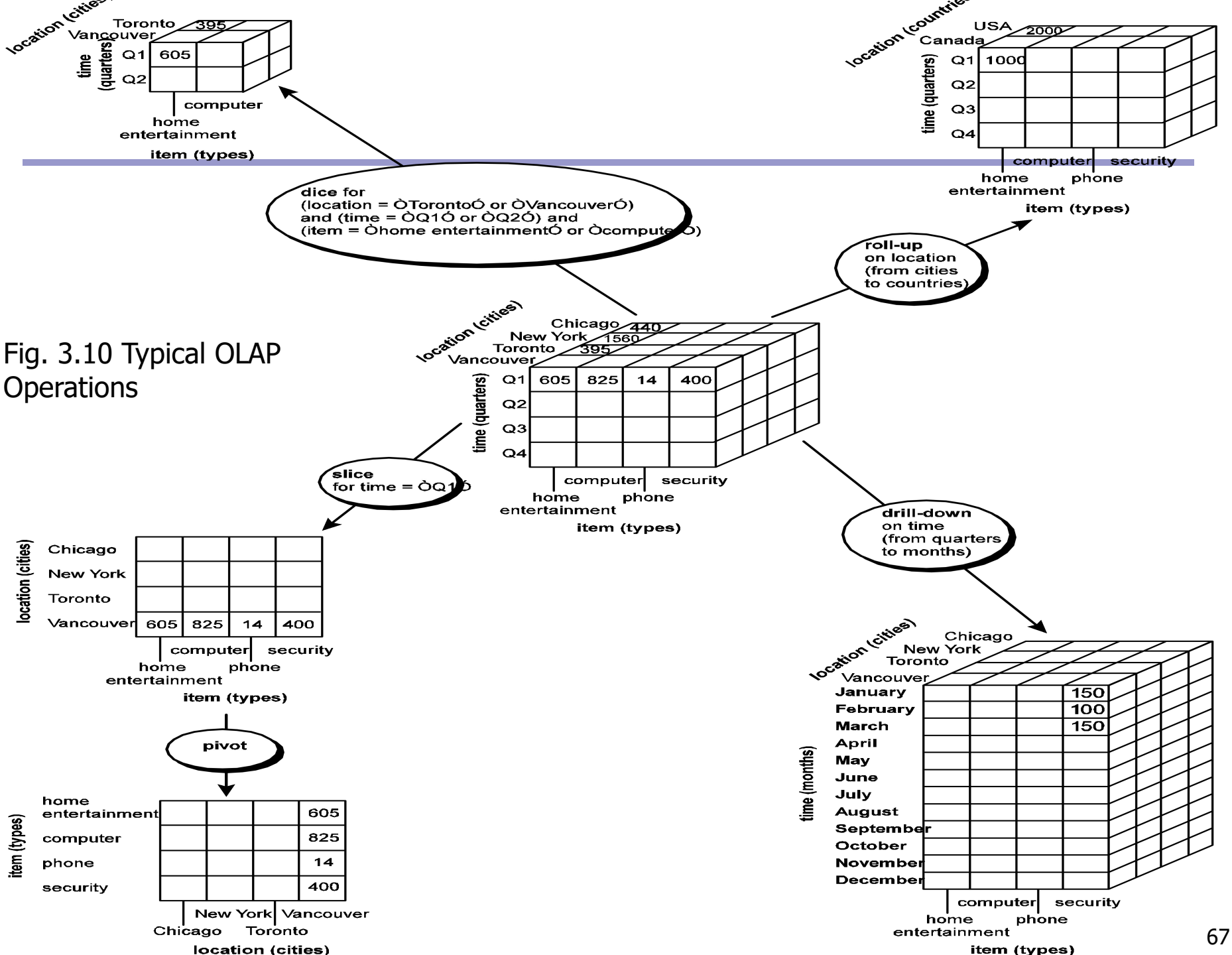
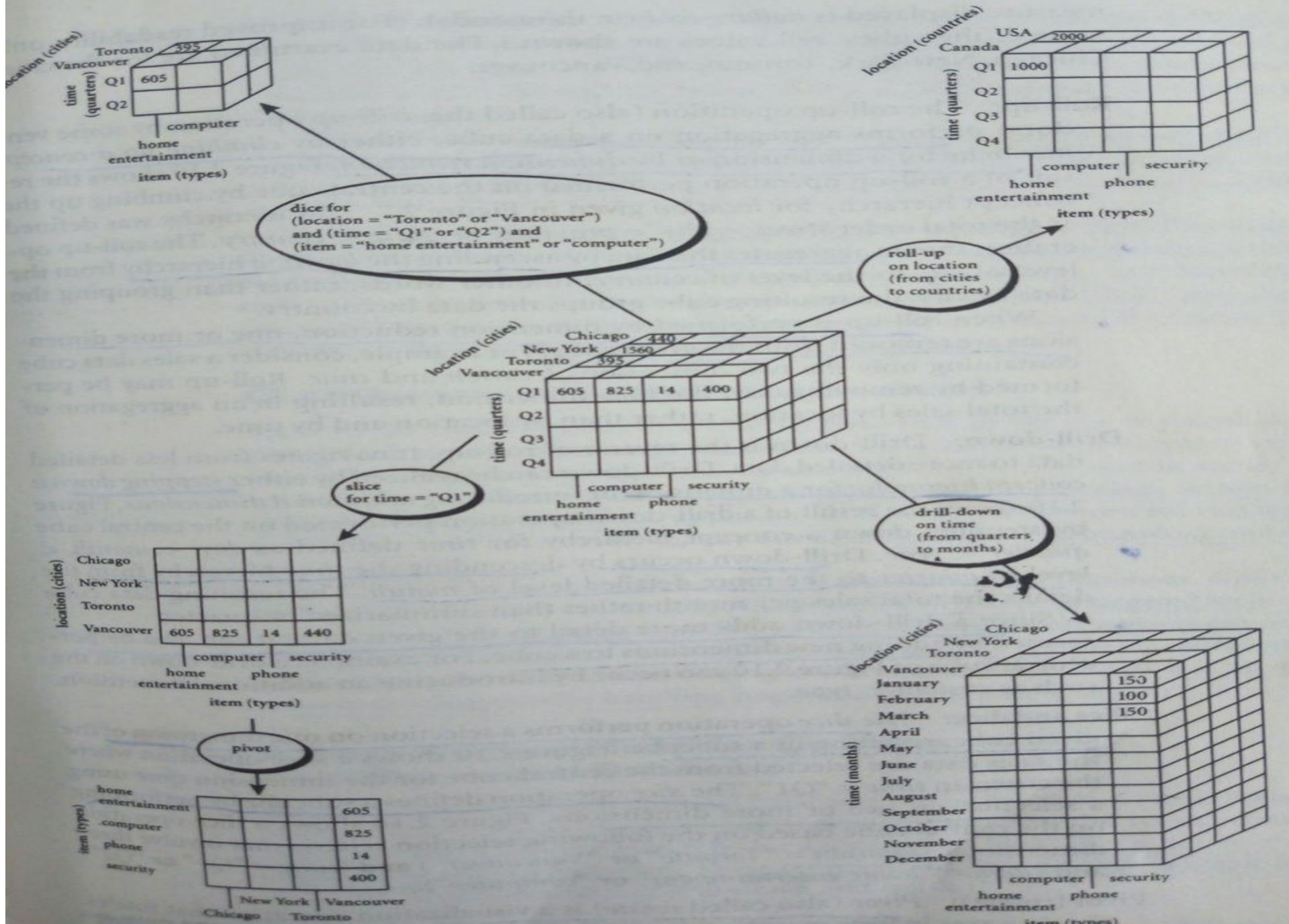


Fig. 3.10 Typical OLAP Operations


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
Q2
605
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
computer
security
home entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
440
computer
security
home entertainment
phone
item (types)
pivot
item (types)
home entertainment
computer
phone
security
605
825
14
400
Chicago
New York
Toronto
Vancouver
location (cities)
Chicago
New York
Toronto
Vancouver
January
February
March
April
May
June
July
August
September
October
November
December
time (months)
150
100
150
computer
security
home entertainment
phone
item (types)