# Data Warehouse

Data Engineering

# Two Broad Types of Database Activity

---

- OLTP - Online Transaction Processing
  - Short transactions
  - Simple queries
  - Touch small portions of data
  - Frequent updates
- OLAP - Online Analytical Processing
  - Long transactions
  - Complex queries
  - Touch large portions of the data
  - Infrequent updates

# More Terminology

---

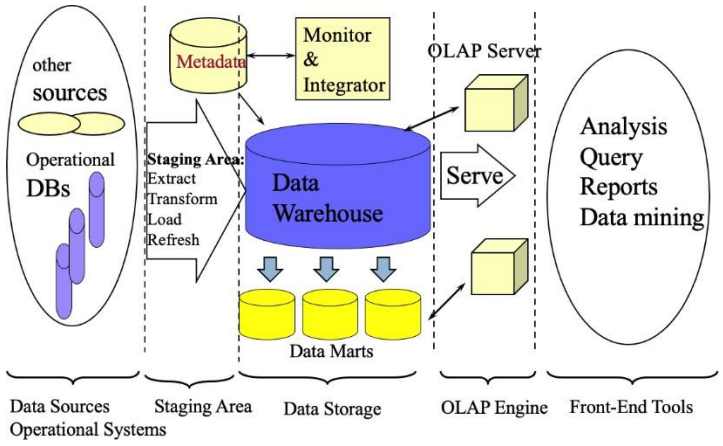
Data warehousing

Bring data from operational (OLTP) sources into a single "warehouse" for (OLAP) analysis

- Decision support system (DSS)

  Infrastructure for data analysis

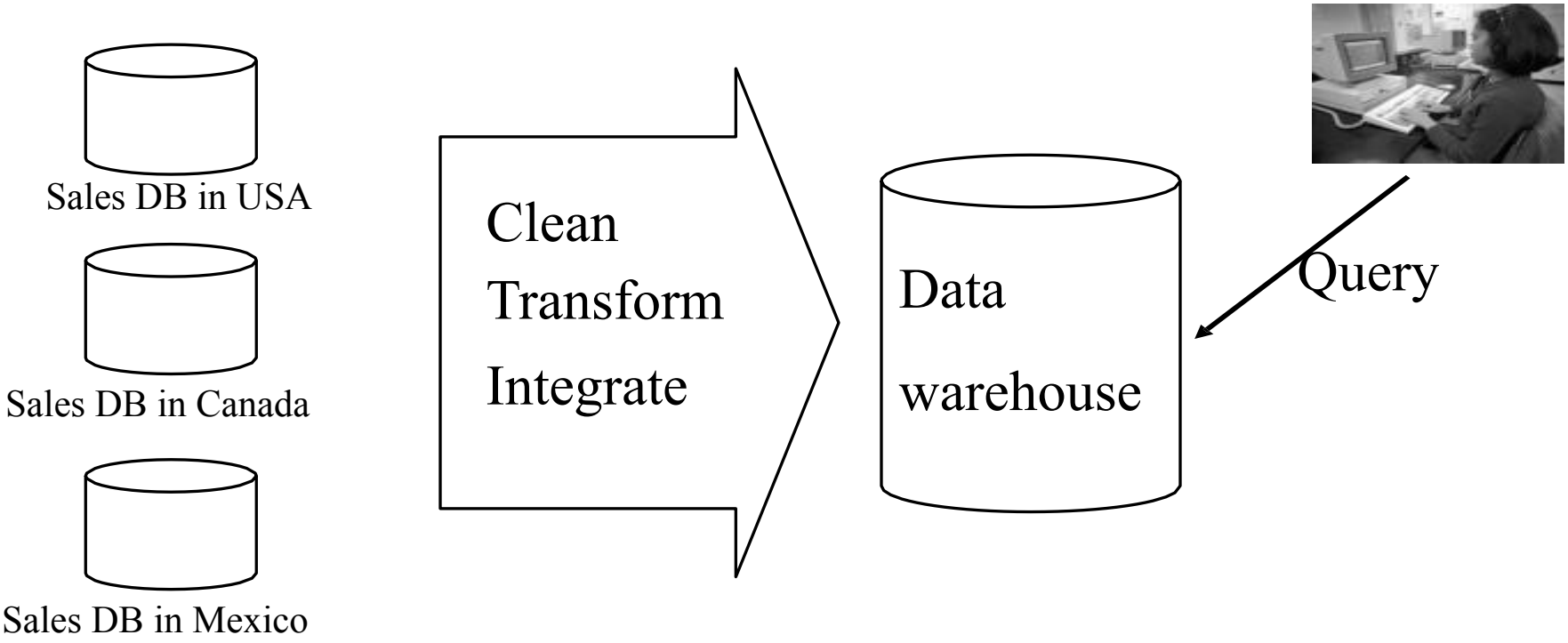
  E.g., data warehouse tuned for OLAP



# Data Warehouse

---

- A repository of multiple heterogeneous data resources (DBs), organized under a unified schema.



# Not in 3NF (Big Data)

---

The purpose of Data warehousing is to provide aggregate data which is in a suitable for decision making

- Data denormalized to 2NF
- Data redundancy
- Need more storage
- But faster data retrieval

# Star Schema

---

- Fact table

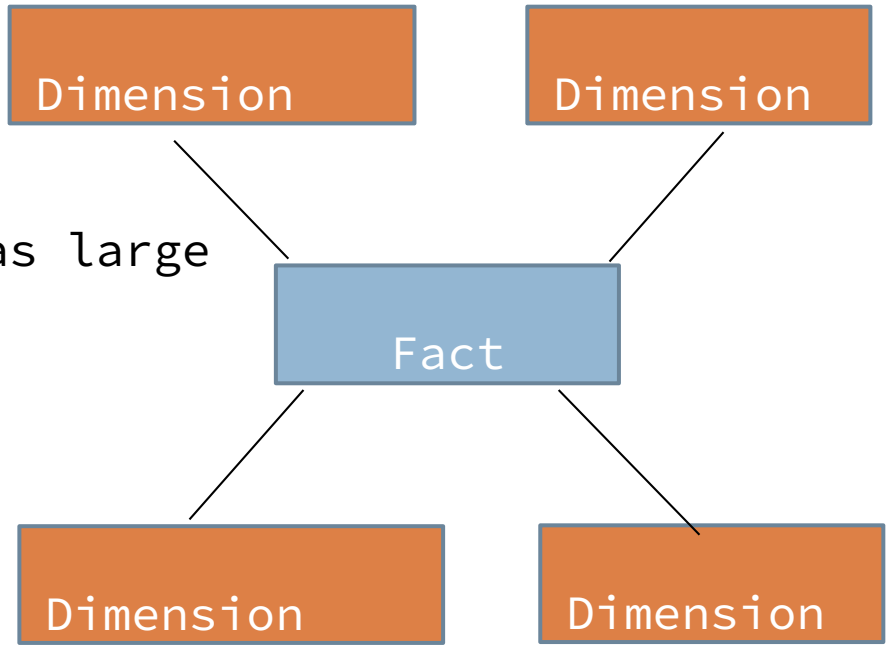
  Updated frequently, often append-only, very large

  - Example:
    - Sales Transaction
    - Course Enrollments
    - Page Views

- Dimension tables

  Updated infrequently, not as large

  - Example:
    - Stores, Items, Customers
    - Students, Courses
    - Web Pages, Users, Advertisers



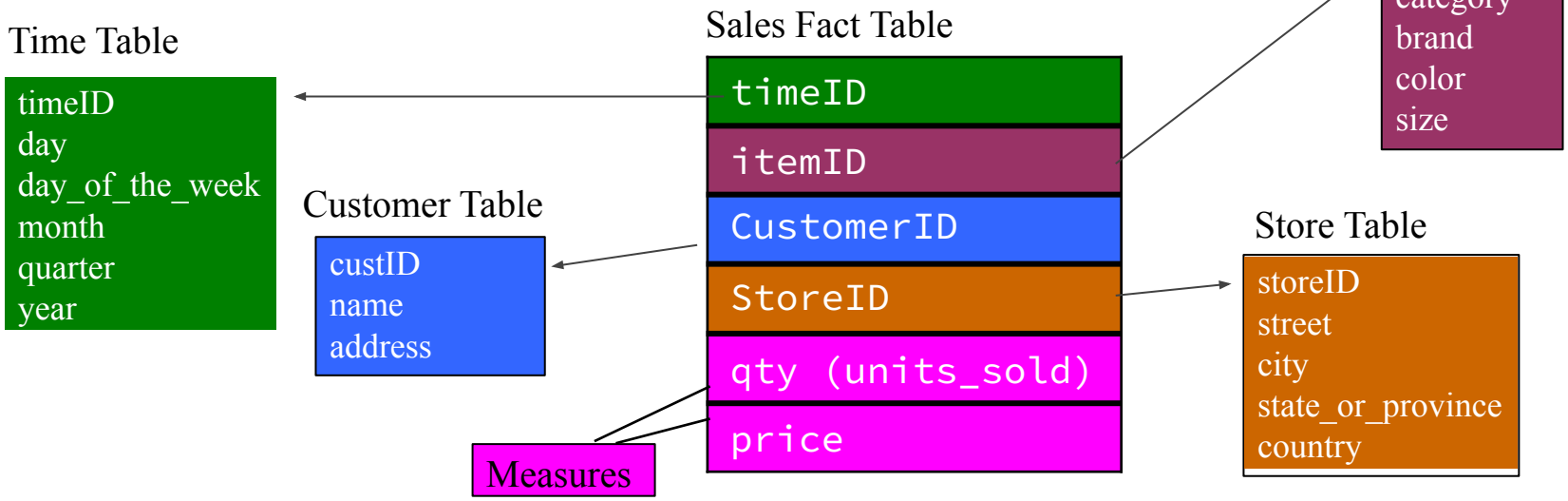
# Star Schema for Sales

--- **Star Schema** - fact table references dimension tables

Dimension attributes  Dependent attributes

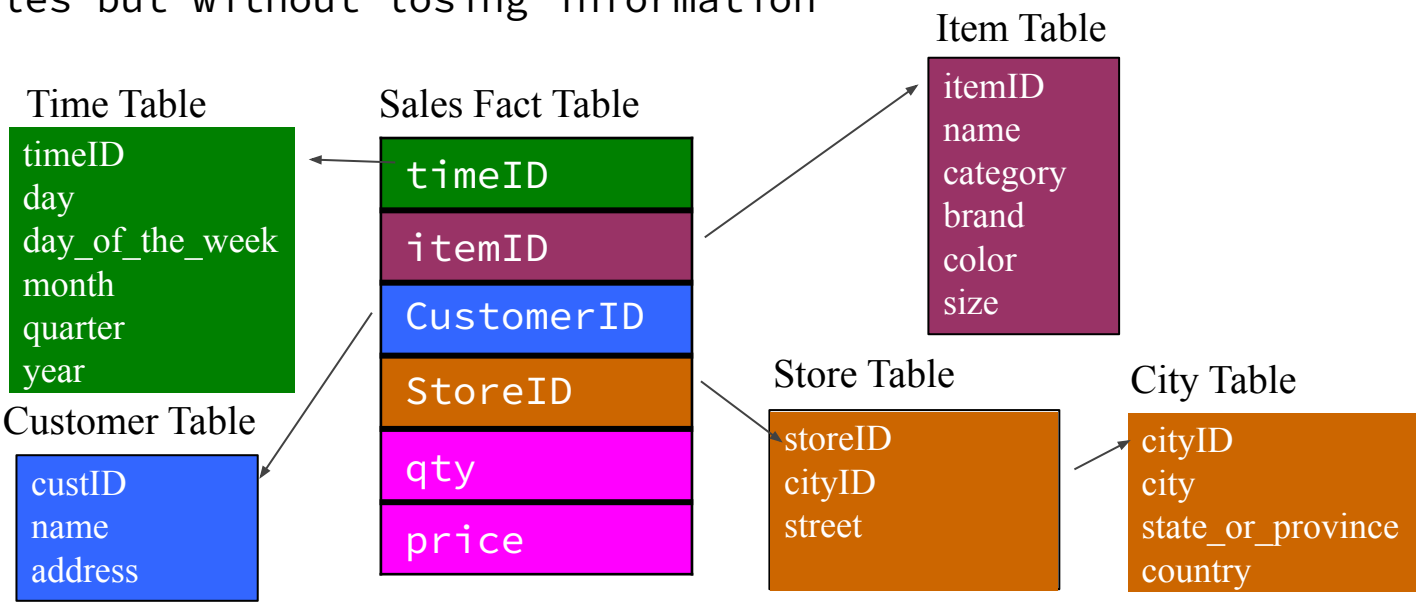
Queries tend to aggregate on dependent attributes

```
Sales(storeID, itemID, custID, timeID, qty, price)
Store(storeID, street, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
Time(timeID,day,day_of_the_week,month,quarter,year)
```



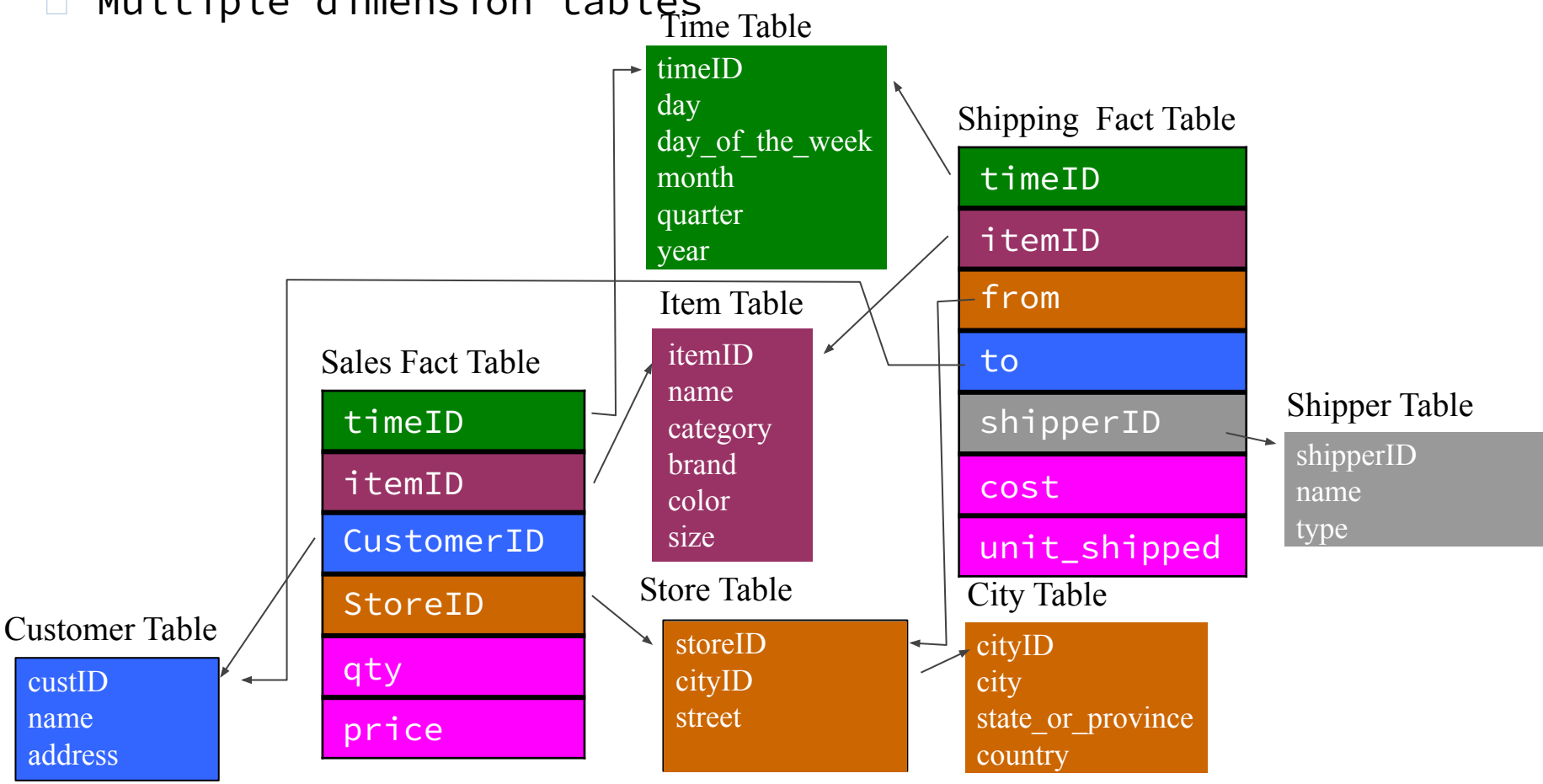
# Snowflake Schema

- Snowflake Schema
  - Some dimension tables in the snowflake schema are normalized
  - Normalization splits up the data into additional tables
- Normalization
  - Database normalization is the process of organizing the attributes and tables of a relational database to minimize data redundancy
  - Normalization involves decomposing a table into less redundant (and smaller) tables but without losing information



# Fact Constellation (Galaxy)

- Galaxy Schema:
  - Multiple fact tables
  - Multiple dimension tables



# OLAP Queries

- OLAP Query Workflow: Join → Filter → Group → Aggregate
  - **Join**: Connect the fact table with relevant dimension tables to enrich facts with descriptive details
  - **Filter**: Narrow down the data based on criteria from dimension attributes or fact table values
  - **Group**: Organize the data by one or more dimension attributes
  - **Aggregate**: Compute summary statistics for each group using aggregate functions

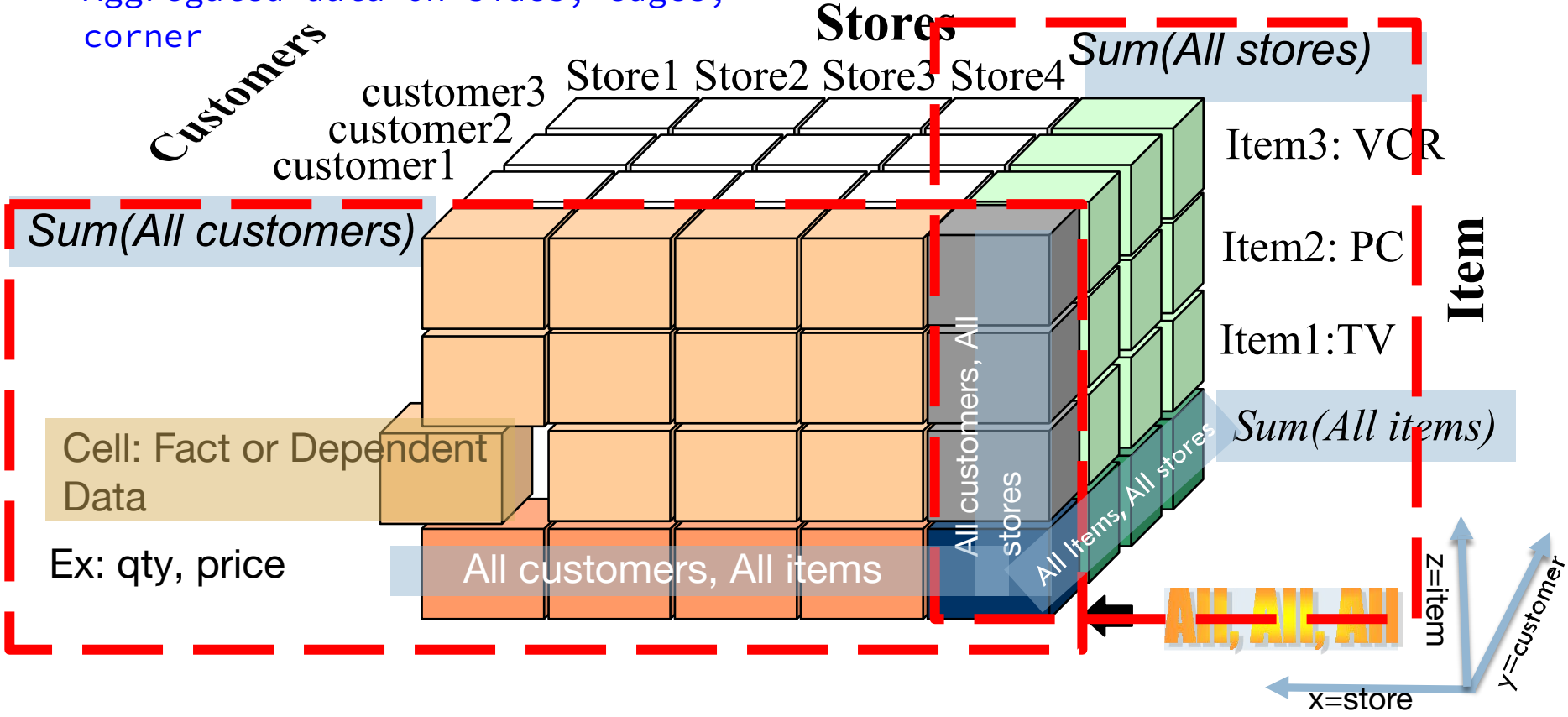
```
Sales(storeID, itemID, custID, qty, price)
Store(storeID, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
```

```
SELECT F.storeID, I.itemID, custID, sum(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID and F.itemID = I.itemID
and state = 'WA' and color = 'red'
GROUP BY F.storeID, I.itemID, custID;
```

# Data Cube (a.k.a Multidimensional OLAP)

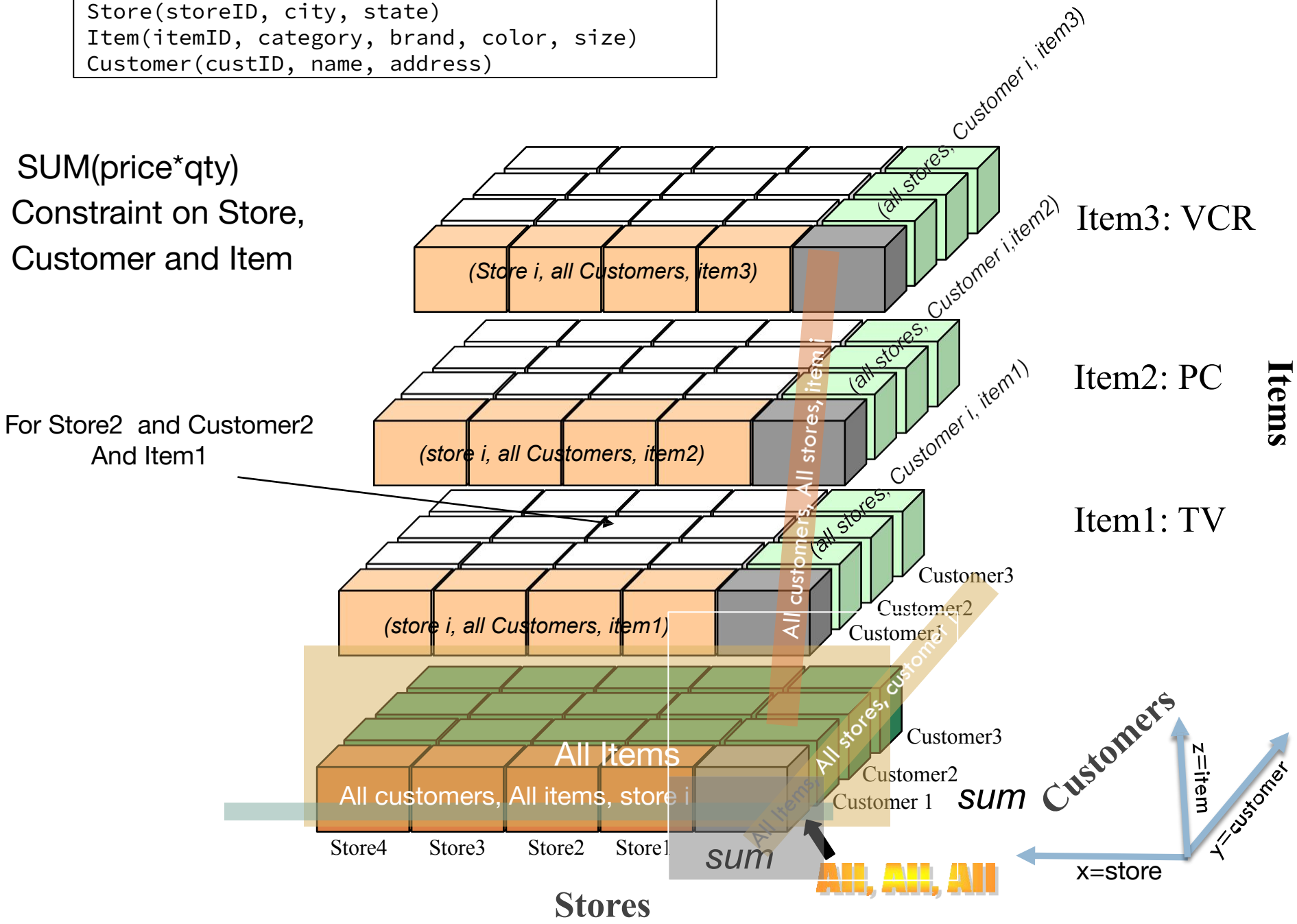
- Dimension data forms axes of "cube"
- Fact (dependent) data in cells
- Aggregated data on sides, edges, corner

```
Sales(storeID, itemID, custID, qty, price)
Store(storeID, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
```



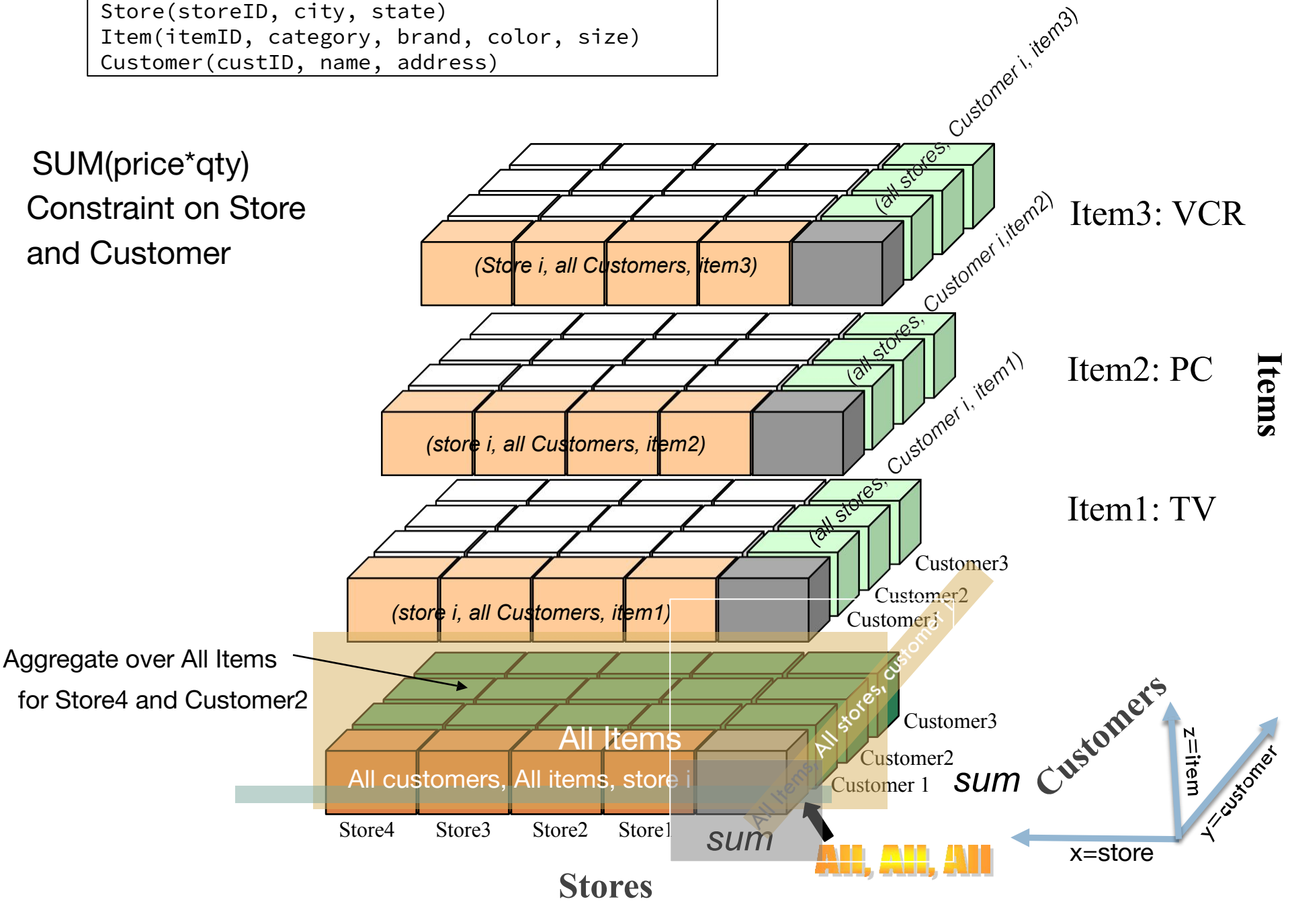
```
Sales(storeID, itemID, custID, qty, price)
Store(storeID, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
```

SUM(price*qty)
Constraint on Store, Customer and Item



```
Sales(storeID, itemID, custID, qty, price)
Store(storeID, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
```

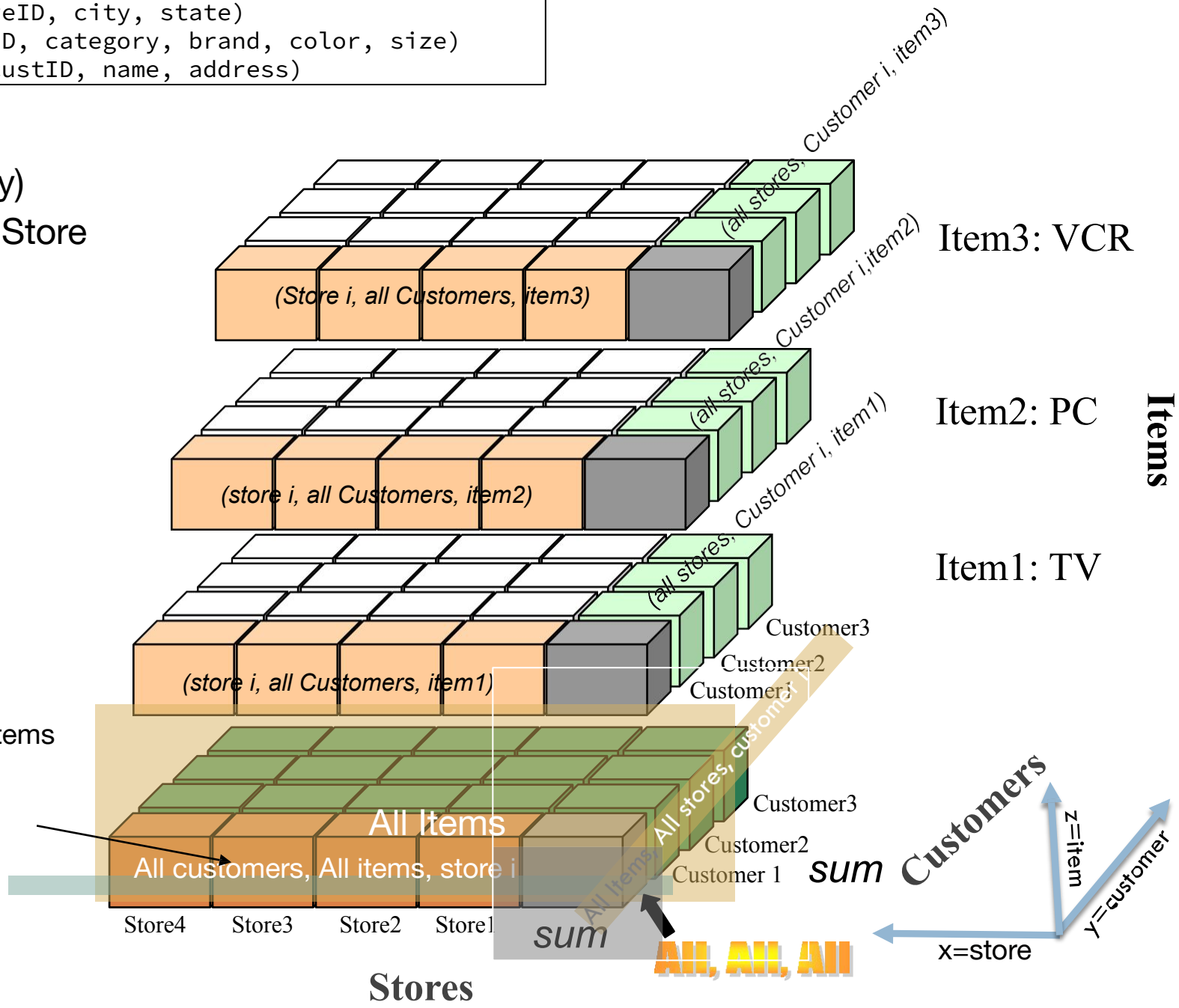
SUM(price*qty)
Constraint on Store and Customer



```
Sales(storeID, itemID, custID, qty, price)
Store(storeID, city, state)
Item(itemID, category, brand, color, size)
Customer(custID, name, address)
```

SUM(price*qty)

Constraint on Store



# Slice

Constraint on one dimension

Constraint on store dimension

Total sales by store, item, and customer for Washington stores only

```
SELECT F.storeID, itemID, custID,
SUM(price) FROM Sales F, Store S
WHERE F.storeID = S.storeID
AND state = ‘WA’
GROUP BY F.storeID, itemID, custID;
```

| storeID | itemID | custID | SUM(price) |
|---|---|---|---|
| store5 | item4 | cust1 | 180.0 |
| store5 | item4 | cust2 | 70.0 |
| store5 | item4 | cust4 | 170.0 |
| store5 | item5 | cust2 | 155.0 |
| store6 | item2 | cust3 | 95.0 |
| store6 | item2 | cust4 | 90.0 |
| store6 | item3 | cust4 | 165.0 |
| store6 | item4 | cust4 | 145.0 |
| store6 | item5 | cust3 | 175.0 |
| store6 | item5 | cust4 | 125.0 |

# Dice

Constraint on store dimension

---

Constraint on two or more dimensions

Total sales by store, item, and customer for Washington stores and red items only

Constraint on Item dimension

```
SELECT F.storeID, I.itemID, custID, sum(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID AND F.itemID = I.itemID
AND state = 'WA' AND color = 'red'
GROUP BY F.storeID, I.itemID, custID;
```

| | storeID | itemID | custID | sum(price) |
|---|---|---|---|---|
| 0 | store5 | item5 | cust2 | 155.0 |
| 1 | store6 | item3 | cust4 | 165.0 |
| 2 | store6 | item5 | cust3 | 175.0 |
| 3 | store6 | item5 | cust4 | 125.0 |

# Drill_down and Roll_up

Examining summary data, break out by dimension attribute

```
SELECT state, color, SUM(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID AND F.itemID = I.itemID
GTOUP BY state, color
```

| state | color | Sum(price) |
|---|---|---|
| CA | blue | 1365 |
| CA | red | 615 |
| WA | blue | 750 |
| WA | red | 620 |

```
SELECT state, color, category, SUM(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID AND F.itemID = I.itemID
GROUP BY state, color, category
```

| state | color | Category | Sum(price) |
|---|---|---|---|
| CA | blue | Tshirt | 135 |
| CA | blue | Jacket | 1230 |
| CA | red | Tshirt | 615 |
| WA | blue | Jacket | 750 |
| WA | red | Jacket | 455 |
| WA | red | Tshirt | 165 |

# Drill_down and Roll_up

- Examining data, summarize by dimension attribute

```
SELECT state, category, SUM(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID AND F.itemID = I.itemID
GROUP BY state, category
```

| state | category | Sum(price) |
|---|---|---|
| CA | Tshirt | 750 |
| CA | Jacket | 1230 |
| WA | Jacket | 1205 |
| WA | Tshirt | 165 |

```
SELECT category, SUM(price)
FROM Sales F, Store S, Item I
WHERE F.storeID = S.storeID AND F.itemID = I.itemID
GROUP BY category
```

| category | Sum(price) |
|---|---|
| Tshirt | 915 |
| Jacket | 2435 |

# SQL Construct WITH ROLLUP

---

With Rollup

```
SELECT dimension-attrs, aggregates FROM tables
WHERE conditions
GROUP BY dimension-attrs (c1, c2, c3 ) WITH ROLLUP
```

The ROLLUP assumes that there is the following hierarchy: c1 > c2 > c3
And it generates the following grouping sets:

- (c1, c2, c3)
- (c1, c2)
- (c1)
- ()

# Example of WITH ROLL UP

WITH ROLLUP on hierarchical grouping attributes

```
SELECT state, county, city, SUM(price)
FROM Sales F, Store S
WHERE F.storeID = S.storeID
GROUP BY state, county, city;
```

| state | county | city | sum(price) |
|---|---|---|---|
| CA | Santa Clara | Palo Alto | 325 |
| CA | Santa Clara | Mountain View | 805 |
| CA | San Mateo | Menlo Park | 625 |
| CA | San Mateo | Belmont | 225 |
| WA | King | Seattle | 575 |
| WA | King | Redmond | 795 |

```
SELECT state, county, city, SUM(price)
FROM Sales F, Store S
WHERE F.storeID = S.storeID
GROUP BY state, county, city WITH ROLLUP;
```

| state | county | city | sum(price) |
|---|---|---|---|
| CA | San Mateo | Belmont | 225 |
| CA | San Mateo | Menlo Park | 625 |
| CA | San Mateo | NULL | 850 |
| CA | Santa Clara | Mountain View | 805 |
| CA | Santa Clara | Palo Alto | 325 |
| CA | Santa Clara | NULL | 1130 |
| CA | NULL | NULL | 1980 |
| WA | King | Redmond | 795 |
| WA | King | Seattle | 575 |
| WA | King | NULL | 1370 |
| WA | NULL | NULL | 1370 |
| NULL | NULL | NULL | 3350 |

Sales in **All** cities in state Ca County San Mateo

Sales in **All** cities in sate Ca County Santa Clara

Sales in **All** cities and **All** Counties in sate Ca

Sales in **All** cities and **All** Counties in **All** states