# 10 – DATA WAREHOUSING

**CS 1656**

Introduction to Data Science

Alexandros Labrinidis – http://labrinidis.cs.pitt.edu
University of Pittsburgh

# What is a Data Warehouse?

- a system used for reporting and data analysis, integrating data from one or more disparate sources and creating a central repository of data, a data warehouse (DW).
  - stores current and historical data and is used for creating trending reports for senior management reporting, such as annual and quarterly comparisons.
  - the data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, etc.)

[ Source: http://en.wikipedia.org/wiki/Data_warehouse ]

- "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

[ Source: Data Mining Concepts and Techniques, 3rd Edition]

# 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, records from online transaction processing (OLTP) systems

- **Data cleaning** and **data integration** techniques 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"

# Data Warehouse: Non-Volatile

- A **physically separate store** of data transformed from the operational environment

- Operational **updates of data do 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*

# OLTP vs OLAP

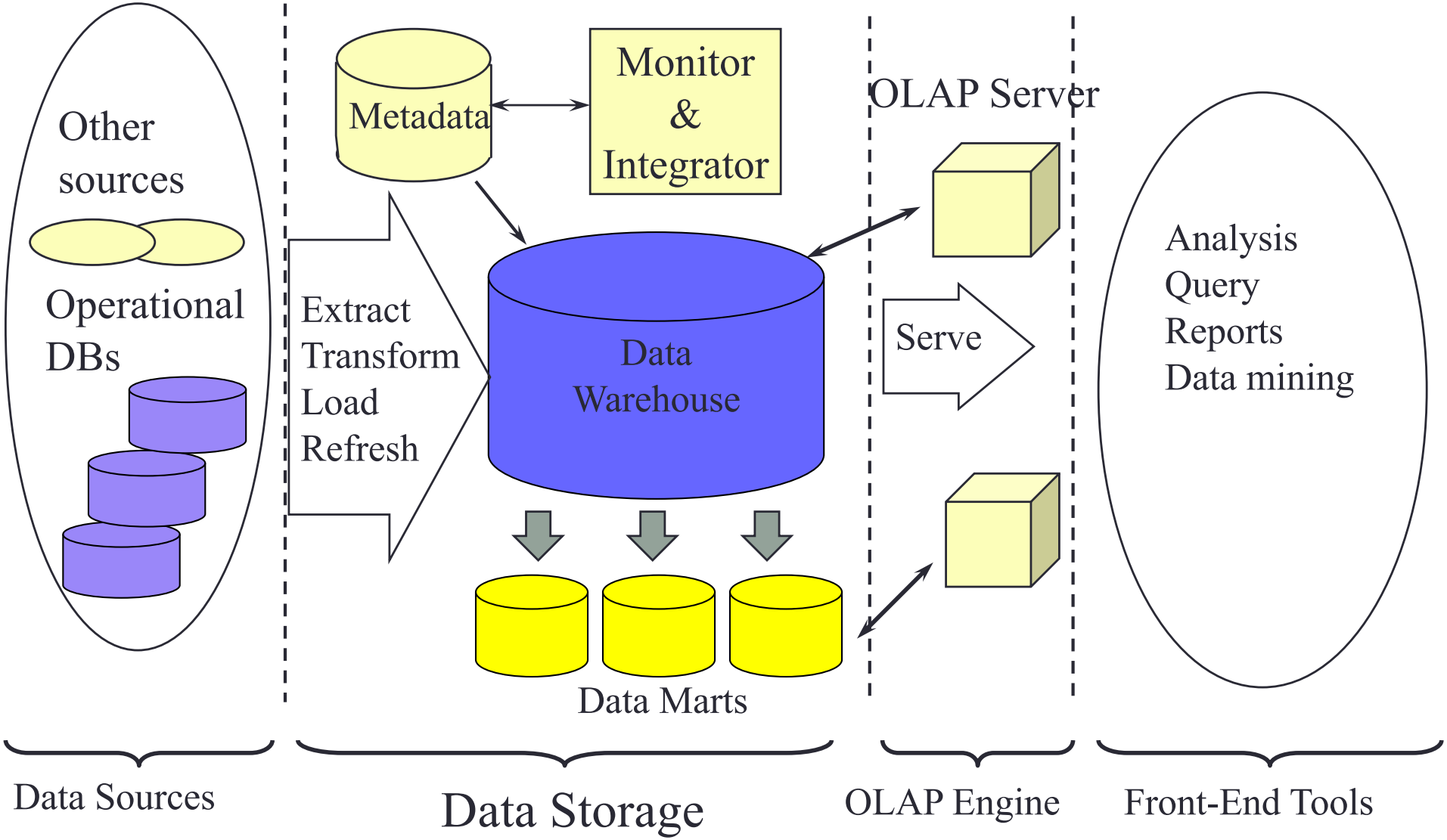
| | OLTP | OLAP |
|---|---|---|
| **users** | clerk, IT professional | knowledge worker |
| **function** | day to day operations | decision support |
| **DB design** | application-oriented | subject-oriented |
| **data** | current, up-to-date<br>detailed, flat relational<br>isolated | historical,<br>summarized, multidimensional<br>integrated, consolidated |
| **usage** | repetitive | ad-hoc |
| **access** | read/write<br>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 |

[ Source: Data Mining Concepts and Techniques, 3rd Edition]

# Why a separate data warehouse?

- High performance for both systems
  - **DBMS**— tuned for **OLTP** (Online Transaction Processing): access methods, indexing, concurrency control, recovery
  - **Warehouse**—tuned for **OLAP** (Online Analytical Processing): 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: DW 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



# Data Lakes

just-go-greece.com

# Data lake vs Data Warehouse

- **Amount of data**
  - Data lake stores ALL or at least most application data
  - Data warehouse stores only data relevant to its purpose
- **Type of data**
  - Data lake stores raw, unprocessed data
  - Data warehouse stores data in specified processed format
- **Purpose**
  - Data lake data can have broad range of applications(data mining, ML, future usage)
  - Data warehouse usually has specific purpose (BI, decision support)
- **Performance**
  - Data lake is fast to store, takes enormous amount of storage
  - Data warehouse has preprocessing, data consume less space

# Extraction, Transformation, 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 indices and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

# SHOW ME THE DATA!

# So, what does the data look like?

http://data.cs1656.org/coffee-chain.xlsx

[Source: http://www.tableausoftware.com]

- **Dimension Attributes:**
  - Date
  - Location (area code, market, state)
  - Attributes about customers (market size)
  - Attributes about products (product, product line, product type, type)

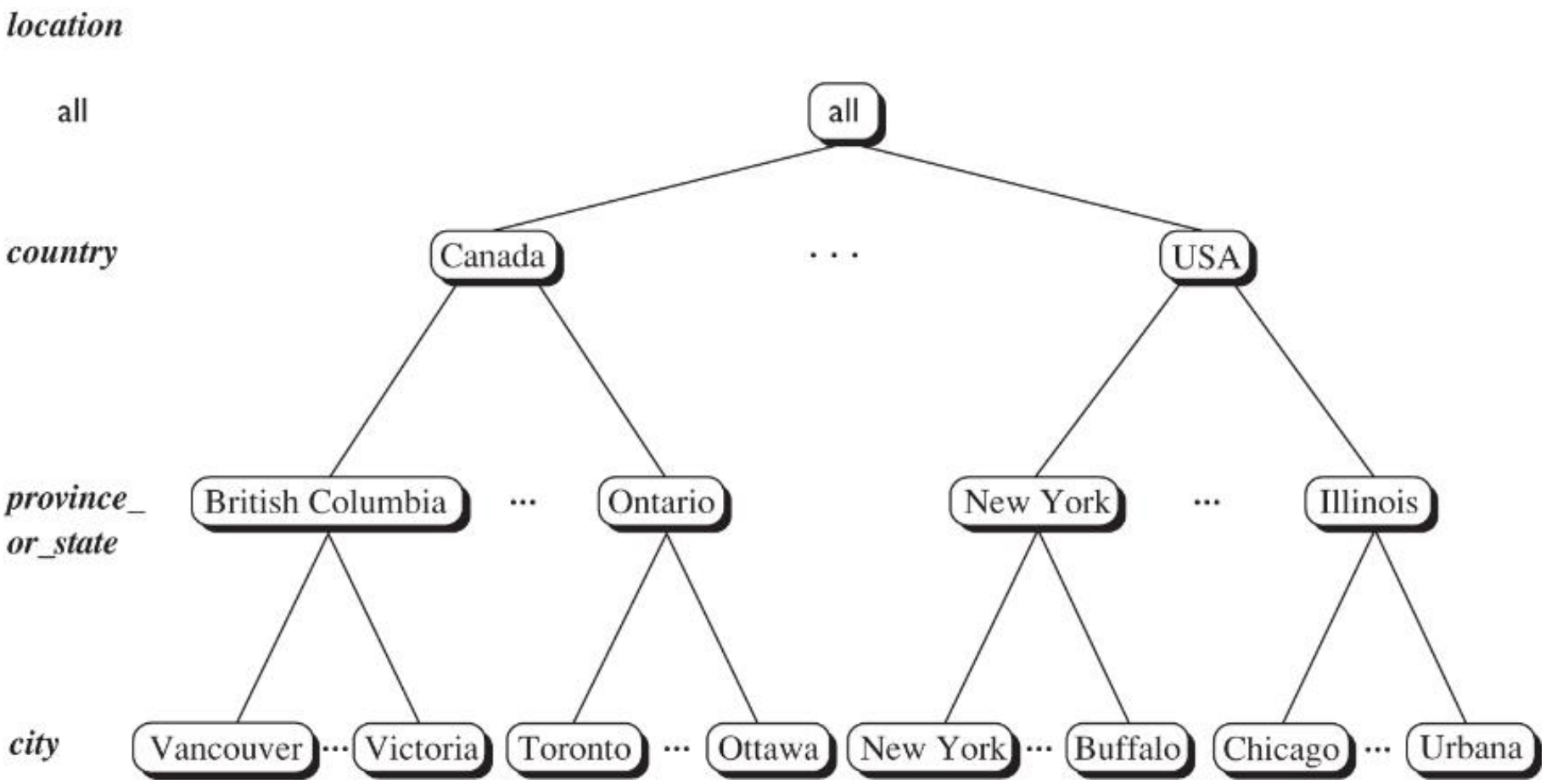
- **Measure Attributes (can be aggregated upon):**
  - Inventory
  - Cost
  - Sales
  - Profit

# Dimension Attributes

- Some dimension attributes could be organized using a **concept hierarchy**
  - i.e., sequence of mappings from low-level concepts to higher-level concepts

- Concept hierarchy examples:
  - Location:
    - City → County → Province or State → Country → All
  - Time:
    - Second → Minute → Hour → **Day** → Month → Quarter → Year → All
    - Second → Minute → Hour → **Day** → Week → Year → All

# Location Hierarchy Example



# Aggregation Function Types

- 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 sub-aggregate.
  - E.g., median(), mode(), rank()

# HOW TO SUMMARIZE

# Crosstab Definition

- A cross-tab is a table where
  - values for one of the dimension attributes form the row headers
  - values for another dimension attribute form the column headers
  - other dimension attributes are listed on top
  - values in individual cells are (aggregates of) the values of the dimension attributes that specify the cell.
  - totals for every row and column are also pre-computed

# Crosstab Example

*size*: **all**

*color*

| *item-name* | dark | pastel | white | Total |
|---|---|---|---|---|
| skirt | 8 | 35 | 10 | 53 |
| dress | 20 | 10 | 5 | 35 |
| shirt | 14 | 7 | 28 | 49 |
| pant | 20 | 2 | 5 | 27 |
| Total | 62 | 54 | 48 | 164 |

# Understanding Question / Table

- Fill out the blanks in the following crosstab:

| | June | July | August | September | All months |
|---|---|---|---|---|---|
| **Blue** | 10 | | | 15 | 45 |
| **Red** | | 15 | | 20 | 65 |
| **Green** | | 5 | 5 | | |
| **All colors** | 30 | 30 | 30 | 45 | 135 |

# Understanding Question / Q1

- **Question**: What is the correct value for the total number of sales of RED cars for the month of AUGUST?

- **Possible Answers**:
  - Fill-in

# Understanding Question / Q1 ANS

- **Question**: What is the correct value for the total number of sales of RED cars for the month of AUGUST?

- **Correct Answer**:
  - 15

| | June | July | August | September | All months |
|---|---|---|---|---|---|
| **Blue** | 10 | | | 15 | 45 |
| **Red** | | 15 | | 20 | 65 |
| **Green** | | 5 | 5 | | |
| **All colors** | 30 | 30 | 30 | 45 | 135 |

# Understanding Question / Q1 ANS

- **Question**: What is the correct value for the total number of sales of RED cars for the month of AUGUST?

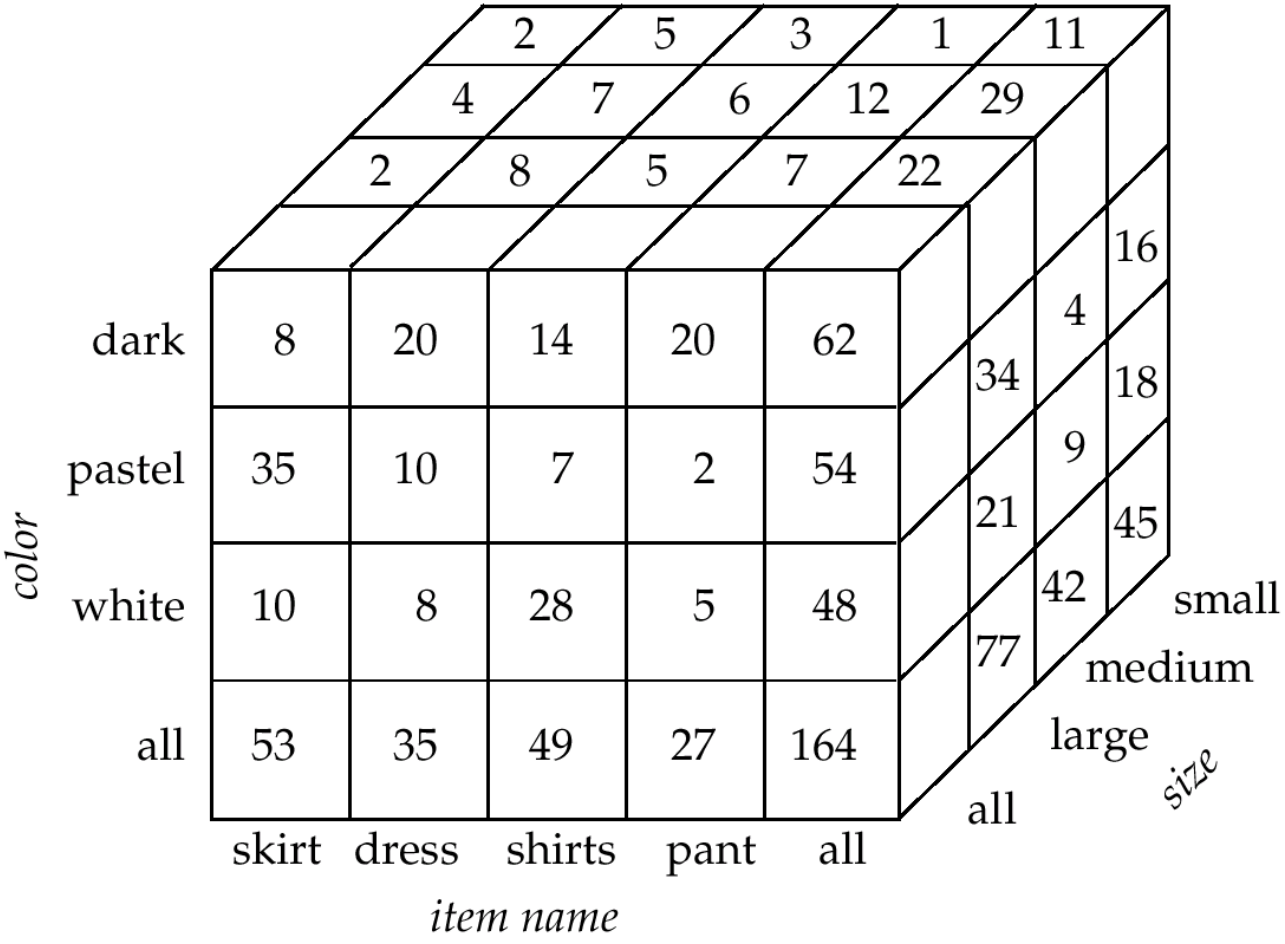
- **Correct Answer**:
  - 15

| | June | July | August | September | All months |
|---|---|---|---|---|---|
| **Blue** | 10 | 10 | 10 | 15 | 45 |
| **Red** | 15 | 15 | **15** | 20 | 65 |
| **Green** | 5 | 5 | 5 | 10 | 25 |
| **All colors** | 30 | 30 | 30 | 45 | 135 |

# Generalization of Crosstab for multiple dimensions

- Generalization of cross-tab for more than two dimensions is a data cube
- 3-dimensional data cubes are “easy” to visualize
- Crosstab can be used as two-dimensional views of any n-dimensional data cube

# Data Cube Example




- Axes represent different **dimension** attributes
  - E.g., color, size, item name
- Cells hold one **measure** attribute
  - E.g., total sales
  - Called facts
- Real data cubes have >> 3 dimensions

# OLAP OPERATIONS

# Typical OLAP Operations

- **Roll-up** – aggregate data further
  - Eliminate a dimension (e.g., eliminate color) OR
  - Climb up a concept hierarchy (e.g., go from months to quarters)
- **Drill-down** – reverse of roll-up – provide more details
  - Introduce additional dimensions OR
  - Climb down a concept hierarchy (e.g., from county to city)
- Slice
  - Select on one dimension of data cube ➔ subcube
- Dice
  - Select on two or more dimension of data cube ➔ subcube
- Pivot (rotate)
  - Rotate data axes to provide alternate visual representation

# Roll-up Example

Number of Autos Sold

| | PA | OH | MD | Total |
|---|---|---|---|---|
| **Jul** | 45 | 33 | 30 | 108 |
| **Aug** | 50 | 36 | 42 | 128 |
| **Sep** | 38 | 31 | 40 | 109 |
| **Total** | 133 | 100 | 112 | 345 |

Roll up by Month

Number of Autos Sold

| PA | OH | MD | Total |
|---|---|---|---|
| 133 | 100 | 112 | 345 |

Roll up by State

| Jul | Aug | Sep | Total |
|---|---|---|---|
| 108 | 128 | 109 | 345 |

# Drill-Down Example 1

Number of Autos Sold

| | PA | OH | MD | Total |
|---|---|---|---|---|
| **Jul** | 45 | 33 | 30 | 108 |
| **Aug** | 50 | 36 | 42 | 128 |
| **Sep** | 38 | 31 | 40 | 109 |
| **Total** | 133 | 100 | 112 | 345 |

Drill Down by Month

Number of Autos Sold

| PA | OH | MD | Total |
|---|---|---|---|
| 133 | 100 | 112 | 345 |

Drill down by State

| Jul | Aug | Sep | Total |
|---|---|---|---|
| 108 | 128 | 109 | 345 |

# Drill-Down Example 2

Number of Autos Sold

| | PA | OH | MD | Total |
|---|---|---|---|---|
| **Jul** | 45 | 33 | 30 | 108 |
| **Aug** | 50 | 36 | 42 | 128 |
| **Sep** | 38 | 31 | 40 | 109 |
| **Total** | 133 | 100 | 112 | 345 |

Roll up by Month

Number of Autos Sold

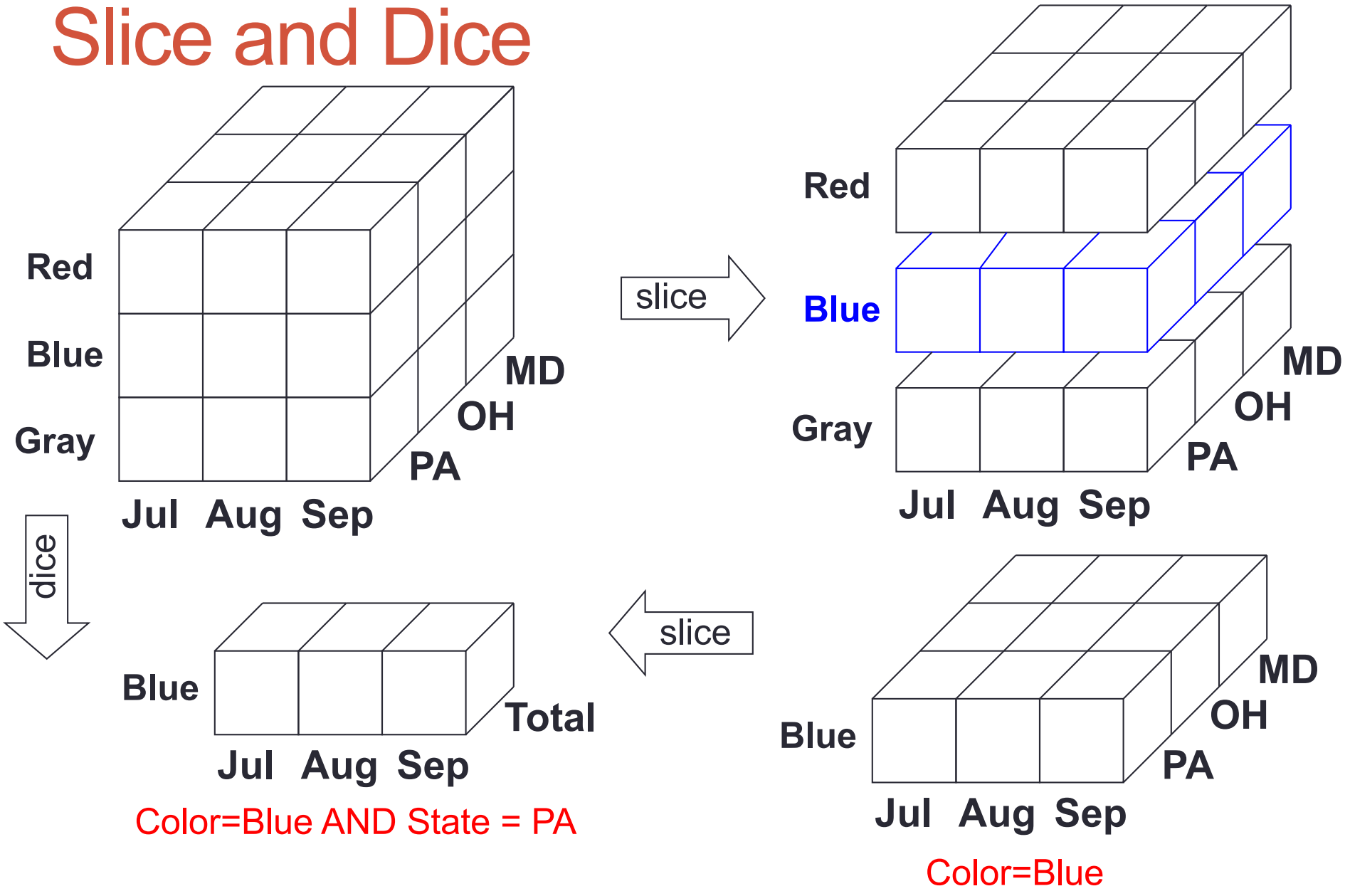
| PA | OH | MD | Total |
|---|---|---|---|
| 133 | 100 | 112 | 345 |

Drill down by Color

Number of Autos Sold

| | PA | OH | MD | Total |
|---|---|---|---|---|
| **Red** | 40 | 29 | 40 | 109 |
| **Blue** | 45 | 31 | 37 | 113 |
| **Gray** | 48 | 40 | 35 | 123 |
| **Total** | 133 | 100 | 112 | 345 |

# Slice and Dice



# Understanding Questions / Q2-Q4

- **Question**: Assume the data cube from the handout and that we perform a **roll-up operation on item-name** and then a **slice operation with the condition size=medium**. Which of the following questions can be answered with the resulting sub-cube?

- **Possible Answers**:
  - Yes/No – Number of dresses sold in medium size
  - Yes/No – Number of clothes items sold in medium size that are pastel in color
  - Yes/No – Number of clothes items sold that are large size