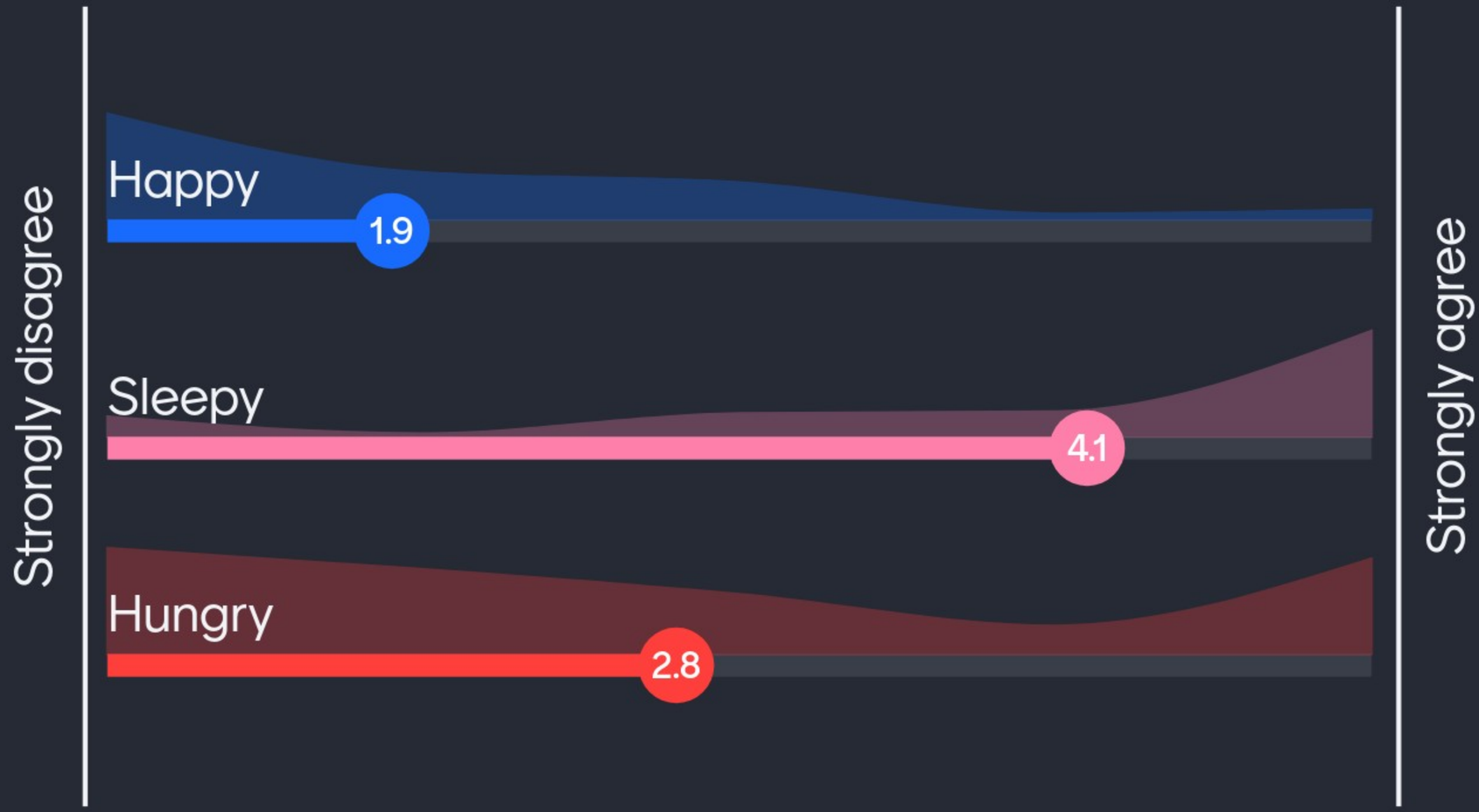
# How are you today?

Mentimeter




111

Mentimeter

# Data Warehouse Concepts 1

STADVDB

Mentimeter

# Outline

- Data Warehouse Overview
- Transactional/Operational vs

  Analytical DB (Data Warehouse)
- Integrating Heterogeneous Databases
- Data Cube
- OLAP Operations

Mentimeter

# How do Organizations gain *Competitive Advantage*?

| Before | Now |
| --- | --- |
| • Efficient and Cost-Effective Customer Service<br>• Systems that automate *Business Processes*<br>• Large volume of *Transactional data*<br>• Stored in *Operational Databases* | • Use Operational Data to support *Decision-Making* Activities<br>• Turn archived data into a Source of Knowledge<br>• Input to *Decision-Support Systems*<br>• Knowledge is Power |

Mentimeter

*But...*

Operational data
may come from
multiple Transactional DB!

Mentimeter

# Data Warehouse

- A subject-oriented, integrated, time-variant, and non-volatile collection of data to help analysts or management to make informed decisions in an organization (Inmon, 1990)
  - Organized around major subjects of the enterprise (e.g., product, customers, sales) rather than the operations (e.g., invoicing, stock control, product sales)
  - Constructed by integrating data from multiple heterogeneous sources (TutorialsPoint.com)
  - Collected data is identified with a particular historical time period
  - Previous data is not erased when new data is added
- Supports analytical reporting, structured and/or adhoc queries and decision-making (TutorialsPoint.com)

Mentimeter

# Types of DW Applications

- Information Processing through traditional querying, basic statistical analysis, reporting using crosstabs, tables, charts or graphs
- Analytical Processing through basic OLAP operations, e.g., slice-and-dice, drill down, drill up and pivoting
- Data Mining or knowledge discovery to find hidden patterns and associations, constructing analytical models, performing classification and prediction
- Example usage:
  - Tuning strategies, such as reposition products or brands
  - Analysis of customer buying preferences
  - Analysis of business operations

http://www.tutorialspoint.com/dwh/dwh_overview.htm

Mentimeter


# Data Mining Classic Case: Diapers & Beers

- study done by Walmart
- men who buy diapers also buy beer
- relocated beers next to diapers
- sales of beers and diapers increased significantly

6

Mentimeter

# Operational DB vs Data Warehouse

| Operational Database | Data Warehouse |
|---|---|
| **Day-to-day** transaction processing | **Historical** analytical processing |
| Used by **operational users** (clerks, DBAs, DB professionals) | Used by **knowledge workers** (analysts, managers, executives) |
| Used to **run** the business | Used to **analyze** the business |
| Narrow, planned and **simple updates and queries** | Broad, adhoc, complex queries and analysis |
| Focuses on **Data In** (read, modify, retrieve) | Focuses on **Information Out** (read only) |
| Based on **Entity Relationship and Relational** Models | Based on **Star, Snowflake and Constellation Schema** |
| Primitive and highly detailed, flat relational view of data | Summarized and consolidated, multidimensional view data |
| DB size: 100MB to 100GB | DB size: 100GB to 100TB |
| Number of users: thousands | Number of users: hundreds |

http://www.tutorialspoint.com/dwh/dwh_overview.htm

Mentimeter

# Integrated Data

- Data Warehouse integrates an organization's application-oriented data from different sources
  - Data are coming from heterogeneous database sources
  - Data may be represented in different format
- A unified view of the heterogeneous data must be presented to the user



Connolly & Begg, 2015

Mentimeter

# Integrating Heterogeneous DB

- Query-driven Approach
  - Wrappers and integrators are built on top of multiple DB
  - During a query, the metadata dictionary is used to translate the query into appropriate form for individual heterogeneous DB
  - Translated queries are mapped and sent to local query processors
  - Results from heterogeneous sites are integrated into global answer set
  - Disadvantages
    - Needs complex integration and filtering processes
    - Inefficient
    - Very expensive for frequent queries
    - Expensive for queries that require aggregations

http://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm

Mentimeter

# Integrating Heterogeneous DB

- Update-driven Approach
  - Information from multiple heterogeneous sources are integrated and stored in a warehouse
  - Information is available for direct querying and analysis
  - Advantages
    - Provide high performance
    - Data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance
    - Query processing does not require an interface to process data at local sources



http://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm

Mentimeter

# Update-Driven Approach
## Design Issue: *When to Gather Data*

- Source Driven Architecture
  - Data sources transmit new information to warehouse, either continuously or periodically
- Destination Driven Architecture
  - Warehouse periodically requests new information from data sources

Keeping warehouse exactly synchronized with data sources is too expensive



Silberschatz, Korth & Sudarshan, 2010

Mentimeter

# Data Cube

- Used to represent data in multiple dimensions
- Defined by dimensions and facts
- Dimensions
  - The entities with respect to which an enterprise preserves the records

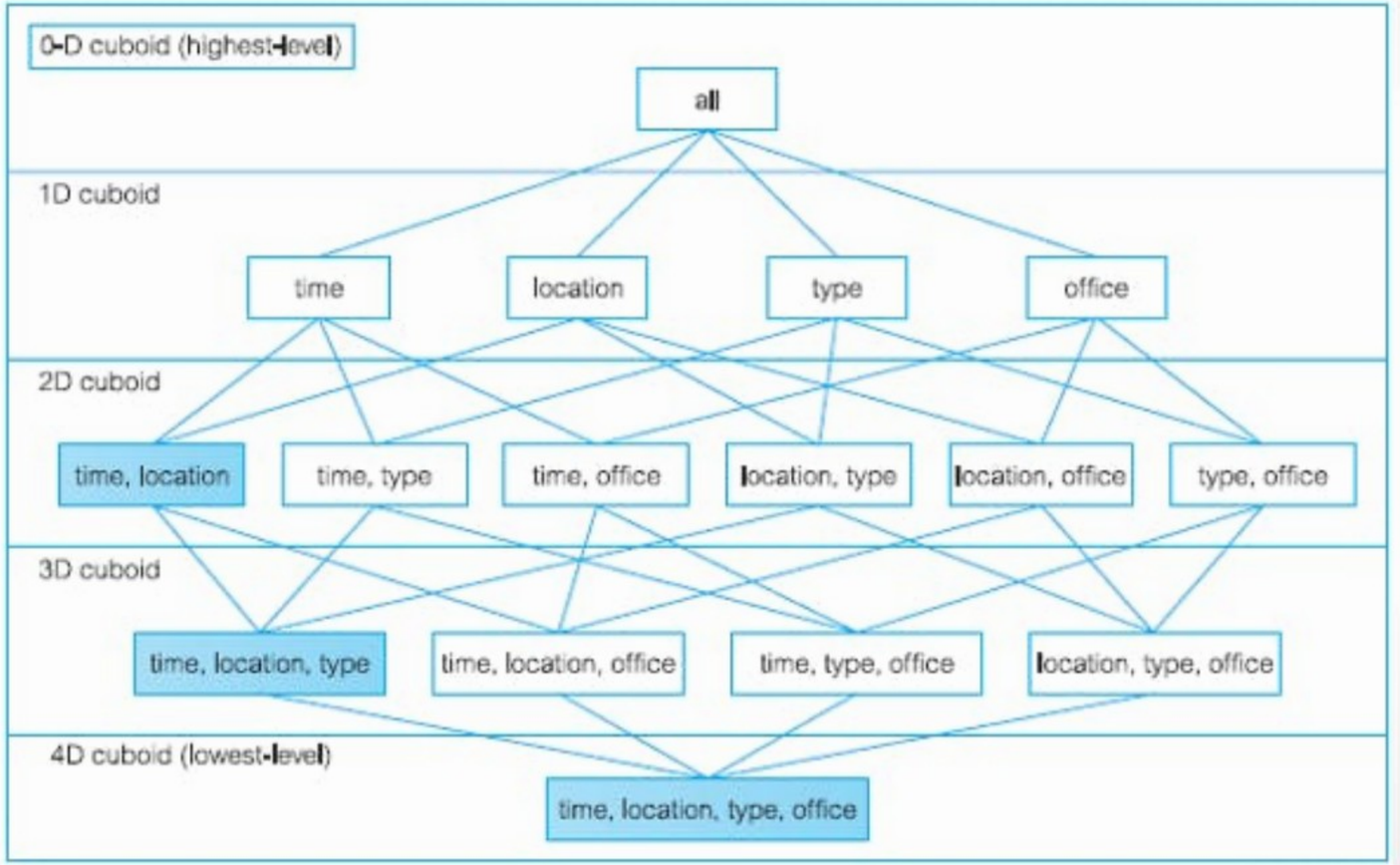


http://www.tutorialspoint.com/dwh/dwh_terminologies.htm

3

Mentimeter

# Data Cube as Lattice of Cuboids



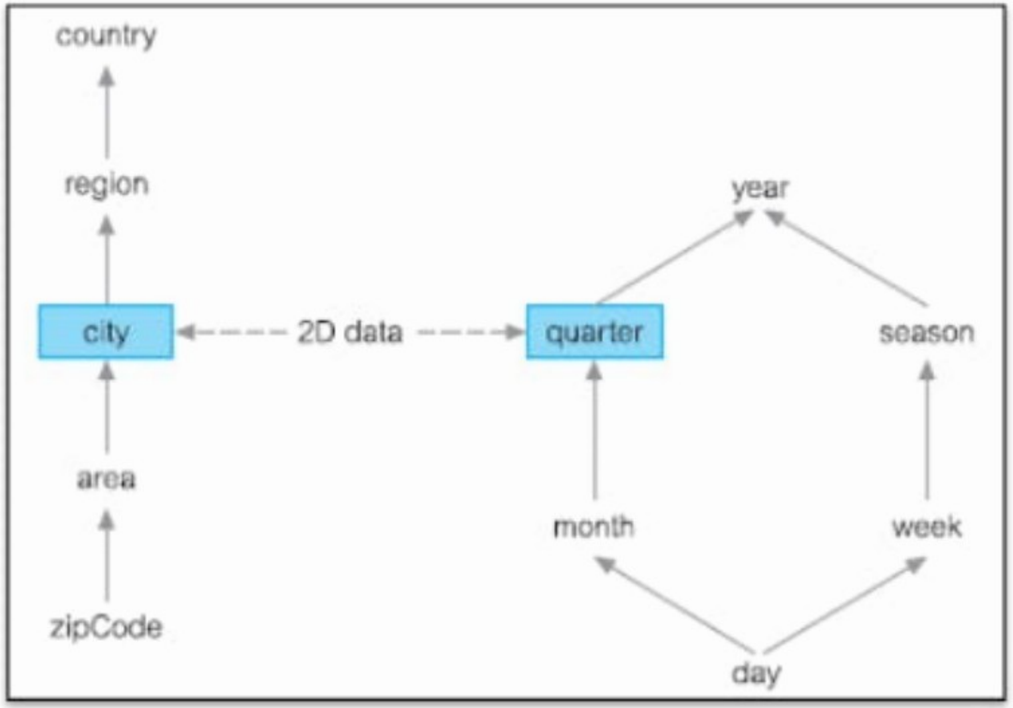

Connolly & Begg, 2015

1

Mentimeter

# Dimensional Hierarchy

- Defines mappings from a set of lower-level concepts to higher-level concepts




Connolly & Begg, 2015

Mentimeter

# Data Cube - Example

- A company wants to keep track of sales records using a sales DW with respect to the following dimensions – *time, item, branch* and *location* – to keep track of *monthly sales and branch where the items were sold*
- Sample 2D view of Sales Data with respect to *time, item* and *location*

| Location="New Delhi" | | | | |
|---|---|---|---|---|
| Time(quarter) | Item(type) | | | |
| | Entertainment | Keyboard | Mobile | Locks |
| Q1 | 500 | 700 | 10 | 300 |
| Q2 | 769 | 765 | 30 | 476 |
| Q3 | 987 | 489 | 18 | 659 |
| Q4 | 666 | 976 | 40 | 539 |

http://www.tutorialspoint.com/dwh/dwh_terminologies.htm



47

Mentimeter

# Data Cube - Example

- Sample 3D view of Sales Data with respect to *time, item, branch* and *location*

| Time | Location="Gurgaon" | | | Location="New Delhi" | | | Location="Mumbai" | | |
|---|---|---|---|---|---|---|---|---|---|
| | Item | | | Item | | | Item | | |
| | Mouse | Mobile | Modem | Mouse | Mobile | Modem | Mouse | Mobile | Modem |
| Q1 | 788 | 987 | 765 | 786 | 85 | 987 | 986 | 567 | 875 |
| Q2 | 678 | 654 | 987 | 659 | 786 | 436 | 980 | 876 | 908 |
| Q3 | 899 | 875 | 190 | 983 | 909 | 237 | 987 | 100 | 1089 |
| Q4 | 787 | 969 | 908 | 537 | 567 | 836 | 837 | 926 | 987 |

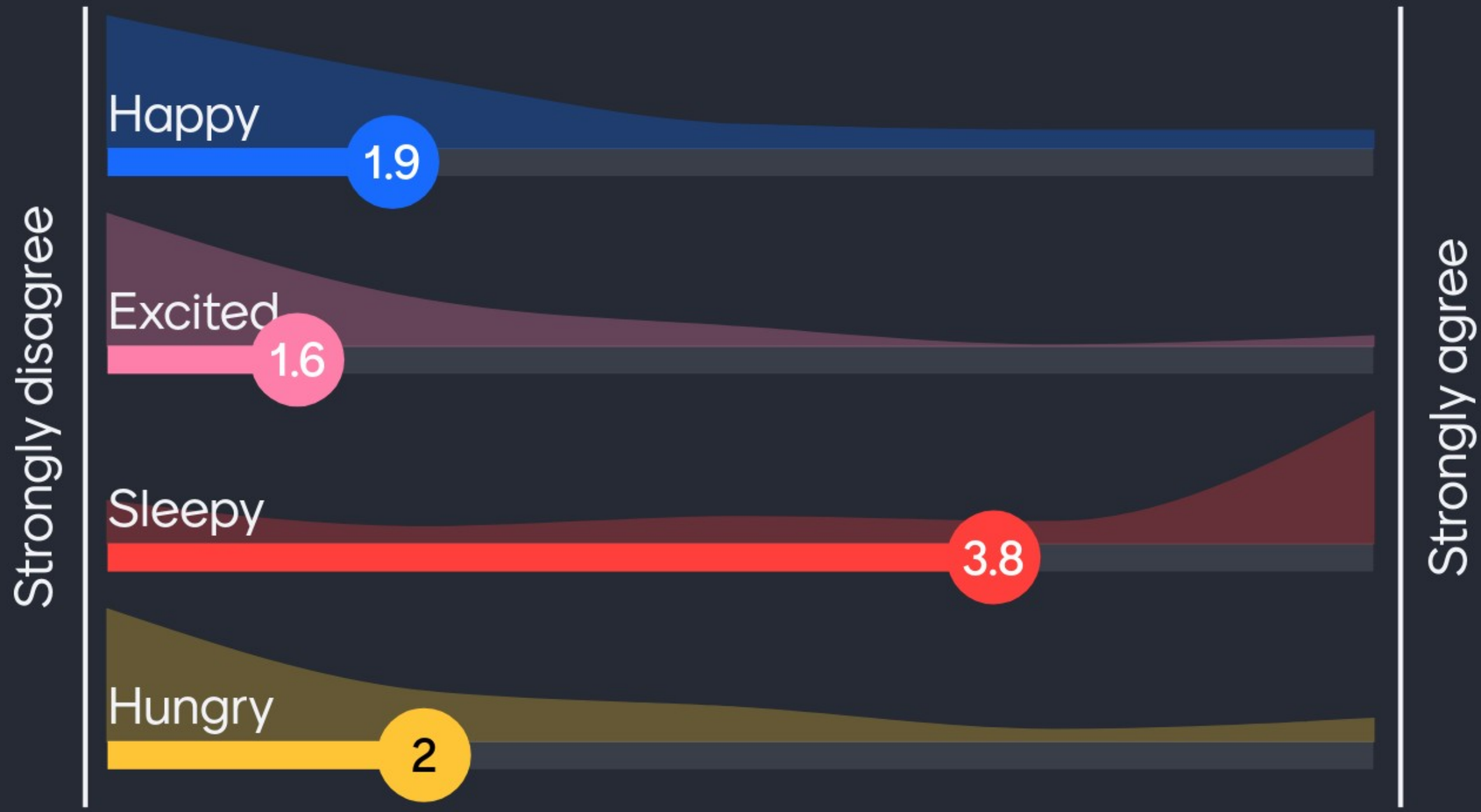
http://www.tutorialspoint.com/dwh/dwh_terminologies.htm



29

# How you doin'?

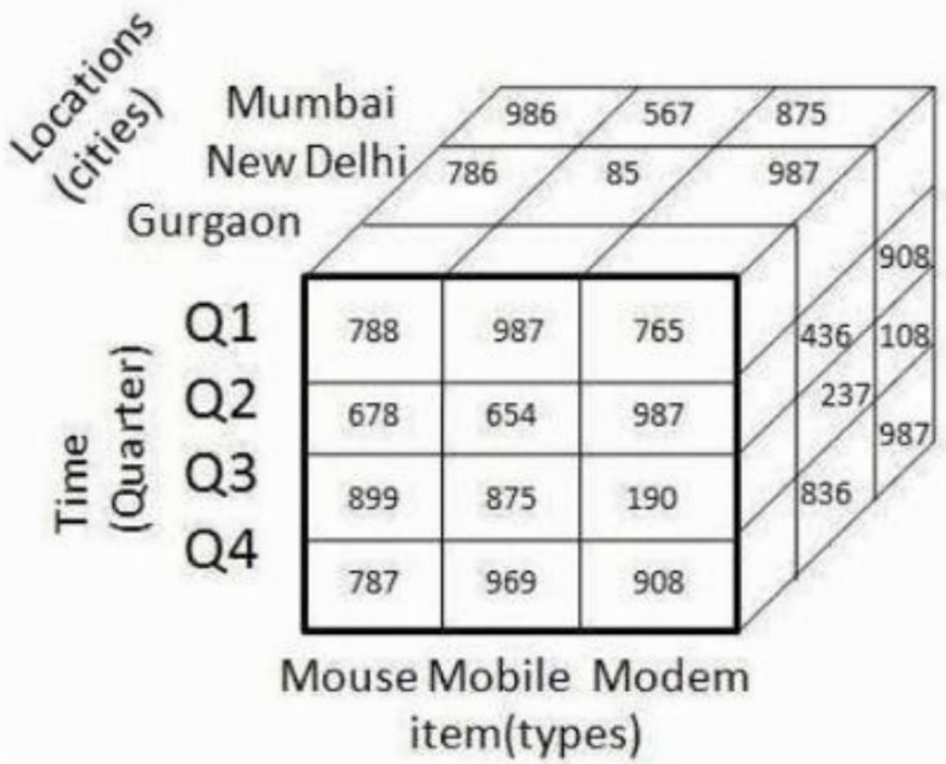
Mentimeter




47

# Data Cube - Example

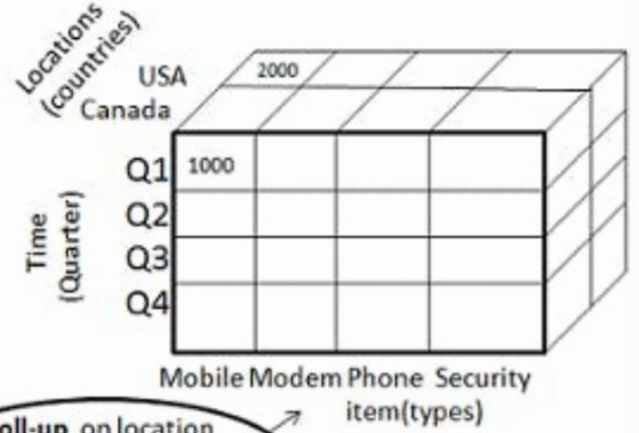
- Sample 3D cube view of Sales Data with respect to *time, item, branch* and *location*



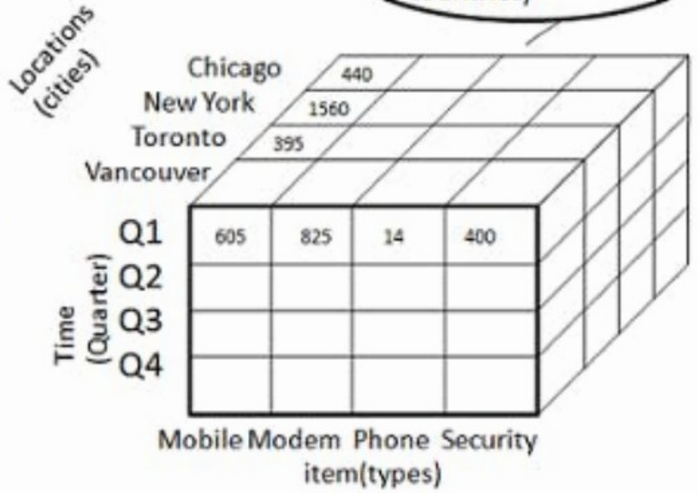

http://www.tutorialspoint.com/dwh/dwh_terminologies.htm

# OLAP Operations

- **Roll-up** – performs aggregation on a data cube by:
  - Climbing up a concept hierarchy for a dimension
  - Reducing >= 1 dimension(s)




roll-up on location (from cities to countries)



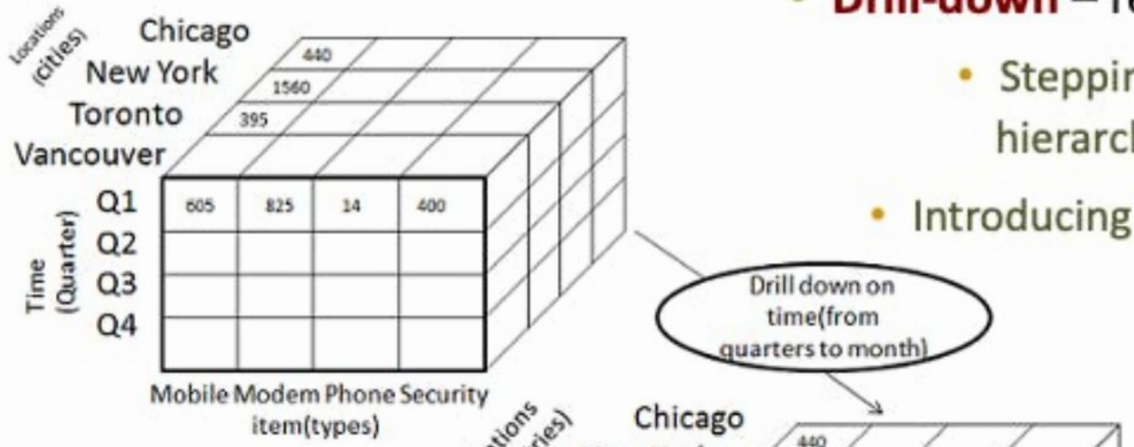

**On Roll-up:**
**Transcend the *location* hierarchy to group data from Cities to Countries**

**Concept Hierarchy for the dimension *Location*:**
**Street < City < Province < Country**

http://www.tutorialspoint.com/dwh/dwh_olap.htm

# OLAP Operations

- **Drill-down** – reverse of roll-up:
  - Stepping down a concept hierarchy for a dimension
  - Introducing >= 1 dimension(s)




Drill down on time(from quarters to month)


**Concept Hierarchy for the dimension *Time*:**

**Day < Month < Quarter < Year**

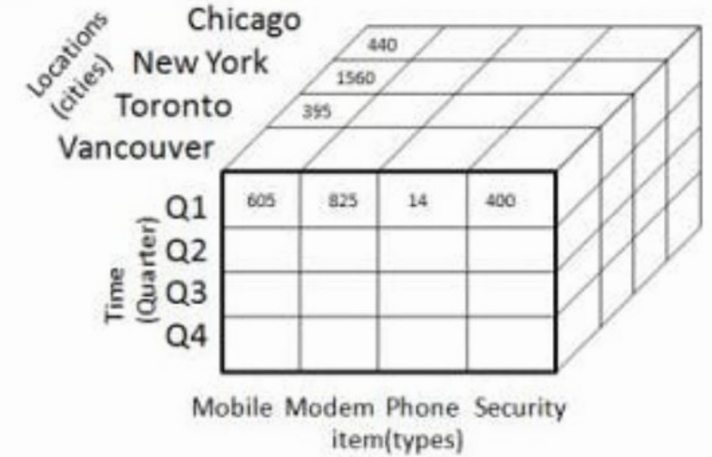
**On Drill-down:**

**Descend the *time* hierarchy to provide highly detailed data from Quarter to Month**

http://www.tutorialspoint.com/dwh/dwh_olap.htm

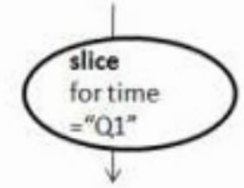
# OLAP Operations

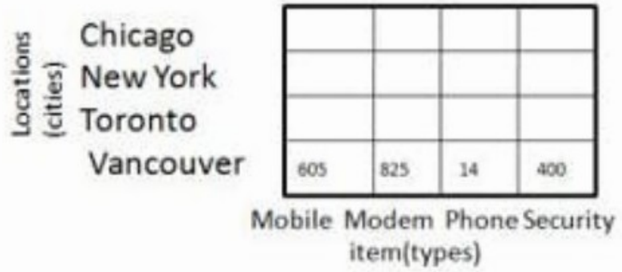
- **Slice** – selects one particular dimension from a given cube and provides a new sub-cube




**Slice is performed for the *Time* dimension using the criterion**

**time = "Q1"**

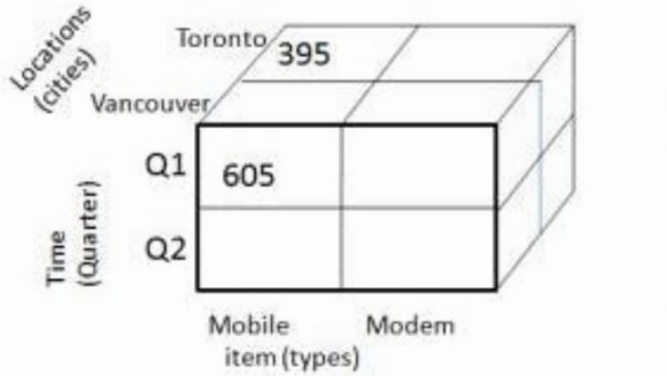



http://www.tutorialspoint.com/dwh/dwh_olap.htm

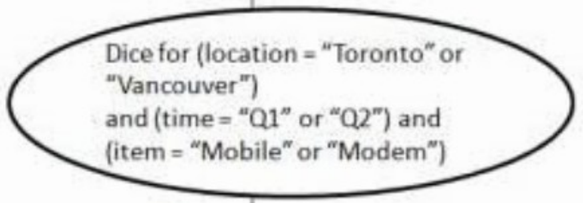
# OLAP Operations

- **Dice** – selects two or more dimensions from a given cube and provides a new sub-cube




Dice is performed for the dimensions *Time*, *Location* and *Item* using the criteria

location = "Toronto" or "Vancover"
time = "Q1" or "Q2"
item = "Mobile" or "Modem"



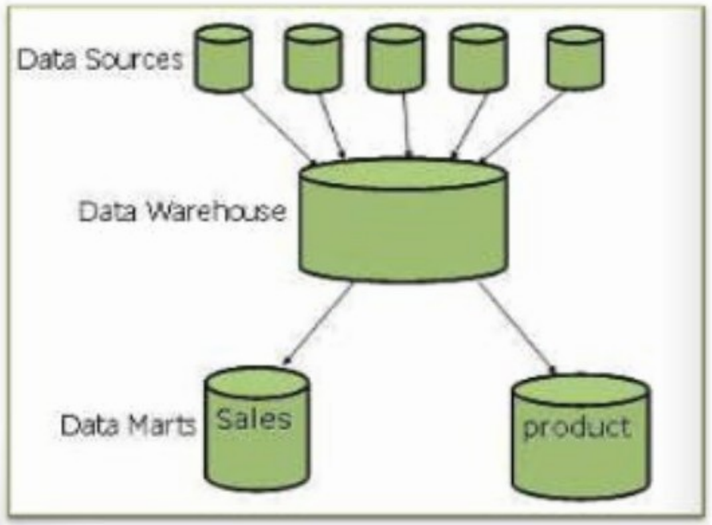


http://www.tutorialspoint.com/dwh/dwh_olap.htm

Mentimeter

# Data Mart

- A database containing a subset of organization-wide data that is valuable to a specific group of people in the organization
- Supports the analytical requirements of a particular business unit (such as the Sales department), or
- Supports users who share the same requirements to analyze a particular business process (such as property sales)




- Benefits
  - Implemented on low-cost servers
  - Small in size → improves end-user response time
  - Customized by department
  - Easier to build



Connolly & Begg, 2015

**TABLE 9-2 Data Warehouse Versus Data Mart**

| Data Warehouse | Data Mart |
|---|---|
| **Scope** | **Scope** |
| • Application independent<br>• Centralized, possibly enterprise-wide<br>• Planned | • Specific DSS application<br>• Decentralized by user area<br>• Organic, possibly not planned |
| **Data** | **Data** |
| • Historical, detailed, and summarized<br>• Lightly denormalized | • Some history, detailed, and summarized<br>• 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<br>• Data oriented<br>• Long life<br>• Large<br>• Single complex structure | • Restrictive<br>• Project oriented<br>• Short life<br>• Start small, becomes large<br>• Multi, semi-complex structures, together complex |

Hoffer, Ramesh & Topi, 2012



23

Mentimeter
5

# References


**Chapter 32: Data Warehouse Design**

Connolly, T. & Begg, C. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management, 6th Edition*. Harlow, Essex: Addison-Wesley

**Chapter 29: Overview of Data Warehousing and OLAP**

Elmasri, R. & Navathe, S. (2011). *Fundamentals of Database Systems, 6th Edition*. Boston: Pearson/Addison Wesley

**Chapter 9: Data Warehousing**

Hoffer, J., Ramesh, V. and Topi, H. (2012). *Modern Database Management, 11th Edition*. Upper Saddle River, N.J.: Pearson/Prentice Hall

**Chapter 20: Data Mining**

Silberschatz, A., Korth, H. & Sudarshan, S. (2010). *Database System Concepts, 6th Edition*. McGraw-Hill Book Co.

Online

**www.tutorialspoint/dwh/index.htm**

TutorialsPoint.com, *Data Warehousing Tutorial*




Mentimeter

14