Chapter 6 – Data WareHouse Part III

INFO401

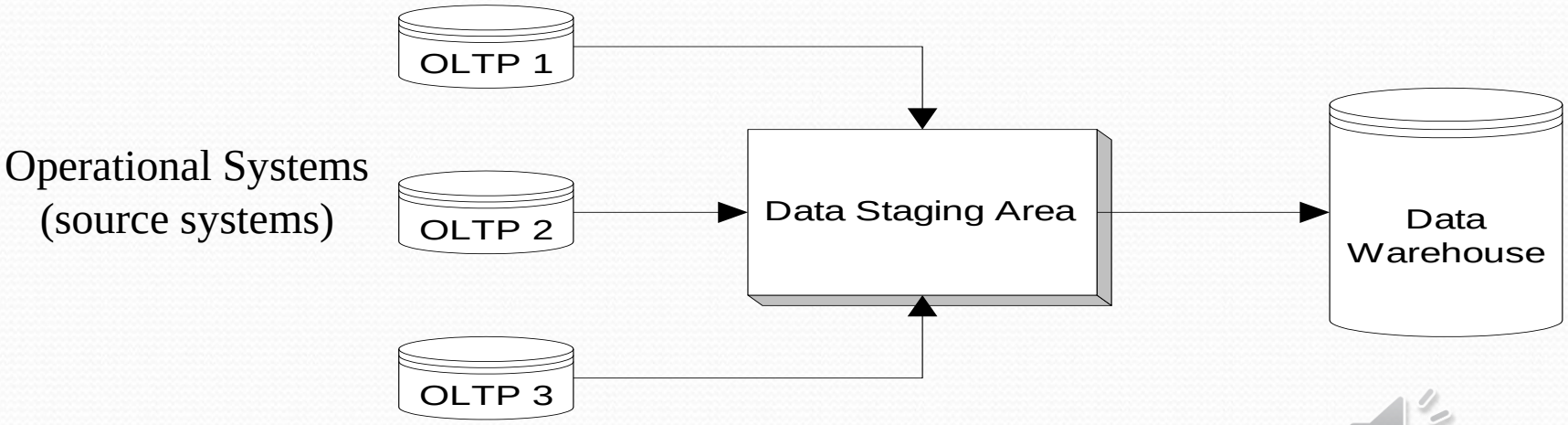
B.Wakim

Outline

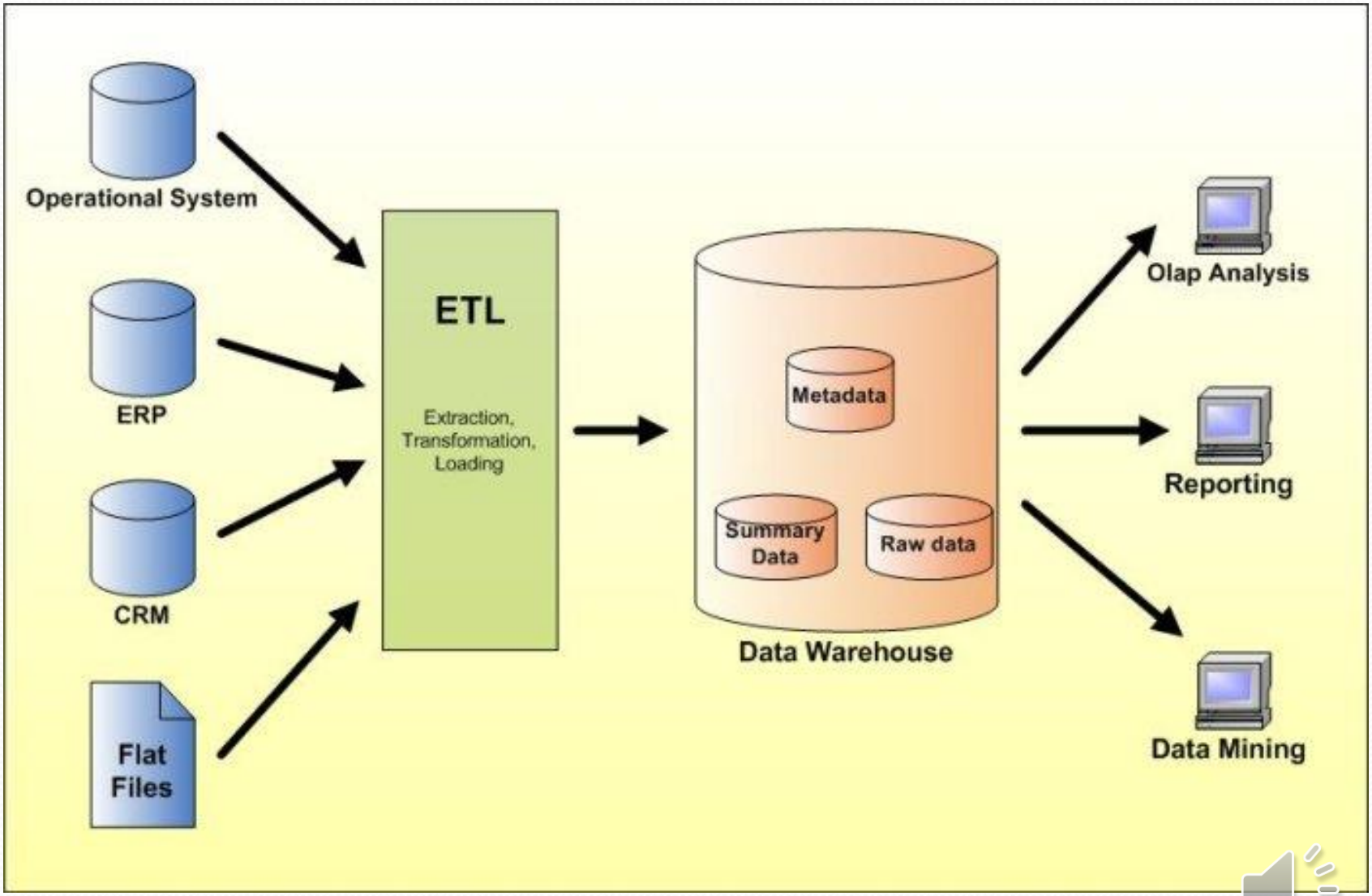
I. Definition and characteristics of a data warehouse
II. Architecture of a data warehouse
III. Lifecycle of a data warehouse
    1. Analysis
    2. Design (Dimensional Modeling)
    3. Import data (ETL)
    4. Install front-end tools
IV. Operational DB Systems (OLTP) vs. Data Warehouse (OLAP)
V. Advantages of the warehousing approach
VI. Data Warehouse vs. Data Mining

III.3- Import Data

- Identify data sources
- Extract the needed data from existing systems to a data staging area
- Transform and Clean the data
  - Resolve data type conflicts
  - Resolve naming and key conflicts
  - Remove, correct, or flag bad data
  - Conform Dimensions
- Load the data into the warehouse



ETL(1)


Operational System
ERP
CRM
Flat Files
ETL
Extraction, Transformation, Loading
Metadata
Summary Data
Raw data
Data Warehouse
Olap Analysis
Reporting
Data Mining

When should we ETL?

- Periodically (e.g., every night, every week) or after significant events
- Refresh policy set by administrator based on user needs and traffic
- Possibly different policies for different sources
- Rarely, on every update (real-time DW)

ETL(3)

- **Extract, Transform and Load** (**ETL**) refers to a process in database usage and especially in data warehousing that:
- Extracts data from homogeneous or heterogeneous data sources
- Transforms the data for storing it in proper format or structure for querying and analysis purpose
- Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)
- Usually all the three phases execute in parallel since the data extraction takes time, so while the data is being pulled another transformation process executes, processing the already received data and prepares the data for loading and as soon as there is some data ready to be loaded into the target, the data loading kicks off without waiting for the completion of the previous phases.

ETL(4)

- Transformation Examples:
  - Selecting only certain columns to load: (or selecting **null** columns not to load). For example, if the source data has three columns, *roll_no, age*, and *salary*, then the selection may take only *roll_no* and *salary*. Or, the selection mechanism may ignore all those records where salary is not present (*salary* = *null*).
  - Translating coded values: (*e.g.*, if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")
  - **Joining** data from multiple sources (*e.g.*, lookup, merge) and **deduplicating** the data

III.4- Install Front-End Tools

- Reporting tools
- Data mining tools
- GIS
- Etc.

IV- Operational DB Systems vs. Data Warehouse

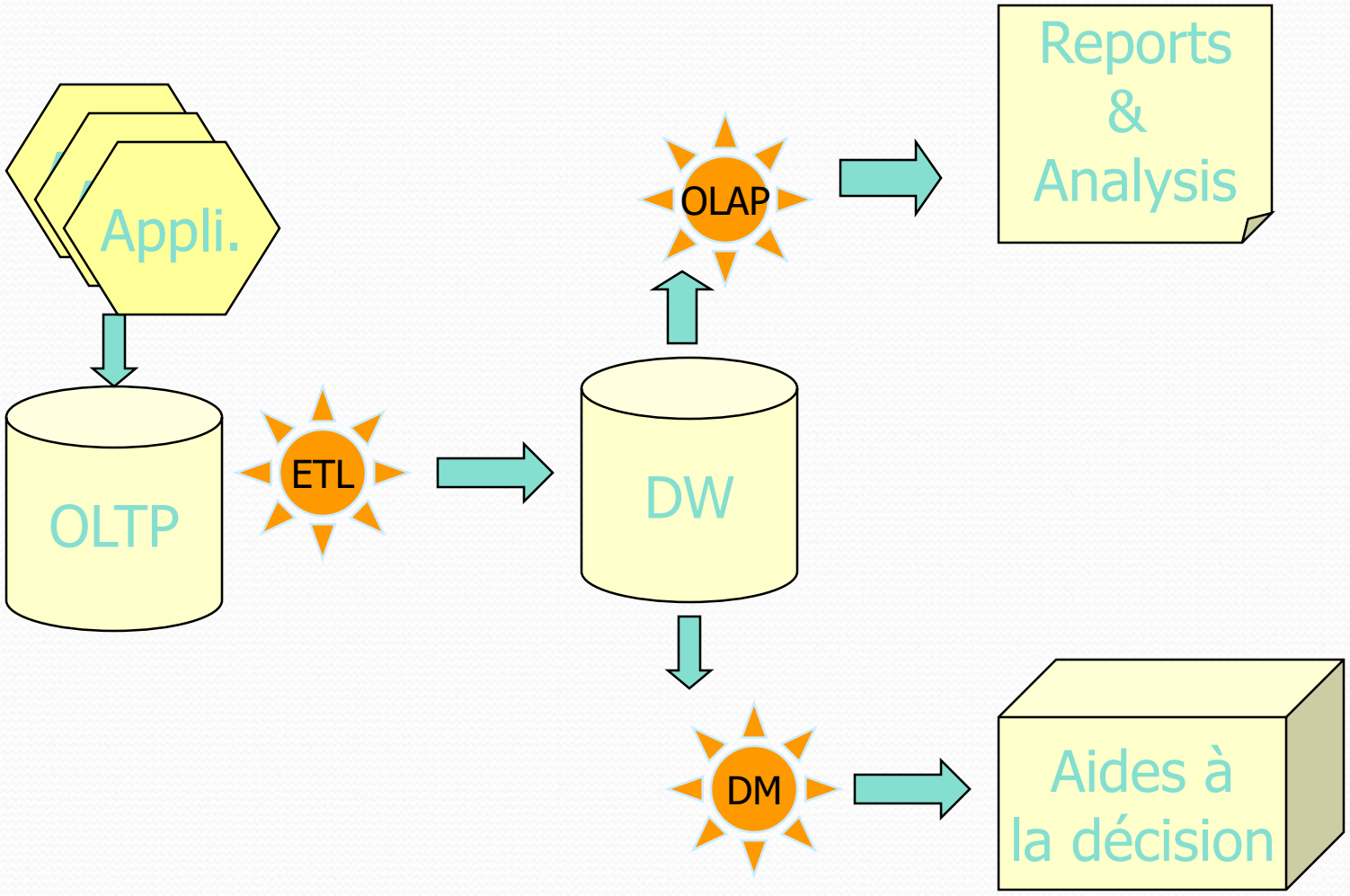
Operational DB Systems (OLTP)

- Transaction oriented
- Mostly updates
- Many small transactions
- Mb - Gb of data
- Current data
- Index on PK
- Raw data
- Few tables, many columns per table
- Thousands of users

Data Warehouse (OLAP)

- *OnLine Analytical Processing*
- Subject oriented
- Mostly reads
- Queries are long and complex
- Gb - Tb of data
- Historic data
- Lots of scans
- Summarized, reconciled data
- Many tables, few columns per table
- Hundreds of users (e.g., decision-makers, analysts)

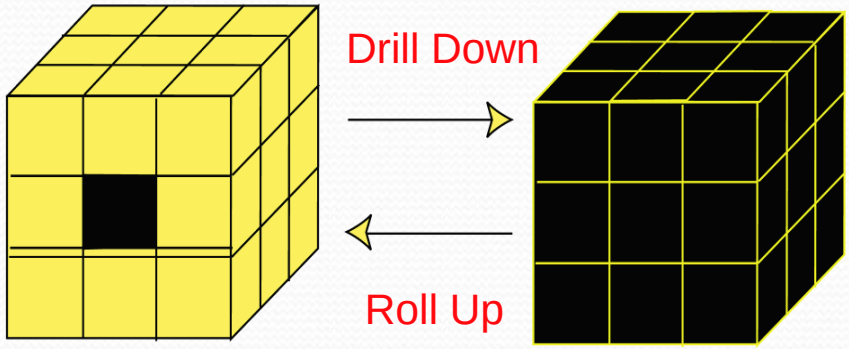
OLTP and OLAP

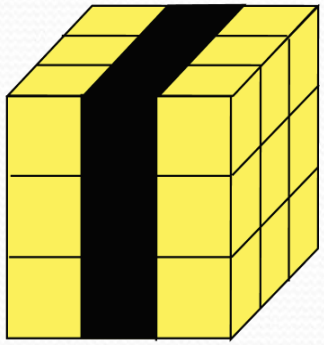


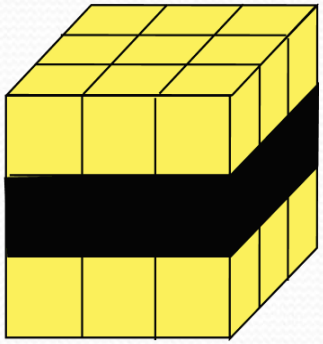
OLAP Operations

- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

OLAP Operations



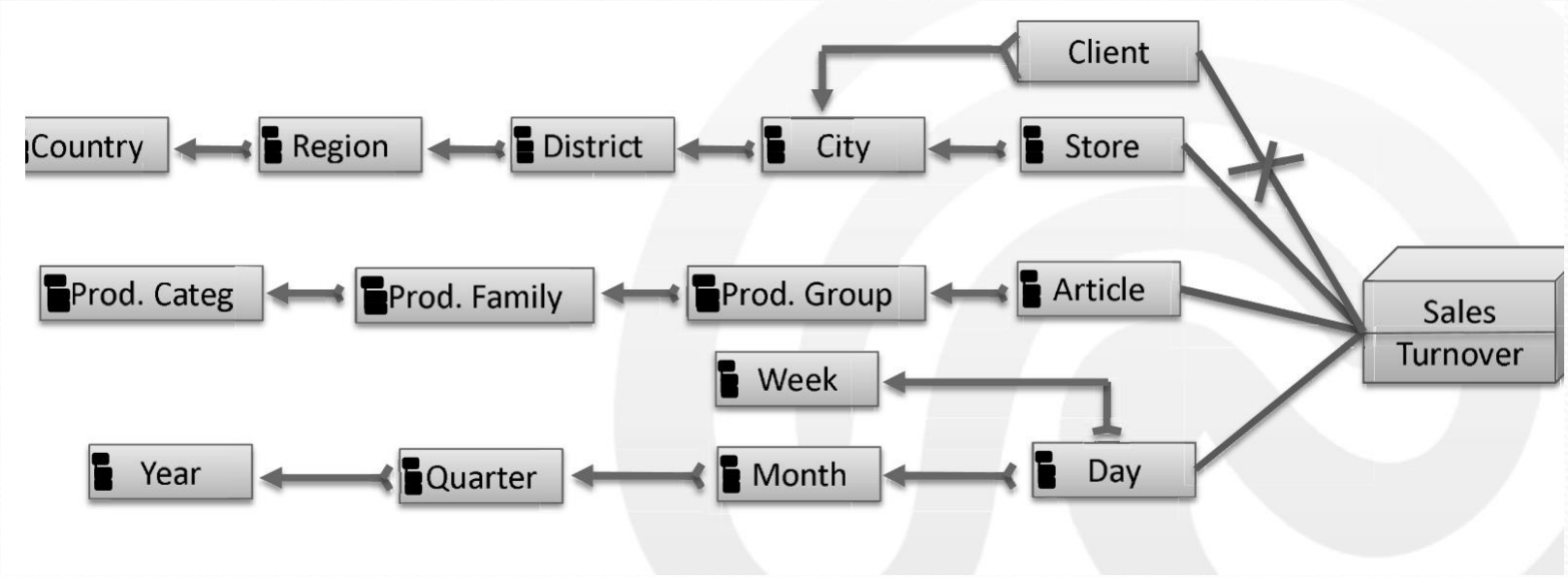




Single Cell **Multiple Cells** **Slice** **Dice**

Dimensional roll-ups

- **Dimensional roll-ups**
- Are done solely on the **fact table** by **dropping** one or more dimensions
- E.g., drop the Client dimension

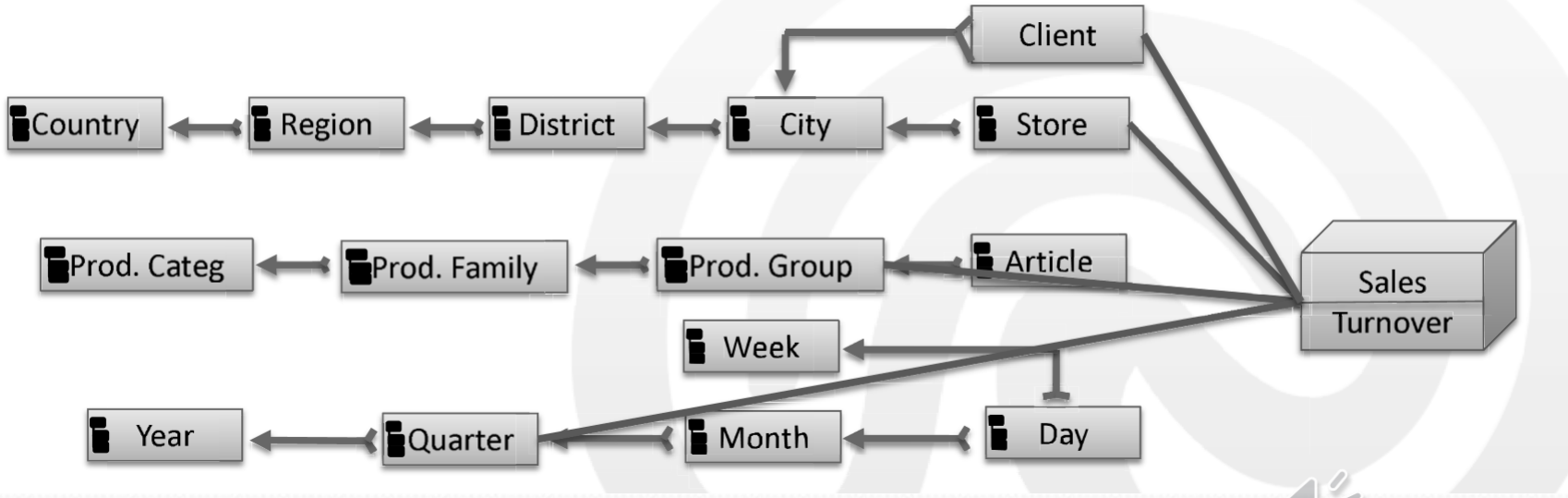


Roll-Up

- Roll-up (drill-up)
- Taking the current aggregation level of fact values and doing a **further aggregation**
- **Summarize data by one dimension**
  - Climbing up hierarchy (hierarchical roll-up)
  - By dimensional reduction
- E.g., from Time.Week to Time.Year

Hierarchical roll-ups

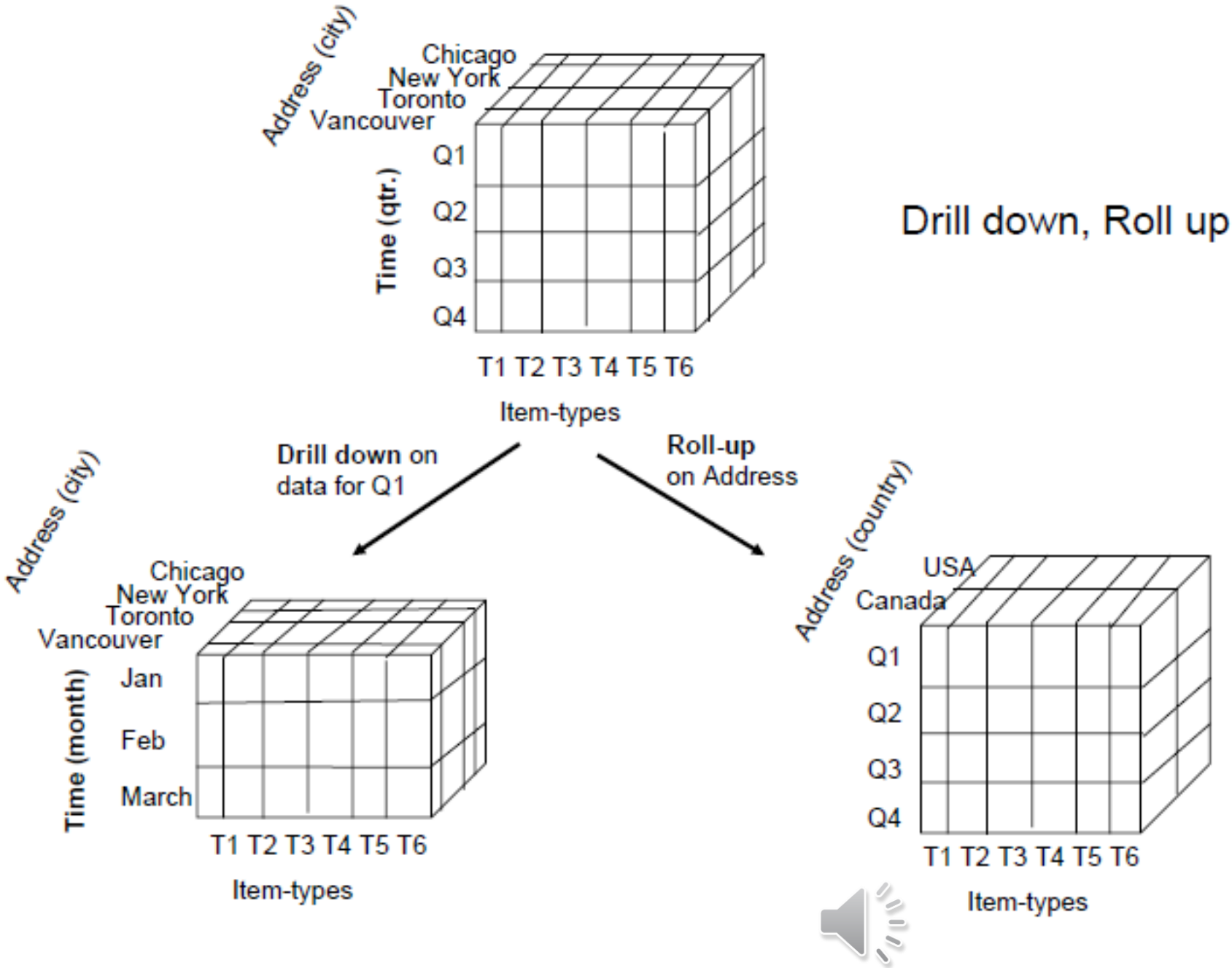
- **Hierarchical roll-ups**
- Performed on the **fact table and some dimension tables** by **climbing up** the attribute hierarchies
- E.g., climbed the **Time** hierarchy to **Quarter** and **Article** hierarchy to **Prod. group**



Drill-down (roll-down)

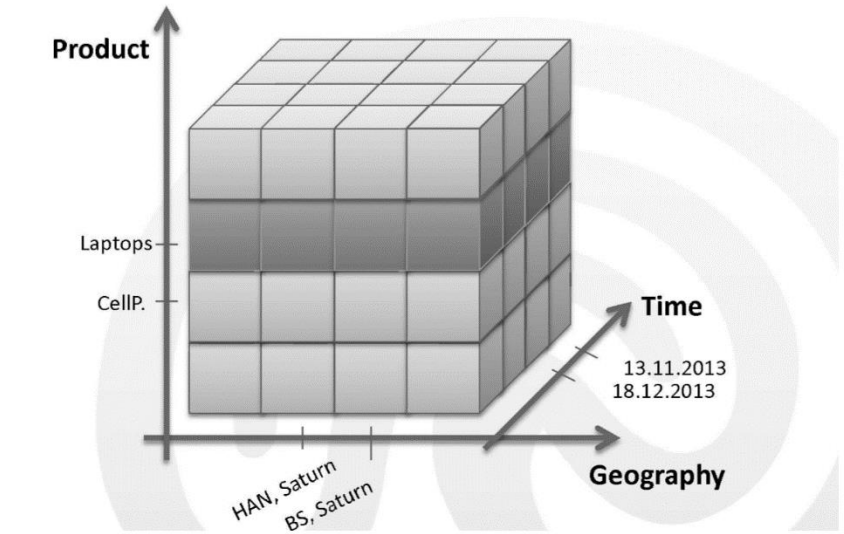
- **Reverse of roll-up**
- Represents a **de-aggregate** operation
- From higher level of summary to lower level of summary – detailed data
- **Drill down** : aggregation according to one dimension.
- E.g: Month ➔ Week

Drill down, Roll up



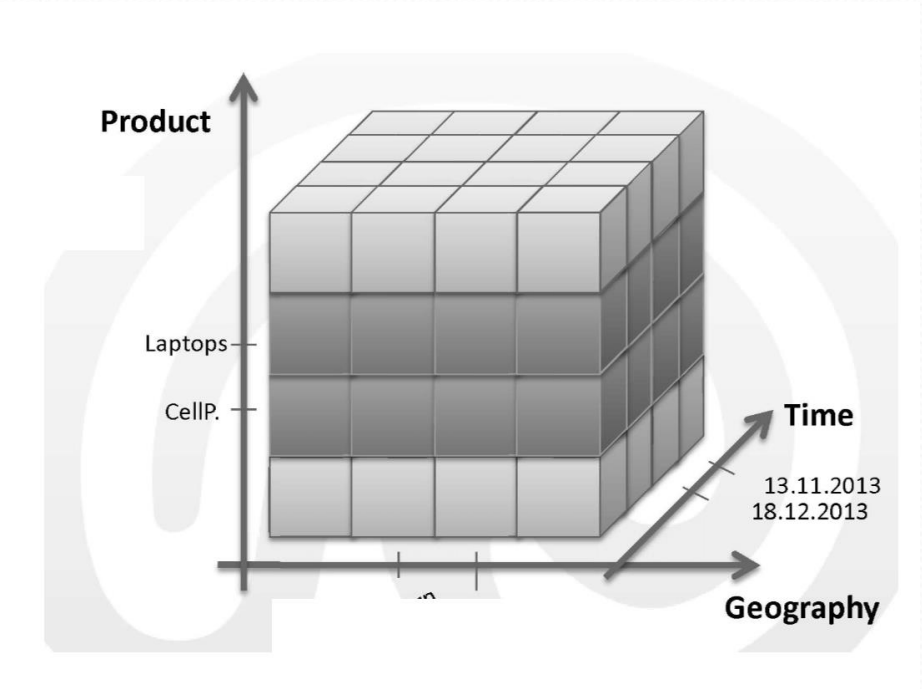
Slide

- Slice: a **subset** of the multi-dimensional array corresponding to **a single value** of one or more dimensions **and projection on the rest** of dimensions
- E.g., project on **Geo (store)** and Time from values corresponding to **Laptops in the product** dimension

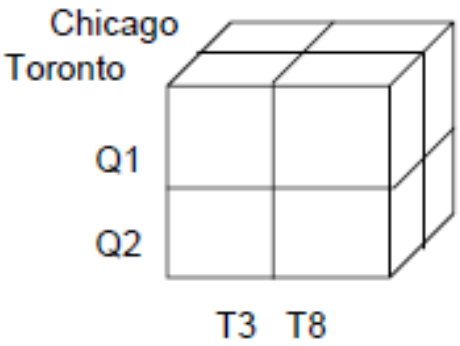


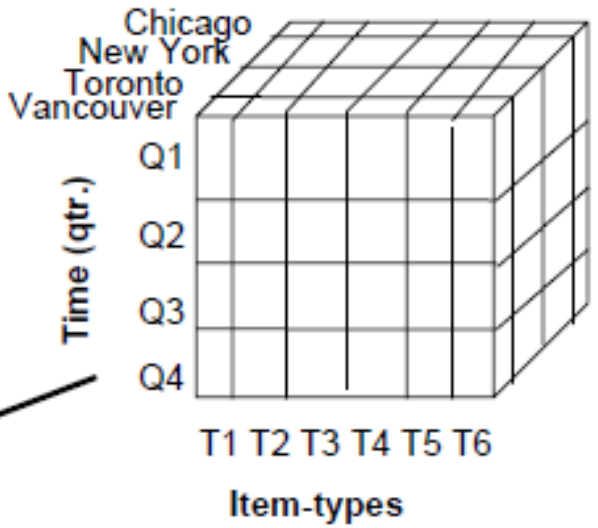
Dice

- Dice: amounts to **range select condition** on one dimension, or to **equality select condition** on **more** than one dimension
- E.g. range SELECT



Dice for
(location in {Chicago, Toronto}
and time in {Q1}
And Item in {T3, T8}





Slice
For Time in {Q1}

Slicing and Dicing

Slice: Selection on one dimension

Dice; Selection on two or more dimensions

Pivot

- Pivot (rotate): re-arranging data for viewing purposes
- The simplest view of pivoting is that it selects two dimensions to **aggregate the measure**
- The aggregated values are often displayed in **a grid** where each point in the (x, y) coordinate system corresponds to an aggregated value of the measure
- The x and y coordinate values are the values of the selected two dimensions
- The result of pivoting is also called **cross–tabulation**

Pivot

- Consider pivoting the following data

Location	
CityId	City
1	Bra..
2	Hann..
3	Ham..

Sales			
CityId	PerId	TimId	Amnt
1	1	1	230
1	1	2	300
1	1	8	310
1	2	7	50
2	3	1	550
2	3	5	100
3	4	6	880
3	5	1	60
3	5	2	60
3	5	4	140

Time	
TimId	Day
1	Mon
2	Tue
3	Wed
4	Thu
5	Fri
6	Sat
7	San
8	Mon

Pivot

- Pivoting on City and Day

	Mon	Tue	Wed	Thu	Fri	Sat	San	SubTotal
Hamburg	60	60	0	140	0	880	0	1140
Hannover	550	0	0	0	100	0	0	650
Braunschweig	540	300	0	0	0	0	50	890
SubTotal	1150	360	0	140	100	880	50	2680

	Hamb..	Han.	Bra..	SubTotal
Mon	60	550	540	1150
Tue	60	0	300	360
Wed	0	0	0	0
Thu	140	0	0	140
Fri	0	100	0	100
Sat	880	0	0	880
San	0	0	50	50
SubTotal	1140	650	890	2680

V- Advantages

- **High query performance**
  - But not necessarily most current information
- **No Interference** with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources
- **Information copied at warehouse**
  - Can modify, annotate, summarize, restructure, etc.
  - Can store historical information

VI- Data Warehouse vs. Data Mining(1)

- **Data warehousing** can be said to be the process of ***centralizing*** or ***aggregating*** data from multiple sources into one common repository.
- **Data mining** is the process of finding patterns in a given data set. These patterns can often provide meaningful and insightful data to whoever is interested in that data.
- Data mining is used today in a wide variety of contexts:
  - in fraud detection,
  - as an aid in marketing campaigns,
  - and even supermarkets use it to study their consumers.

Data Warehouse vs. Data Mining(2)

- **Data Mining Example:**
  - If you've ever used a credit card, then you may know that credit card companies will alert you when they think that your credit card is being fraudulently used by someone other than you.
  - This is a perfect example of data mining – credit card companies have a history of your purchases from the past and know geographically where those purchases have been made.
  - If all of a sudden some purchases are made in a city far from where you live, the credit card companies are put on alert to a possible fraud since their data mining shows that you don't normally make purchases in that city.
  - Then, the credit card company can disable your card for that transaction or just put a flag on your card for suspicious activity.

Data Warehouse vs. Data Mining(3)

- **Data Warehousing Example:**
- A great example of data warehousing that everyone can relate to is what Facebook does. Facebook basically gathers all of your data – your friends, your likes, who you stalk, etc. – and then stores that data into one central repository.
- Even though Facebook most likely stores your friends, your likes, etc., in separate databases, they do want to take the most relevant and important information and put it into one central aggregated database.
- Why would they want to do this? For many reasons – they want to make sure that you see the most relevant ads that you're most likely to click on, they want to make sure that the friends that they suggest are the most relevant to you, etc. – keep in mind that this is the data mining phase, in which meaningful data and patterns are extracted from the aggregated data.
- But, underlying all these motives is the main motive: to make more money – after all, Facebook is a business.

Data Warehouse vs. Data Mining(4)

- Data warehousing is a process that must occur before any data mining can take place.
  - Data warehousing is the process of compiling and organizing data into one common database,
  - Data mining is the process of extracting meaningful data from that database.
  - The data mining process relies on the data compiled in the data warehousing phase in order to detect meaningful patterns.