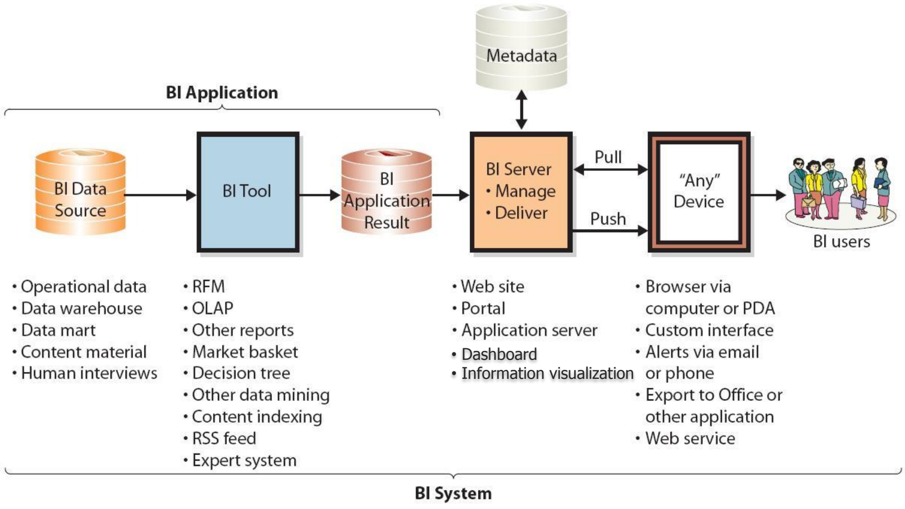
Data Warehouse and OLAP

IT 4153 Advanced Database

J.G. Zheng
Spring 2012

SPSU SOUTHERN POLYTECHNIC STATE UNIVERSITY

Introduction: BI System



BI Application

- BI Data Source
- BI Tool
- BI Application Result
- Metadata
- BI Server • Manage • Deliver
- Pull
- Push
- "Any" Device
- BI users

BI Data Source:
- Operational data
- Data warehouse
- Data mart
- Content material
- Human interviews

BI Tool:
- RFM
- OLAP
- Other reports
- Market basket
- Decision tree
- Other data mining
- Content indexing
- RSS feed
- Expert system

BI Server:
- Web site
- Portal
- Application server
- Dashboard
- Information visualization

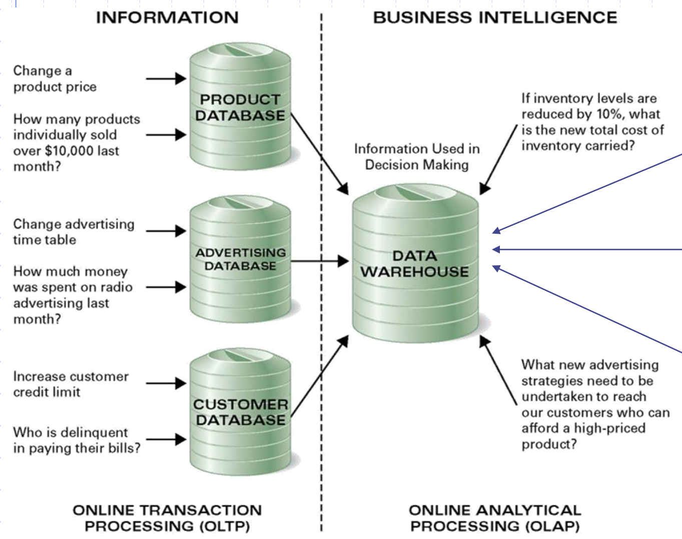
"Any" Device:
- Browser via computer or PDA
- Custom interface
- Alerts via email or phone
- Export to Office or other application
- Web service

BI System

Data Warehouse

- Data warehouse is a special kind of database that support data analysis and decision making
- Traditional (operational) databases facilitate data management and transaction processing. They have two limitations for data analysis and decision support
  - Performance
    - They are transaction oriented (data insert, update, move, etc.)
    - Not optimized for complex data analysis
    - Usually do not hold historical data
  - Heterogeneity
    - Individual databases usually manage data in very different ways, even in the same organization (not to mention external data sources which may be dramatically different).

Data Warehousing Supports Analytical Processing


INFORMATION
BUSINESS INTELLIGENCE
Change a product price
How many products individually sold over $10,000 last month?
PRODUCT DATABASE
Change advertising time table
How much money was spent on radio advertising last month?
ADVERTISING DATABASE
Increase customer credit limit
Who is delinquent in paying their bills?
CUSTOMER DATABASE
Information Used in Decision Making
DATA WAREHOUSE
If inventory levels are reduced by 10%, what is the new total cost of inventory carried?
What new advertising strategies need to be undertaken to reach our customers who can afford a high-priced product?
ONLINE TRANSACTION PROCESSING (OLTP)
ONLINE ANALYTICAL PROCESSING (OLAP)

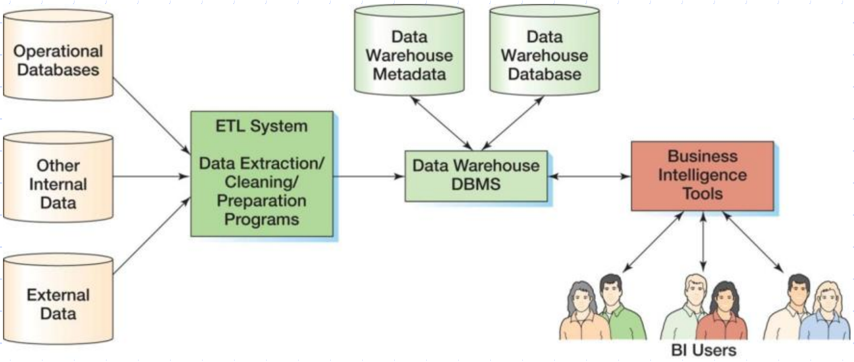
What is the reason for a decrease of total sales this year? (reasoning)

How do advertising activities affect sales of different products bought by different type of customers, in different regions? (synthesizing)

Should we invest more on our e-business? (fuzzy question → need high level analysis for decision making)

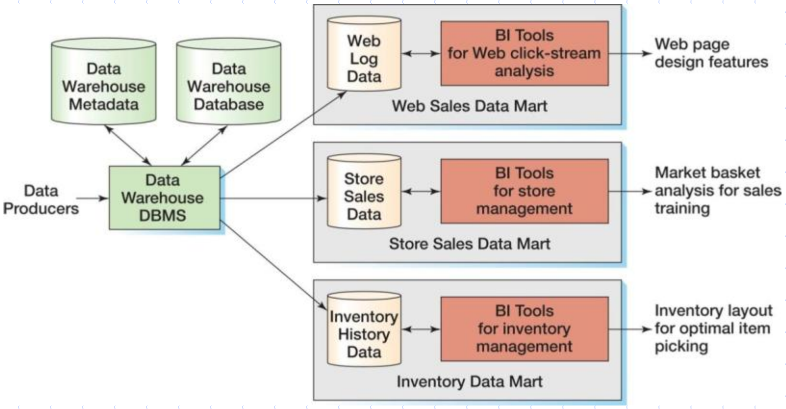
Data Sources

- Data warehouses extract data from many data sources, including operational (or transactional) databases



Data Mart

- Data mart is a small data warehouse focusing on certain type of analysis

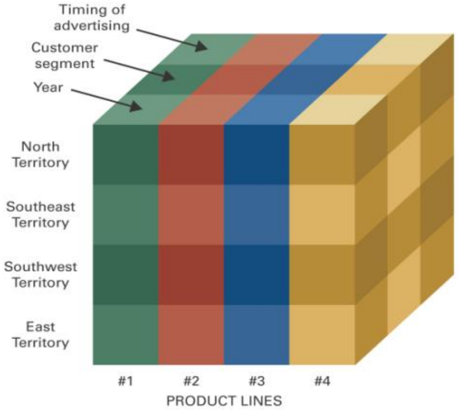


Comparison of Data

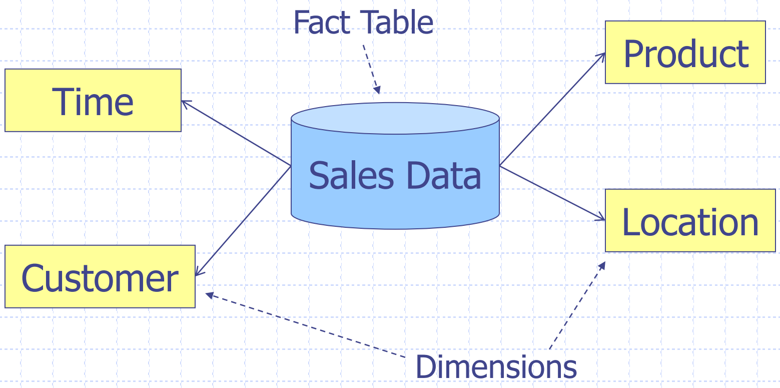
Data Warehouse Data		Operational Database Data
Subject oriented	Data are stored with a subject orientation that facilitates multiple views of the data and decision making. E.g., sales may be recorded by product, by division, by manager, or by region.	Data are stored with a functional orientation. E.g., data may be stored for invoices, payments, credit amounts, and so on.
Integrated	Provide a unified view of all data elements with a common definition and representation for all business units.	Similar data may be represented differently in different databases (either structure or format)
Time-variant	Data are recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.	Data are recorded as current transactions. E.g., the sales data may be the sales of a product on a given date, such as $342.78 on 12-May-2004.
Non-volatile	Data is not updated in real time but is refreshed from operational systems on a regular basis. Data structure is not optimized for updates – redundancy is not the major issue	Data updates are frequent and common. E.g., an inventory amount changes with each sale. New data is added as a replacement to the database.

Data Warehouse Structure

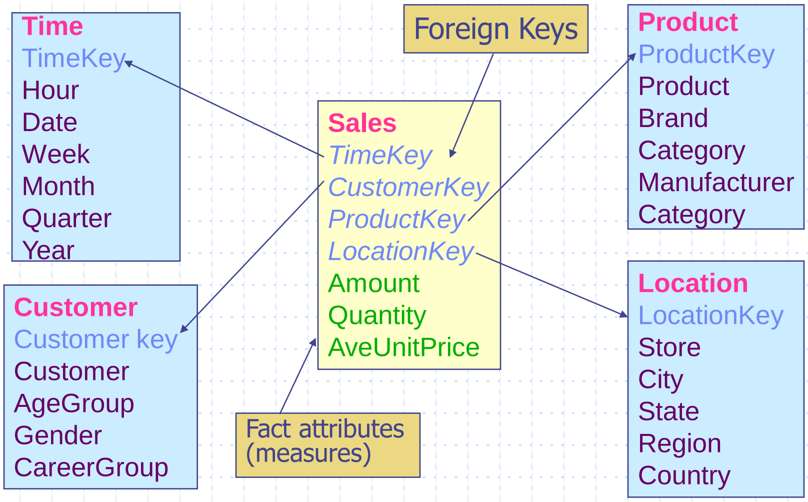
- Data warehouse is (often) multi-dimensional
  - The view of data is usually called a "cube", although it can have more than 3 dimensions



Star Schema



Star Schema in the Relational Model



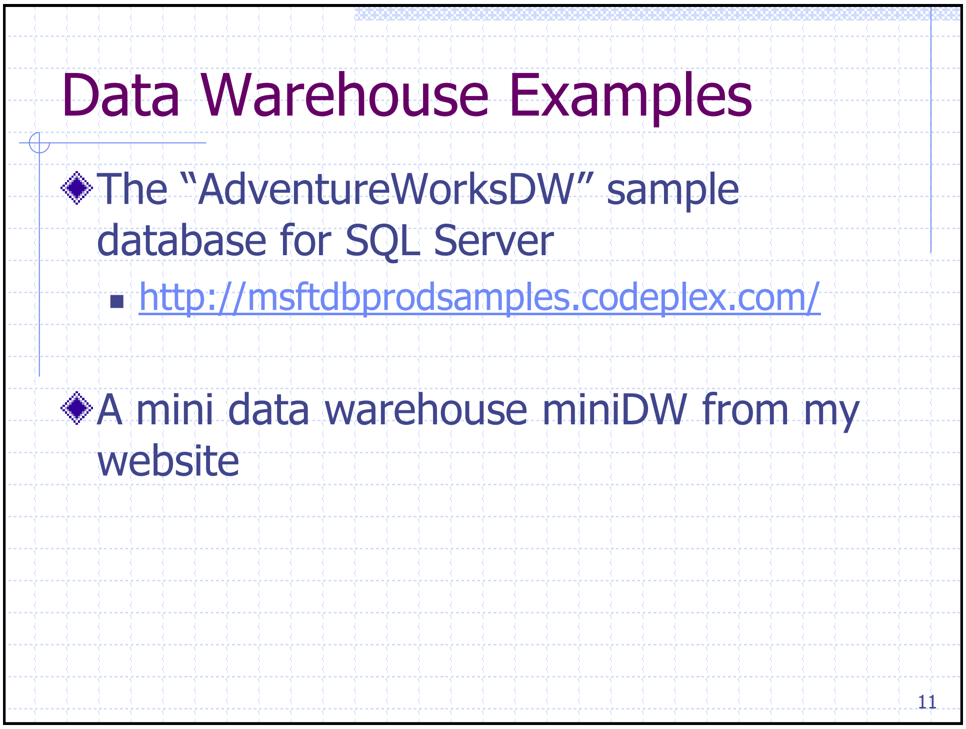
Time
TimeKey
Hour
Date
Week
Month
Quarter
Year

Customer
Customer key
Customer
AgeGroup
Gender
CareerGroup

Sales
TimeKey
CustomerKey
ProductKey
LocationKey
Amount
Quantity
AveUnitPrice

Product
ProductKey
Product
Brand
Category
Manufacturer
Category

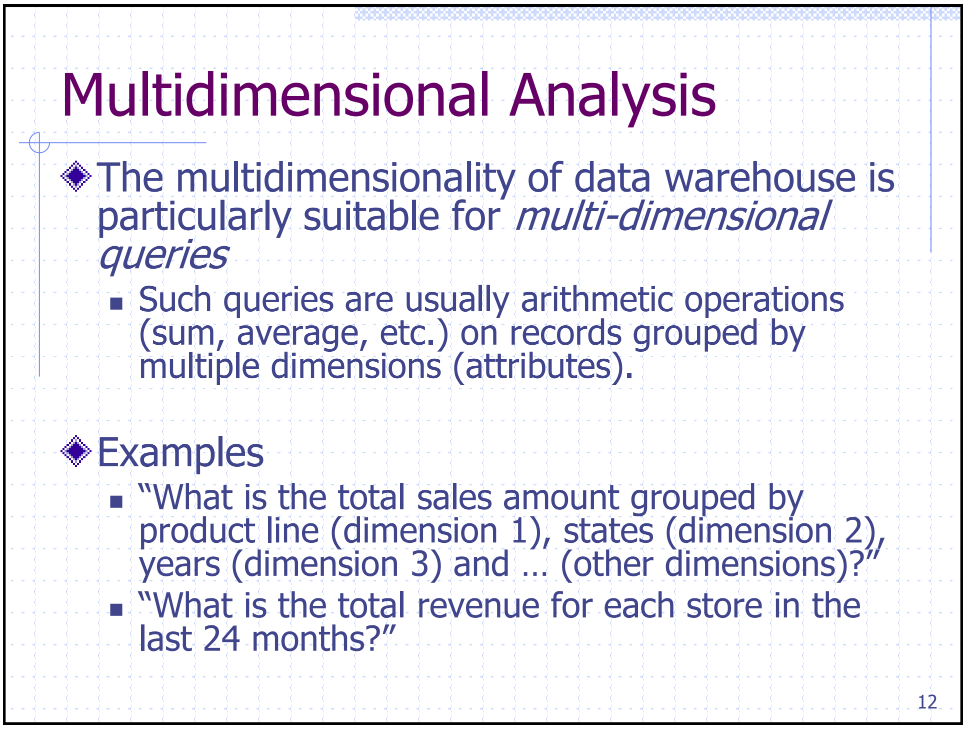
Location
LocationKey
Store
City
State
Region
Country



Data Warehouse Examples

- The “AdventureWorksDW” sample database for SQL Server
  - http://msftdbprodsamples.codeplex.com/
- A mini data warehouse miniDW from my website

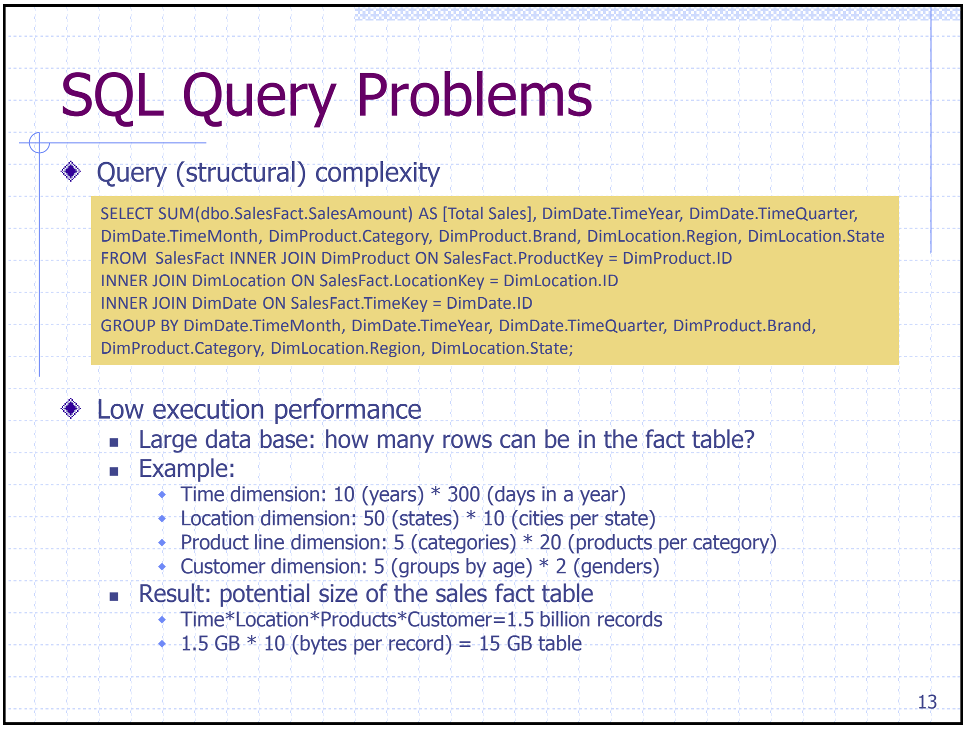
11



Multidimensional Analysis

- The multidimensionality of data warehouse is particularly suitable for *multi-dimensional queries*
  - Such queries are usually arithmetic operations (sum, average, etc.) on records grouped by multiple dimensions (attributes).
- Examples
  - “What is the total sales amount grouped by product line (dimension 1), states (dimension 2), years (dimension 3) and … (other dimensions)?”
  - “What is the total revenue for each store in the last 24 months?”

12

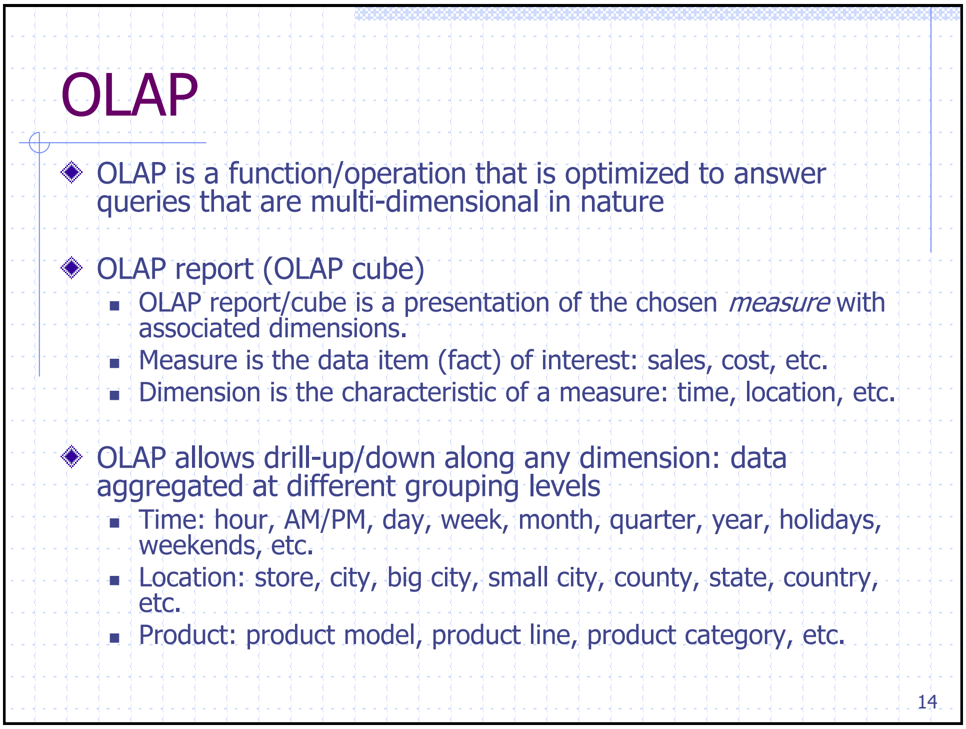


SQL Query Problems

- Query (structural) complexity

```
SELECT SUM(dbo.SalesFact.SalesAmount) AS [Total Sales], DimDate.TimeYear, DimDate.TimeQuarter,
DimDate.TimeMonth, DimProduct.Category, DimProduct.Brand, DimLocation.Region, DimLocation.State
FROM  SalesFact INNER JOIN DimProduct ON SalesFact.ProductKey = DimProduct.ID
INNER JOIN DimLocation ON SalesFact.LocationKey = DimLocation.ID
INNER JOIN DimDate ON SalesFact.TimeKey = DimDate.ID
GROUP BY DimDate.TimeMonth, DimDate.TimeYear, DimDate.TimeQuarter, DimProduct.Brand,
DimProduct.Category, DimLocation.Region, DimLocation.State;
```

- Low execution performance
  - Large data base: how many rows can be in the fact table?
  - Example:
    - Time dimension: 10 (years) * 300 (days in a year)
    - Location dimension: 50 (states) * 10 (cities per state)
    - Product line dimension: 5 (categories) * 20 (products per category)
    - Customer dimension: 5 (groups by age) * 2 (genders)
  - Result: potential size of the sales fact table
    - Time*Location*Products*Customer=1.5 billion records
    - 1.5 GB * 10 (bytes per record) = 15 GB table



OLAP

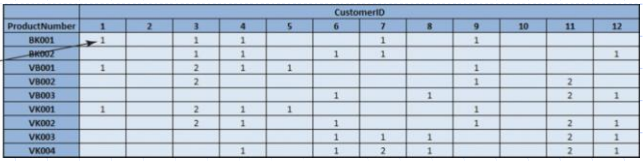
- OLAP is a function/operation that is optimized to answer queries that are multi-dimensional in nature

- OLAP report (OLAP cube)
  - OLAP report/cube is a presentation of the chosen *measure* with associated dimensions.
  - Measure is the data item (fact) of interest: sales, cost, etc.
  - Dimension is the characteristic of a measure: time, location, etc.

- OLAP allows drill-up/down along any dimension: data aggregated at different grouping levels
  - Time: hour, AM/PM, day, week, month, quarter, year, holidays, weekends, etc.
  - Location: store, city, big city, small city, county, state, country, etc.
  - Product: product model, product line, product category, etc.

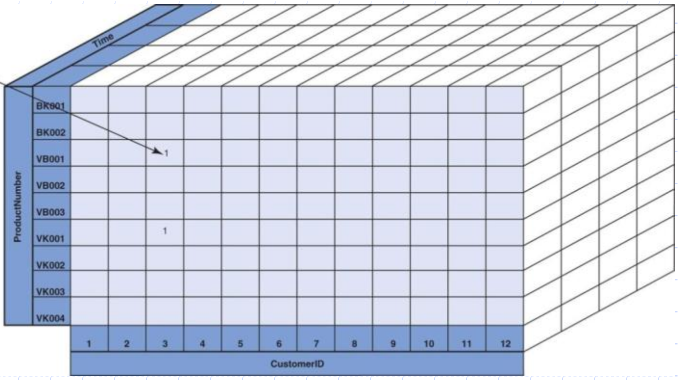
OLAP Report View

Each cell shows the total quantity of each product that has been purchased by each customer



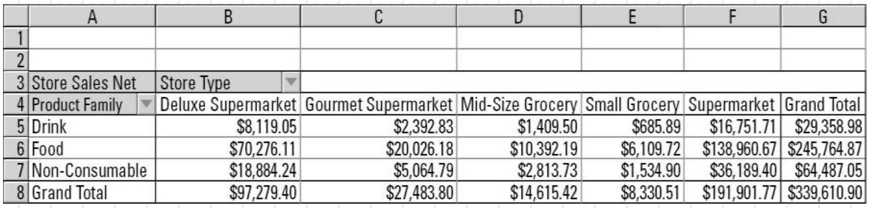
	CustomerID											
ProductNumber	1	2	3	4	5	6	7	8	9	10	11	12
BK001	1		1	1			1		1			
BK002			1	1		1	1					1
VB001	1		2	1	1				1			
VB002			2						1		2	
VB003						1		1			2	1
VK001	1		2	1	1				1			
VK002			2	1		1			1		2	1
VK003						1	1	1			2	1
VK004				1		1	2	1			2	1

Each cell will show the total quantity of each product that has been purchased by each customer on a specific date



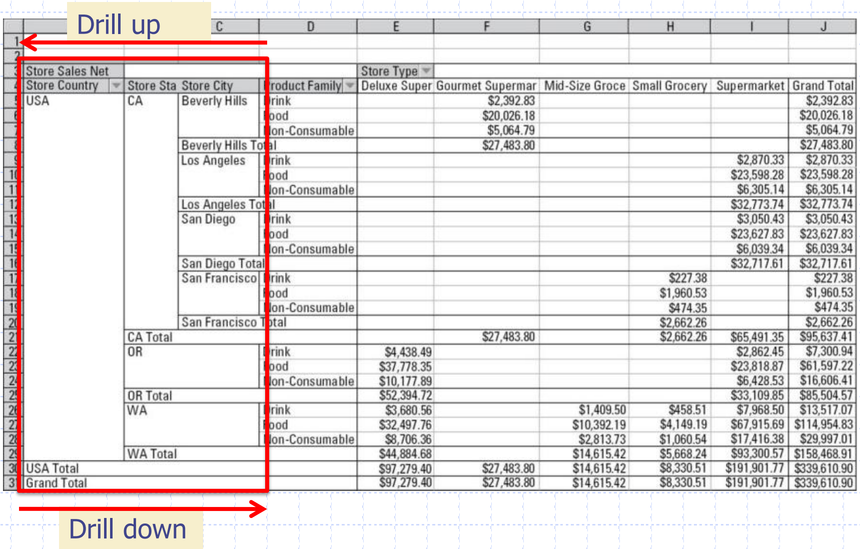
OLAP Report and Pivot Table

- OLAP results are often presented in a way similar to pivot tables in spreadsheets applications.
- Example: a pivot table in MS Excel



	A	B	C	D	E	F	G
1							
2							
3	Store Sales Net	Store Type ▼					
4	Product Family ▼	Deluxe Supermarket	Gourmet Supermarket	Mid-Size Grocery	Small Grocery	Supermarket	Grand Total
5	Drink	$8,119.05	$2,392.83	$1,409.50	$685.89	$16,751.71	$29,358.98
6	Food	$70,276.11	$20,026.18	$10,392.19	$6,109.72	$138,960.67	$245,764.87
7	Non-Consumable	$18,884.24	$5,064.79	$2,813.73	$1,534.90	$36,189.40	$64,487.05
8	Grand Total	$97,279.40	$27,483.80	$14,615.42	$8,330.51	$191,901.77	$339,610.90

OLAP Drill Up/Down



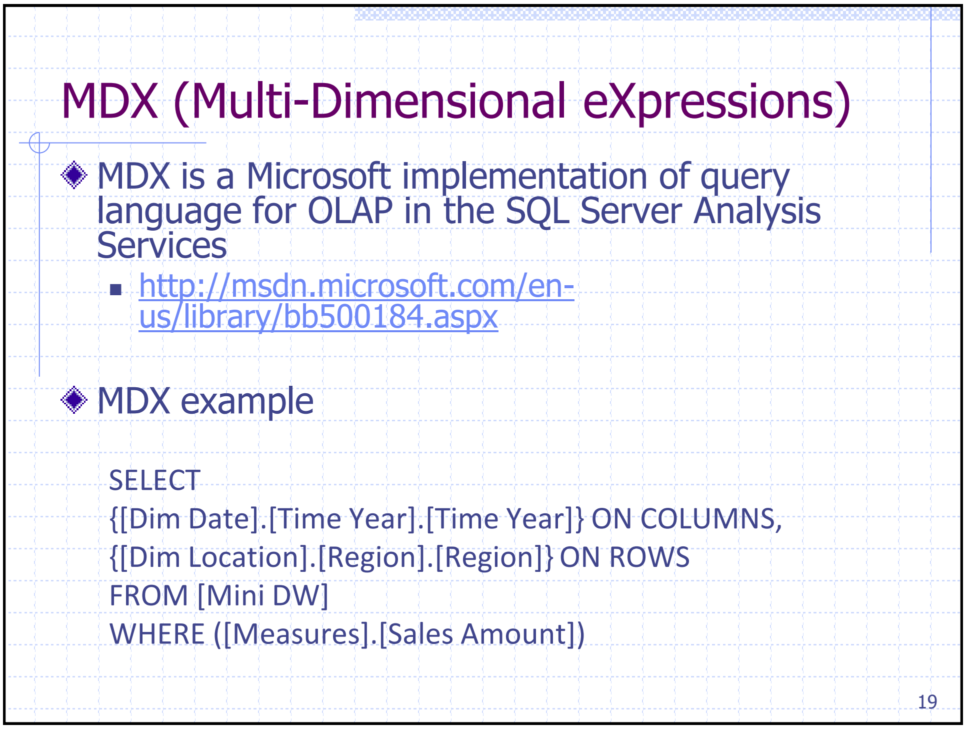
Drill up

	Store Sales Net				Store Type					
	Store Country	Store Sta	Store City	Product Family	Deluxe Super	Gourmet Supermar	Mid-Size Groce	Small Grocery	Supermarket	Grand Total
5	USA	CA	Beverly Hills	Drink		$2,392.83				$2,392.83
6				Food		$20,026.18				$20,026.18
7				Non-Consumable		$5,064.79				$5,064.79
8			Beverly Hills Total			$27,483.80				$27,483.80
9			Los Angeles	Drink					$2,870.33	$2,870.33
10				Food					$23,598.28	$23,598.28
11				Non-Consumable					$6,305.14	$6,305.14
12			Los Angeles Total						$32,773.74	$32,773.74
13			San Diego	Drink					$3,050.43	$3,050.43
14				Food					$23,627.83	$23,627.83
15				Non-Consumable					$6,039.34	$6,039.34
16			San Diego Total						$32,717.61	$32,717.61
17			San Francisco	Drink				$227.38		$227.38
18				Food				$1,960.53		$1,960.53
19				Non-Consumable				$474.35		$474.35
20			San Francisco Total					$2,662.26		$2,662.26
21		CA Total				$27,483.80		$2,662.26	$65,491.35	$95,637.41
22		OR		Drink	$4,438.49				$2,862.45	$7,300.94
23				Food	$37,778.35				$23,818.87	$61,597.22
24				Non-Consumable	$10,177.89				$6,428.53	$16,606.41
25		OR Total			$52,394.72				$33,109.85	$85,504.57
26		WA		Drink	$3,680.56		$1,409.50	$458.51	$7,968.50	$13,517.07
27				Food	$32,497.76		$10,392.19	$4,149.19	$67,915.69	$114,954.83
28				Non-Consumable	$8,706.36		$2,813.73	$1,060.54	$17,416.38	$29,997.01
29		WA Total			$44,884.68		$14,615.42	$5,668.24	$93,300.57	$158,468.91
30	USA Total				$97,279.40	$27,483.80	$14,615.42	$8,330.51	$191,901.77	$339,610.90
31	Grand Total				$97,279.40	$27,483.80	$14,615.42	$8,330.51	$191,901.77	$339,610.90

Drill down

OLAP/Pivot Table in Action

- Connecting to a SQL Server Database Engine
  - Use SQL to generate data for analysis
  - Use Microsoft Excel Pivot Table analysis
  - Use Visio 2010 Pivot Diagram
- Utilizing the SQL Server 2008 Analysis Service
  - Use the SQL Server BI Development Studio to design, deploy and browse an OLAP cube
  - Use the SQL Server Management Studio to browse an OLAP cube
  - Use the Excel Pivot Table and Visio Pivot Diagram to browse an OLAP cube

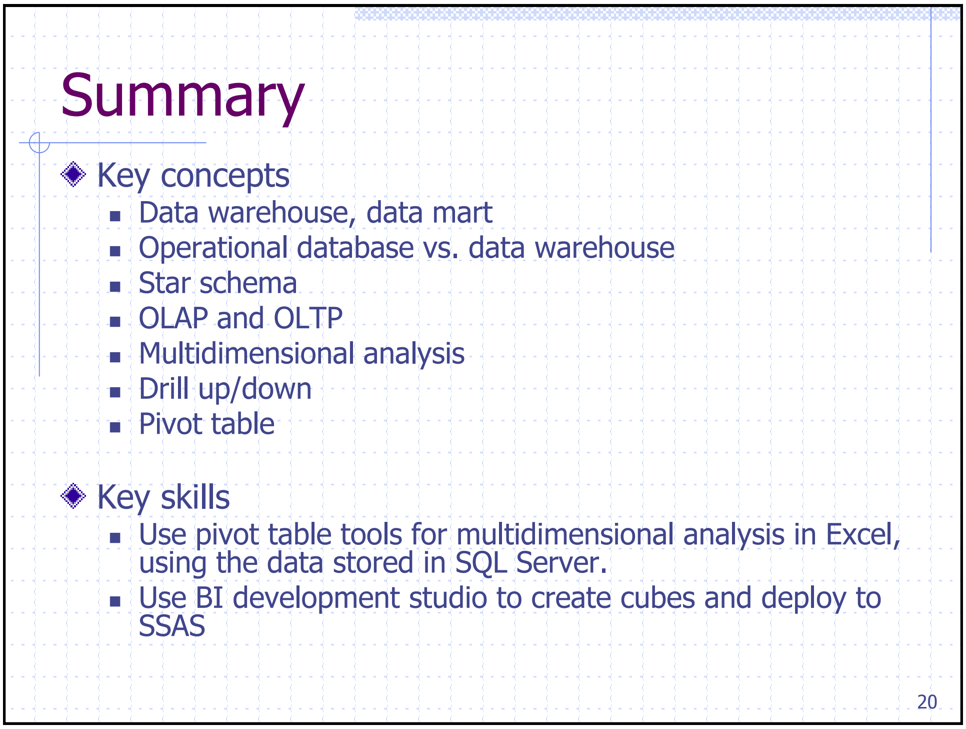


MDX (Multi-Dimensional eXpressions)

- MDX is a Microsoft implementation of query language for OLAP in the SQL Server Analysis Services
  - http://msdn.microsoft.com/en-us/library/bb500184.aspx
- MDX example

```
SELECT
{[Dim Date].[Time Year].[Time Year]} ON COLUMNS,
{[Dim Location].[Region].[Region]} ON ROWS
FROM [Mini DW]
WHERE ([Measures].[Sales Amount])
```

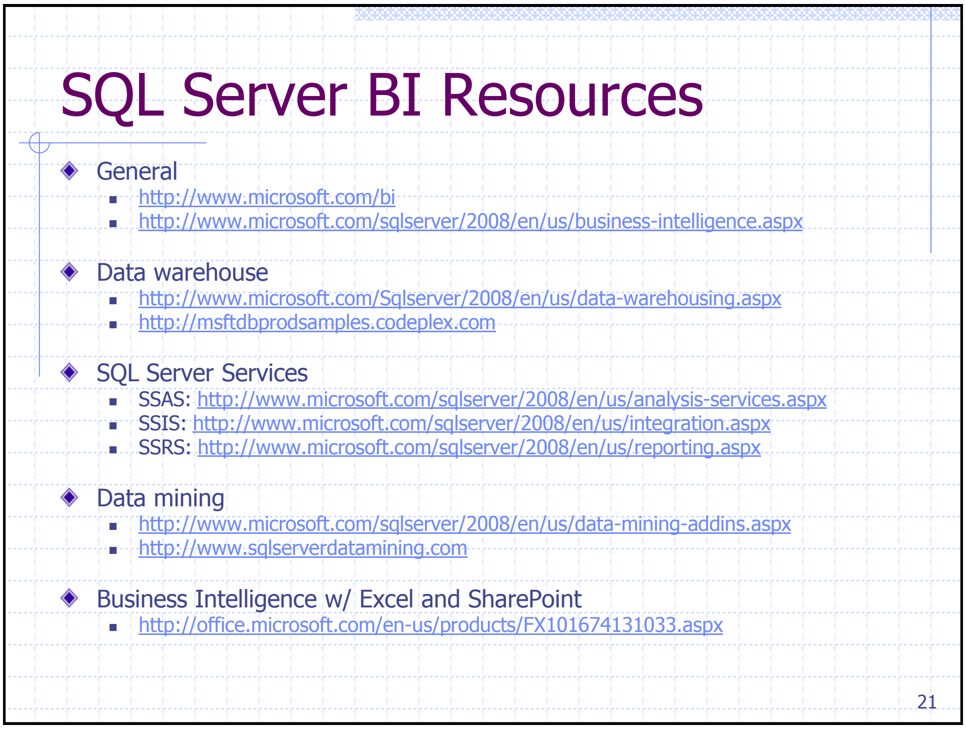
19



Summary

- Key concepts
  - Data warehouse, data mart
  - Operational database vs. data warehouse
  - Star schema
  - OLAP and OLTP
  - Multidimensional analysis
  - Drill up/down
  - Pivot table
- Key skills
  - Use pivot table tools for multidimensional analysis in Excel, using the data stored in SQL Server.
  - Use BI development studio to create cubes and deploy to SSAS

20



SQL Server BI Resources

- General
  - http://www.microsoft.com/bi
  - http://www.microsoft.com/sqlserver/2008/en/us/business-intelligence.aspx
- Data warehouse
  - http://www.microsoft.com/Sqlserver/2008/en/us/data-warehousing.aspx
  - http://msftdbprodsamples.codeplex.com
- SQL Server Services
  - SSAS: http://www.microsoft.com/sqlserver/2008/en/us/analysis-services.aspx
  - SSIS: http://www.microsoft.com/sqlserver/2008/en/us/integration.aspx
  - SSRS: http://www.microsoft.com/sqlserver/2008/en/us/reporting.aspx
- Data mining
  - http://www.microsoft.com/sqlserver/2008/en/us/data-mining-addins.aspx
  - http://www.sqlserverdatamining.com
- Business Intelligence w/ Excel and SharePoint
  - http://office.microsoft.com/en-us/products/FX101674131033.aspx

21