Topic 11: Databases for Business Intelligence

ICT285 Databases

Dr Danny Toohey

About this topic

- In today’s topic we look at how databases can be used to add value beyond their role in operational transaction processing, by supporting management analysis, planning, and decision making.

Topic learning outcomes

- **After completing this topic you should be able to:**
- Explain the difference between BI systems and operational systems
- Describe the benefits and issues associated with data warehousing
- Describe the major components of a data warehouse
- Explain why ‘dirty data’ is a significant issue for data warehouses, and ways in which dirty data can be cleaned
- Explain how a dimensional model differs from an entity relationship model
- Describe some reporting systems for BI
- Explain the concept of ‘drill down’ in an OLAP cube
- Explain what data mining is and what it can be used for

Resources for this topic

READING

- Text, Chapter 12: Big Data, Data Warehouses, and Business Intelligence Systems – omit sections from ‘Distributed Database Processing’ onwards

Other resources:

- An overview of data mining in Oracle:
- https://docs.oracle.com/database/121/DMCON/GUID-8232ABAD-E6B9-4C70-B227-E00738040932.htm#DMCON002
- And data warehousing: https://docs.oracle.com/database/121/DWHSG/toc.htm
- Lab 11 demonstrates some of Oracle’s analytic functions

Lab 11 – Oracle’s analytic functions

- In this lab we will take a very brief look at some of the features in Oracle that extend its application into Business Intelligence. Oracle contains many functions that allow us to manipulate and examine data in ways beyond the simple table/record structure we have been used to. For example, we can summarise the records in a crosstab type structure using the ‘pivot’ feature, or use Oracle’s analytic functions to look at relationships between records in a sequence.

Topic Outline

1. Introduction

2. Data for BI systems

3. Data warehouses and data marts

•4. Designing data warehouses

•5. Reporting systems: RFM and OLAP

6. Data mining

Introduction

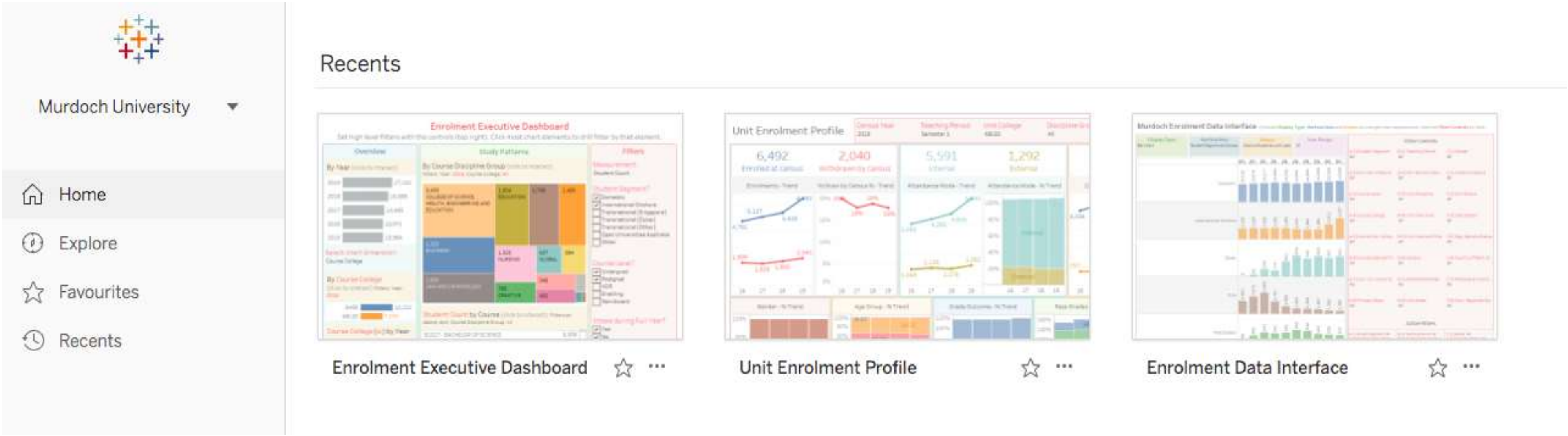
Introduction

- Earlier topics have discussed the value of data to an organisation and how it needs to be managed as any other valuable resource
- In this topic the focus changes to *adding value* to the data resource

Business Intelligence (BI) Systems

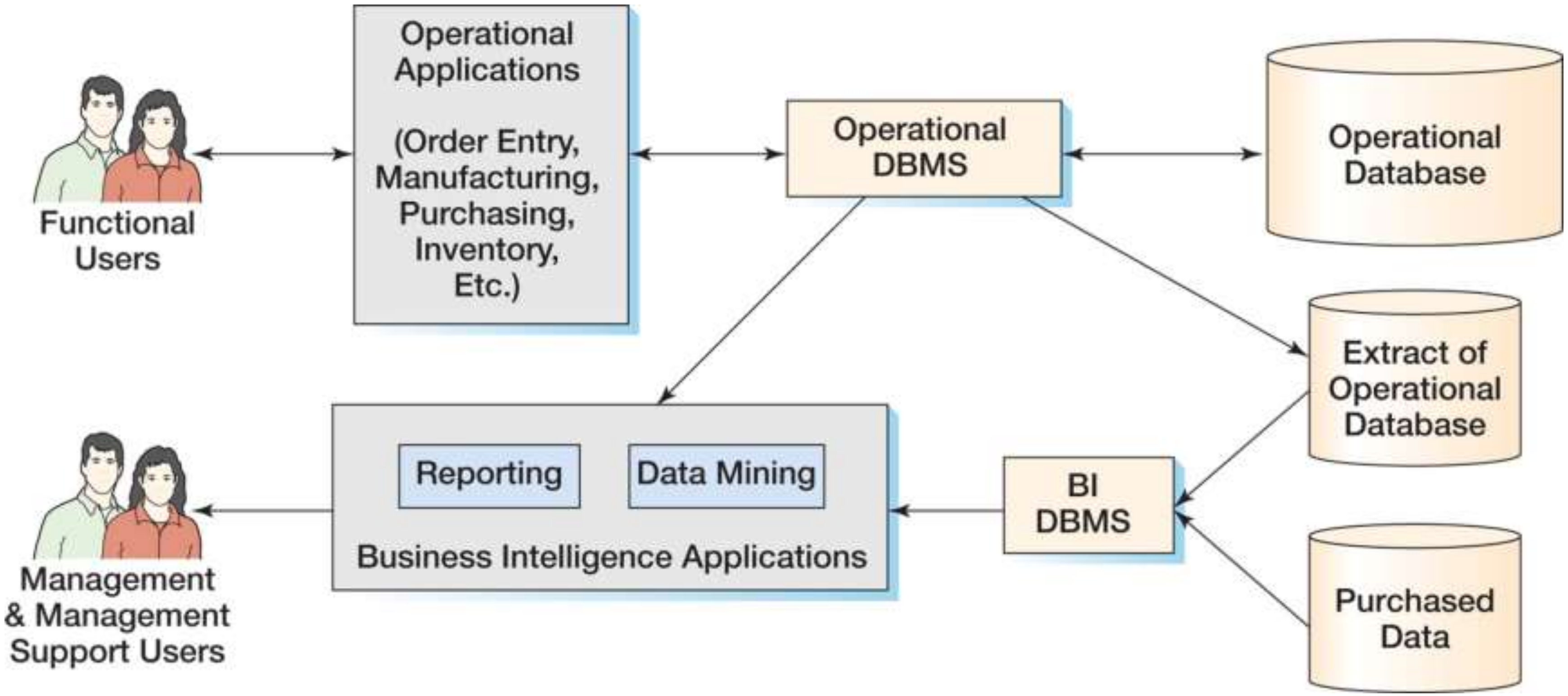
- Business Intelligence (BI) systems aims to retrieve the ideal data and give it to the right people in the correct format at the right time for the purpose of assisting management decision making (Add value to an organization data)
- The management decisions for this system is generally longer-term in nature and may include analysis of current as well as past activities or some prediction of what might happen in the future EG-
    - Capacity planning: how are we going to make sure we have tutorial rooms…
    - Product development: where will we build our new campus…
    - Outlet locations
    - Product promotions
- The data the BI system extract come from operational database and external brought data (REFER TO Next slides for data)

Example BI System…



Relationship between Operational and BI Systems

Daily activites


Functional Users
Operational Applications
(Order Entry, Manufacturing, Purchasing, Inventory, Etc.)
Operational DBMS
Operational Database
Extract of Operational Database
Reporting
Data Mining
Business Intelligence Applications
BI DBMS
Purchased Data
Management & Management Support Users

KROENKE AND AUER - DATABASE PROCESSING, 12th Edition
© 2012 Pearson Prentice Hall

Two broad categories of BI systems

Reporting Systems- What managers need to know to do daily activity

- Look at filtering, sorting, summarizing current status and calcuations
- Past and current comparisons
- Deliver regular report

Data Mining Systems

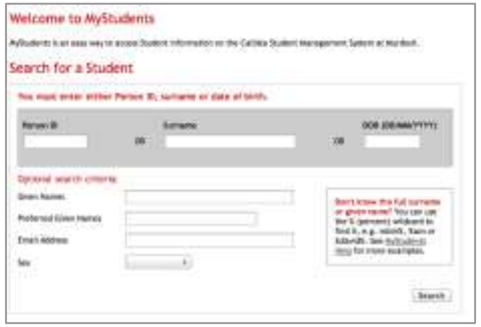
- Often utilize statistical and mathematical tools to look at-
  - what-if, predictions, and assist with decision-making
- Results generally are integrated into some other system or report

Operational systems

- Operational systems (OLTP) helps organisations to finish their daily business activities efficiently
  - EG- The student records system

Operational system are designed and optimised to support transactions

The management decisions for this system generally includes business process and are often short term in focus



Welcome to MyStudents

MyStudents is an easy way to access Student information on the Callista Student Management System at Murdoch.

Search for a Student

You must enter either Person ID, surname or date of birth.

Don't know the full surname or given name? You can use the % (percent) wildcard to find it, e.g. mitch%, %am or %dam%. See MyStudents Help for more examples.

The takeaways…

- Business Intelligence (BI) systems are a way of adding value to an organisation’s data by supporting management analysis and decision making
- BI systems differ from operational (OLTP) systems as they do not support the primary business activities, instead using extracts of the operational databases together with external purchased data
- BI tools fall into two main categories, reporting and data mining
- A data warehouse usually underpins an organisation’s BI processing

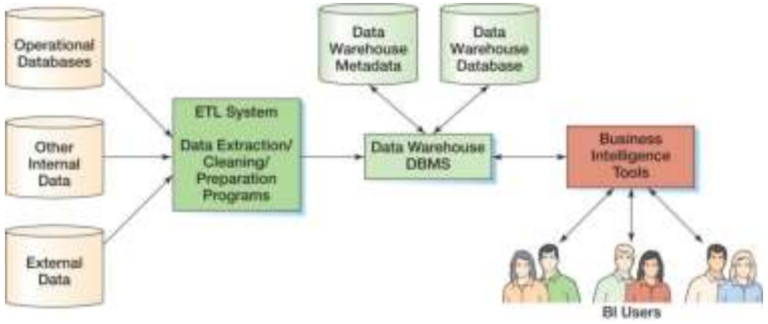
Data for BI systems

Data for BI Systems

Data for BI systems will come from **many different sources**

- Think about a BI system for the University. What data sources could you identify?
- **Internal data/Operational databases →**
  - Student record system
  - Human resources system

  External brought data →
  - Census data for capacity planning
  - Department of education



Problems with integrating operational data

- Dirty data
- Missing values
- Inconsistent data
- Data not integrated
- Wrong format: too fine
- Wrong format: not fine enough
- Too much data: too many attributes
- Too much data: too much volume

Dirty Data

- Data that has errors such as
  - Misspelled, inaccurate, not in range data
  - Not in the domain
  - Nonsense (e.g. made up just to fill a field)

- ~~Note that some definitions (including the textbook) use 'dirty' only to mean erroneous data, while others use it to describe **all** data that is problematic when integrated~~

  - ~~(Note also that this type of 'dirty data' is not the same as we discussed in the context of concurrency management)~~

Sources of dirt…

Poor integration

- The external sources not optimised to be integrated thus resulting in problems arising with e.g. structure of data structures, physical representation and Varying keys.

User requirements changing

Data warehouses that are very old

- Which results in changes with the business structures/rules. This creates dirt. For example

  e.g. Changes in course codes and structures at Murdoch make it problematic to compare enrolment figures in courses over time

More sources of dirt…

~~**Changing expectations of the data warehouse**~~

- ~~Timeliness of analysis, currency of data~~

~~–~~

~~**Unclean legacy systems**~~

- ~~Problems in legacy systems are often only discovered in the integration process~~
- ~~Some fields not being required for data entry, but required for complete analysis~~

Ways to clean dirty data

Clean **in the legacy environment**

- May be an issue because it may disrupt the legacy environment itself and may impact on the applications that access the environment also may be expensive because the environment may not be well known

Clean **at the point of integration:**

- Usually involves the following activities:
  - Transformation of data to a common format
  - Reading the data based on some common model
  - ~~Summation of data to common level of granularity~~
  - ~~Movement of data from disparate environments (DBMS/Operating/hardware)~~

Ways to clean dirty data

Cleaning at and after the point of loading

- Because while inside the system data can still get ‘dirty’
- Activities usually involve:
  - Keeping track of ‘data cleanliness’
  - Clean data when needed

ETL: Extract, Transform, Load

Take data from the operational databases as well as other sources and move them to the BI system through the ETL process (For BI Systems): Components

- EXTRACT
- TRANSFORM
- LOAD

•And eventually

- FEED

ETL: Extraction

- Getting data from different external data sources and operational systems
- Issues with extraction may be:
  - Finding out what data are there
  - Who owns the data (cos Human resources may not allow data to be revealed or may take lots of steps to get their data)

ETL: Transformation

- This is where data that we get is transformed into a common format because the data that we get may not have a common data dictionary. Dirty data is also cleaned here. For example
  - Domains may differ EG- HD -80 to 100 but other uni 85-100
  - Scale may differ
  - Attributes representing the same thing may be named differently
  - Attributes representing different things may be named the same

EG Murdoch – course if a collection of units but other uni course is a unit

ETL: Load

Once data has been cleaned and transformed we load the data into the data warehouse

- Data is able to be loaded as a batch at a specific time
- Data is also able to be loaded in real time from the operational databases

Feed

- Once data is loaded it is able to be feeded to the user of the data

- There are now many business intelligence products that support this; e.g. see list at http://www.capterra.com/business-intelligence-software/

The takeaways…

- Data from BI sources comes from many different sources including operational systems and external data sources
- The data is moved to the BI system through the **ETL** process:
  - **Extract** from the source data systems
  - **Transform** into a common format

    Cleaning (cleansing) problematic or '**dirty data**' is a significant issue here
  - **Load** into the warehouse

Data warehouses and data marts

What is a Data Warehouse?

- A data warehouse usually underpins the organisation’s BI strategy
- A database system with application, data and personnel specialised for BI
- Organisation can use a data from data warehouses rather than data directly from operational systems (REFER TO Next slide for data)

The traditional definition of a data warehouse is that it is a database that is (components/characteristics) :

- *Integrated*
- *Subject-oriented*
- *Time-variant*
- *Non volatile*

…what do all these terms mean?

Problems associated with data warehouses

- Resource intensive such as lots of storage needed
- Difficulty of integration since problems may arise with integration
- Needs to be maintained regularly

  - Organisational reorganisation

- Data ownership
- Long implementation time

Problems associated with data warehouses

- Issues with sources where we get data
  - Fix in the source system or in the data warehouse??
- Greater demands from the end-users
  - I want one too!
- Don’t acquire the data we need for analysis because it may not be captured
- Underestimation of resources for data ETL

Integrated

Integrates data from multiple possibly heterogeneous sources in a centralised, consolidated database

- The real data in real organisations tend to be inconsistent
- Needs to be planned out and organised
- Provides assistance in making decisions and better understanding of strategic opportunities because data is centralised

- Integration allows a consistent view of the data to be presented to the users

Subject-oriented

- A data warehouse is based around data that is summarised and organised by topic
  - e.g. sales, customers, products

- Operational systems are based around processes and functions (transactions)
  - e.g. invoicing, stock control, or student records

Time-variant

- Data warehouses are focused on the movement of data over time
- Operational systems focus on transactions that are processing now
- As new data is uploaded to data warehouses, if we have aggregations they are recalculated

Non-volatile

Once data goes into the data warehouses they tend to stay there

- DW must be able to work with many data (i.e., multi-terabytes)

- The data are not updated in real-time, but refreshed from operational systems on a regular basis

Data Mart

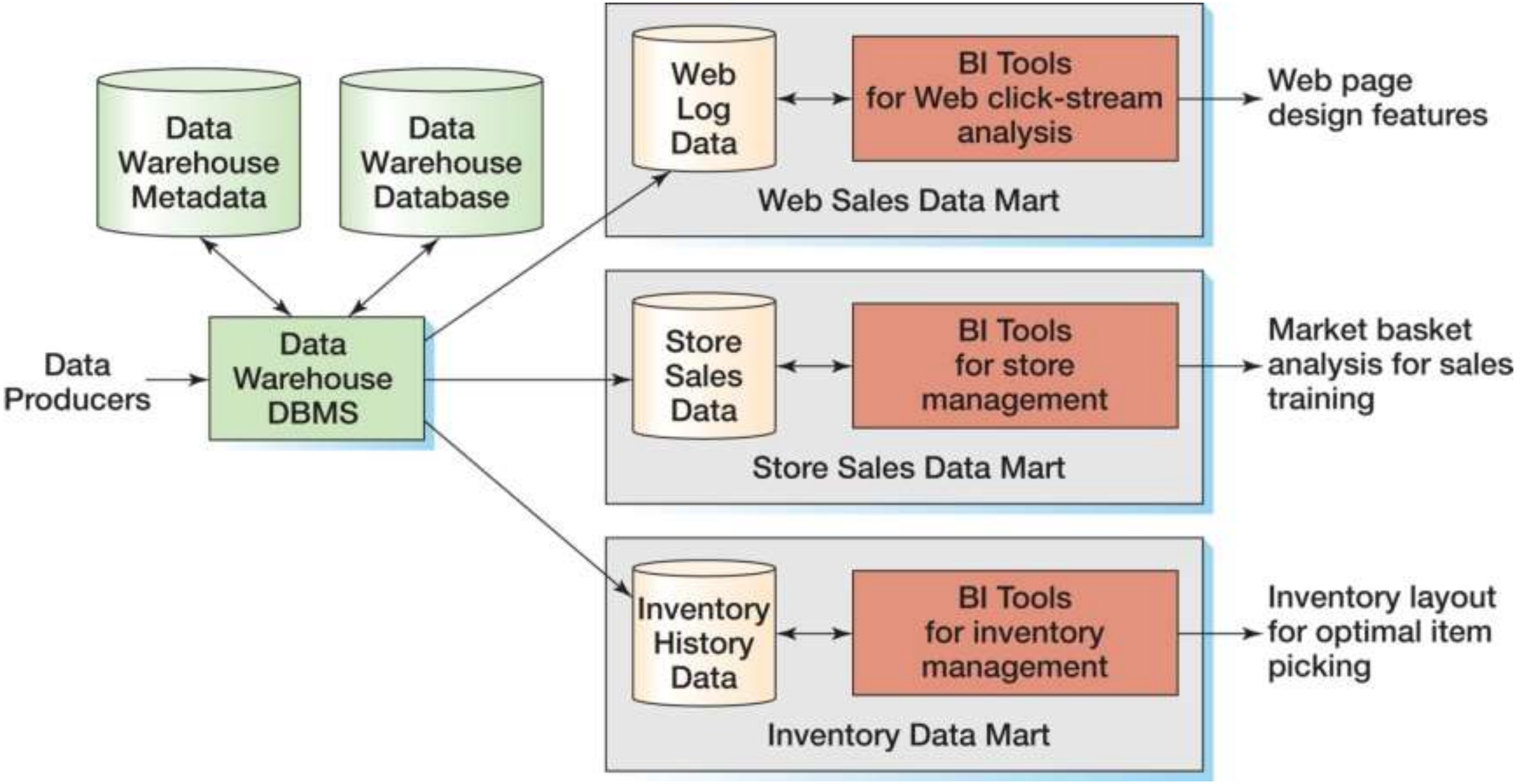
- Smaller than the data warehouse and looks at a specific functional area of business or component

Usually a subset of a data warehouse

Characteristics include:

- Focus on requirements on one part of an organisation or a business function
- Easy to navigate through and understand since smaller in scope
- Doesn’t contain operational data at the same level of detail as with the data warehouse

Enterprise Data Warehouse and Data Marts:



KROENKE AND AUER - DATABASE PROCESSING, 12th Edition © 2012 Pearson Prentice Hall

Reasons for creating a data mart

- Provides better end-user response time since less data needs to be accessed
- Cheaper to build a data mart compared to a data warehouse
- Give people access to data that they need to analyse often
- Give people data in a format that makes sense to their organisation

- Building a data mart is simpler compared with establishing a corporate data warehouse
- To provide data in a form that matches the collective view of the data by a group of users in a department or business function area
- To provide appropriately structured data as dictated by the requirements of the end-user access tools
- The potential users of a data mart are more clearly defined and can be more easily targeted to obtain support for a data mart project rather than a corporate data warehouse project

The takeaways…

- A data warehouse houses the data and systems in preparation for BI processing
- The data warehouse is a centralised, consolidated database that integrates data from multiple sources and systems
- A data mart is a smaller collection of data (perhaps an extract from a larger data warehouse) that addresses a particular component or functional area of the business

Designing a data warehouse

Designing Data Warehouses

~~The requirements collection and analysis stage of a data warehouse project involves:~~

- ~~Interviewing appropriate members of staff (such as marketing users, finance users, and sales users) to enable the identification of a prioritised set of **requirements** that the data warehouse must meet~~
- ~~Interviews with members of staff responsible for operational systems to identify which data sources can **provide** clean, valid, and consistent data that will remain supported over the next few years.~~

Designing Data Warehouses

~~To begin a data warehouse project, need to find answers for questions such as:~~

- ~~Which user requirements are most important?~~
- ~~Which data should be considered first?~~
- ~~Should the project be scaled down into something more manageable?~~
- ~~Should the infrastructure for a scaled down project be capable of ultimately delivering a full-scale enterprise-wide data warehouse?~~

Designing Data Warehouses

- Interviews provide the necessary information for the top-down view (user requirements) and the bottom-up view (which data sources are available) of the data warehouse
- The database component of a data warehouse is then described using a technique called **dimensionality modelling**

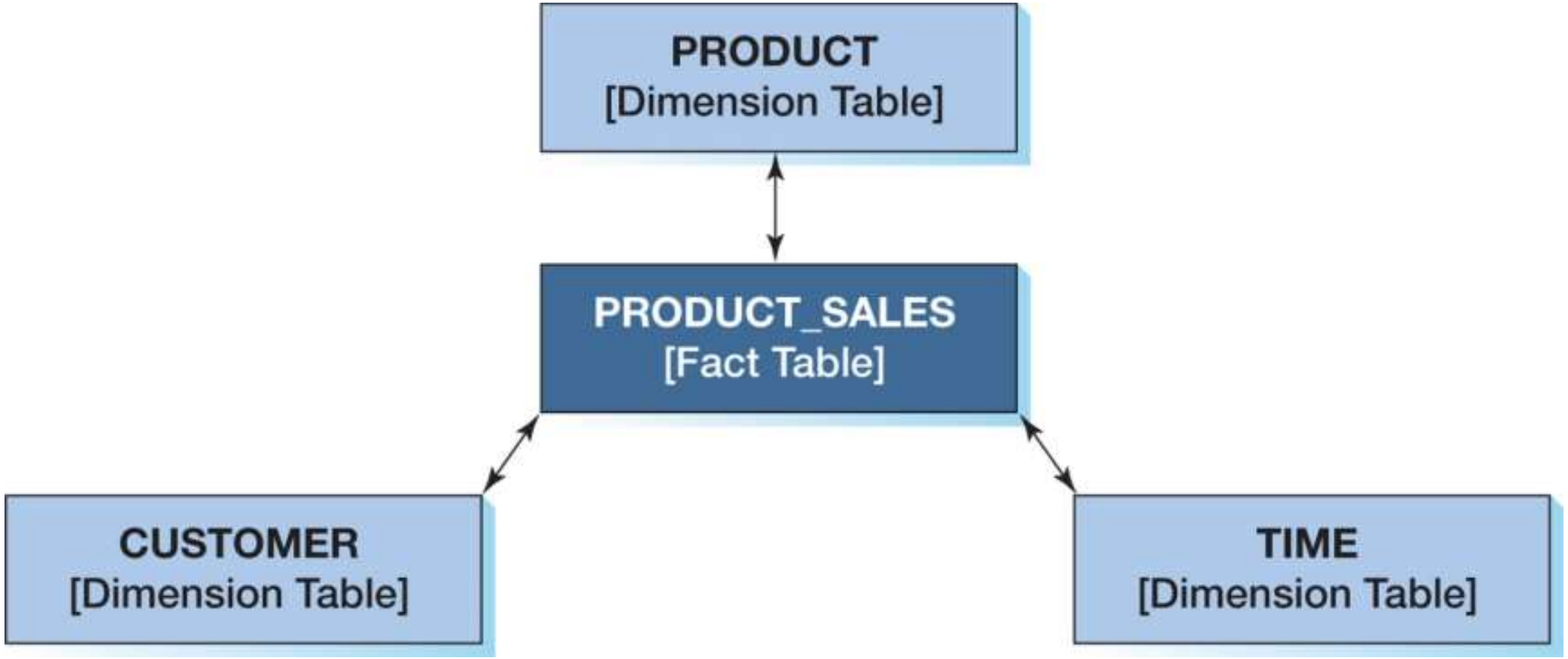
Dimensionality modelling

A logical design technique that aims to present the data in a standard form that allows for high-performance access

Uses the concepts of Entity-Relationship modelling with some important properties:

- Dimensional model consist of one **fact table** (with a composite primary key) and multiple **dimension tables** (smaller tables)
- A dimension table has a simple (non-composite) primary key that matches to foreign key/component of composite of fact table
- The structure is a star and called **star schema**
- **Models** structure of data warehouse

The Star Schema



KROENKE AND AUER - DATABASE PROCESSING, 12th Edition © 2012 Pearson Prentice Hall

Dimensionality modelling schema

- **Star schema** structure has a single centralised fact table consisting of factual data with dimension tables surrounding it containing reference data. Can be denormalised

- **Snowflake schema** is a variant of the star schema where each dimension tables is also normalised
- **Starflake schema** is a hybrid structure that contains a mixture of star (denormalised) and snowflake (normalised) schemas
  - Allows dimensions to be present in both forms to cater for different query requirements.

Dimensionality modelling advantages

The predictable and standard form of the underlying dimensional model offers important advantages:

- Predictable query processing because our queries are similar and accessing similar part of database
- Provides a way to model common business scenarios
- Better efficiency because of the way we built the model
- Ability to handle changing requirements
- Extensibility

Dimensionality modelling

Fact tables (single)

- Most data belonging to data warehouse is in fact tables, which can be massive
- Facts are created by past events
- Attributes are considered read only reference data that will not change over time

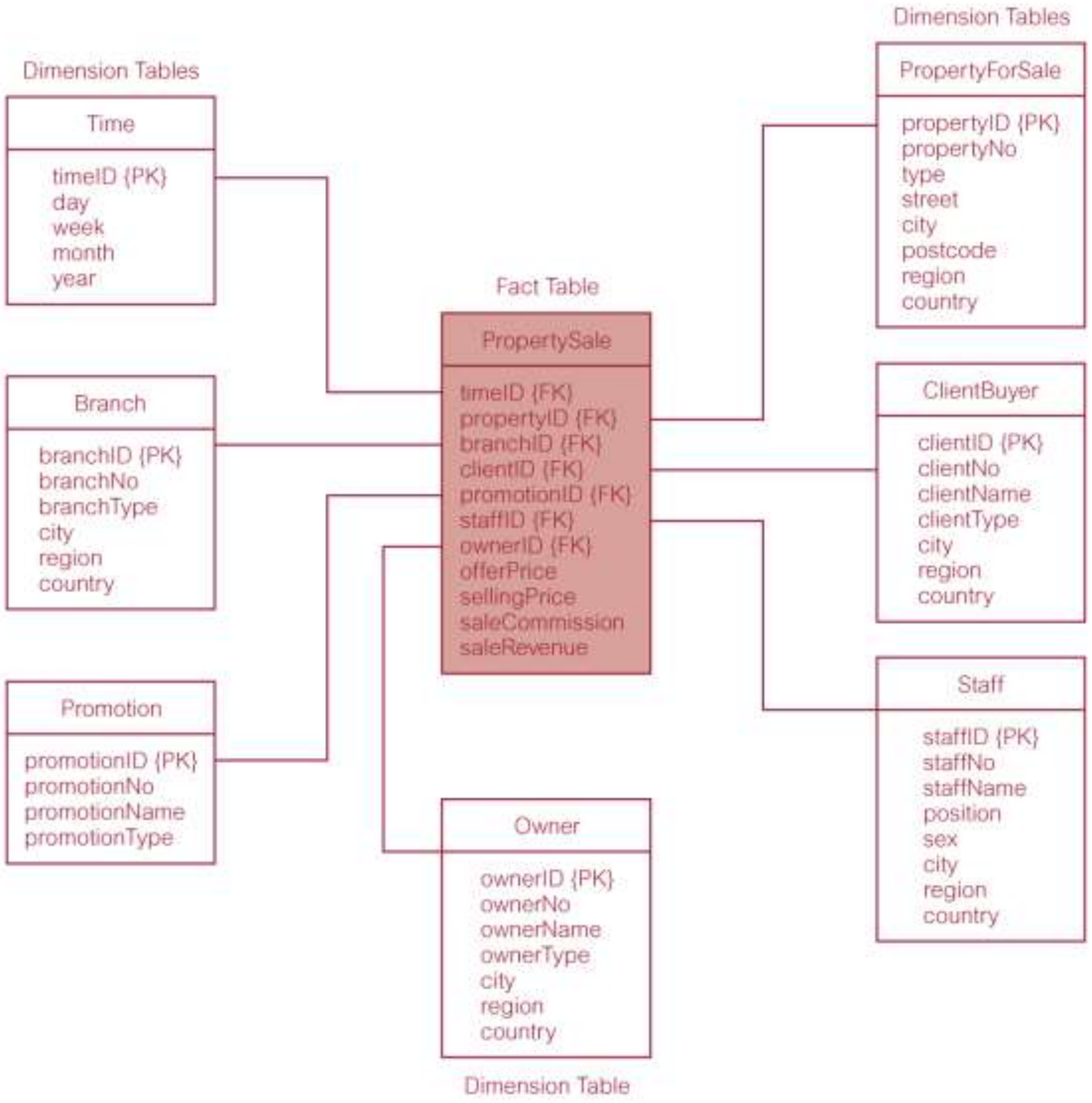
- Most useful fact tables contain one or more numerical measures, or ‘facts’ that occur for each record and are numeric and additive.

Dimension tables (multiple) (think lookup table)

- Dimension tables usually consist of long descriptive text information

- Dimension attributes are used as the constraints in data warehouse queries

Example: Star Schema for Dream Home real estate



Dimension Tables

Time

timeID {PK}
day
week
month
year

Branch

branchID {PK}
branchNo
branchType
city
region
country

Promotion

promotionID {PK}
promotionNo
promotionName
promotionType

Fact Table

PropertySale

timeID {FK}
propertyID {FK}
branchID {FK}
clientID {FK}
promotionID {FK}
staffID {FK}
ownerID {FK}
offerPrice
sellingPrice
saleCommission
saleRevenue

Owner

ownerID {PK}
ownerNo
ownerName
ownerType
city
region
country

Dimension Table

Dimension Tables

PropertyForSale

propertyID {PK}
propertyNo
type
street
city
postcode
region
country

ClientBuyer

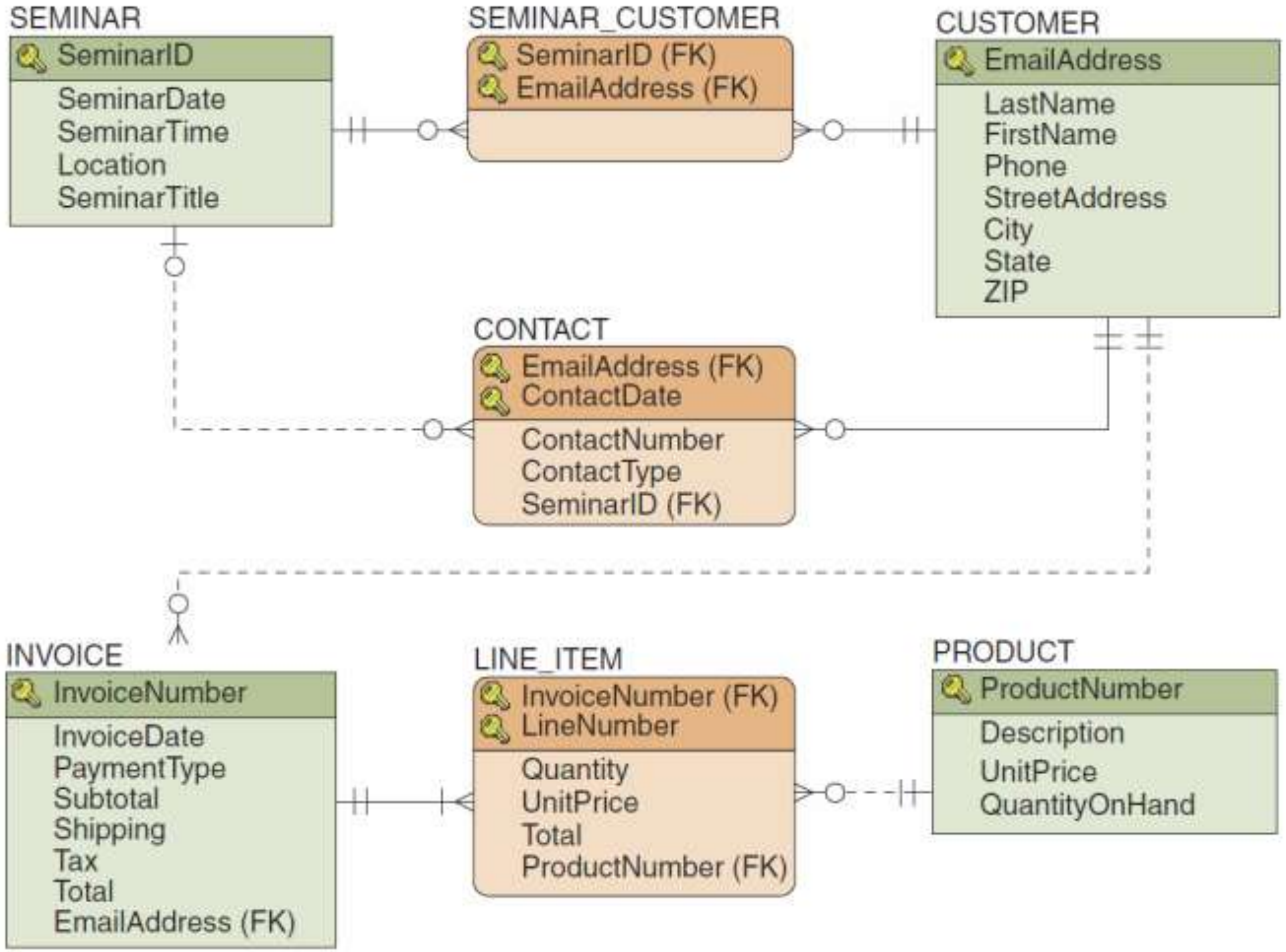
clientID {PK}
clientNo
clientName
clientType
city
region
country

Staff

staffID {PK}
staffNo
staffName
position
sex
city
region
country

C&B: Fig 32.1 p. 1184

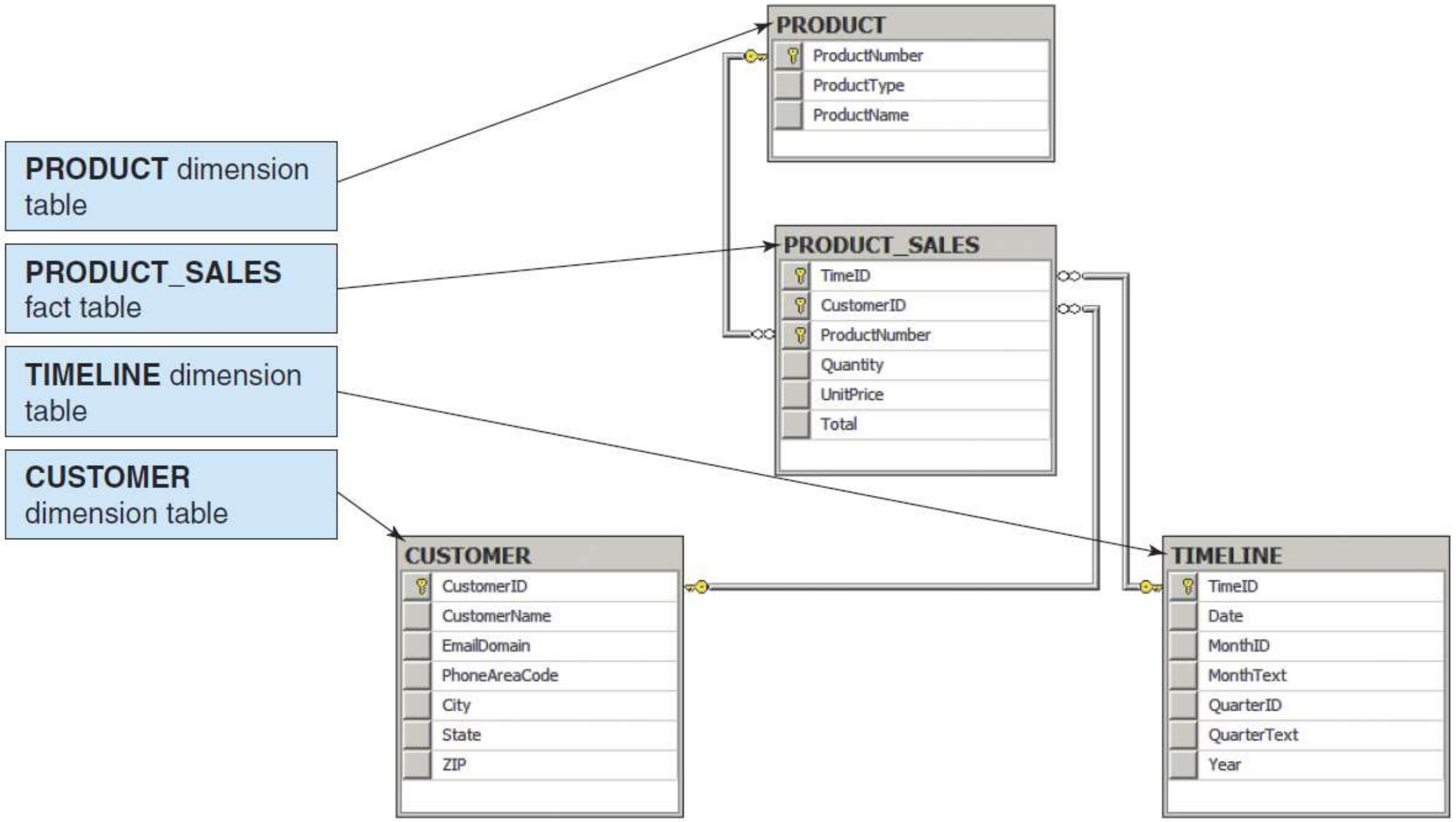
The HSD Database Design (textbook example)



This is the design for the operational database

KROENKE AND AUER - DATABASE PROCESSING, 12th Edition © 2012 Pearson Prentice Hall

The HSD-DW Star Schema



PRODUCT dimension table

PRODUCT_SALES fact table

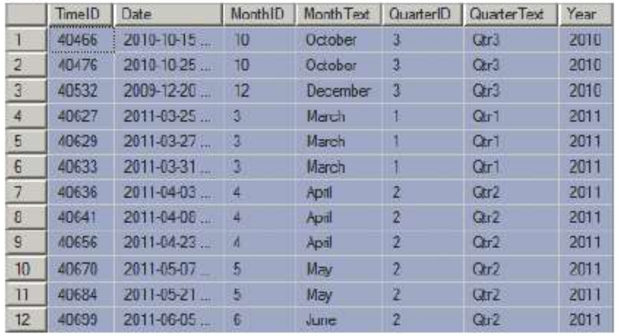
TIMELINE dimension table

CUSTOMER dimension table

KROENKE AND AUER - DATABASE PROCESSING, 12th Edition © 2012 Pearson Prentice Hall

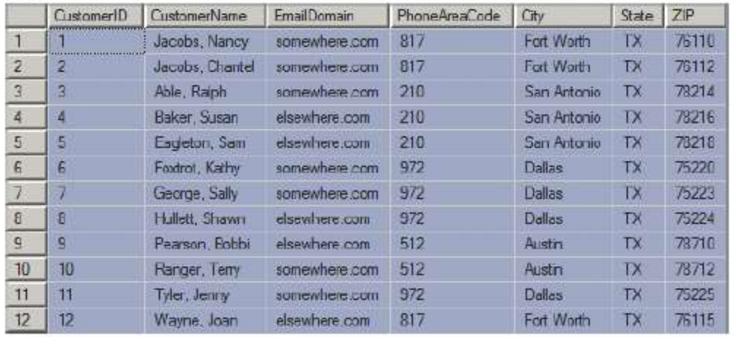
The HSD-DW Table Data

(a) TIMELINE Dimension Table



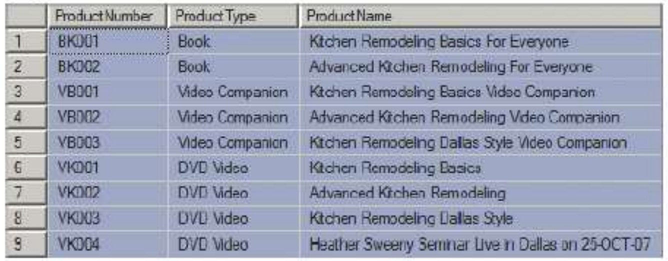
	TimeID	Date	MonthID	MonthText	QuarterID	QuarterText	Year
1	40466	2010-10-15 ...	10	October	3	Qtr3	2010
2	40476	2010-10-25 ...	10	October	3	Qtr3	2010
3	40532	2009-12-20 ...	12	December	3	Qtr3	2010
4	40627	2011-03-25 ...	3	March	1	Qtr1	2011
5	40629	2011-03-27 ...	3	March	1	Qtr1	2011
6	40633	2011-03-31 ...	3	March	1	Qtr1	2011
7	40636	2011-04-03 ...	4	April	2	Qtr2	2011
8	40641	2011-04-08 ...	4	April	2	Qtr2	2011
9	40656	2011-04-23 ...	4	April	2	Qtr2	2011
10	40670	2011-05-07 ...	5	May	2	Qtr2	2011
11	40684	2011-05-21 ...	5	May	2	Qtr2	2011
12	40699	2011-06-05 ...	6	June	2	Qtr2	2011

(b) CUSTOMER Dimension Table



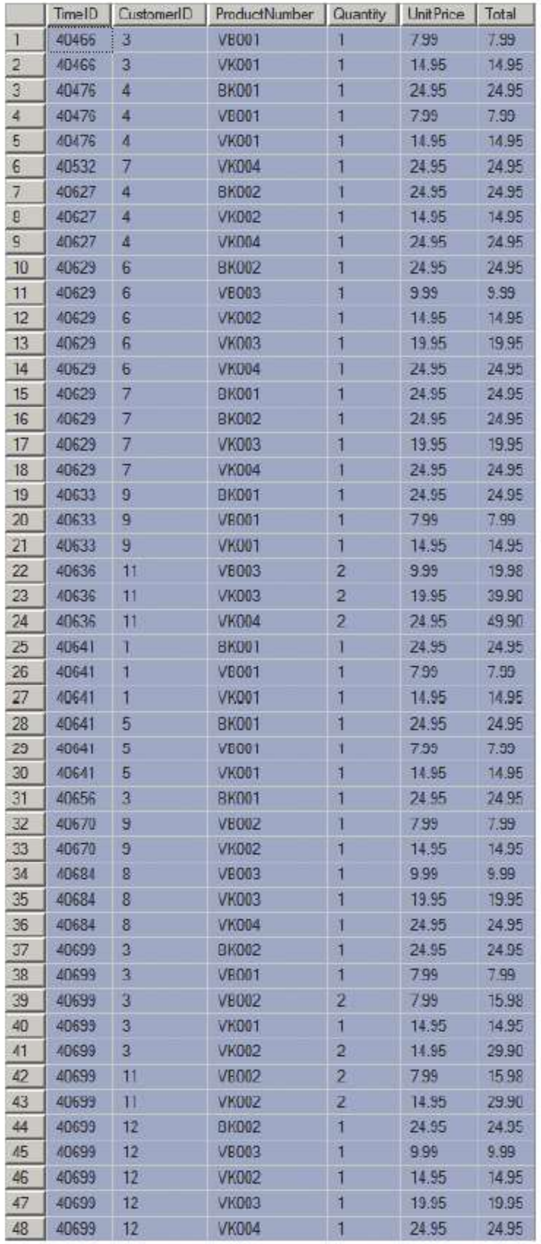
	CustomerID	CustomerName	EmailDomain	PhoneAreaCode	City	State	ZIP
1	1	Jacobs, Nancy	somewhere.com	817	Fort Worth	TX	76110
2	2	Jacobs, Chantel	somewhere.com	817	Fort Worth	TX	76112
3	3	Able, Ralph	somewhere.com	210	San Antonio	TX	78214
4	4	Baker, Susan	elsewhere.com	210	San Antonio	TX	78216
5	5	Eagleton, Sam	elsewhere.com	210	San Antonio	TX	78218
6	6	Foxtrot, Kathy	somewhere.com	972	Dallas	TX	75220
7	7	George, Sally	somewhere.com	972	Dallas	TX	75223
8	8	Hullett, Shawn	elsewhere.com	972	Dallas	TX	75224
9	9	Pearson, Bobbi	elsewhere.com	512	Austin	TX	78710
10	10	Ranger, Terry	somewhere.com	512	Austin	TX	78712
11	11	Tyler, Jenny	somewhere.com	972	Dallas	TX	75225
12	12	Wayne, Joan	elsewhere.com	817	Fort Worth	TX	76115

(c) PRODUCT Dimension Table



	ProductNumber	ProductType	ProductName
1	BK001	Book	Kitchen Remodeling Basics For Everyone
2	BK002	Book	Advanced Kitchen Remodeling For Everyone
3	VB001	Video Companion	Kitchen Remodeling Basics Video Companion
4	VB002	Video Companion	Advanced Kitchen Remodeling Video Companion
5	VB003	Video Companion	Kitchen Remodeling Dallas Style Video Companion
6	VK001	DVD Video	Kitchen Remodeling Basics
7	VK002	DVD Video	Advanced Kitchen Remodeling
8	VK003	DVD Video	Kitchen Remodeling Dallas Style
9	VK004	DVD Video	Heather Sweeny Seminar Live in Dallas on 25-OCT-07

(d) PRODUCT_SALES Fact Table



	TimeID	CustomerID	ProductNumber	Quantity	UnitPrice	Total
1	40466	3	VB001	1	7.99	7.99
2	40466	3	VK001	1	14.95	14.95
3	40476	4	BK001	1	24.95	24.95
4	40476	4	VB001	1	7.99	7.99
5	40476	4	VK001	1	14.95	14.95
6	40532	7	VK004	1	24.95	24.95
7	40627	4	BK002	1	24.95	24.95
8	40627	4	VK002	1	14.95	14.95
9	40627	4	VK004	1	24.95	24.95
10	40629	6	BK002	1	24.95	24.95
11	40629	6	VB003	1	9.99	9.99
12	40629	6	VK002	1	14.95	14.95
13	40629	6	VK003	1	19.95	19.95
14	40629	6	VK004	1	24.95	24.95
15	40629	7	BK001	1	24.95	24.95
16	40629	7	BK002	1	24.95	24.95
17	40629	7	VK003	1	19.95	19.95
18	40629	7	VK004	1	24.95	24.95
19	40633	9	BK001	1	24.95	24.95
20	40633	9	VB001	1	7.99	7.99
21	40633	9	VK001	1	14.95	14.95
22	40636	11	VB003	2	9.99	19.98
23	40636	11	VK003	2	19.95	39.90
24	40636	11	VK004	2	24.95	49.90
25	40641	1	BK001	1	24.95	24.95
26	40641	1	VB001	1	7.99	7.99
27	40641	1	VK001	1	14.95	14.95
28	40641	5	BK001	1	24.95	24.95
29	40641	5	VB001	1	7.99	7.99
30	40641	5	VK001	1	14.95	14.95
31	40656	3	BK001	1	24.95	24.95
32	40670	9	VB002	1	7.99	7.99
33	40670	9	VK002	1	14.95	14.95
34	40684	8	VB003	1	9.99	9.99
35	40684	8	VK003	1	19.95	19.95
36	40684	8	VK004	1	24.95	24.95
37	40699	3	BK002	1	24.95	24.95
38	40699	3	VB001	1	7.99	7.99
39	40699	3	VB002	2	7.99	15.98
40	40699	3	VK001	1	14.95	14.95
41	40699	3	VK002	2	14.95	29.90
42	40699	11	VB002	2	7.99	15.98
43	40699	11	VK002	2	14.95	29.90
44	40699	12	BK002	1	24.95	24.95
45	40699	12	VB003	1	9.99	9.99
46	40699	12	VK002	1	14.95	14.95
47	40699	12	VK003	1	19.95	19.95
48	40699	12	VK004	1	24.95	24.95

KROENKE AND AUER - DATABASE PROCESSING, 12th Edition © 2012 Pearson Prentice Hall

Using a dimensional database

- ~~**Reporting systems** analyse the dimensional data from the data warehouse using fairly simple operations such as sorting, filtering, grouping etc~~
  - ~~RFM analysis~~
  - ~~OLAP~~
- ~~**Data mining systems** use more advanced statistical and mathematical techniques for what-if analysis and prediction~~

The takeaways…

- The structure of a data warehouse is unlike that of a normalised transactional database
- Instead, the basic *dimensional model* is of a single **Fact** table with multiple **Dimensions** or descriptors
- Dimensional models include the Star Schema and variations
- The advantages of dimensional modelling for data warehouses include efficiency and flexibility

Reporting systems RFM and OLAP

RFM Analysis

- RFM analysis is about of analysing and ranking customers based on their purchasing trends. Enables organisations to target each customer appropriately (marketing tool) It considers:
  - - How **recently** (**R** score) a customer ordered;
  - - How **frequently** (**F** score) they order;
  - - How much **money** (**M** score) they spend per order

- ~~Typically scores are from 1(highest)-5 and the RFM for a particular customer would be written e.g. {2,4,3}~~
- ~~*Note that most of these slides are from those supplied with the text book. You can follow through the example there.*~~

OnLine Analytical Processing (OLAP)

- OLAP allows you to do aggregation on the data to generate OLAP reports or OLAP cubes

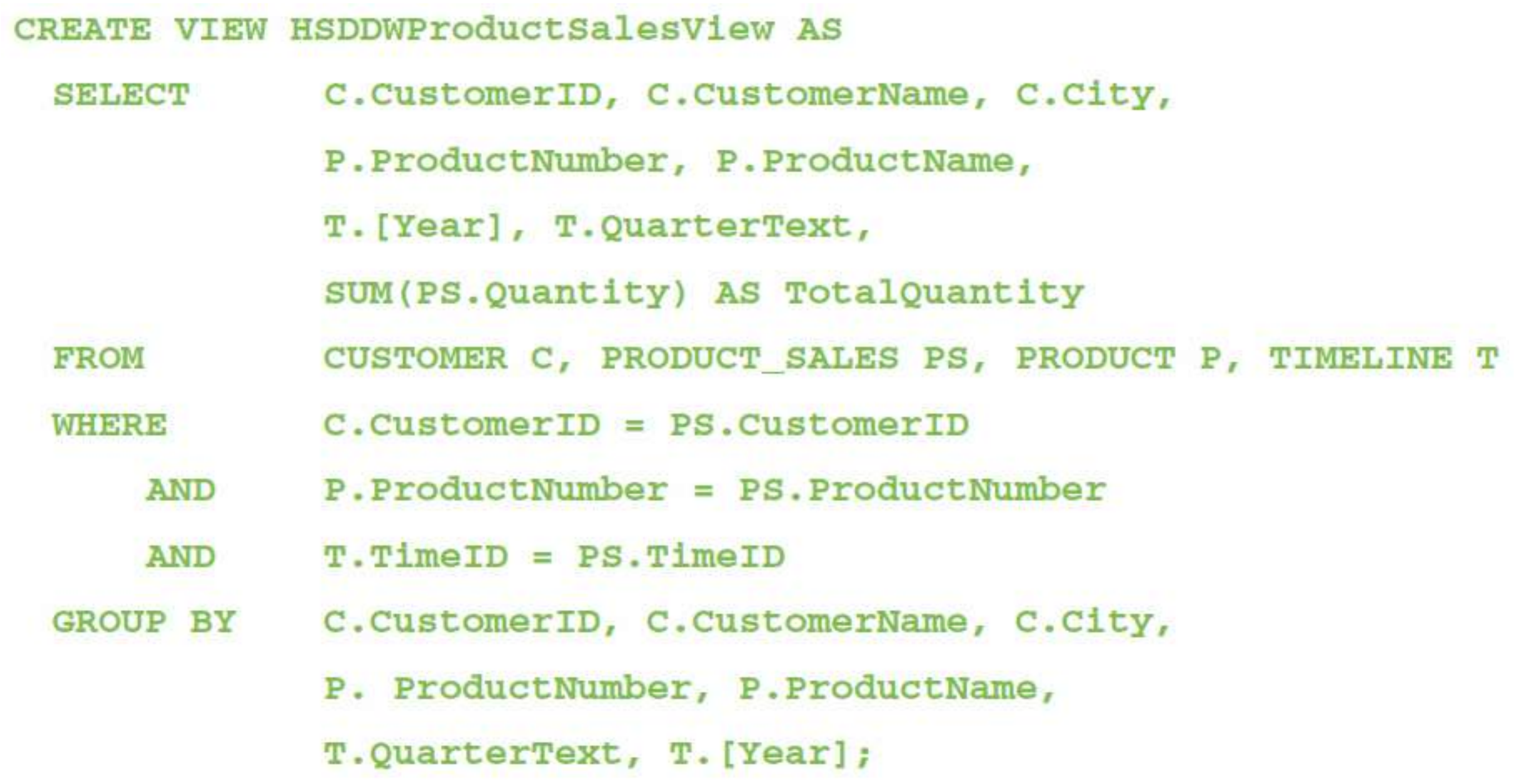
- An **OLAP report** consist of:
  - **Measure** - a data item of interest
  - **Dimension** - a feature of data item

- **OLAP cube**- a presentation of a measure with associated dimensions.
  - An OLAP cube can have *any* number of axes.
  - The terms OLAP cube and OLAP report are synonymous
- OLAP allows **drill-down** - a further division of the data into more detail
- OLAP reports can often be displayed effectively using an Excel **pivot table**

- ~~OLAP uses the dimensional model where the *measure* is the fact that is summed (etc) in the OLAP report~~

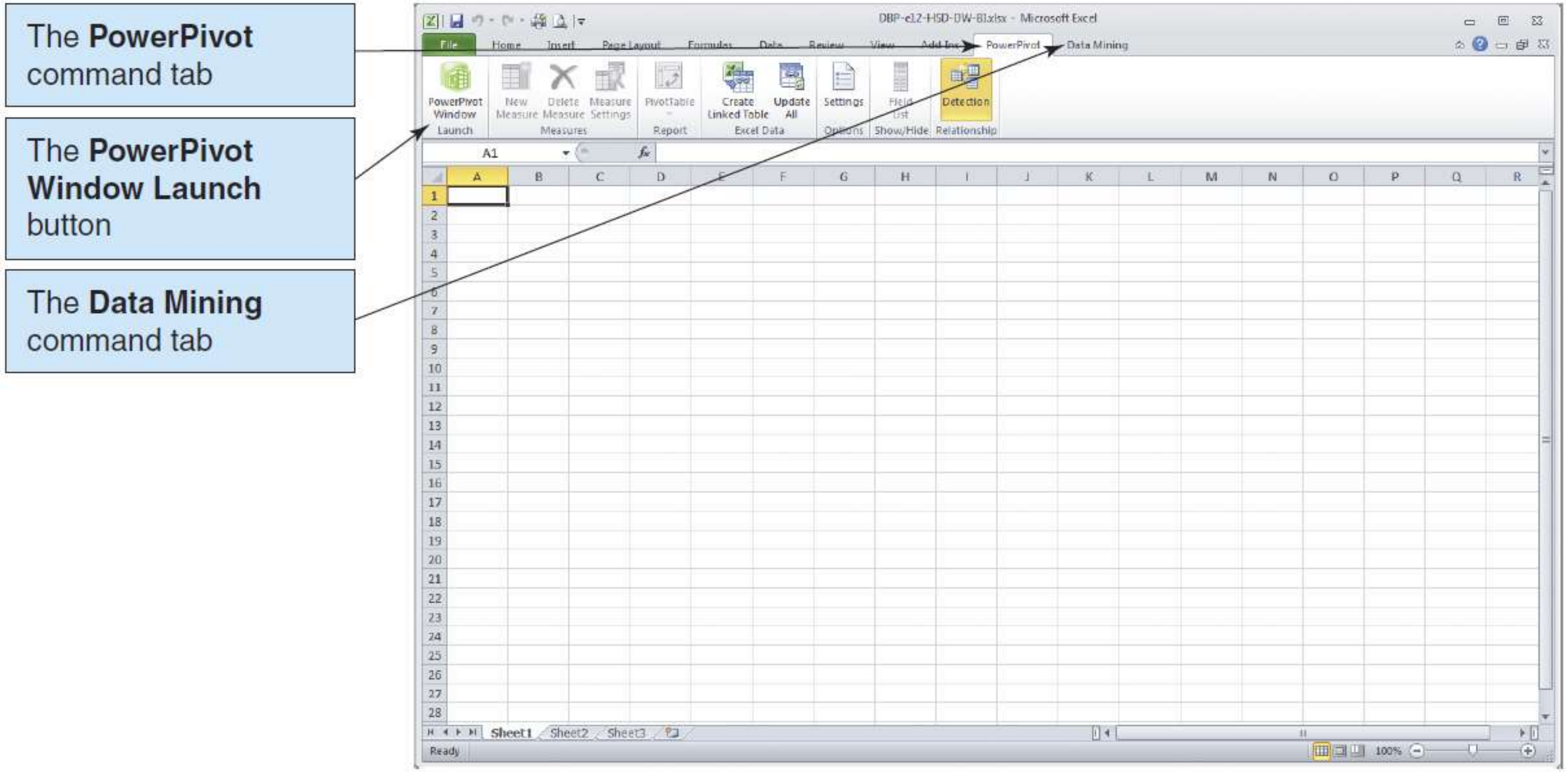
- ~~The *dimension* is a characteristic of the measure or fact, such as the time dimension~~

Reporting Systems:
OLAP Reports I



```
CREATE VIEW HSDDWProductSalesView AS
  SELECT      C.CustomerID, C.CustomerName, C.City,
              P.ProductNumber, P.ProductName,
              T.[Year], T.QuarterText,
              SUM(PS.Quantity) AS TotalQuantity
  FROM        CUSTOMER C, PRODUCT_SALES PS, PRODUCT P, TIMELINE T
  WHERE       C.CustomerID = PS.CustomerID
      AND     P.ProductNumber = PS.ProductNumber
      AND     T.TimeID = PS.TimeID
  GROUP BY    C.CustomerID, C.CustomerName, C.City,
              P. ProductNumber, P.ProductName,
              T.QuarterText, T.[Year];
```

Reporting Systems:
OLAP Reports II



The **PowerPivot** command tab

The **PowerPivot Window Launch** button

The **Data Mining** command tab

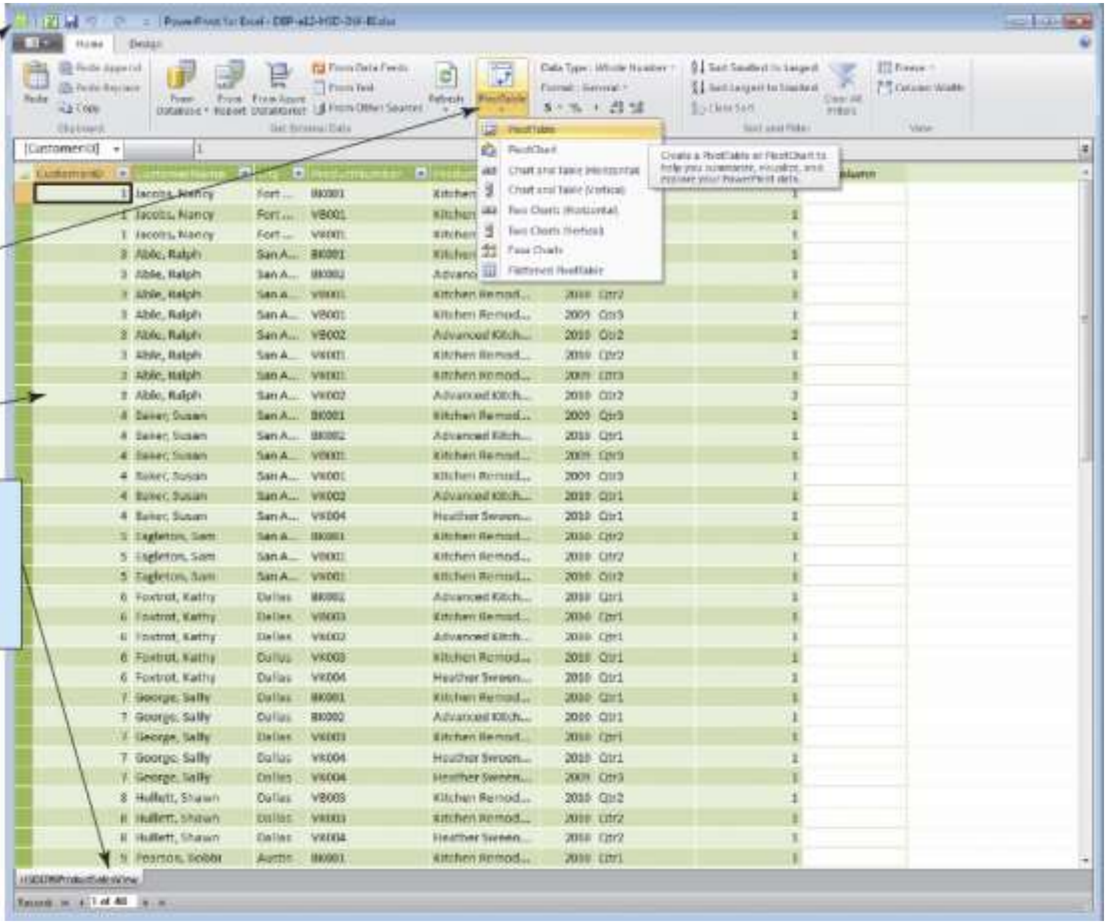
Reporting Systems:
OLAP Reports III

The **PivotTable for Excel** window for the DBP-e12-HSD-DW-BI.xlsx workbook

The **PivotTable** button showing the various options for displaying the data

The PowerPivot data table

The data table is based on the **HSDDWProductSalesView** in the HSD-DW database



Create a PivotTable or PivotChart to help you summarize, visualize, and explore your PowerPivot data.

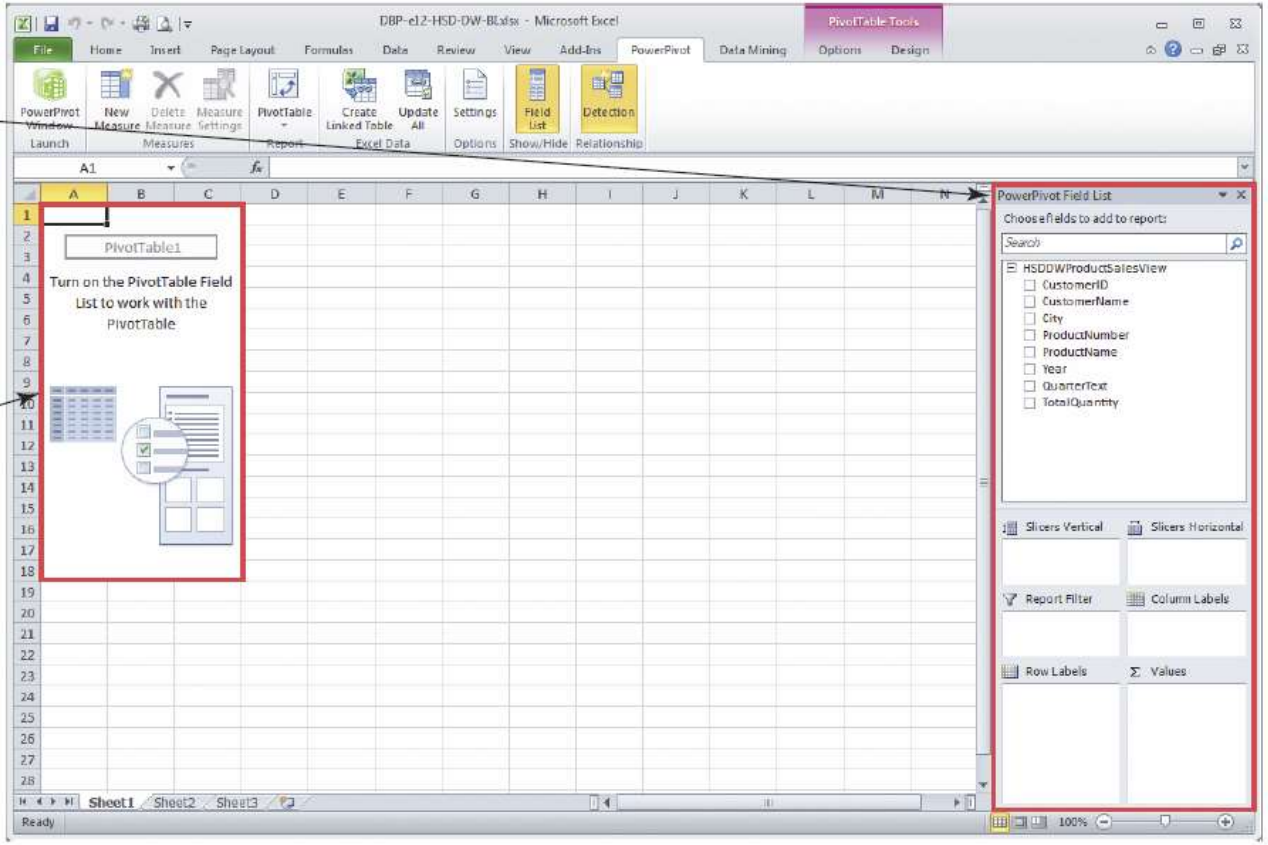
[CustomerID] 1

CustomerID	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1	Jacobs, Nancy	Fort ...	BK001	Kitchen...	[illegible]	[illegible]	[illegible]
1	Jacobs, Nancy	Fort ...	VB001	Kitchen...	[illegible]	[illegible]	1
1	Jacobs, Nancy	Fort ...	VK001	Kitchen...	[illegible]	[illegible]	1
3	Able, Ralph	San A...	BK001	Kitchen...	[illegible]	[illegible]	1
3	Able, Ralph	San A...	BK002	Advanc...	[illegible]	[illegible]	1
3	Able, Ralph	San A...	VK001	Kitchen Remod...	2010	Qtr2	1
3	Able, Ralph	San A...	VB001	Kitchen Remod...	2009	Qtr3	1
3	Able, Ralph	San A...	VB002	Advanced Kitch...	2010	Qtr2	2
3	Able, Ralph	San A...	VK001	Kitchen Remod...	2010	Qtr2	1
3	Able, Ralph	San A...	VK001	Kitchen Remod...	2009	Qtr3	1
3	Able, Ralph	San A...	VK002	Advanced Kitch...	2010	Qtr2	3
4	Baker, Susan	San A...	BK001	Kitchen Remod...	2009	Qtr3	1
4	Baker, Susan	San A...	BK002	Advanced Kitch...	2010	Qtr1	1
4	Baker, Susan	San A...	VB001	Kitchen Remod...	2009	Qtr3	1
4	Baker, Susan	San A...	VK001	Kitchen Remod...	2009	Qtr3	1
4	Baker, Susan	San A...	VK002	Advanced Kitch...	2010	Qtr1	1
4	Baker, Susan	San A...	VK004	Heather Sweeney...	2010	Qtr1	1
5	Eagleton, Sam	San A...	BK001	Kitchen Remod...	2010	Qtr2	1
5	Eagleton, Sam	San A...	VB001	Kitchen Remod...	2010	Qtr2	1
5	Eagleton, Sam	San A...	VK001	Kitchen Remod...	2010	Qtr2	1
6	Foxtrot, Kathy	Dallas	BK002	Advanced Kitch...	2010	Qtr1	1
6	Foxtrot, Kathy	Dallas	VB003	Kitchen Remod...	2010	Qtr1	1
6	Foxtrot, Kathy	Dallas	VK002	Advanced Kitch...	2010	Qtr1	1
6	Foxtrot, Kathy	Dallas	VK003	Kitchen Remod...	2010	Qtr1	1
6	Foxtrot, Kathy	Dallas	VK004	Heather Sweeney...	2010	Qtr1	1
7	George, Sally	Dallas	BK001	Kitchen Remod...	2010	Qtr1	1
7	George, Sally	Dallas	BK002	Advanced Kitch...	2010	Qtr1	1
7	George, Sally	Dallas	VK003	Kitchen Remod...	2010	Qtr1	1
7	George, Sally	Dallas	VK004	Heather Sweeney...	2010	Qtr1	1
7	George, Sally	Dallas	VK004	Heather Sweeney...	2009	Qtr3	1
8	Hullett, Shawn	Dallas	VB003	Kitchen Remod...	2010	Qtr2	1
8	Hullett, Shawn	Dallas	VK003	Kitchen Remod...	2010	Qtr2	1
8	Hullett, Shawn	Dallas	VK004	Heather Sweeney...	2010	Qtr2	1
9	Pearson, Bobbi	Austin	BK001	Kitchen Remod...	2010	Qtr1	1

Reporting Systems:
OLAP Reports IV

The **PowerPivot Field List** pane—select the report elements to be displayed here

The PivotTable report area—the PivotTable will be displayed in this area, which can be expanded as necessary to accomodate the PivotTable



Turn on the PivotTable Field List to work with the PivotTable

PowerPivot Field List

Choose fields to add to report:

Search

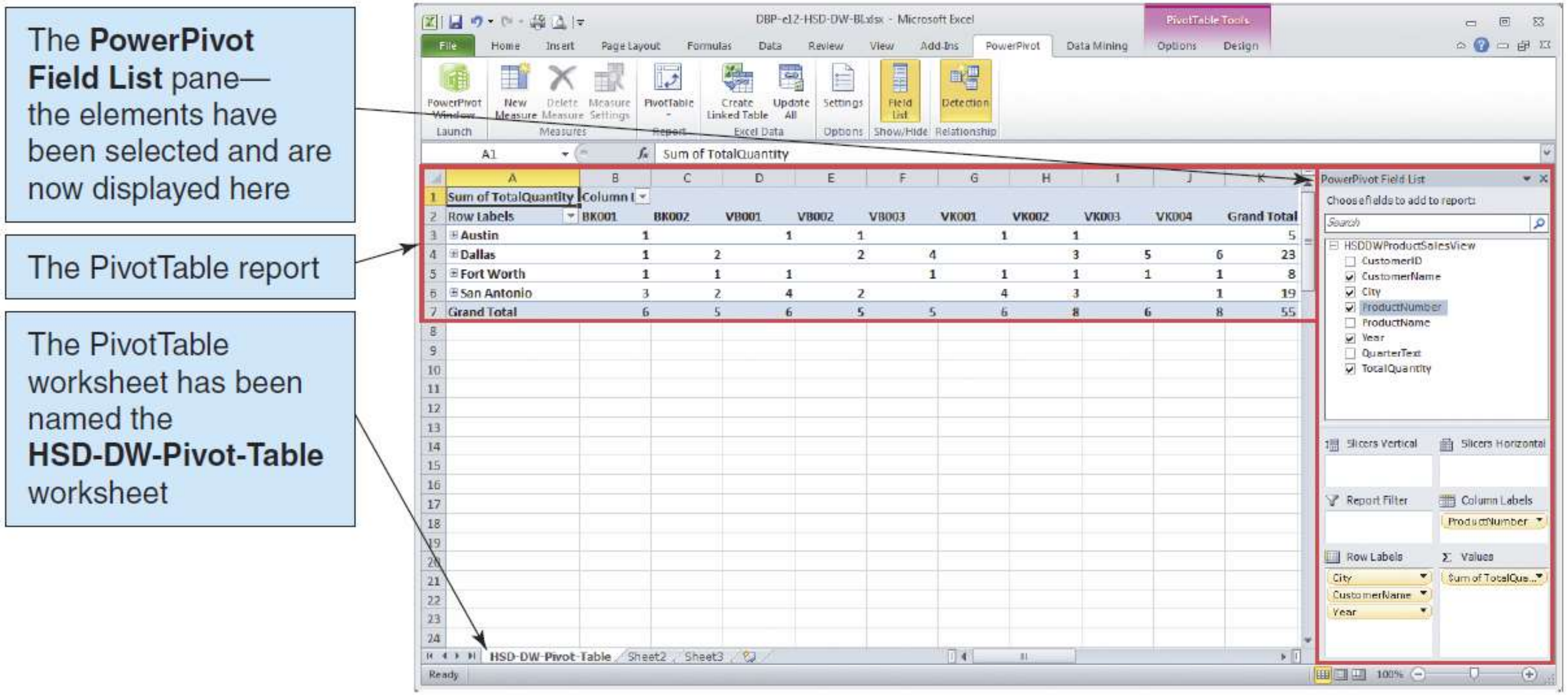
⊟ HSDDWProductSalesView
- ☐ CustomerID
- ☐ CustomerName
- ☐ City
- ☐ ProductNumber
- ☐ ProductName
- ☐ Year
- ☐ QuarterText
- ☐ TotalQuantity

Slicers Vertical | Slicers Horizontal

Report Filter | Column Labels

Row Labels | Σ Values

Reporting Systems:
OLAP Reports V



The **PowerPivot Field List** pane—the elements have been selected and are now displayed here

The PivotTable report

The PivotTable worksheet has been named the **HSD-DW-Pivot-Table** worksheet

Sum of TotalQuantity

Sum of TotalQuantity	Column L									
Row Labels	BK001	BK002	VB001	VB002	VB003	VK001	VK002	VK003	VK004	Grand Total
⊞ Austin	1		1	1		1	1			5
⊞ Dallas	1	2		2	4		3	5	6	23
⊞ Fort Worth	1	1	1		1	1	1	1	1	8
⊞ San Antonio	3	2	4	2		4	3		1	19
Grand Total	6	5	6	5	5	6	8	6	8	55

HSD-DW-Pivot-Table | Sheet2 | Sheet3

PowerPivot Field List

Choose fields to add to report:

Search

⊟ HSDDWProductSalesView
- □ CustomerID
- ☑ CustomerName
- ☑ City
- ☑ ProductNumber
- □ ProductName
- ☑ Year
- □ QuarterText
- ☑ TotalQuantity

Slicers Vertical

Slicers Horizontal

Report Filter

Column Labels: ProductNumber

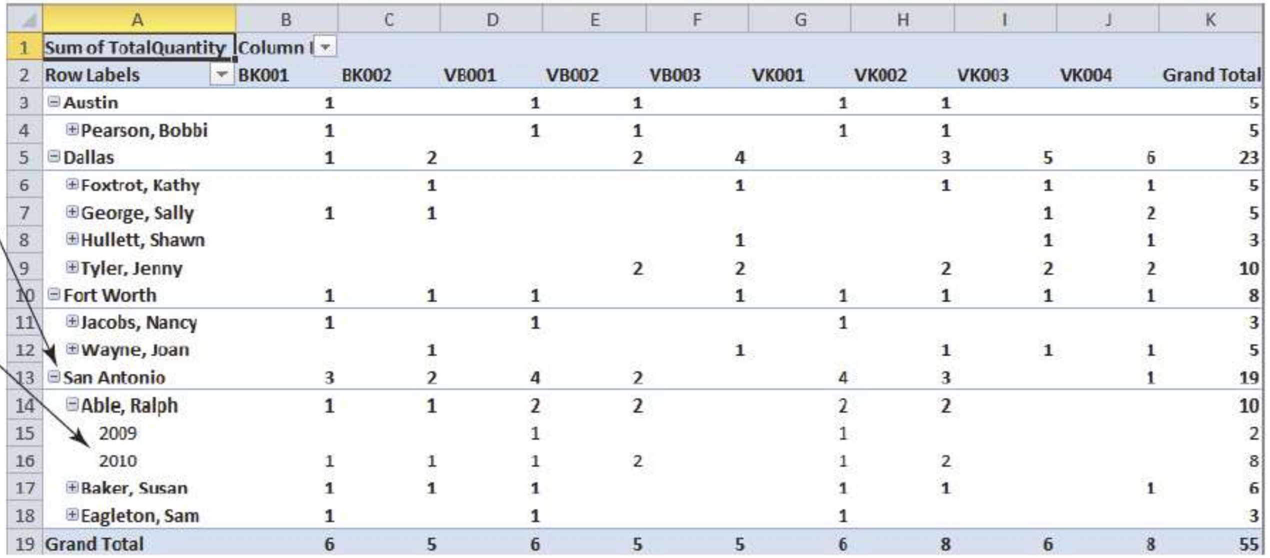
Row Labels: City, CustomerName, Year

Values: Sum of TotalQua...

Reporting Systems:
OLAP Drill Down I

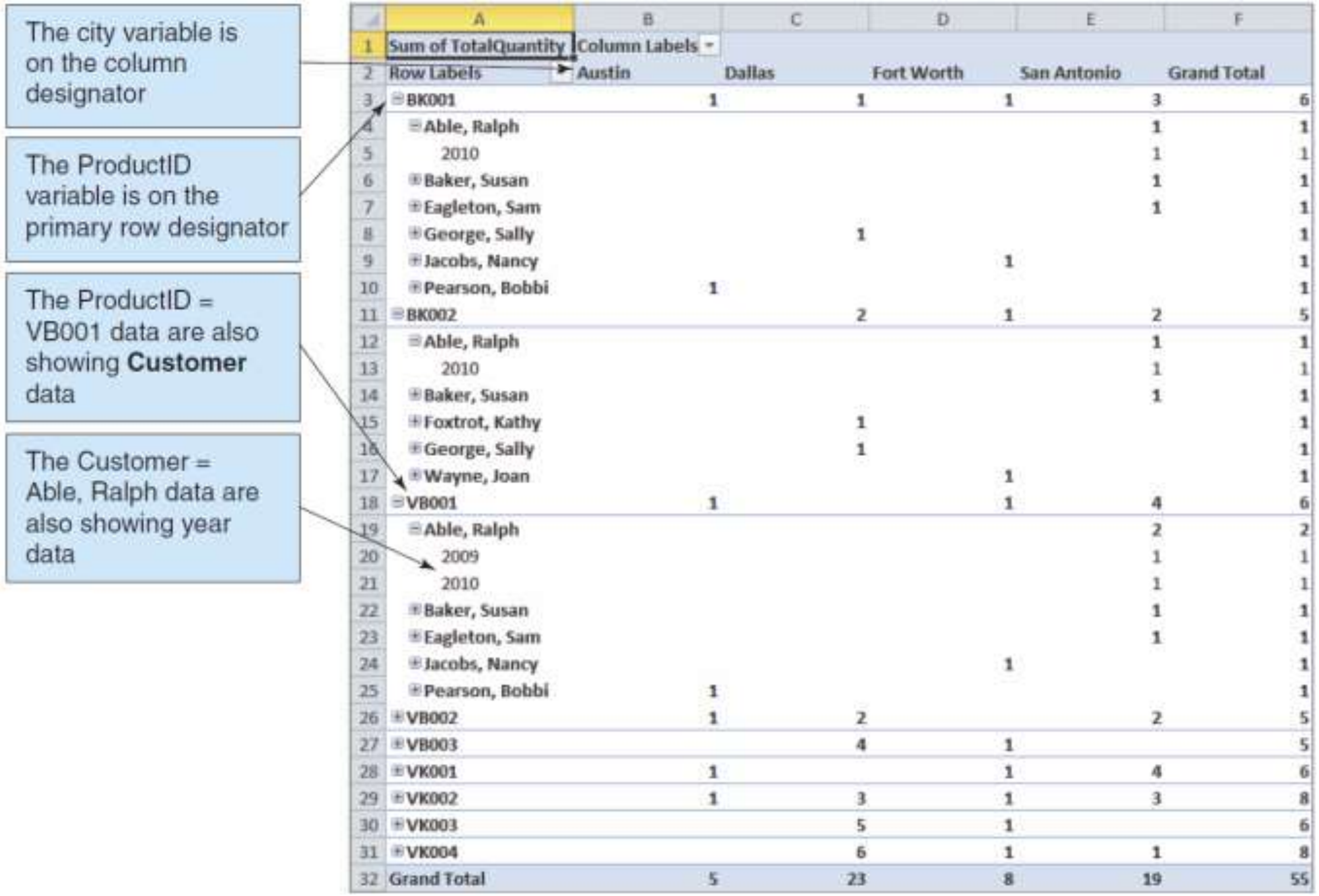
The City = San Antonio data are also showing customer data

The Customer = Able, Ralph data are also showing year data



	A	B	C	D	E	F	G	H	I	J	K
1	Sum of TotalQuantity	Column l									
2	Row Labels	BK001	BK002	VB001	VB002	VB003	VK001	VK002	VK003	VK004	Grand Total
3	⊟Austin	1		1	1		1	1			5
4	⊞Pearson, Bobbi	1		1	1		1	1			5
5	⊟Dallas	1	2		2	4		3	5	6	23
6	⊞Foxtrot, Kathy		1			1		1	1	1	5
7	⊞George, Sally	1	1						1	2	5
8	⊞Hullett, Shawn					1			1	1	3
9	⊞Tyler, Jenny				2	2		2	2	2	10
10	⊟Fort Worth	1	1	1		1	1	1	1	1	8
11	⊞Jacobs, Nancy	1		1			1				3
12	⊞Wayne, Joan		1			1		1	1	1	5
13	⊟San Antonio	3	2	4	2		4	3		1	19
14	⊟Able, Ralph	1	1	2	2		2	2			10
15	2009			1			1				2
16	2010	1	1	1	2		1	2			8
17	⊞Baker, Susan	1	1	1			1	1		1	6
18	⊞Eagleton, Sam	1		1			1				3
19	Grand Total	6	5	6	5	5	6	8	6	8	55

Reporting Systems:
OLAP Drill Down II



The city variable is on the column designator

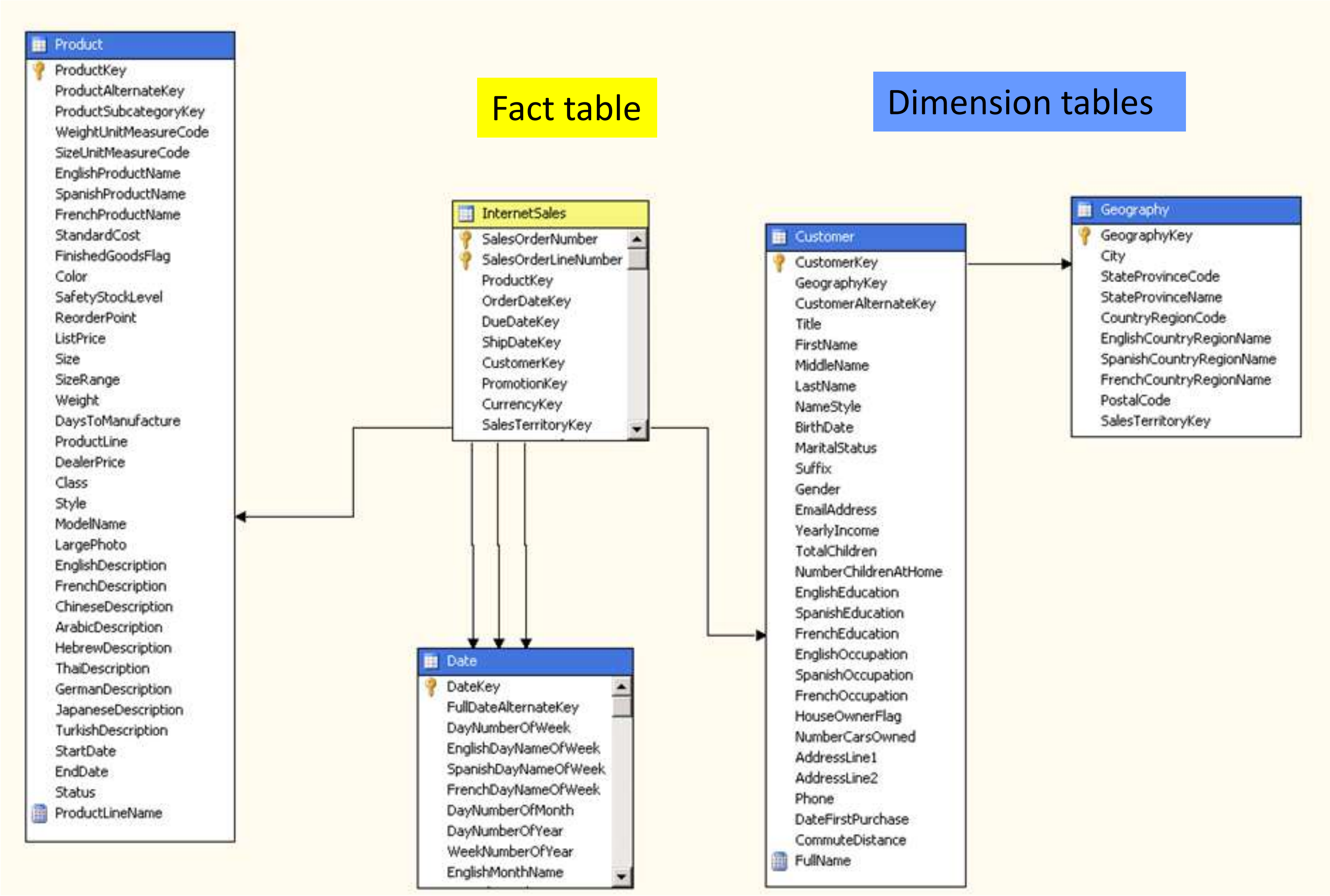
The ProductID variable is on the primary row designator

The ProductID = VB001 data are also showing **Customer** data

The Customer = Able, Ralph data are also showing year data

	A	B	C	D	E	F
1	Sum of TotalQuantity	Column Labels				
2	Row Labels	Austin	Dallas	Fort Worth	San Antonio	Grand Total
3	⊟BK001	1	1	1	3	6
4	⊟Able, Ralph				1	1
5	2010				1	1
6	⊞Baker, Susan				1	1
7	⊞Eagleton, Sam				1	1
8	⊞George, Sally		1			1
9	⊞Jacobs, Nancy			1		1
10	⊞Pearson, Bobbi	1				1
11	⊟BK002		2	1	2	5
12	⊟Able, Ralph				1	1
13	2010				1	1
14	⊞Baker, Susan				1	1
15	⊞Foxtrot, Kathy		1			1
16	⊞George, Sally		1			1
17	⊞Wayne, Joan			1		1
18	⊟VB001	1		1	4	6
19	⊟Able, Ralph				2	2
20	2009				1	1
21	2010				1	1
22	⊞Baker, Susan				1	1
23	⊞Eagleton, Sam				1	1
24	⊞Jacobs, Nancy			1		1
25	⊞Pearson, Bobbi	1				1
26	⊞VB002	1	2		2	5
27	⊞VB003		4	1		5
28	⊞VK001	1		1	4	6
29	⊞VK002	1	3	1	3	8
30	⊞VK003		5	1		6
31	⊞VK004		6	1	1	8
32	Grand Total	5	23	8	19	55

Adventure Works cube



Fact table

Dimension tables

Product

- ProductKey
- ProductAlternateKey
- ProductSubcategoryKey
- WeightUnitMeasureCode
- SizeUnitMeasureCode
- EnglishProductName
- SpanishProductName
- FrenchProductName
- StandardCost
- FinishedGoodsFlag
- Color
- SafetyStockLevel
- ReorderPoint
- ListPrice
- Size
- SizeRange
- Weight
- DaysToManufacture
- ProductLine
- DealerPrice
- Class
- Style
- ModelName
- LargePhoto
- EnglishDescription
- FrenchDescription
- ChineseDescription
- ArabicDescription
- HebrewDescription
- ThaiDescription
- GermanDescription
- JapaneseDescription
- TurkishDescription
- StartDate
- EndDate
- Status
- ProductLineName

InternetSales

- SalesOrderNumber
- SalesOrderLineNumber
- ProductKey
- OrderDateKey
- DueDateKey
- ShipDateKey
- CustomerKey
- PromotionKey
- CurrencyKey
- SalesTerritoryKey

Date

- DateKey
- FullDateAlternateKey
- DayNumberOfWeek
- EnglishDayNameOfWeek
- SpanishDayNameOfWeek
- FrenchDayNameOfWeek
- DayNumberOfMonth
- DayNumberOfYear
- WeekNumberOfYear
- EnglishMonthName

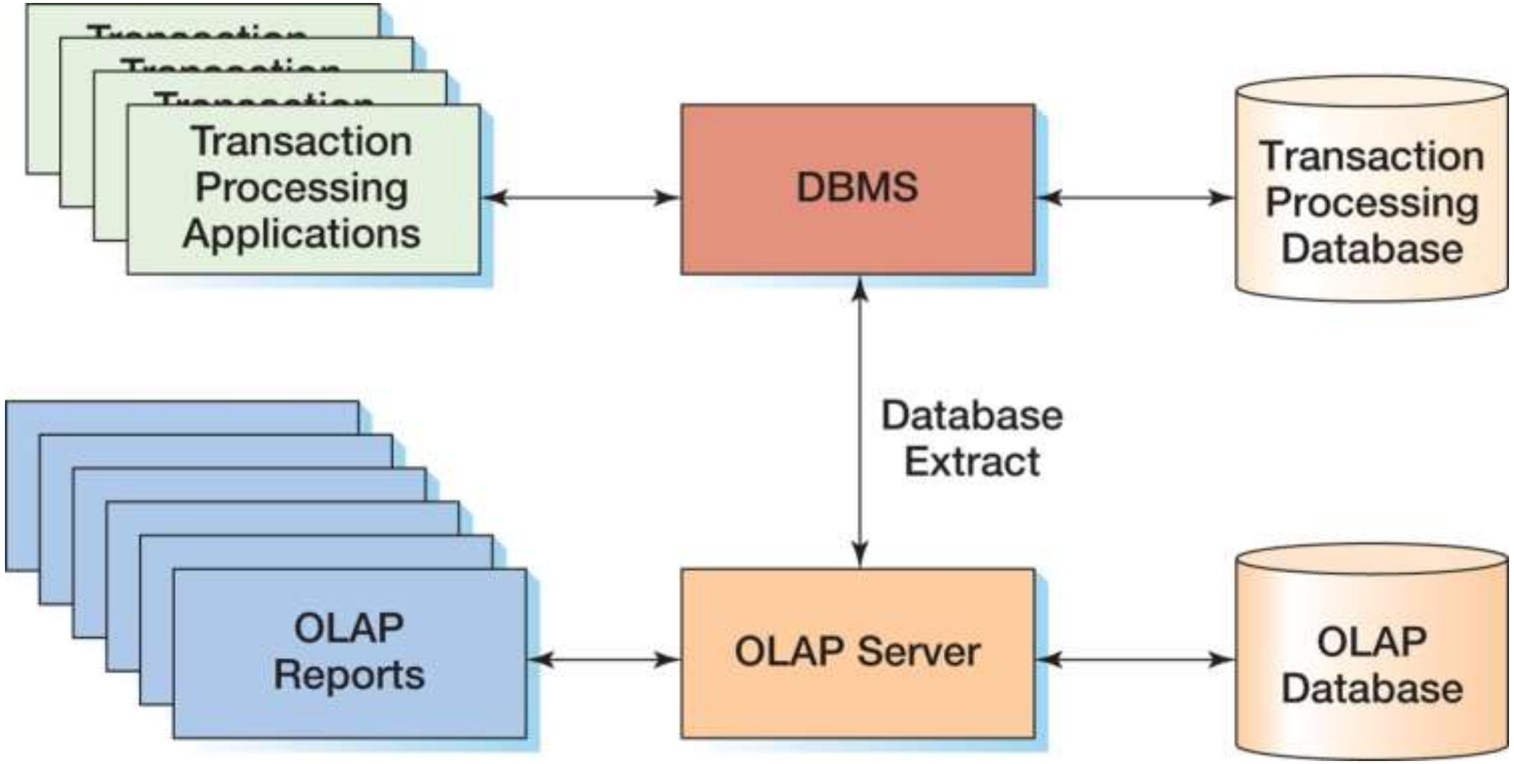
Customer

- CustomerKey
- GeographyKey
- CustomerAlternateKey
- Title
- FirstName
- MiddleName
- LastName
- NameStyle
- BirthDate
- MaritalStatus
- Suffix
- Gender
- EmailAddress
- YearlyIncome
- TotalChildren
- NumberChildrenAtHome
- EnglishEducation
- SpanishEducation
- FrenchEducation
- EnglishOccupation
- SpanishOccupation
- FrenchOccupation
- HouseOwnerFlag
- NumberCarsOwned
- AddressLine1
- AddressLine2
- Phone
- DateFirstPurchase
- CommuteDistance
- FullName

Geography

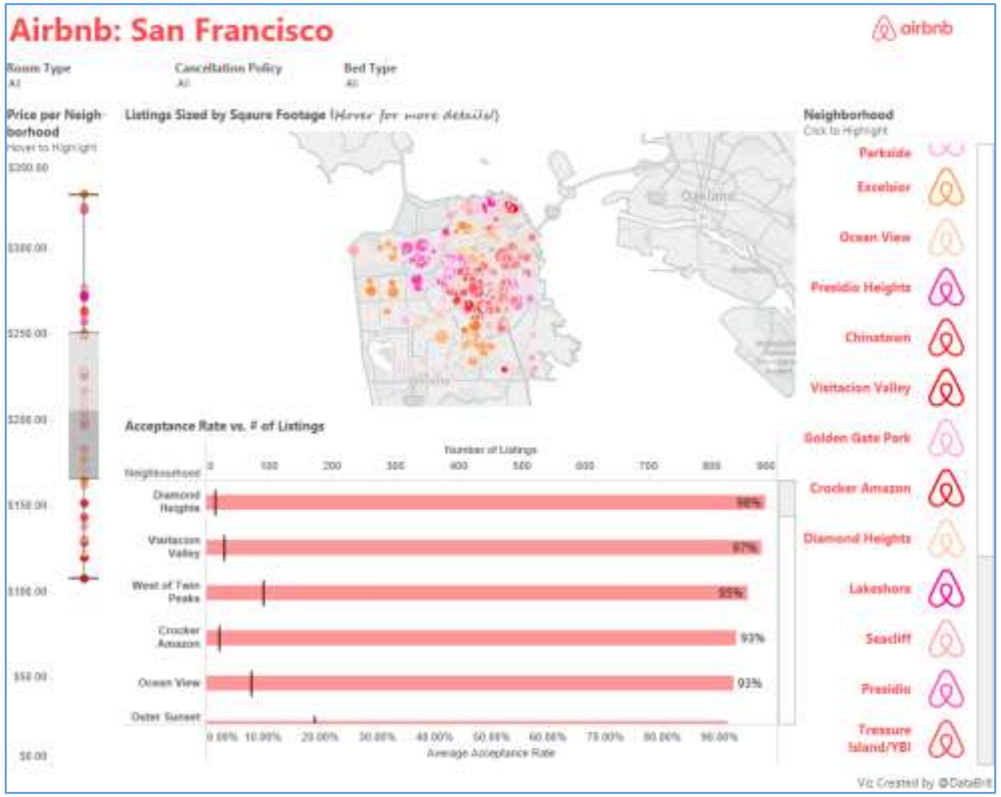
- GeographyKey
- City
- StateProvinceCode
- StateProvinceName
- CountryRegionCode
- EnglishCountryRegionName
- SpanishCountryRegionName
- FrenchCountryRegionName
- PostalCode
- SalesTerritoryKey

Reporting Systems:
OLAP Servers and OLAP Databases



Visualisation

- ~~As data becomes more complex and difficult to understand effective **visualisation** becomes necessary to understand it and ask meaningful questions of it~~
- ~~e.g. **Tableau** http://www.tableau.com/ is an interactive data visualisation tool that works with any data set~~

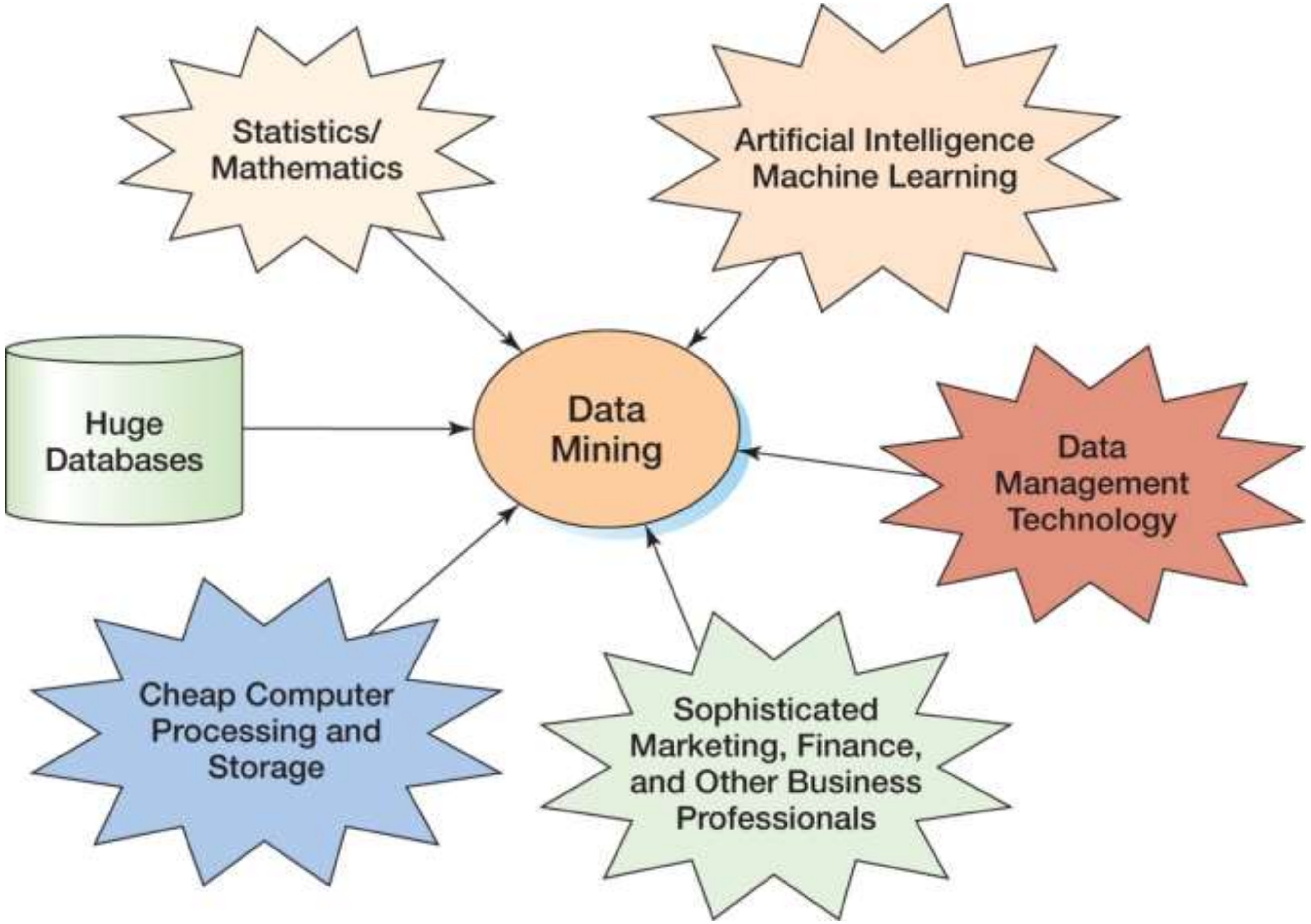


https://public.tableau.com/en-us/gallery/?tab=viz-of-the-day&type=viz-of-the-day

https://public.tableau.com/en-us/s/gallery/airbnb-prices-san-francisco

Data mining

Data Mining Applications:
The Convergence of the Disciplines


Statistics/ Mathematics
Artificial Intelligence Machine Learning
Huge Databases
Data Mining
Data Management Technology
Cheap Computer Processing and Storage
Sophisticated Marketing, Finance, and Other Business Professionals

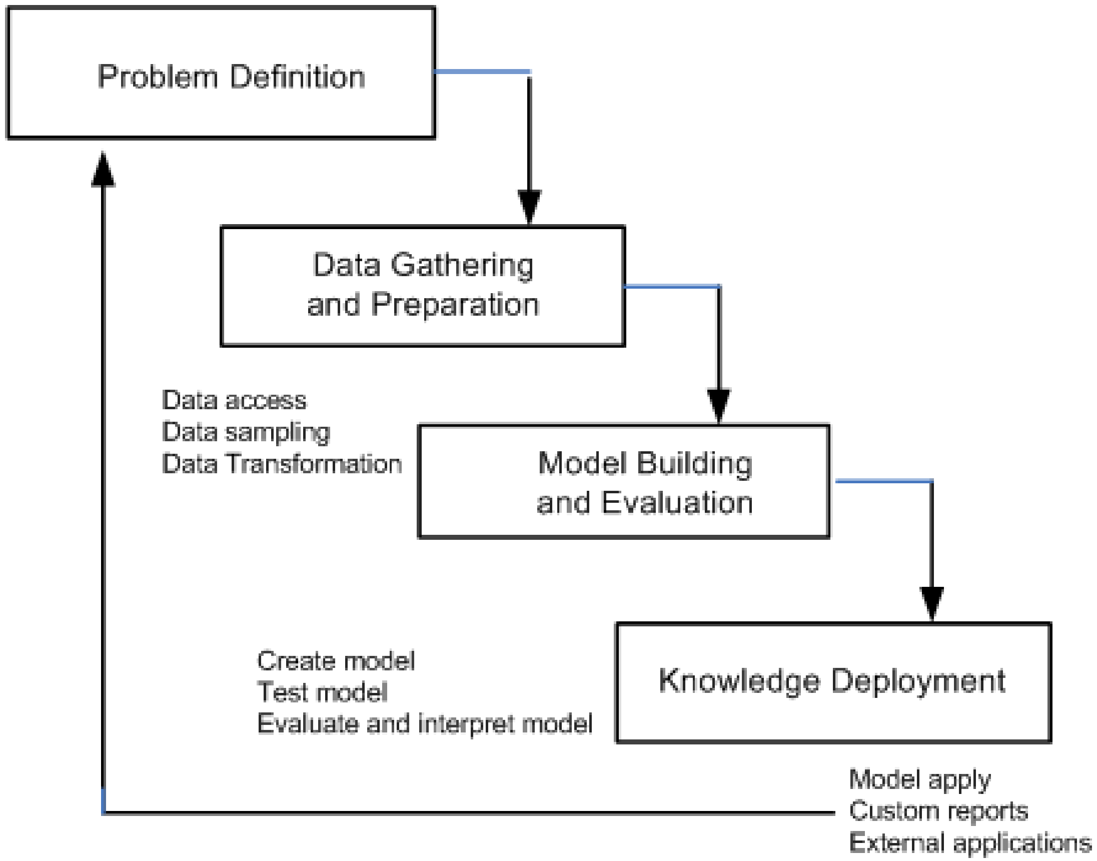
Data mining applications

Data mining software utilizes mathematical methods and statistics to identify unsuspected relationship, patterns and trends that can be used to classify data and predict. Usually identify patterns we might not have seen

- **Unsupervised data mining**
  - Have statistical techniques to find collection of tables with similar features
    - e.g. Cluster Analysis
- **Supervised data mining:**
  - Model is created and we use statistics methods to roughly calculate the parameter values of the model
    - EG- Regression analysis

Data mining process

Figure 1-1 The Data Mining Process

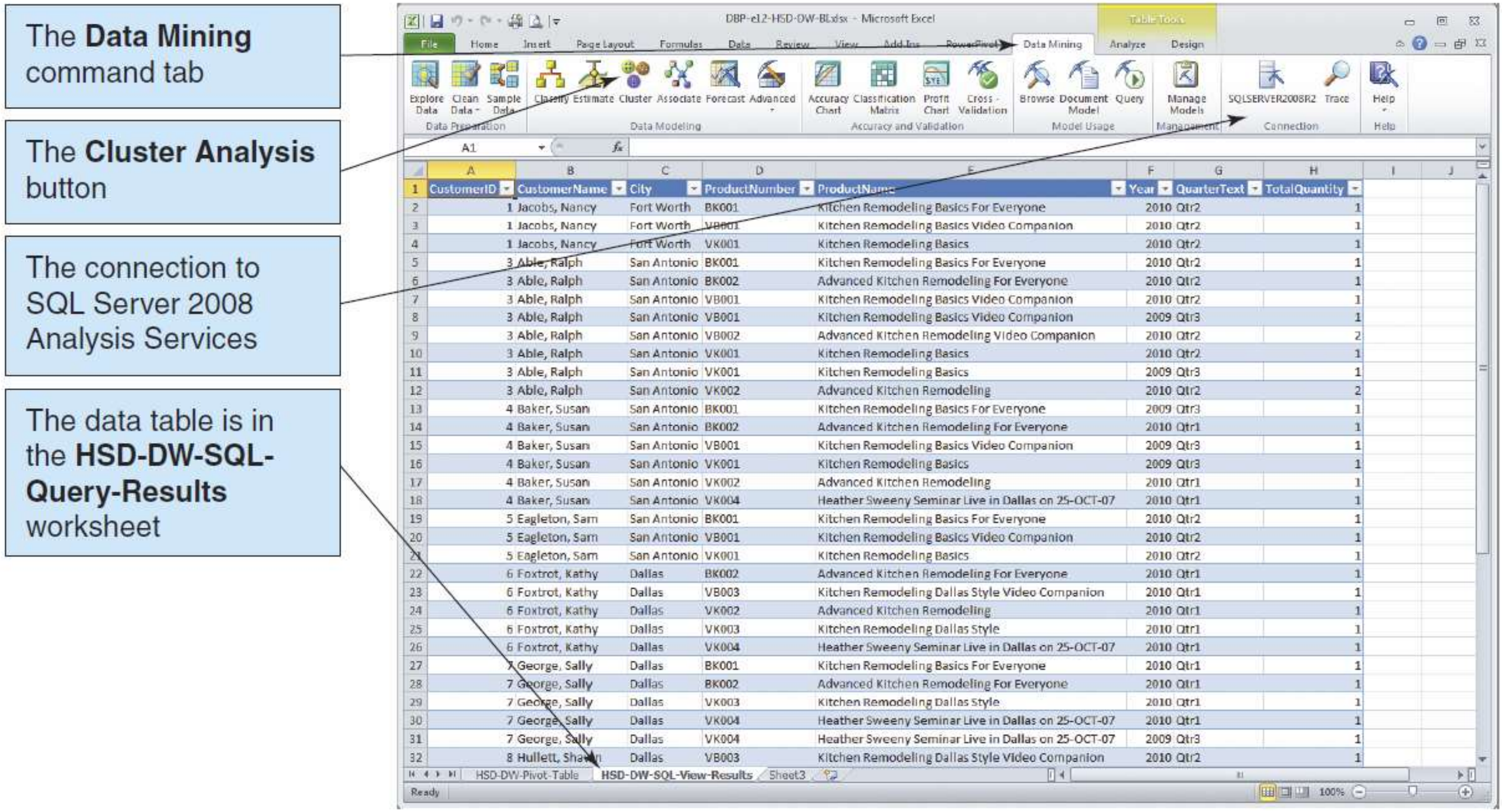


See Description of "Figure 1-1 The Data Mining Process"

https://docs.oracle.com/cd/E11882_01/datamine.112/e16808/process.htm#DMCON126

~~Data mining is an *iterative* process where the results trigger new questions that feed back into improved models~~

Cluster Analysis I



The **Data Mining** command tab

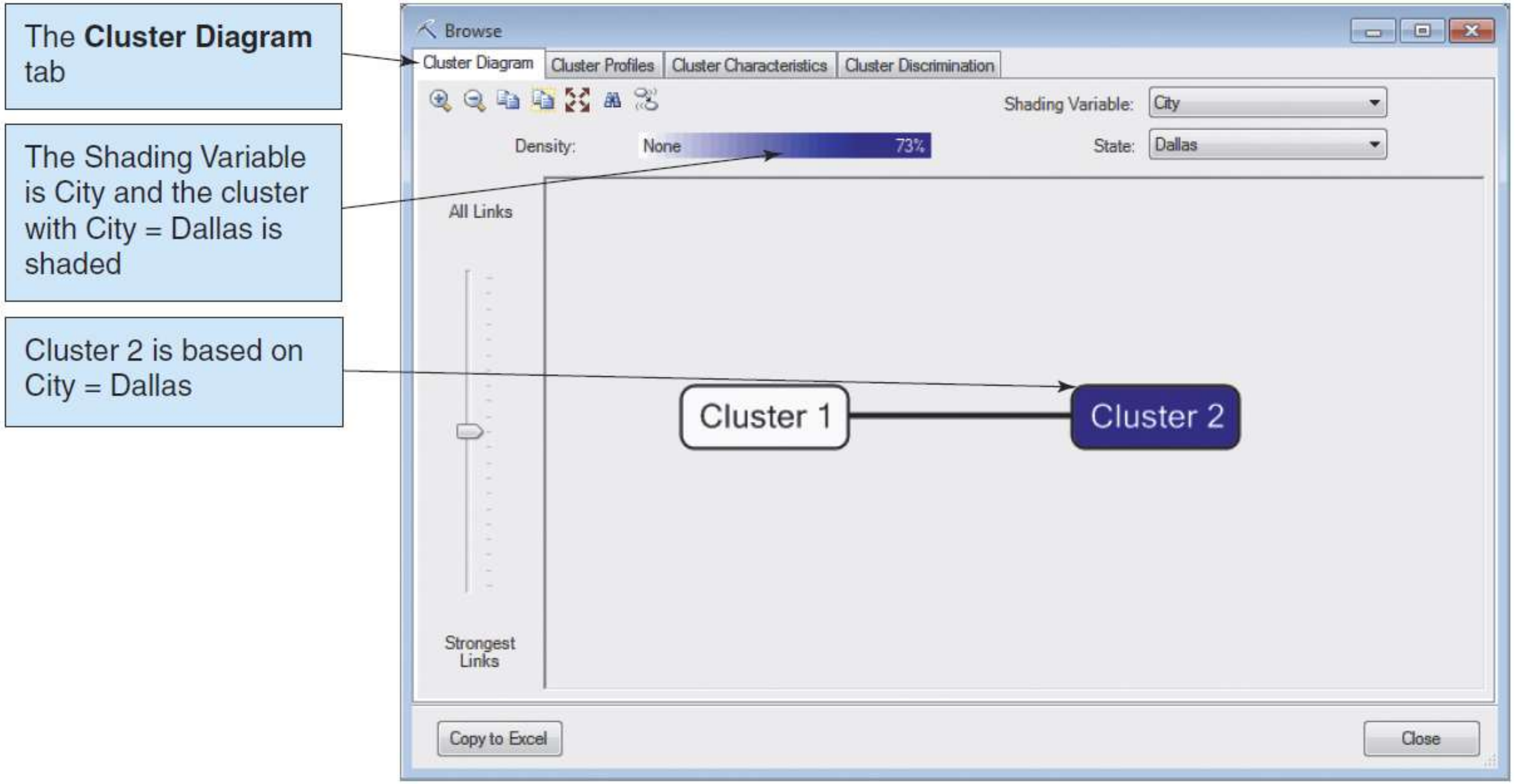
The **Cluster Analysis** button

The connection to SQL Server 2008 Analysis Services

The data table is in the **HSD-DW-SQL-Query-Results** worksheet

	CustomerID	CustomerName	City	ProductNumber	ProductName	Year	QuarterText	TotalQuantity
2	1	Jacobs, Nancy	Fort Worth	BK001	Kitchen Remodeling Basics For Everyone	2010	Qtr2	1
3	1	Jacobs, Nancy	Fort Worth	VB001	Kitchen Remodeling Basics Video Companion	2010	Qtr2	1
4	1	Jacobs, Nancy	Fort Worth	VK001	Kitchen Remodeling Basics	2010	Qtr2	1
5	3	Able, Ralph	San Antonio	BK001	Kitchen Remodeling Basics For Everyone	2010	Qtr2	1
6	3	Able, Ralph	San Antonio	BK002	Advanced Kitchen Remodeling For Everyone	2010	Qtr2	1
7	3	Able, Ralph	San Antonio	VB001	Kitchen Remodeling Basics Video Companion	2010	Qtr2	1
8	3	Able, Ralph	San Antonio	VB001	Kitchen Remodeling Basics Video Companion	2009	Qtr3	1
9	3	Able, Ralph	San Antonio	VB002	Advanced Kitchen Remodeling Video Companion	2010	Qtr2	2
10	3	Able, Ralph	San Antonio	VK001	Kitchen Remodeling Basics	2010	Qtr2	1
11	3	Able, Ralph	San Antonio	VK001	Kitchen Remodeling Basics	2009	Qtr3	1
12	3	Able, Ralph	San Antonio	VK002	Advanced Kitchen Remodeling	2010	Qtr2	2
13	4	Baker, Susan	San Antonio	BK001	Kitchen Remodeling Basics For Everyone	2009	Qtr3	1
14	4	Baker, Susan	San Antonio	BK002	Advanced Kitchen Remodeling For Everyone	2010	Qtr1	1
15	4	Baker, Susan	San Antonio	VB001	Kitchen Remodeling Basics Video Companion	2009	Qtr3	1
16	4	Baker, Susan	San Antonio	VK001	Kitchen Remodeling Basics	2009	Qtr3	1
17	4	Baker, Susan	San Antonio	VK002	Advanced Kitchen Remodeling	2010	Qtr1	1
18	4	Baker, Susan	San Antonio	VK004	Heather Sweeny Seminar Live in Dallas on 25-OCT-07	2010	Qtr1	1
19	5	Eagleton, Sam	San Antonio	BK001	Kitchen Remodeling Basics For Everyone	2010	Qtr2	1
20	5	Eagleton, Sam	San Antonio	VB001	Kitchen Remodeling Basics Video Companion	2010	Qtr2	1
21	5	Eagleton, Sam	San Antonio	VK001	Kitchen Remodeling Basics	2010	Qtr2	1
22	6	Foxtrot, Kathy	Dallas	BK002	Advanced Kitchen Remodeling For Everyone	2010	Qtr1	1
23	6	Foxtrot, Kathy	Dallas	VB003	Kitchen Remodeling Dallas Style Video Companion	2010	Qtr1	1
24	6	Foxtrot, Kathy	Dallas	VK002	Advanced Kitchen Remodeling	2010	Qtr1	1
25	6	Foxtrot, Kathy	Dallas	VK003	Kitchen Remodeling Dallas Style	2010	Qtr1	1
26	6	Foxtrot, Kathy	Dallas	VK004	Heather Sweeny Seminar Live in Dallas on 25-OCT-07	2010	Qtr1	1
27	7	George, Sally	Dallas	BK001	Kitchen Remodeling Basics For Everyone	2010	Qtr1	1
28	7	George, Sally	Dallas	BK002	Advanced Kitchen Remodeling For Everyone	2010	Qtr1	1
29	7	George, Sally	Dallas	VK003	Kitchen Remodeling Dallas Style	2010	Qtr1	1
30	7	George, Sally	Dallas	VK004	Heather Sweeny Seminar Live in Dallas on 25-OCT-07	2010	Qtr1	1
31	7	George, Sally	Dallas	VK004	Heather Sweeny Seminar Live in Dallas on 25-OCT-07	2009	Qtr3	1
32	8	Hullett, Shawn	Dallas	VB003	Kitchen Remodeling Dallas Style Video Companion	2010	Qtr2	1

Cluster Analysis II



The **Cluster Diagram** tab

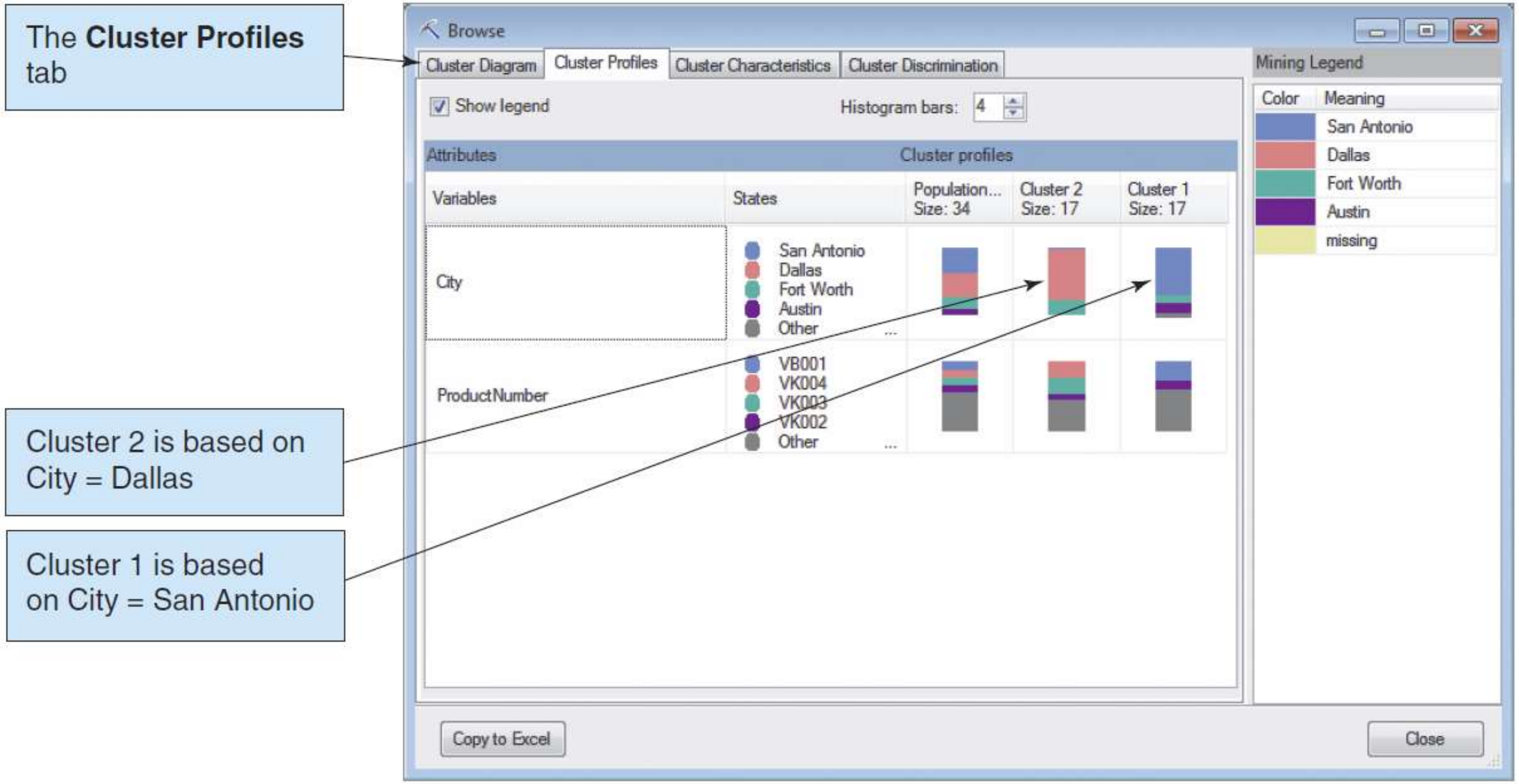
The Shading Variable is City and the cluster with City = Dallas is shaded

Cluster 2 is based on City = Dallas

Shading Variable: City

State: Dallas

Cluster Analysis III



The **Cluster Profiles** tab

Cluster 2 is based on City = Dallas

Cluster 1 is based on City = San Antonio

☑ Show legend

Histogram bars: 4

Attributes | Cluster profiles

Variables	States	Population... Size: 34	Cluster 2 Size: 17	Cluster 1 Size: 17
City	● San Antonio ● Dallas ● Fort Worth ● Austin ● Other ...			
ProductNumber	● VB001 ● VK004 ● VK003 ● VK002 ● Other ...			

Mining Legend

Color	Meaning
🟦	San Antonio
🟥	Dallas
🟩	Fort Worth
🟪	Austin
🟨	missing

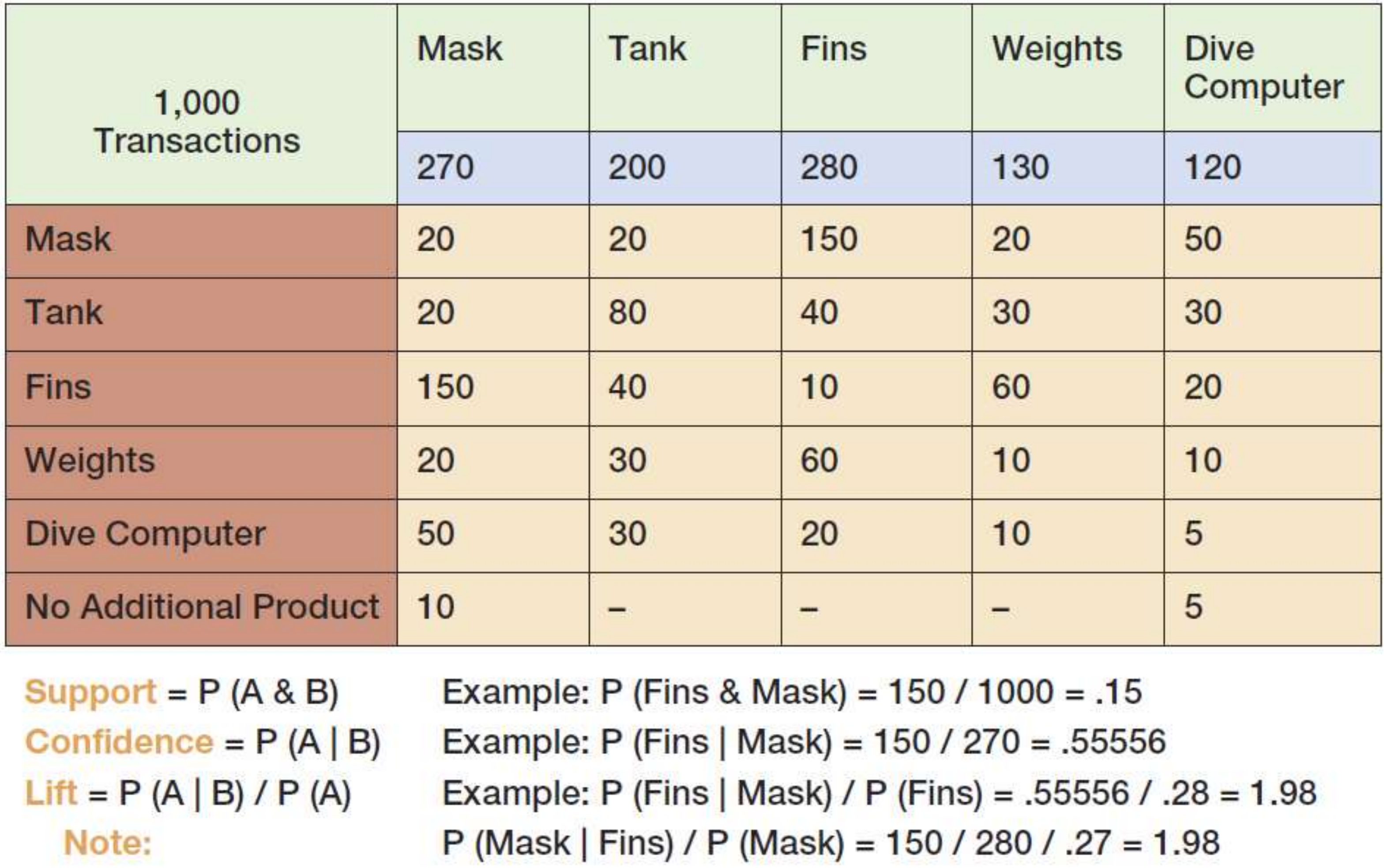
Data Mining Applications:
Popular Data Mining Techniques

- ~~**Decision tree analysis**—classifies entities into groups based on past history~~
- ~~**Logistic regression**—produces equations that offer probabilities that certain events will occur~~
- ~~**Neural Networks**—complex statistical prediction techniques~~
- ~~**Market Basket Analysis**—determines patterns of associated buying behavior~~

Data Mining Applications:
Market Basket Analysis

- ~~**Support**—the probability that two items will be purchased together~~
- ~~**Confidence**—the probability that an item will be purchased given the fact that the customer has already purchased another particular item~~
- ~~**Lift**—the ration of confidence to the basic probability that a particular item will be purchased~~

Data Mining Applications:
Market Basket Analysis



1,000 Transactions	Mask	Tank	Fins	Weights	Dive Computer
	270	200	280	130	120
Mask	20	20	150	20	50
Tank	20	80	40	30	30
Fins	150	40	10	60	20
Weights	20	30	60	10	10
Dive Computer	50	30	20	10	5
No Additional Product	10	–	–	–	5

Support = P (A & B) Example: P (Fins & Mask) = 150 / 1000 = .15

Confidence = P (A | B) Example: P (Fins | Mask) = 150 / 270 = .55556

Lift = P (A | B) / P (A) Example: P (Fins | Mask) / P (Fins) = .55556 / .28 = 1.98

Note: P (Mask | Fins) / P (Mask) = 150 / 280 / .27 = 1.98

Data Mining Applications:
SQL for Market Basket Analysis

```
CREATE VIEW  TwoItemBasket AS
   SELECT    T1.ItemID as FirstItem,
             T2.ItemID as SecondIem
   FROM             TRANS_DATA T1 JOIN TRANS_DATA T2
       ON    T1.TransactionID = T2.TransactionID
       AND   T1.ItemID <> T2.ItemID;

CREATE VIEW  ItemSupport AS
   SELECT    FirstItem, SecondItem,
             Count(*) AS SupportCount
   FROM             TwoItemBasket
   GROUP BY  FirstItem, SecondItem;
```

Data mining in Oracle

Oracle’s data mining features are summarised here: https://docs.oracle.com/cd/E11882_01/datamine.112/e16808/intro_concepts.htm#DMCON001

e.g. Supervised functions:

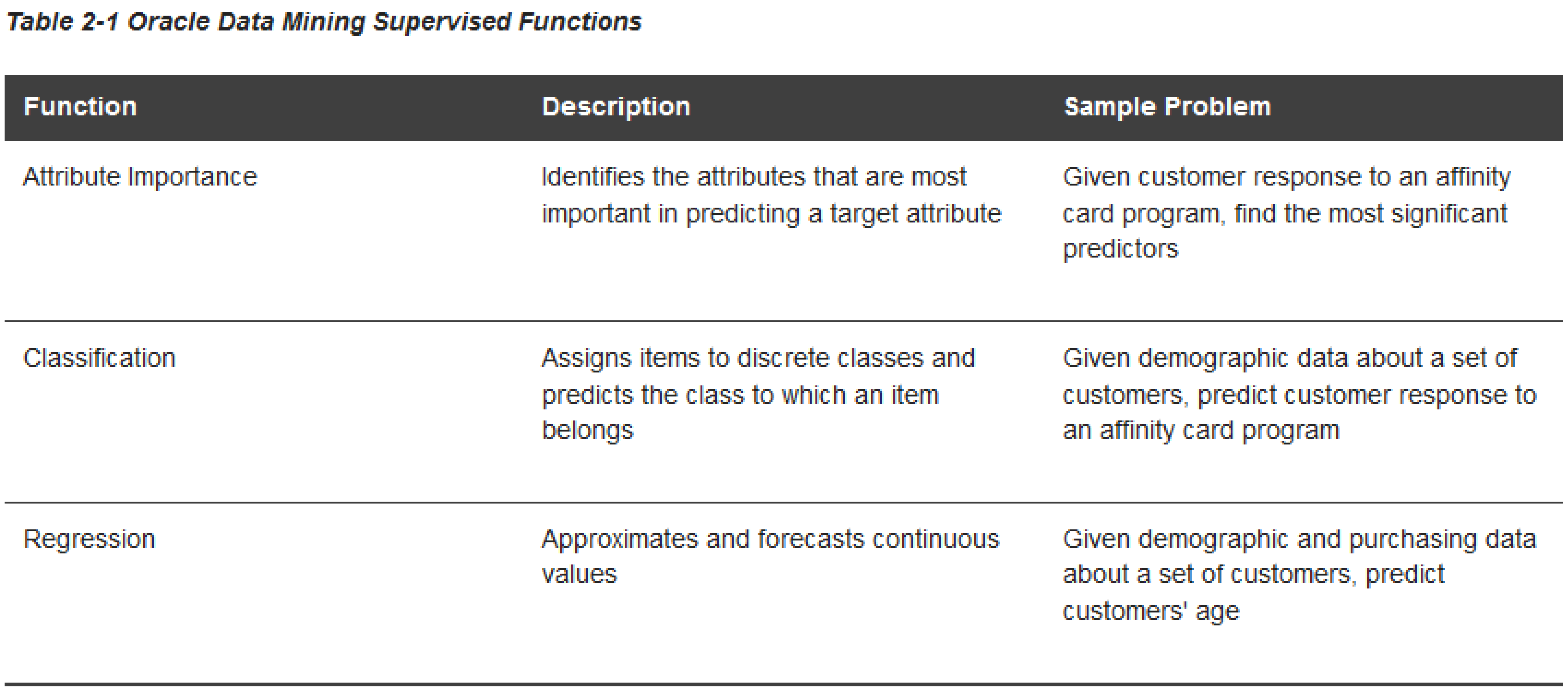


Table 2-1 Oracle Data Mining Supervised Functions

Function	Description	Sample Problem
Attribute Importance	Identifies the attributes that are most important in predicting a target attribute	Given customer response to an affinity card program, find the most significant predictors
Classification	Assigns items to discrete classes and predicts the class to which an item belongs	Given demographic data about a set of customers, predict customer response to an affinity card program
Regression	Approximates and forecasts continuous values	Given demographic and purchasing data about a set of customers, predict customers' age

Data mining in Oracle

e.g. Unsupervised functions:

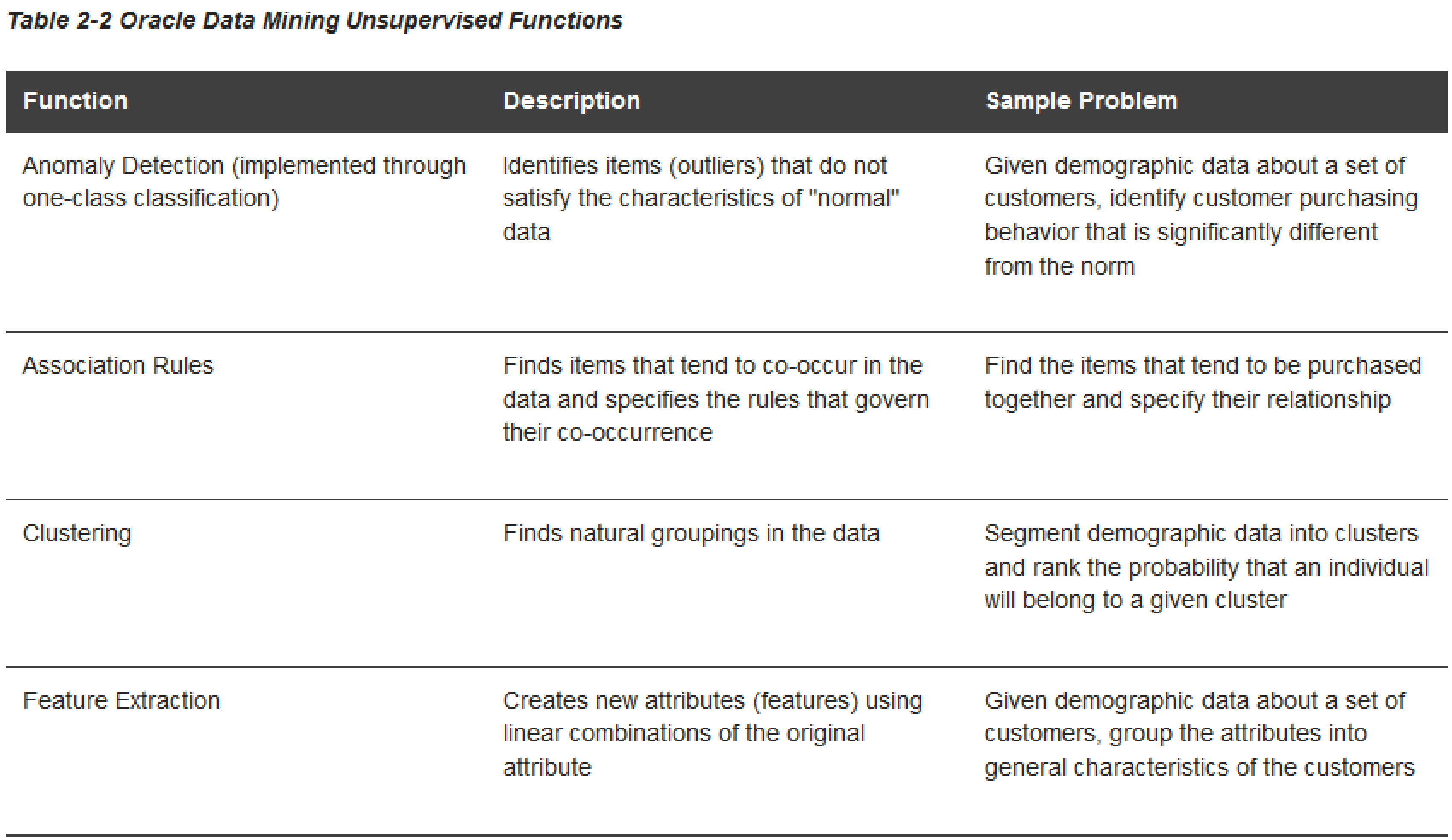


Table 2-2 Oracle Data Mining Unsupervised Functions

Function	Description	Sample Problem
Anomaly Detection (implemented through one-class classification)	Identifies items (outliers) that do not satisfy the characteristics of "normal" data	Given demographic data about a set of customers, identify customer purchasing behavior that is significantly different from the norm
Association Rules	Finds items that tend to co-occur in the data and specifies the rules that govern their co-occurrence	Find the items that tend to be purchased together and specify their relationship
Clustering	Finds natural groupings in the data	Segment demographic data into clusters and rank the probability that an individual will belong to a given cluster
Feature Extraction	Creates new attributes (features) using linear combinations of the original attribute	Given demographic data about a set of customers, group the attributes into general characteristics of the customers

The takeaways…

- Data mining is the process of automatically searching large stores of data to discover patterns and trends
- Data mining goes beyond the basic calculations used in OLTP systems to use sophisticated mathematical and statistical techniques
- ‘Market basket’ analysis is a typical technique to find what products are purchased together

Learning outcomes revisited

After completing this topic you should be able to:

- Explain the difference between BI systems and operational systems
- Describe the benefits and issues associated with data warehousing
- Describe the major components of a data warehouse
- Explain why ‘dirty data’ is a significant issue for data warehouses, and ways in which dirty data can be cleaned
- Explain the difference between a data warehouse and a data mart
- Explain how a dimensional model differs from an entity relationship model
- Describe some reporting systems for BI
- Explain what data mining is and what it can be used for

What’s next?

- In the final topic of the unit we will look briefly at NoSQL databases and their role in today’s world of Big Data and massively distributed processing.