DATA MANAGEMENT AND STORAGE DESIGN (CH 9)

Systems Analysis and Design

Dr. Joycelyn Streator

Reminders

- Exam 3 is next Wednesday. Open book, take home format.

Learning Objectives

- Become familiar with several file and database formats.
- Describe several goals of data storage.
- Be able to revise a logical ERD into a physical ERD.
- Be able to optimize a relational database for data storage and data access.
- Become familiar with indexes.
- Be able to estimate the size of a database.

Data Models and Storage in The Real World

At Western Digital Corporation (maker of hard drives), one hard drive is associated with 1,000 attributes.

3M had to track in their data warehouse: 500,000 finished goods, 300,000 order related items per day, and over 1 million customer account numbers. The logical data model contains more than 1,000 tables that are pulled in from hundreds of source systems. The data warehouse has 3,000 data elements.

The IRS has an accounts receivable table with 1 billion rows.

Blue Cross and Blue Shield of North Carolina has a data warehouse. The database is 500GB. There are 200 tables and 1000 fields.

Medicaid in New York State answers more than 200 million medical claims each year. A database is set up in the Office of the Attorney General that has 850 million claims taking up 4 Terabytes on the server.

Key Concepts

- Data storage function: how data is stored and handled by programs that run the system.

- Data storage design:
  - select the data storage format
  - convert the logical data model into a *physical* data model to reflect implementation decisions
  - ensure that DFDs and ERDs balance
  - design the selected data storage format to optimize its processing efficiency.

DATA STORAGE FORMATS

Selecting the best storage option

DATA STORAGE FORMATS

- Types of data storage formats:
  - *Files*:
  - Electronic lists of data, optimized to perform a particular transaction.

  - *Database*:
  - A collection of groupings of information that are related to each other in some way.

- *Database Management System (DBMS)*:
  - software that creates and manipulates the databases.

Files

- *Data file:* an electronic list of information that is formatted for a particular transaction.
- Sequential organization is typical.
- Record associations with other records created by pointers.
- Also called *linked lists* because of the way the records are linked together using pointers.

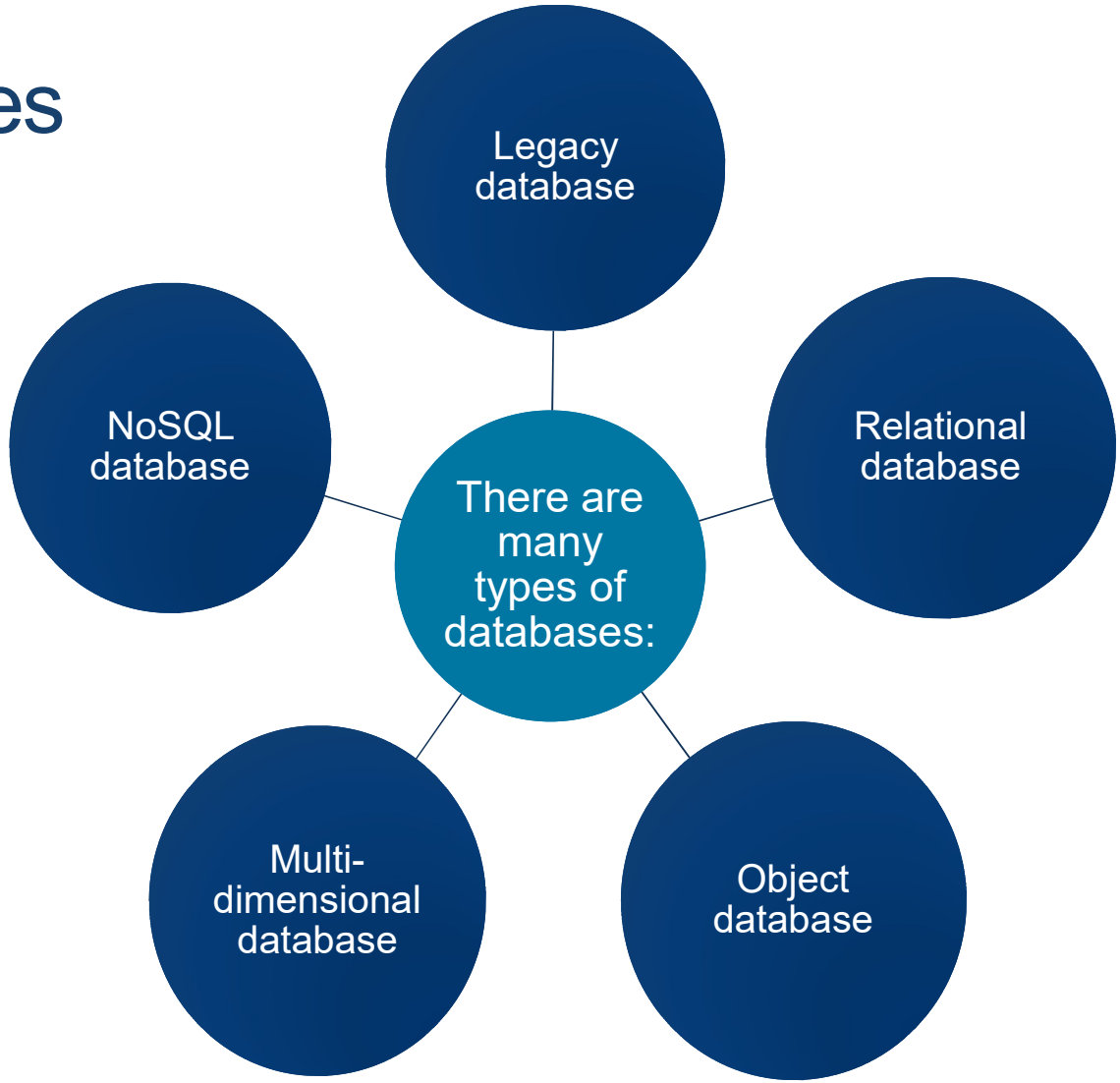
Types of Files

Master files	• store core information that is important to the application.
Look-up files	• contain static values.
Transaction files	• store information that can be used to update a master file.
Audit files	• record “before” and “after” images of data as the data is altered.
History files (or archive files)	• store past transactions.

Appointment File Example

Appointment Date	Appointment Time	Duration	Reason	Patient ID	First Name	Last Name	Phone Number	Doctor ID	Doctor Last Name
11/23/2015	2:30	0.25 hour	Flu	758843	Patrick	Dennis	548-9456	V524625587	Vroman
11/23/2015	2:30	1 hour	Physical	136136	Adelaide	Kin	548-7887	T445756225	Tantalo
11/23/2015	2:45	0.25 hour	Shot	544822	Chris	Pullig	525-5464	V524625587	Vroman
11/23/2015	3:00	1 hour	Physical	345344	Felicia	Marston	548-9333	B544742245	Brousseau
11/23/2015	3:00	0.5 hour	Migraine	236454	Thomas	Bateman	667-8955	V524625587	Vroman
11/23/2015	3:30	0.5 hour	Muscular	887777	Ryan	Nelson	525-4772	V524625587	Vroman
11/23/2015	3:30	0.25 hour	Shot	966233	Peter	Todd	667-2325	T445756225	Tantalo
11/23/2015	3:45	0.75 hour	Muscular	951657	Mike	Morris	663-8944	T445756225	Tantalo
11/23/2015	4:00	1 hour	Physical	223238	Ellen	Whitener	525-8874	B544742245	Brousseau
11/23/2015	4:00	0.5 hour	Flu	365548	Jerry	Starsia	548-9887	V524625587	Vroman
11/23/2015	4:30	1 hour	Minor surg	398633	Susan	Perry	525-6632	V524625587	Vroman
11/23/2015	4:30	0.5 hour	Migraine	222577	Elizabeth	Gray	667-8400	T445756225	Tantalo
11/24/2015	8:30	0.25 hour	Shot	858756	Elias	Awad	663-6364	T445756225	Tantalo
11/24/2015	8:30	1 hour	Minor surg	232158	Andy	Ruppel	525-9888	V524625587	Vroman
11/24/2015	8:30	0.25 hour	Flu	244875	Rick	Grenci	548-2114	B544742245	Brousseau
11/24/2015	8:45	0.5 hour	Muscular	655683	Eric	Meier	667-0254	T445756225	Tantalo
11/24/2015	8:45	1 hour	Physical	447521	Jane	Pace	548-0025	B544742245	Brousseau
11/24/2015	9:30	0.5 hour	Flu	554263	Trey	Maxham	663-8547	V524625587	Vroman

Databases


There are many types of databases:
Legacy database
Relational database
Object database
Multi-dimensional database
NoSQL database

Legacy Databases

- Databases which are based on older technology; seldom used to develop new applications.

- Two major types:
  - *Hierarchical databases* use hierarchies, or inverted trees, to represent relationships.

  - *Network databases* are collections of records that are related to each other through *pointers*.

Relational Databases

- Tables are related to each other by the placing the primary key from one table into the related table as a *foreign key*.
- Based on collections of *tables*, each of which has a *primary key*.
- The most popular kind of database for application development today.

Relational Databases Continued

- Most relational database management systems (RDBMS) support *referential integrity*

  Referential integrity ensures that values linking the tables together are valid and correctly synchronized.

- *Structured Query Language (SQL)* is the standard language for accessing the data in the tables.

Appointment Database

Appointment Date	Appointment Time	Duration	Reason	Patient ID	Doctor ID
11/23/2015	2:30	0.5 hour	Flu	758843	V524625587
11/23/2015	2:30	1 hour	Physical	136136	T445756225
11/23/2015	2:45	0.25 hour	Shot	544822	V524625587
11/23/2015	3:00	1 hour	Physical	345344	B544742245
11/23/2015	3:00	0.5 hour	Migraine	236454	V524625587
11/23/2015	3:30	0.5 hour	Muscular	887777	V524625587
11/23/2015	3:30	0.25 hour	Shot	966233	T445756225
11/23/2015	3:45	0.75 hour	Muscular	951657	T445756225
11/23/2015	4:00	1 hour	Physical	223238	B544742245
11/23/2015	4:00	0.5 hour	Flu	365548	V524625587
11/23/2015	4:30	1 hour	Minor surg	398633	V524625587
11/23/2015	4:30	0.5 hour	Migraine	222577	T445756225
11/24/2015	8:30	0.25 hour	Shot	858756	T445756225
11/24/2015	8:30	1 hour	Minor surg	232158	V524625587
11/24/2015	8:30	0.25 hour	Flu	244875	B544742245
11/24/2015	8:45	0.5 hour	Muscular	655683	T445756225
11/24/2015	8:45	1 hour	Physical	447521	B544742245
11/24/2015	9:30	0.5 hour	Flu	554263	V524625587

Tables related by patient ID

Tables related by doctor ID

Patient ID	First Name	Last Name	Phone Number
136136	Adelaide	Kin	548-7887
222577	Elizabeth	Gray	667-8400
223238	Ellen	Whitener	525-8874
232158	Andy	Ruppel	525-9888
236454	Thomas	Bateman	667-8955
244875	Rick	Grenci	548-2114
345344	Felicia	Marston	548-9333
365548	Jerry	Starsia	548-9887
398633	Susan	Perry	525-6632
447521	Jane	Pace	548-0025
544822	Chris	Pullig	525-5464
554263	Trey	Maxham	663-8547
655683	Eric	Meier	667-0254
758843	Patrick	Dennis	548-9456
858756	Elias	Awad	663-6364
887777	Ryan	Nelson	525-4772
951657	Mike	Morris	663-8944
966233	Peter	Todd	667-2325

Doctor ID	Last Name
B544742245	Brousseau
T445756225	Tantalo
V524625587	Vroman

Object Databases

- Based on object orientation: that all things should be treated as *objects* that have both data (attributes) and processes (behaviors).
- *Object-oriented database management system (OODBMS)* are mainly used to support multimedia applications or systems that involve complex data.
- Play a minor role in the DBMS market at this time.

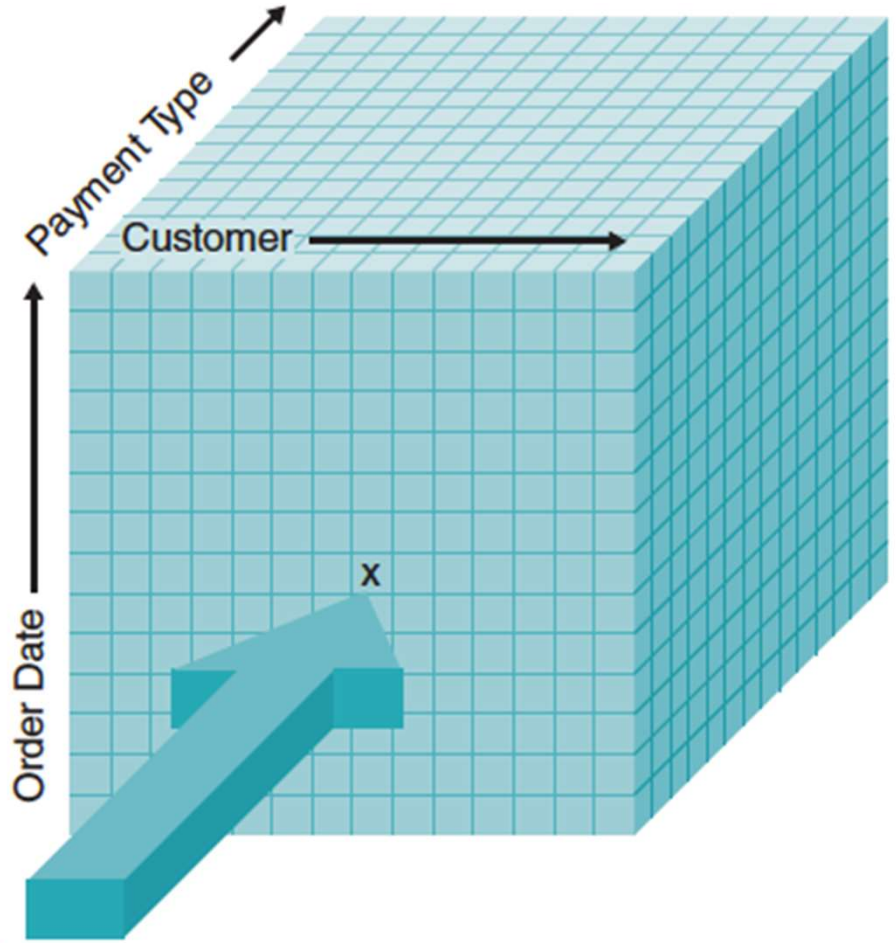
Multidimensional Databases

- A type of relational database used extensively in data warehousing.
- *Data warehousing* is the practice of taking and storing data in a data warehouse (i.e., a large database) that supports *business intelligence (BI)* systems.
- *Data marts* are smaller databases based on data warehouse data; support BI for specific departments or functional areas of the organization.

Multidimensional Databases, con't.

- Stores data to support *aggregations* of data on multiple dimensions.
- When the data are first loaded into a multidimensional database, the database precalculates the data across the multiple dimensions and stores the answers for fast access.

Multidimensional Database



Last quarter, how many customers placed more than one order, using an American Express card?

What is Intelligence

Intelligence:

The ability to understand the interrelationships of presented facts in such a way as to guide action towards a desired goal

-
- - Hans Peter Luhn, IBM, 1958
- "A Business Intelligence System”

Business Intelligence

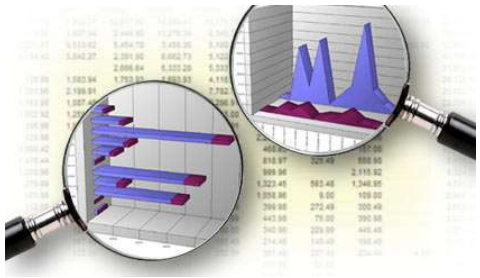
Declining cost of computing

Declining cost of storage

Increasing interconnectivity of devices and systems

More data and information available to businesses

Business opportunities: Results of data analyses could be used to generate and sustain competitive advantage

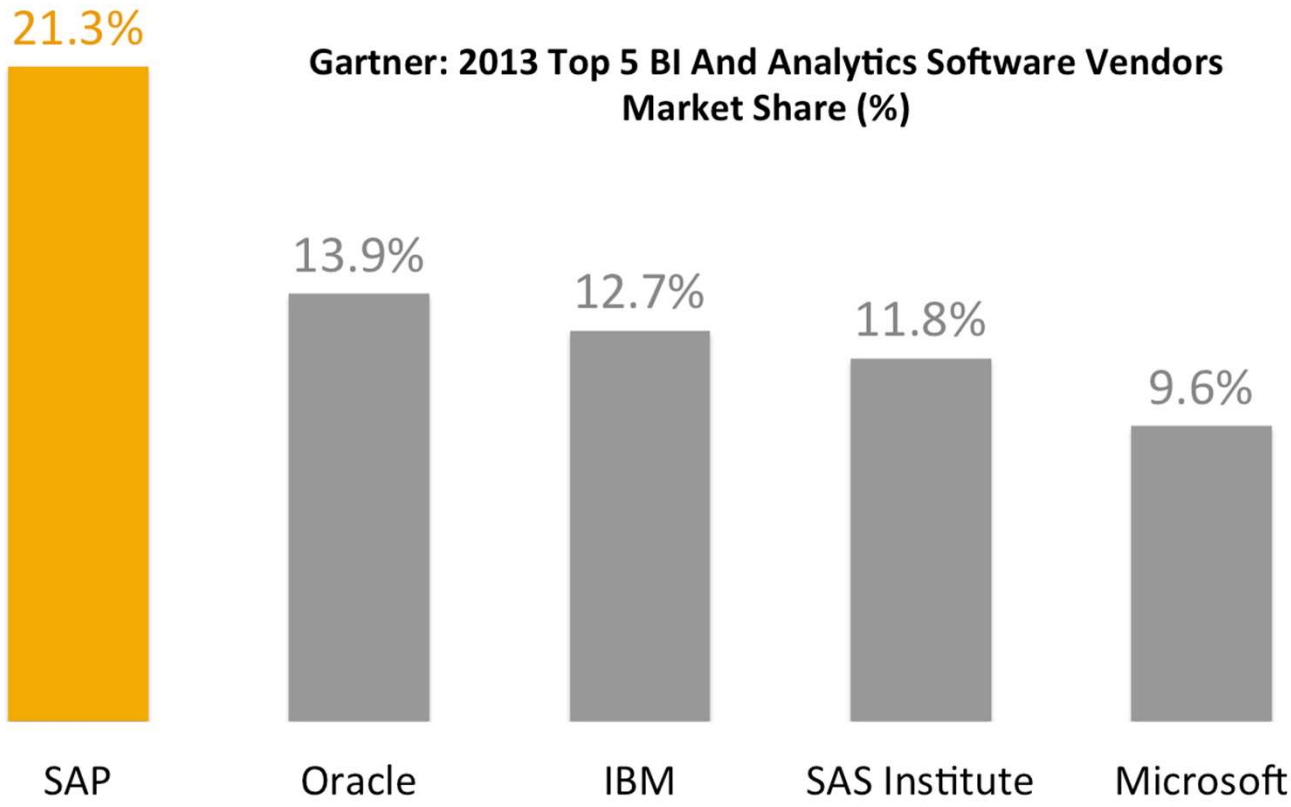


Business Intelligence

- The ability to gather and make sense of information in the context of a business.
  - Set of techniques
  - Processes
  - Technologies

- Purpose:
  - Gain superior insight and understanding of the
  - business and its ecosystem
    - Understand the past and the present
    - Predict the future
  - Make better decisions

Business Intelligence



BI System Components: Data Warehouse / Data Mart

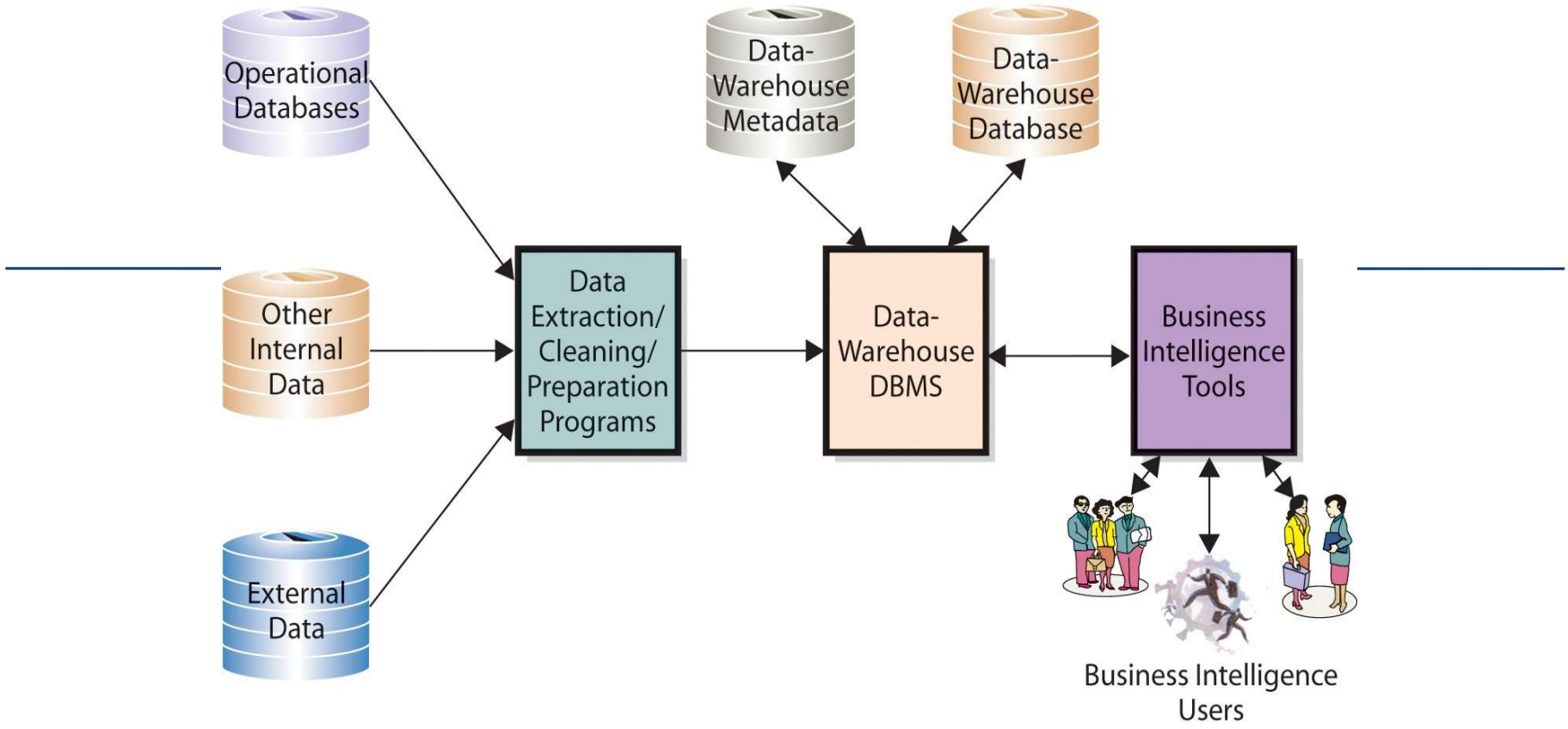
Data Warehouse:

- Consolidates and integrates multiple data sources (both internal and external)
- May include also metadata – data about one or more aspects of data in the warehouse (creation time, purpose, method of creation, author, etc)
- Large size and scope
- Designed for analytics
- Data managed (stored, edited, accessed) through a database management system (DBMS – e.g. Microsoft Access, Oracle Database 11g, etc.)

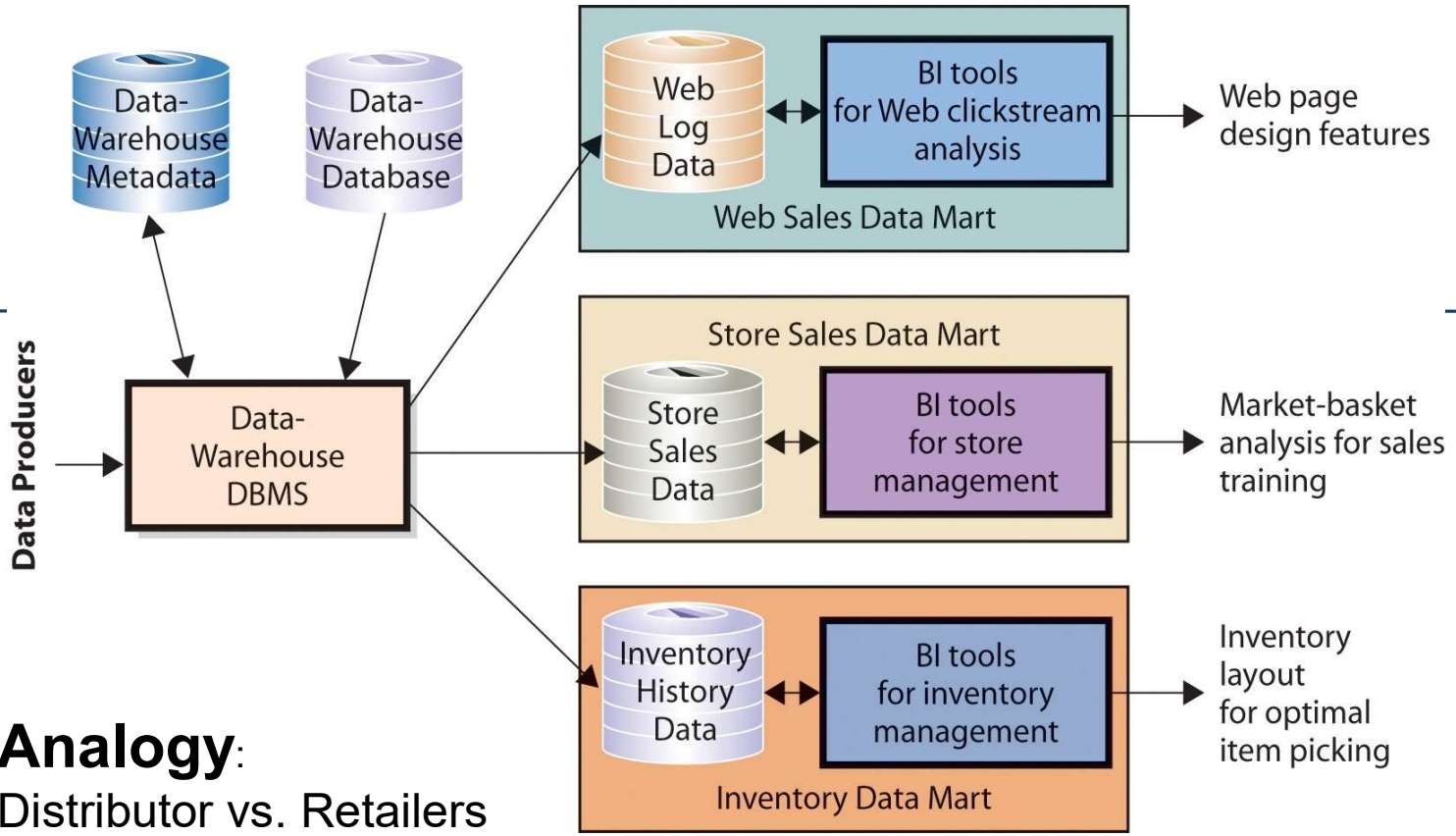
Data Mart

- scaled down version of a data warehouse
- created to address / approach / serve
  - a specific audience / business function
  - an opportunity
  - a problem

DATA WAREHOUSE



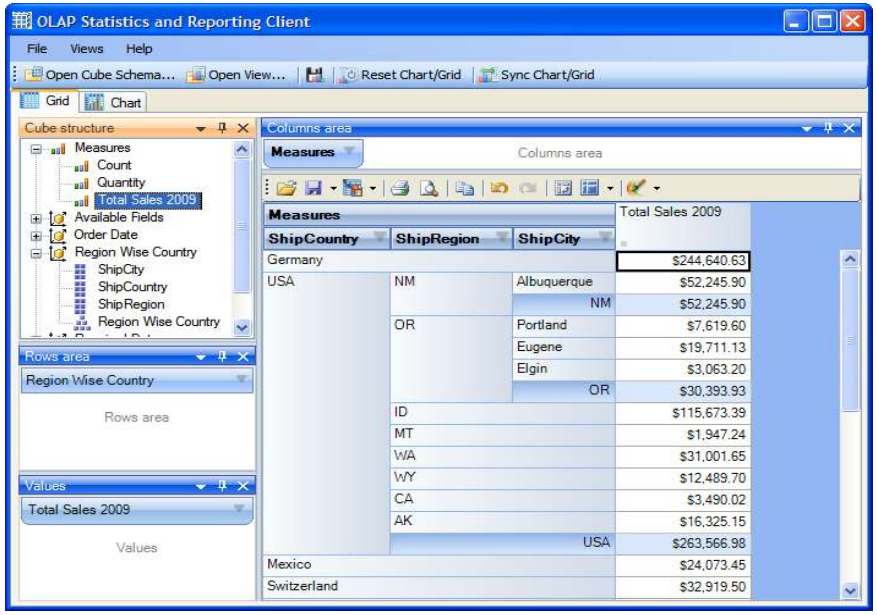
DATA MART EXAMPLES



Analogy:
Distributor vs. Retailers

BI System Components: Reporting Tools

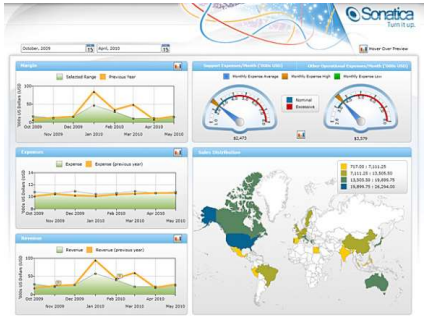
- **Online Analytical Processing (OLAP) tools:**
  - Software that extracts and views data from different perspectives (drill down, pivot, slice, etc.)



video

BI System Components: Reporting Tools

- **Business Performance Management (dashboard)** tools provide an interface to data for end users



BI System Components: Data Mining Tools

- **Data Mining:** The automated search for non-obvious patterns in large databases

- Examples of patterns of interest:

  - **associations** – one event is correlated to another event
  - **sequences** – one event leads to another event
  - **classification / clustering** – categories are generated from data
  - **forecasting** – data trends used to forecast future events

- Example: Google scans Gmail emails in order to offer Gmail users targeted advertising (… and identify viruses, spam, etc).

Evolution of BI

- **Real time** BI
- **Mobile** BI
- Oracle Business Intelligence Mobile Demo: video (2:53)

NoSQL Databases

- Newest database approach; not based on the relational model or SQL.
- Rapid processing on replicated database servers in the cloud.
- Various types include:
  - Document-oriented databases: manage collection of documents of varying forms and structures (e.g., Mongo DB)
  - Wide column databases: store data in records holding very large numbers of dynamic columns (potentially billions of columns). E.g., Bigtable, Cassandra, Dynamo
  - Graph databases: a collection of nodes and edges using graph theory to store, map, and query relationships.

Selecting a Storage Format

- Each of the file and database data storage format has its strengths and weaknesses.

- Factors to consider in selecting a storage format:
  - Data Types
  - Type of Application System
  - Existing Storage Formats
  - Future Needs

Comparing Storage Formats

Data Storage Formats

	Files	Legacy databases	Relational databases	Object databases	Multidimensional databases	NoSQL databases
Major strengths	Files can be designed for fast performance; good for short-term data storage.	Very mature products	Leader in the database market; can handle fast updating and querying needs	Able to handle complex data	Configured to answer business intelligence questions quickly	Designed for huge, varied data sets
Major weaknesses	Redundant data; data must be updated using programs.	Not able to store data as efficiently; limited future	Cannot handle complex data	Limited market acceptance; skills are hard to find.	Highly specialized use; skills are hard to find	New in the market, highly specialized use; skills are hard to find
Data types supported	Simple	*Not recommended for new systems*	Simple	Complex (e.g., video, audio, images)	Aggregated	Mixed data sets with structured and unstructured components
Application system types supported	Transaction processing	*Not recommended for new systems*	Transaction processing and decision making	Transaction processing	Business intelligence	Business intelligence; finding patterns and relationships in mixed data
Existing data formats	Organization dependent	Organization dependent	Organization dependent	Organization dependent	Organization dependent	Organization dependent
Future needs	Limited future prospects	Poor future prospects	Good future prospects	Uncertain future prospects	Uncertain future prospects	New, uncertain future prospects

MOVING FROM A LOGICAL TO PHYSICAL DATA MODEL

Adding implementation details

Transforming the Logical Data Model

- The *logical* ***entity relationship diagrams*** *(ERD)* depicts the "business view" of the data; omits implementation details.

- Having determined the data storage format, *physical data models* are created to show implementation details and to explain more about the "how" of the final system.

The Physical ERD

- The physical ERD includes entities, relationships, and attributes.
- Adds references to how data will be stored.
- Much more metadata is defined.

Steps to Create the Physical ERD

Step	Explanation
1. Change entities to tables or files.	Beginning with the logical ERD, change the entities to tables or files and update the metadata.
2. Change attributes to fields.	Convert the attributes to fields and update the metadata.
3. Add primary keys.	Assign primary keys to all entities.
4. Add foreign keys.	Add foreign keys to represent the relationships among entities.
5. Add system-related components.	Add system-related tables and fields.

Example

- The Campus Housing Service helps students find apartments. Owners of apartments fill in information forms about the rental units they have available (e.g., location, number of bedrooms, monthly rent).

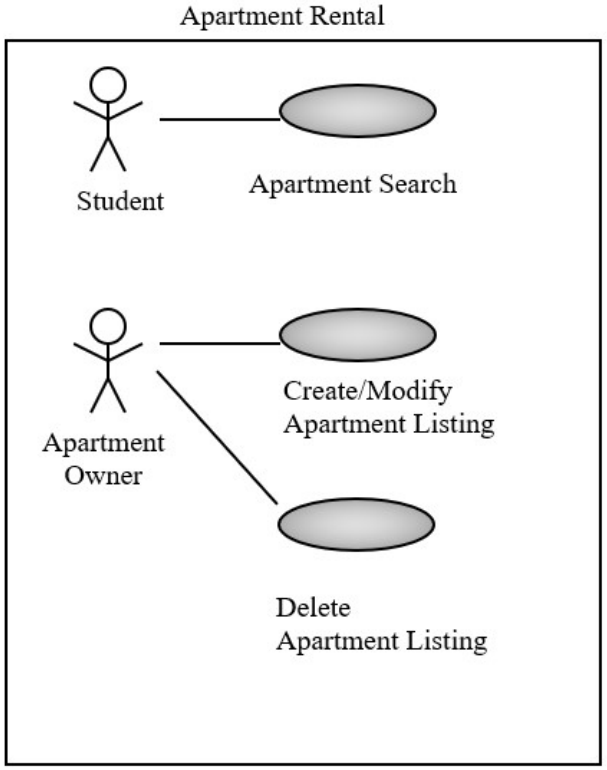
- Students who register with the service can search the rental information to find apartments that meet their needs (e.g., a two-bedroom apartment for $800 or less per month within 1/2 mile of campus). They then contact the apartment owners directly to see the apartment and, possibly, rent it.

- Apartment owners call the service to delete their listing when they have rented their apartment(s)

Recall the Use Case Diagram

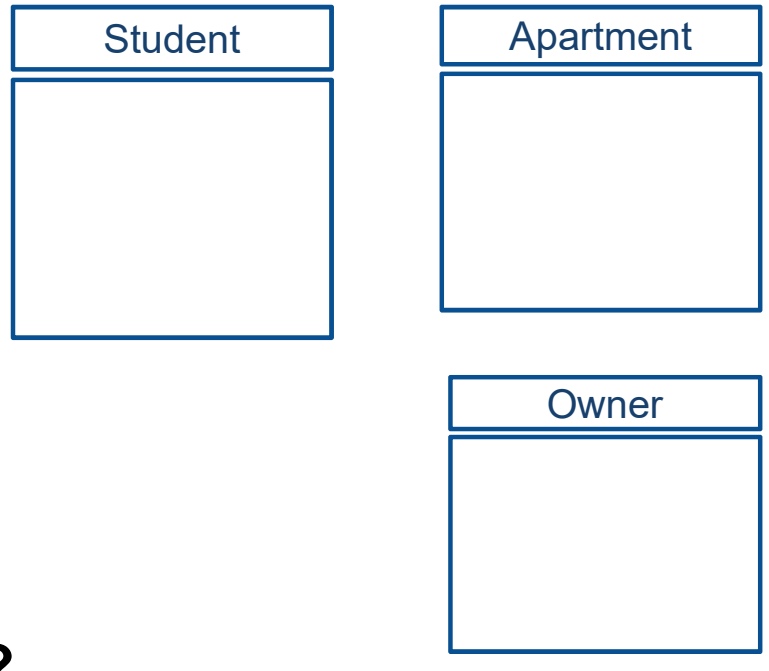
- The Campus Housing Service helps students find apartments. Owners of apartments fill in information forms about the rental units they have available (e.g., location, number of bedrooms, monthly rent).
- Students who register with the service can search the rental information to find apartments that meet their needs (e.g., a two-bedroom apartment for $800 or less per month within 1/2 mile of campus). They then contact the apartment owners directly to see the apartment and, possibly, rent it.
- Apartment owners call the service to delete their listing when they have rented their apartment(s)

What are the entities?



Recall the Use Case Diagram

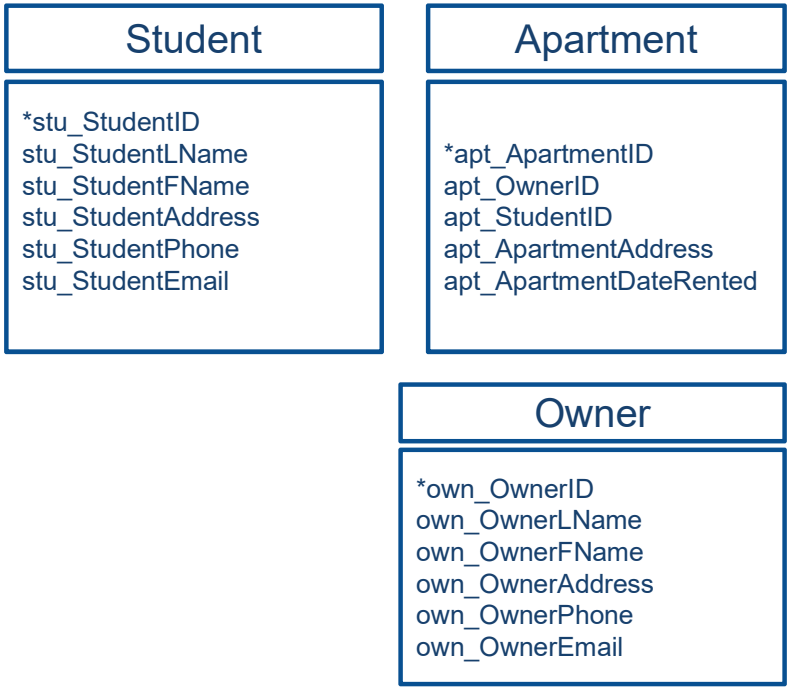
- The Campus Housing Service helps students find apartments. Owners of apartments fill in information forms about the rental units they have available (e.g., location, number of bedrooms, monthly rent).
- Students who register with the service can search the rental information to find apartments that meet their needs (e.g., a two-bedroom apartment for $800 or less per month within 1/2 mile of campus). They then contact the apartment owners directly to see the apartment and, possibly, rent it.
- Apartment owners call the service to delete their listing when they have rented their apartment(s)



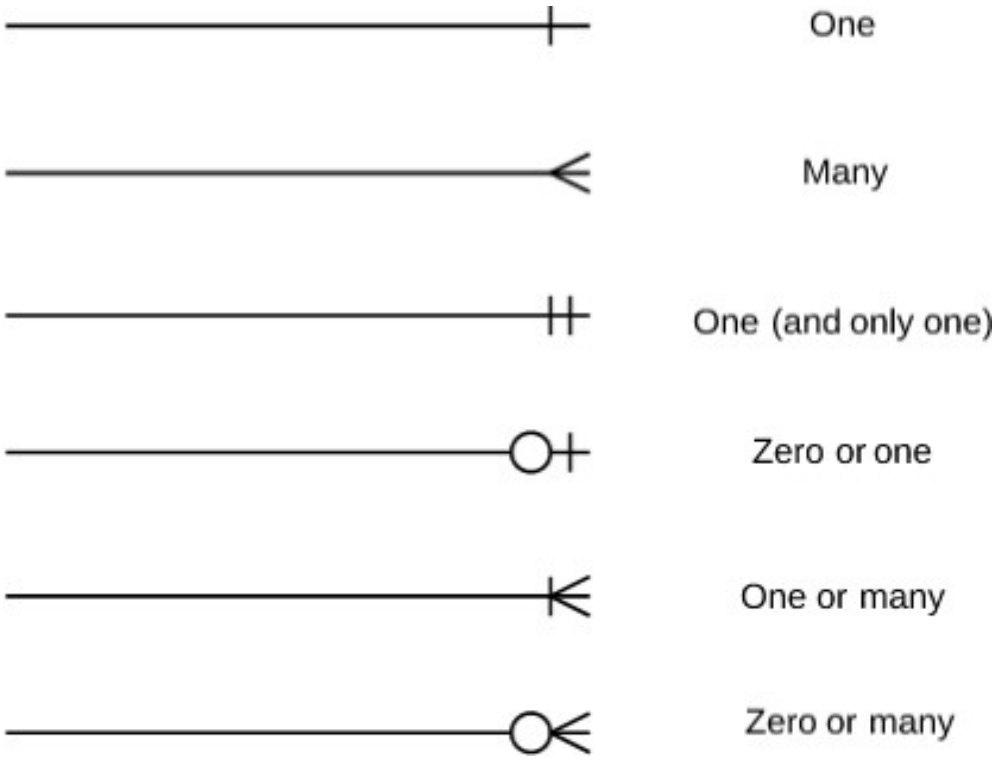
What are the attributes for each entity?
What are the identifiers for the entities

- The Campus Housing Service helps students find apartments. Owners of apartments fill in information forms about the rental units they have available (e.g., location, number of bedrooms, monthly rent).
- Students who register with the service can search the rental information to find apartments that meet their needs (e.g., a two-bedroom apartment for $800 or less per month within 1/2 mile of campus). They then contact the apartment owners directly to see the apartment and, possibly, rent it.
- Apartment owners call the service to delete their listing when they have rented their apartment(s)

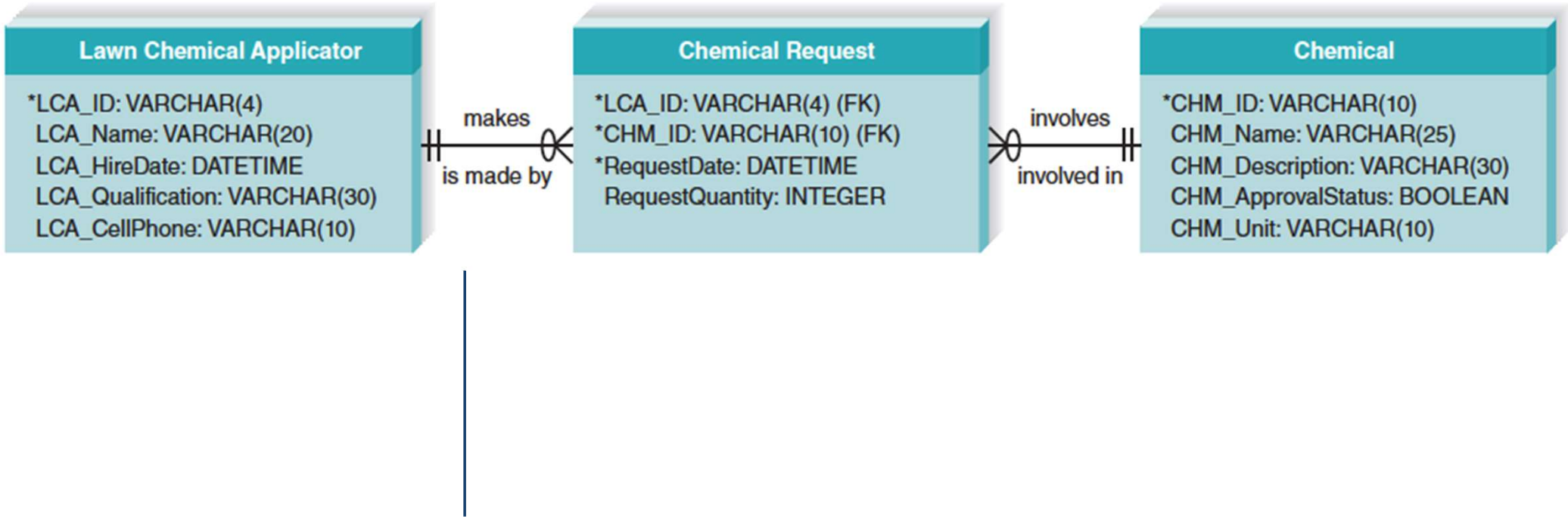
What are the relationships?



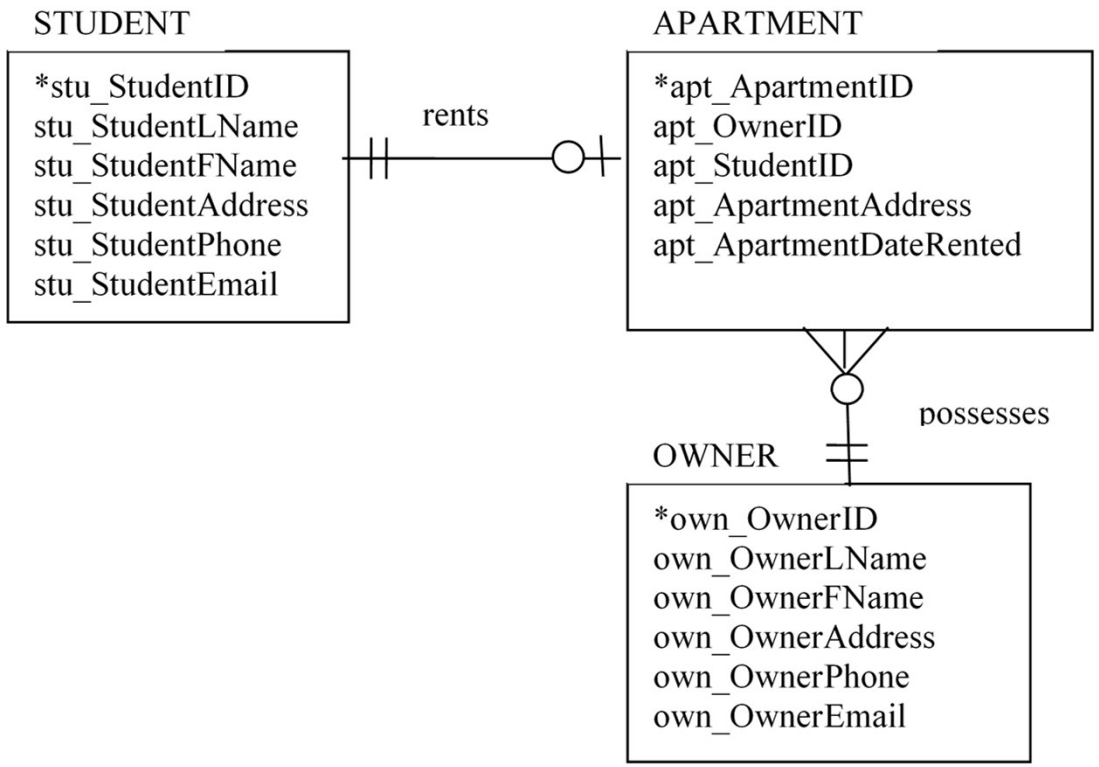
ERD Diagram Symbols



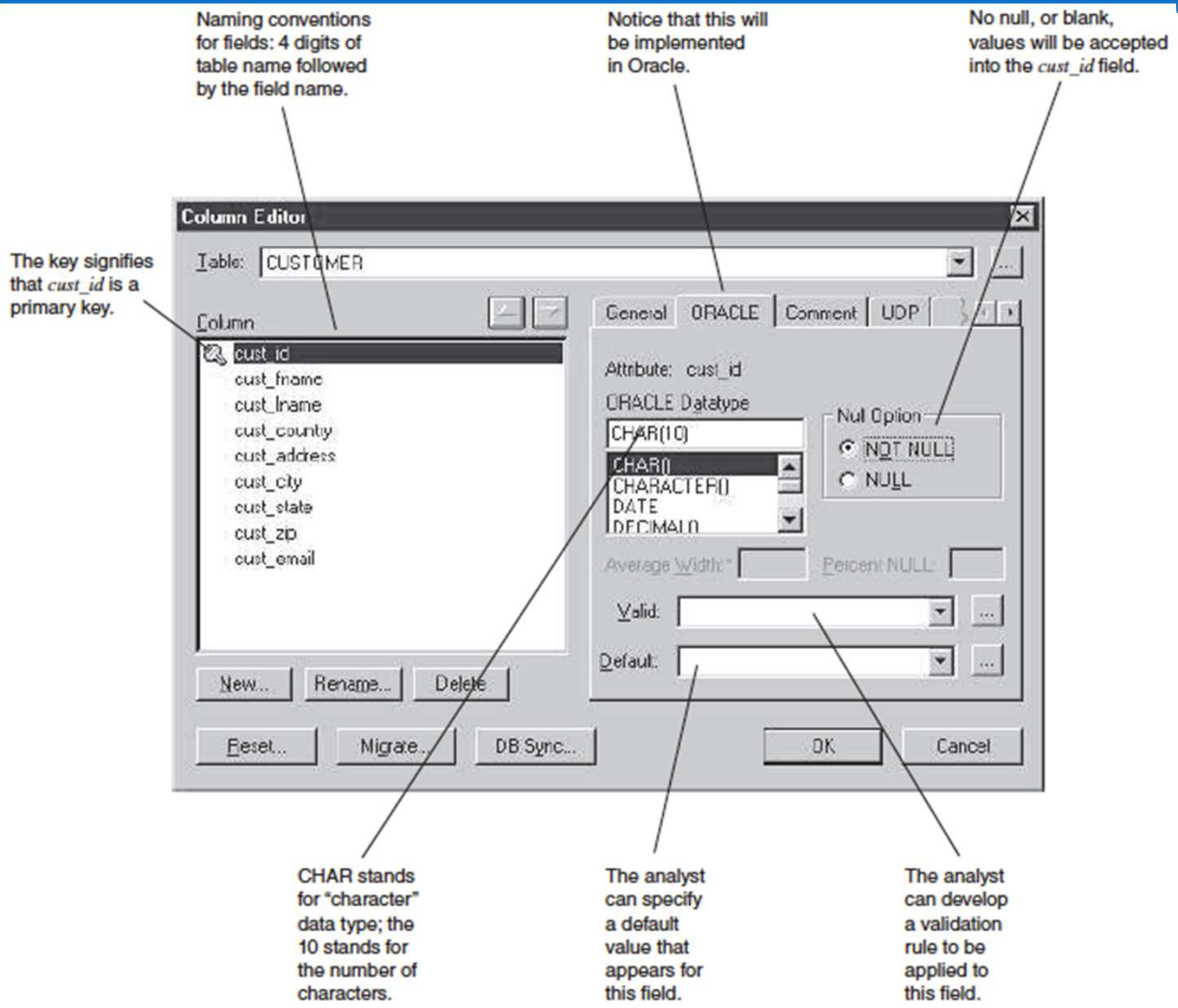
Example Physical ERD



Final ERD


STUDENT
*stu_StudentID
stu_StudentLName
stu_StudentFName
stu_StudentAddress
stu_StudentPhone
stu_StudentEmail
rents
APARTMENT
*apt_ApartmentID
apt_OwnerID
apt_StudentID
apt_ApartmentAddress
apt_ApartmentDateRented
possesses
OWNER
*own_OwnerID
own_OwnerLName
own_OwnerFName
own_OwnerAddress
own_OwnerPhone
own_OwnerEmail

Physical Aspects of Data Element in Metadata



OPTIMIZING DATA STORAGE

Enhance processing efficiency

OPTIMIZING DATA STORAGE

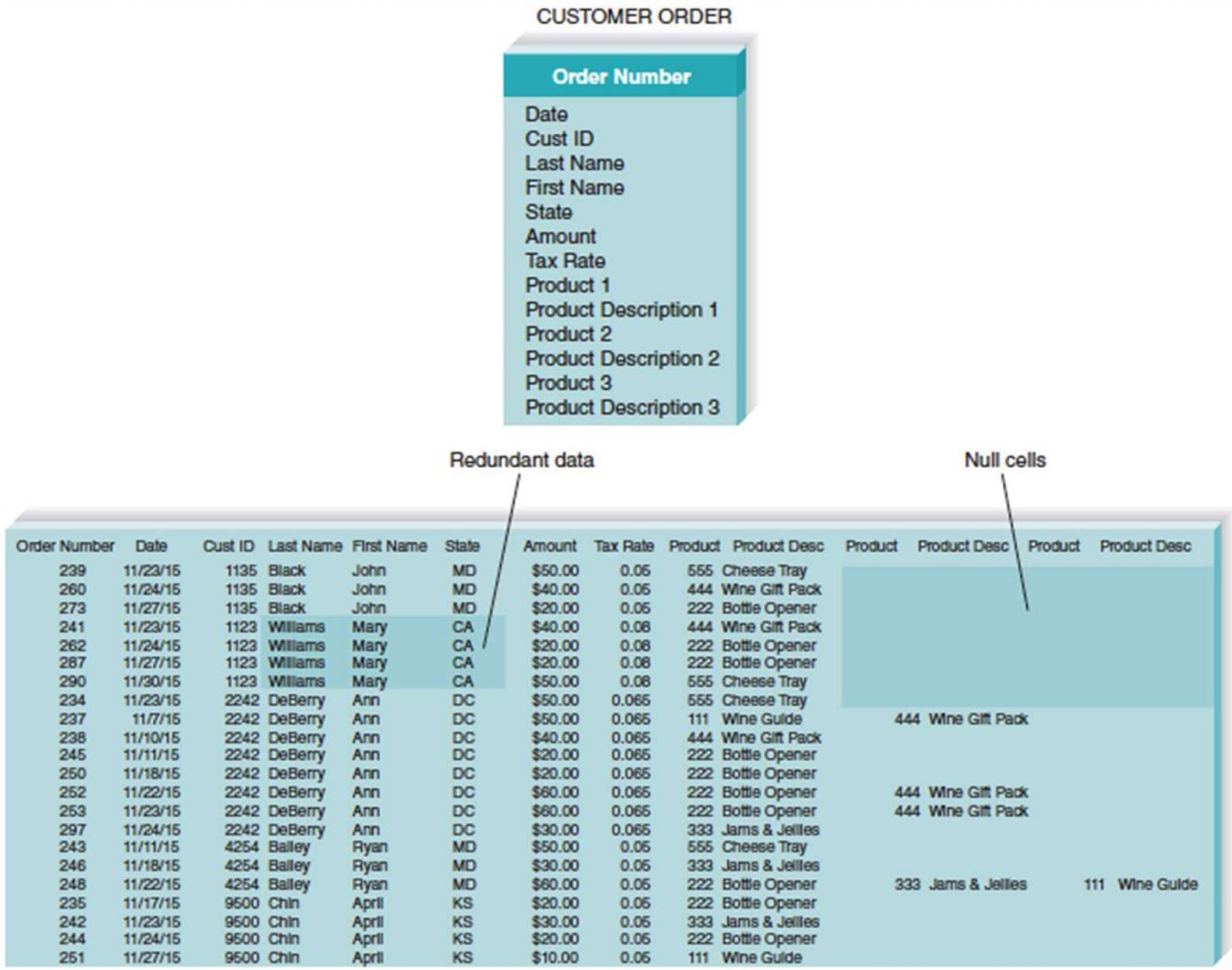
- Next, the data storage format is \optimized for processing efficiency.

- Two primary dimensions:
  - Storage efficiency.
  - Speed of access.

- Limit data redundancy; very few null values.

- Best way to achieve efficiency is normalization.

Optimizing Data Storage

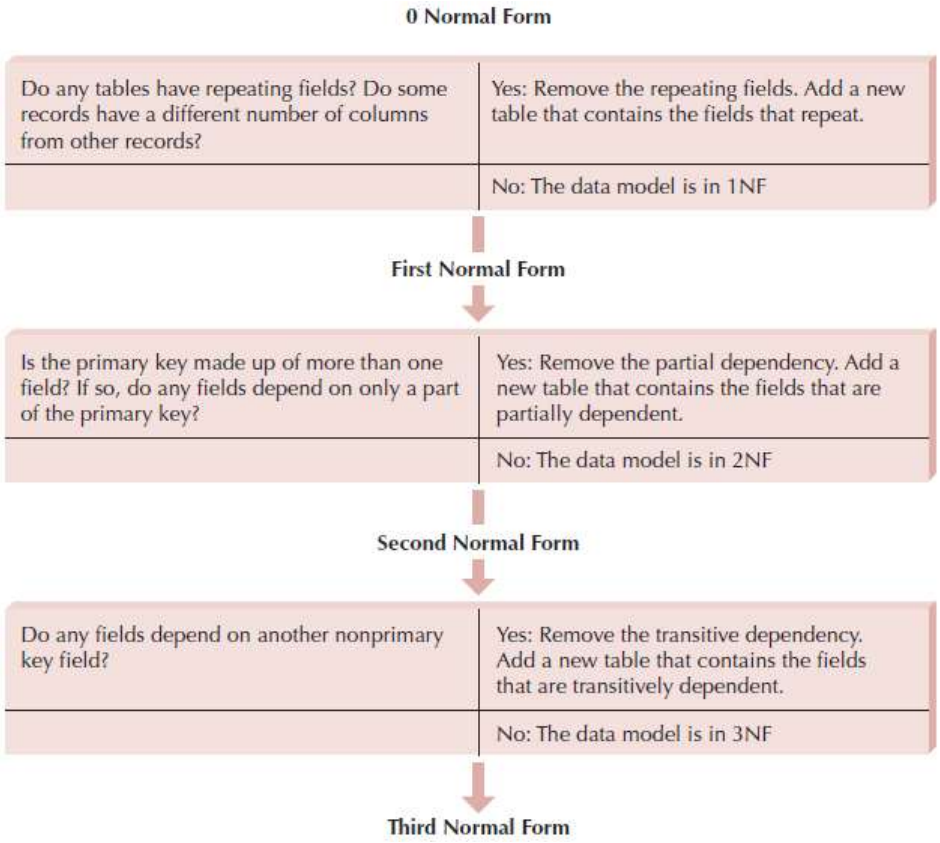


Order Number	Date	Cust ID	Last Name	First Name	State	Amount	Tax Rate	Product	Product Desc	Product	Product Desc	Product	Product Desc
239	11/23/15	1135	Black	John	MD	$50.00	0.05	555	Cheese Tray				
260	11/24/15	1135	Black	John	MD	$40.00	0.05	444	Wine Gift Pack				
273	11/27/15	1135	Black	John	MD	$20.00	0.05	222	Bottle Opener				
241	11/23/15	1123	Williams	Mary	CA	$40.00	0.08	444	Wine Gift Pack				
262	11/24/15	1123	Williams	Mary	CA	$20.00	0.08	222	Bottle Opener				
287	11/27/15	1123	Williams	Mary	CA	$20.00	0.08	222	Bottle Opener				
290	11/30/15	1123	Williams	Mary	CA	$50.00	0.08	555	Cheese Tray				
234	11/23/15	2242	DeBerry	Ann	DC	$50.00	0.065	555	Cheese Tray				
237	11/7/15	2242	DeBerry	Ann	DC	$50.00	0.065	111	Wine Guide	444	Wine Gift Pack		
238	11/10/15	2242	DeBerry	Ann	DC	$40.00	0.065	444	Wine Gift Pack				
245	11/11/15	2242	DeBerry	Ann	DC	$20.00	0.065	222	Bottle Opener				
250	11/18/15	2242	DeBerry	Ann	DC	$20.00	0.065	222	Bottle Opener				
252	11/22/15	2242	DeBerry	Ann	DC	$60.00	0.065	222	Bottle Opener	444	Wine Gift Pack		
253	11/23/15	2242	DeBerry	Ann	DC	$60.00	0.065	222	Bottle Opener	444	Wine Gift Pack		
297	11/24/15	2242	DeBerry	Ann	DC	$30.00	0.065	333	Jams & Jellies				
243	11/11/15	4254	Bailey	Ryan	MD	$50.00	0.05	555	Cheese Tray				
246	11/18/15	4254	Bailey	Ryan	MD	$30.00	0.05	333	Jams & Jellies				
248	11/22/15	4254	Bailey	Ryan	MD	$60.00	0.05	222	Bottle Opener	333	Jams & Jellies	111	Wine Guide
235	11/17/15	9500	Chin	April	KS	$20.00	0.05	222	Bottle Opener				
242	11/23/15	9500	Chin	April	KS	$30.00	0.05	333	Jams & Jellies				
244	11/24/15	9500	Chin	April	KS	$20.00	0.05	222	Bottle Opener				
251	11/27/15	9500	Chin	April	KS	$10.00	0.05	111	Wine Guide				

Normalization

- Store each data fact only once in the database
- Reduces data redundancies and chances of errors
- First four levels of normalization are
  - 0 Normal Form: normalization rules not applied
  - 1 Normal Form: no multi-valued attributes (each cell has only a single value)
  - 2 Normal Form: no partial dependencies (non-key fields depend on the entire primary key, not just part of it)
  - 3 Normal Form: no transitive dependencies (non-key fields do not depend on other non-key fields)

Steps of Normalization

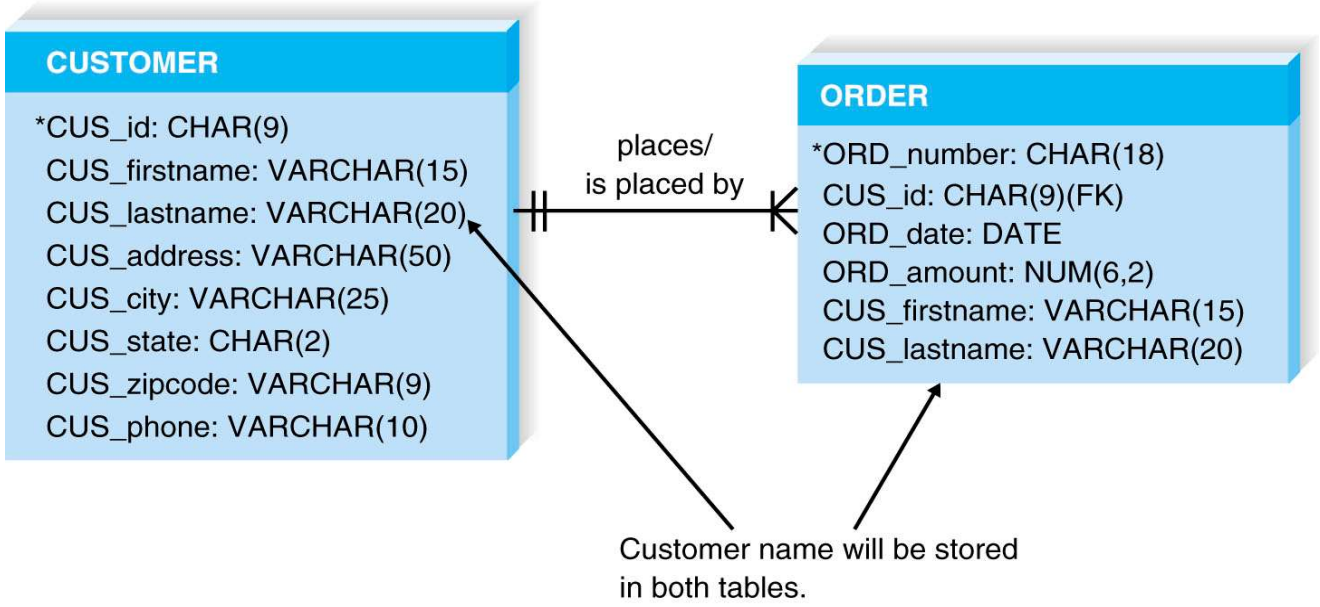


Optimizing Access Speed

- After optimizing for data storage efficiency, data are spread out across a number of tables
- For a large relational database, it is necessary to optimize access speed.
- Techniques of optimizing access speed:
  - Denormalization
  - Clustering
  - Indexing
  - Estimating the size of data for hardware planning

Denormalization

- Add redundancy back into the design.
- Reduce the number of joins required during processing to enhance data access speed.



Clustering

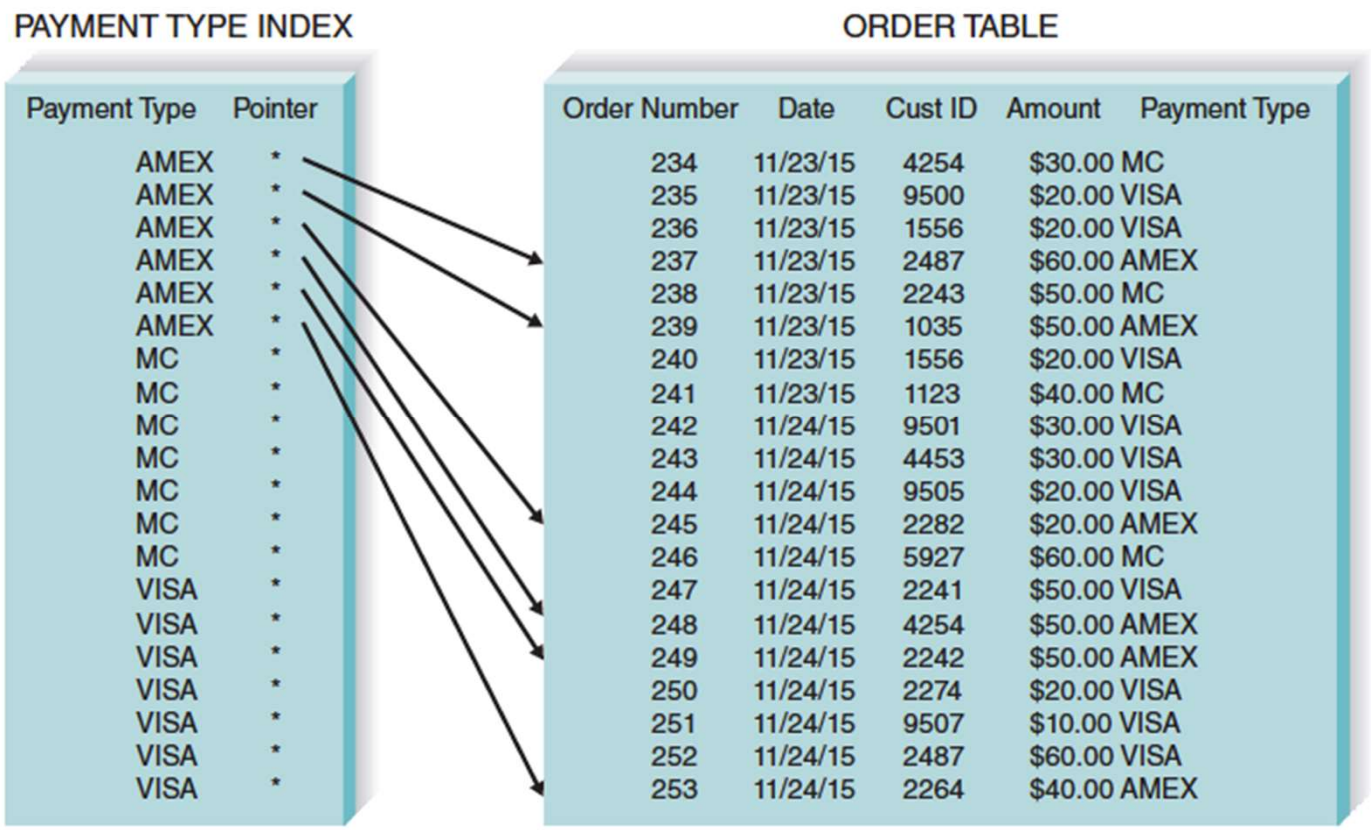
- Placing records together physically so that like records are stored close together.
- *Intrafile clustering:* Similar records in the table are stored together.
- *Interfile clustering:* Combining records from more that one table that typically are retrieved together.

Indexing

- A data storage *index* is a minitable (similar to an index of a book) containing values from one or more columns in a table and the location of the values within the table.

- Indexes require overhead; they take up storage space.

Index Illustration



Indexing Guidelines

- Use indexes sparingly for transaction systems.
- Use many indexes to improve response times in business intelligence systems.
- For each table, create a unique index that is based on the primary key.
- For each table, create an index that is based on the foreign key to improve the performance of joins.
- Create an index for fields that are used frequently for grouping, sorting, or criteria.

Estimating Storage Size

- ***Volumetrics*** – technique of estimating the amount of data that the hardware must support.

1. Calculate the amount of *raw data* - all the data stored within the database tables.

2. Calculate the *overhead* requirements based on the DBMS vendor's recommendations.

3. Record the number of initial records loaded into the table, as well as the expected growth per month.

Sample Volumetrics Calculation

Field	Average Size (Characters)
Order number	8
Date	7
Cust ID	4
Last name	13
First name	9
State	2
Amount	4
Tax rate	2
Record size	49
Overhead	30%
Total record size	63.7
Initial table size	50,000
Initial table volume	3,185,000
Growth rate/month	1000
Table volume @ 3 years	5,478,200

EXERCISE

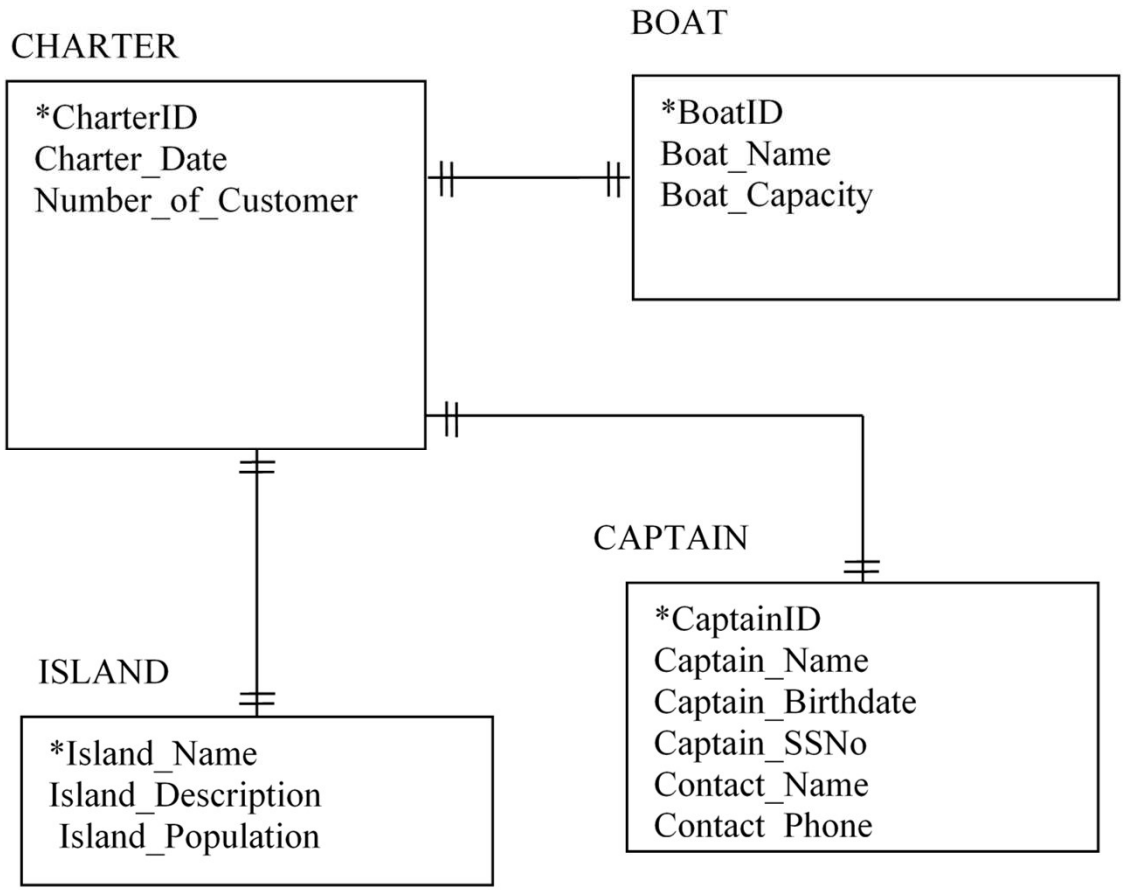
A charter company owns boats that are used for charter trips to islands. The company has created a computer system to track the boats it owns, including each boat's ID number, name, and seating capacity.

The company also tracks information about the various islands, such as their names and population. Every time a boat is chartered, it is important to know the date that the trip is to take place and the number of people on the trip.

The company also keeps information about each captain, such as Social Security number, name, birthdate, and contact information for next of kin. Boats travel to only one island per visit.

Create a data model. Include entities, attributes, identifiers, and relationships.

Solution


CHARTER
*CharterID
Charter_Date
Number_of_Customer
BOAT
*BoatID
Boat_Name
Boat_Capacity
CAPTAIN
*CaptainID
Captain_Name
Captain_Birthdate
Captain_SSNo
Contact_Name
Contact_Phone
ISLAND
*Island_Name
Island_Description
Island_Population