BZAN 6354

Lecture 1

January 22, 2024

Dr. Mark Grimes, Ph.D.
gmgrimes@bauer.uh.edu

UNIVERSITY of HOUSTON
C. T. BAUER COLLEGE of BUSINESS
Department of Decision & Information Sciences

Welcome to BZAN 6354!

Tonight's Agenda:

- Course Overview
- About me
- About you
- About the class (syllabus and what not)

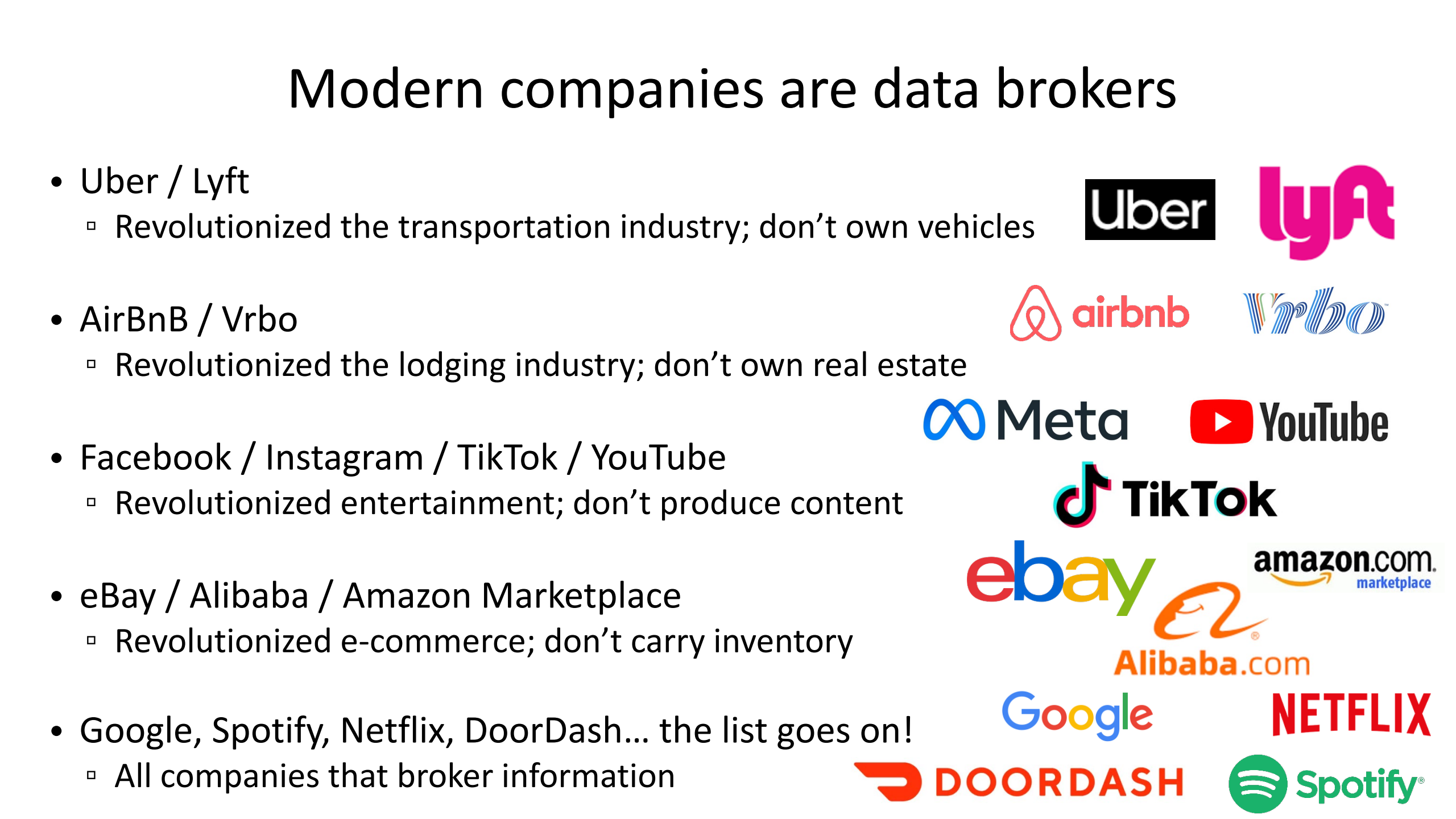
- Break
  - 10 minutes

- Content
  - Module 1.1: Data, Information, and Metadata
  - Module 1.2: Data Management
  - Module 1.3: Limitations of File Systems
  - Module 1.4: Three Schema Architecture
  - Module 1.5: Characteristics of Database Systems
  - Module 1.6: Data Models

A deep question to start…

Modern companies are data brokers

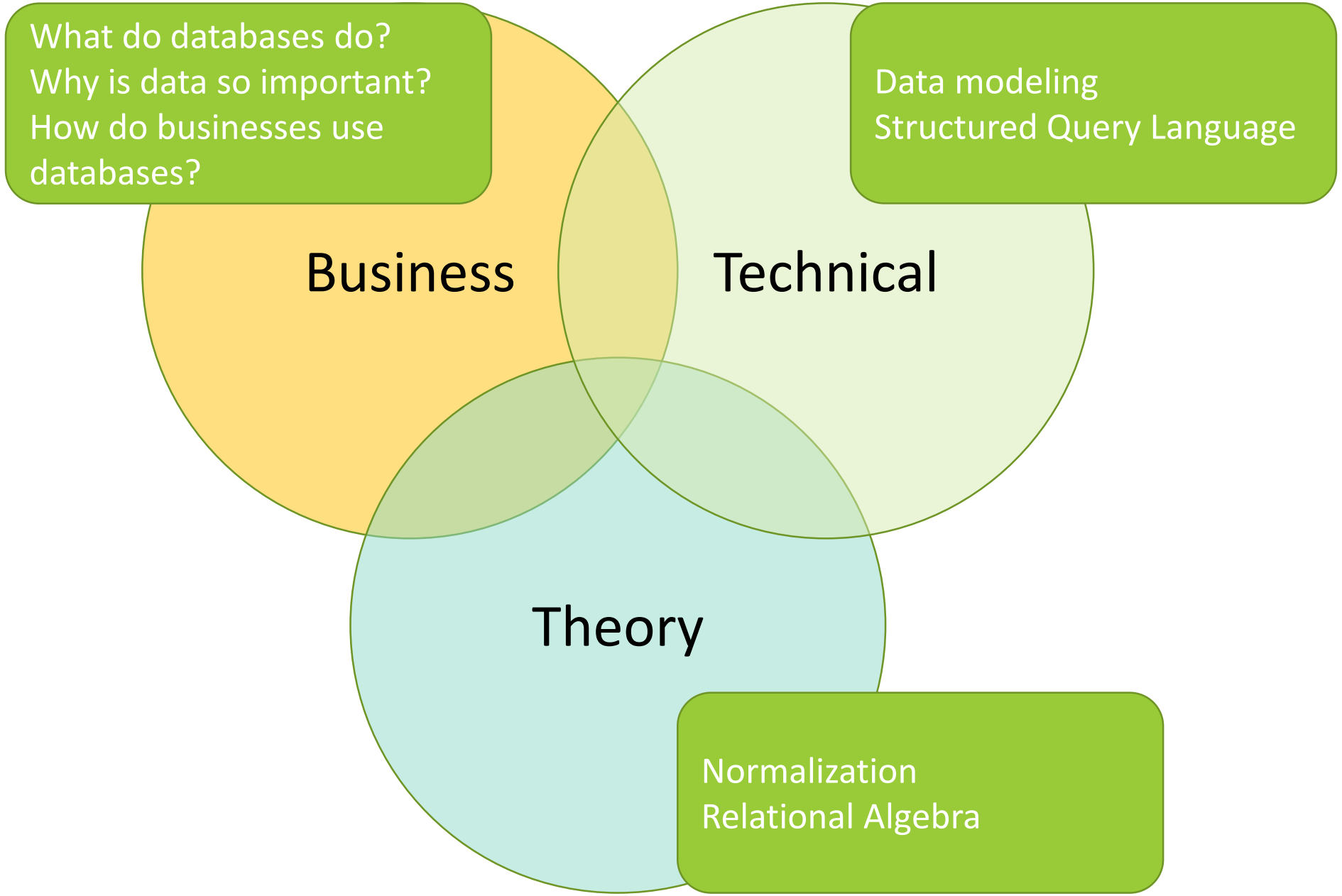
- Uber / Lyft
  - Revolutionized the transportation industry; don't own vehicles
- AirBnB / Vrbo
  - Revolutionized the lodging industry; don't own real estate
- Facebook / Instagram / TikTok / YouTube
  - Revolutionized entertainment; don't produce content
- eBay / Alibaba / Amazon Marketplace
  - Revolutionized e-commerce; don't carry inventory
- Google, Spotify, Netflix, DoorDash… the list goes on!
  - All companies that broker information



Modern companies are data brokers

- Even companies that are not traditionally “data” companies need good data management strategies to compete in a modern world
  - Data to facilitate supply chains
  - Analytics to understand customers
  - Predicting demand
  - Predicting equipment failures
  - Understanding employee workload / resource utilization
  - Automation / AI

- Data was “nice to have” twenty years ago – it is a “must have” today


What do databases do?
Why is data so important?
How do businesses use databases?
Business
Technical
Data modeling
Structured Query Language
Theory
Normalization
Relational Algebra

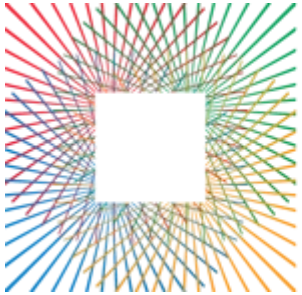
About you (get ready)

- 30 seconds
  - First and Last Name (as it is in Canvas)
  - Where are you from?
  - What program are you in (BZAN, MIS, or something else?)
  - Something interesting about yourself

Who I am

- Began working in IT in 1996
  - Small ISP in Mississippi

- Prior to academia, worked in industry for nine years
  - IT Infrastructure Architect for BNY Mellon



The BANK of NEW YORK

2007

BNY MELLON

- Ph.D. from University of Arizona in 2015
  - Research in Credibility Assessment, Information Systems Security, and Human-Computer Interactions (HCI)


BACKUP01
ACRUX
3.2GB
IBMEMERGENCY
WINDOWS
ALDEBARAN NEW
IBMEMERGENCY20
VALKANOS
OXIA
IBMEMERGENCY2
CALLISTO
DANZIG


Google
google.com/datacenters

About you

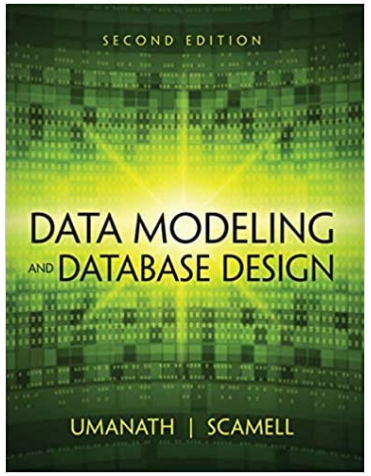
- 30 seconds
  - First and Last Name (as it is in Canvas)
    - Loud and clear please, as I'm frantically searching for you on my roll!
  - What program are you in?
    - BZAN, MIS, etc…
  - Where are you from?
  - Something interesting about yourself

Smile!

…or not, it’s up to you

Syllabus and expectations

- Everything will be posted to Canvas
  - http://canvas.uh.edu



- **Optional** Textbook
  - Data Modeling and Database Design 2nd Edition
    N. S. Umanath and R. W. Scamell, ISBN: 1-285-08525-6

- Cheating
  - Don’t do it – you will be caught, and that’s no fun for anyone
  - Some assignments allow working in teams – it is to your benefit to actually DO some of the work!

- Threatening/disruptive actions
  - Also don’t do this

Course Structure: Resources

- Canvas: https://canvas.uh.edu
- YouTube: https://YouTube.com/@ProfessorMarkGrimes
- Kahoot!: https://kahoot.it
- Office hours:
  - Face to face (Melcher 290H): Monday 10:00 – 11:00 or by appointment

Course Structure: Schedule

- Tentative schedule

Date	Topic	Notes
1: 2024-01-22	Intro / Database Concepts	Tonight!
2: 2024-01-29	Conceptual Data Modeling Introduction to Data Definition Language (DDL) and Structured Query Language (SQL)	Assignment 1 Assigned
3: 2024-02-05	Relationships	
4: 2024-02-12	Entity Relationship Modeling	Assignment 2 Assigned
5: 2024-02-19	Relational Data Modeling	
6: 2024-02-26	Relational Data Modeling	
7: 2024-03-04	Exam 1	In class
8: 2024-03-11	Spring Break (No class!)	

Course Structure: Schedule

- Tentative schedule

Date	Topic	Notes
9: 2024-03-18	Relational Algebra Structured Query Language	
10: 2024-03-25	Structured Query Language	Assignment 3 Assigned
11: 2024-04-01	Normalization	
12: 2024-04-08	Normalization	Assignment 4 Assigned
13: 2024-04-15	Advanced / Applied SQL	
14: 2024-04-22	Wrap up / Review	
15: 2024-04-29	Exam 2	In class

Course Structure: Grading

- 50% Exams
  - Two @ 25% each
- 20% Progress Quizzes
  - 10 quizzes on Canvas, almost every week
- 10% Assignments
  - Four @ 2.5% each
- 10% SQL Assignment
  - One large multi-part assignment
- 10% Professionalism
  - Do not turn in excessively messy work
  - Do not cheat/mislead/misrepresent yourself
  - Do not ask for your grade to be increased

Grade Allocation	
A	90-100%
B	80-89%
C	70-79%
D	60-69%
F	< 60%

Exams (50%)

- In this room during class time

- Will be approximately:
  - 1/3: Multiple Choice
  - 1/3: Short answer / Fill in blank / Matching / etc.
  - 1/3: ER Diagram (exam 1), SQL/Normalization (exam 2)

- I do not provide a “study guide” however:
  - The Learning Objectives at the start of each module could serve as such
  - The Progress Quizzes are made up of questions very similar to what you will see on the exams

Progress Quizzes (20%)

- Progress Quizzes will be on Canvas almost every week

- Due by 5:00 PM on Friday
  - You have up to three attempts on the quiz
  - No makeups or late submissions allowed… The point is to keep you current on the material!

- We will go over the questions on the Progress Quiz in class BEFORE the quiz is due! (so… come to class and the quiz will be easy!)

- Using Kahoot!

Assignments (10%)

- For each assignment there will be a “walk through” video
  - Link will be in the assignment file

- You should:
  1. Attempt the assignment yourself
  2. When you feel either a) **Lost** or b) **Successful**, watch the video
  3. Correct your mistakes based on the video
  4. Write a short summary of your mistakes, how you fixed them, what you learned, etc.
     - Between two and five sentences
  5. Submit your assignment

- If you follow this process you will hopefully:
  1. Correctly learn the material
  2. Generate a submission that should earn close to full marks

Assignments (10%)

- Assignments (each worth 25 points, or 2.5% of your grade) will be rapidly assessed and receive one of four grades:
  - 25: No immediately obvious errors
  - 23: Some minor errors, but overall good work
  - 15: Glaring errors, messy or incomplete work, etc.
  - 0: Did not submit, or submission was not reasonable

- If you want detailed feedback, contact me and I will more carefully assess your work and respond back.

Professionalism (10%)

- While this class is largely technical in nature, being able to interact effectively with others is an important part of using data in business.

- To this end, points may be deducted for:
  - Excessively messy work
  - Disruptive or disrespectful behavior
  - Unethical behavior including but not limited to asking for your grade to be increased, academic integrity infractions
  - Other unscrupulous behaviors

- PLEASE NOTE – I am 100% OK with you asking to review your grades, discussing your exams, etc…

Learning Objectives

- These are the high level questions you should be able to answer
- Great study guide for the exams
- Database design is as much an art as a science
  - There is often not one “right” answer – but some answers are better than others (and some are wrong…)

This is a course about relational databases

- This is a course about relational databases, but what does that mean?
  - Relational databases are made up of “relations” which are collections (or sets) of related pieces of data

- Relational databases are foundationally built on the branch of mathematics known as Set Theory

- Relational databases have been around since 1970
  and are a huge driving force in business

Non-Relational Databases

- An emerging database paradigm is “non-relational” databases
  - Column Family Databases (HBase, Cassandra)
  - Document Databases (MongoDB, CouchDB)
  - Graph Databases (Neo4j, Amazon Neptune)
  - Key-Value Databases (Redis, Amazon DynamoDB)

- If you find this content intriguing, consider taking BZAN 6356 next semester!

Any questions?

10 minute break

- Since this is a 3 hour class, we will have a 10 minute break in the middle of class each week – please be back promptly!

Welcome Back!

Let’s do this!

- Module 1.1: Data, Information, and Metadata
- Module 1.2: Data Management
- Module 1.3: Limitations of File Systems
- Module 1.4: Three Schema Architecture
- Module 1.5: Characteristics of Database Systems
- Module 1.6: Data Models

Module 1.1
Data, Information, and Metadata

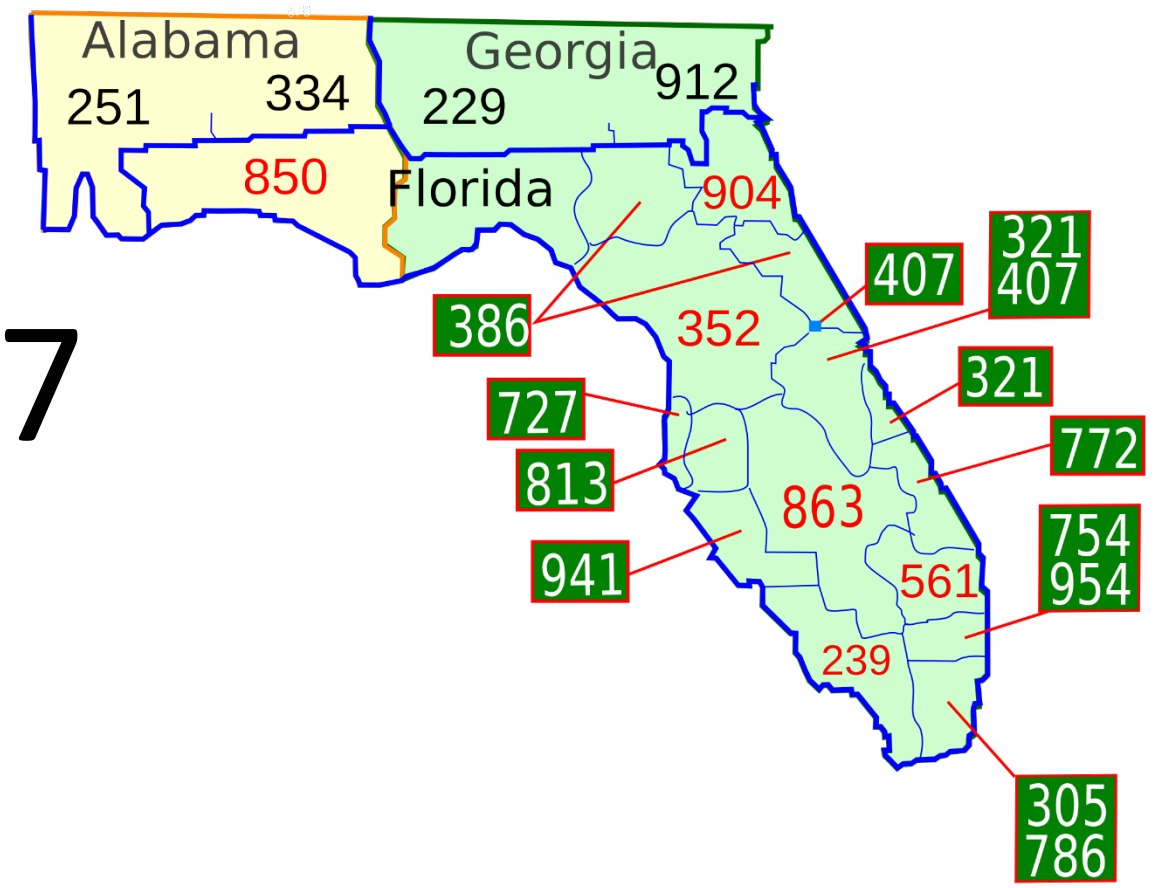
- What is the difference in data and information?
- What is metadata?

What does this number mean?

2397111317

Now?

(239) 711-1317



Now?

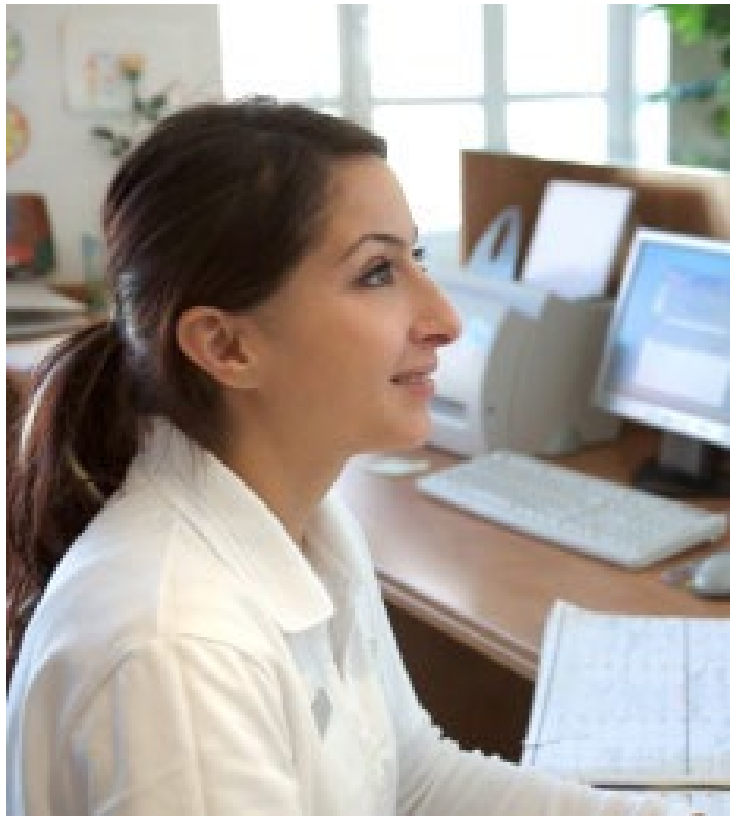
23.971°,11.317°

Now?

2, 397-11-1317

1 = Male

2 = Female



The number means nothing without context!

Data = Raw, unorganized facts

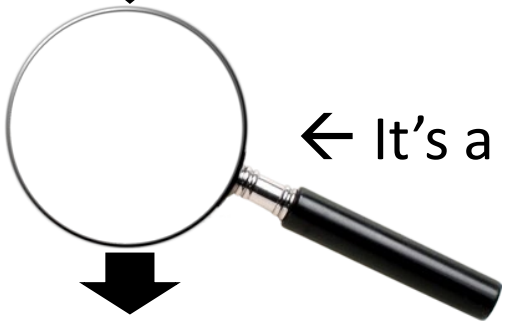
Information = Organized data with context

Metadata

- Metadata is the “lens” we use to understand what data means

2397111317

Metadata “Lens” →



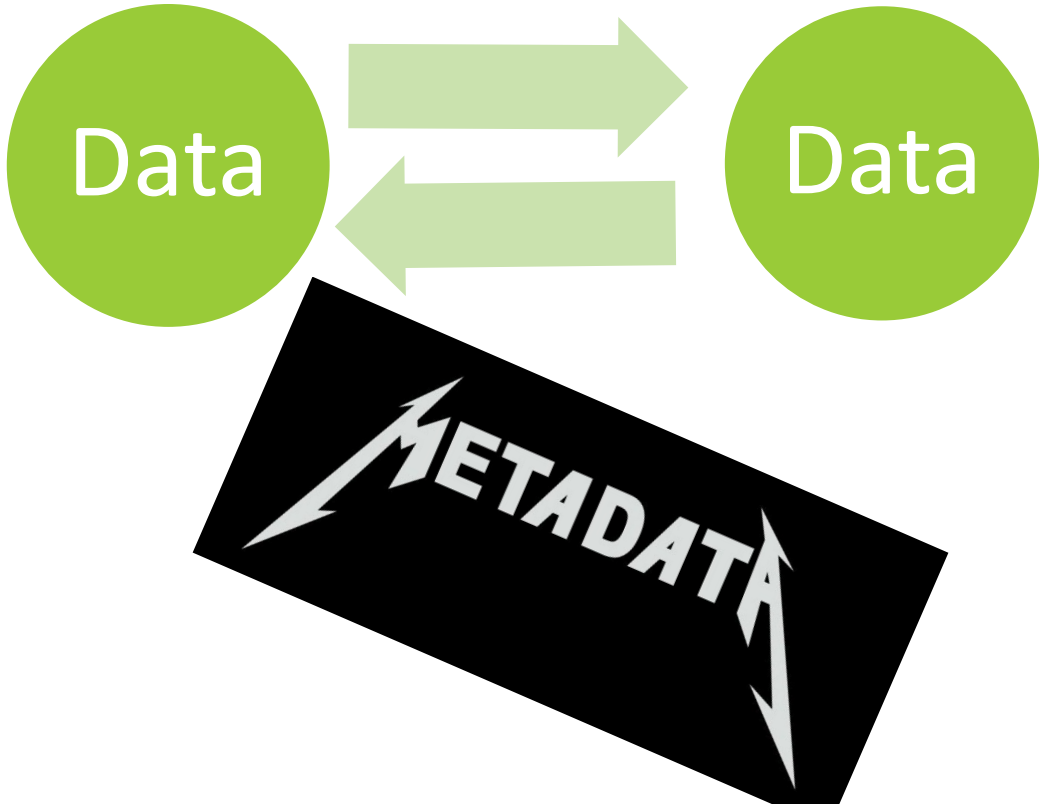
← It’s a phone number!

(239) 711-1317

- Describes properties of data so we can *infer* information

Metadata

Metadata is DATA that describes your DATA



What is all this?

12012012,345844475,2295,2213,140223
12012012,345844475,1245,25100,115123
12012012,427658847,1154,885,57625
12052012,345844475,3011,754,114369
12062012,427658847,9584,10001,47624
12082012,427658847,2295,2523,45101
12122012,345844475,9584,12245,101217
12152012,345844475,1154,1300,99917
12192012,345844475,1154,907,113462
12192012,427658847,2224,1085,44016
12192012,427658847,1154,975,43041
12222012,427658847,2224,1085,41956
12231012,427658847,3030,122,41834
12262012,427658847,2295,1850,39984
12272012,427658847,1199,1925,38059
12272012,427658847,2224,1085,36974
12292012,427658847,9999,2000,34974

Raw data becomes information

Date	Cust_ID	Vend_ID	Charge	Balance
12-01-2012	345-84-4475	2295	$22.13	$1,402.23
12-01-2012	345-84-4475	1245	$251.00	$1,151.23
12-01-2012	427-65-8847	1154	$8.85	$576.25
12-05-2012	345-84-4475	3011	$7.54	$1,143.69
12-06-2012	427-65-8847	9584	$100.01	$476.24
12-08-2012	427-65-8847	2295	$25.23	$451.01
12-12-2012	345-84-4475	9584	$122.45	$1,012.17
12-15-2012	345-84-4475	1154	$13.00	$999.17
12-19-2012	345-84-4475	1154	$9.07	$1,134.62
12-19-2012	427-65-8847	2224	$10.85	$440.16
12-19-2012	427-65-8847	1154	$9.75	$430.41
12-22-2012	427-65-8847	2224	$10.85	$419.56
12-23-1012	427-65-8847	3030	$1.22	$418.34
12-26-2012	427-65-8847	2295	$18.50	$399.84
12-27-2012	427-65-8847	1199	$19.25	$380.59
12-27-2012	427-65-8847	2224	$10.85	$369.74
12-29-2012	427-65-8847	9999	$20.00	$349.74

Raw data becomes information

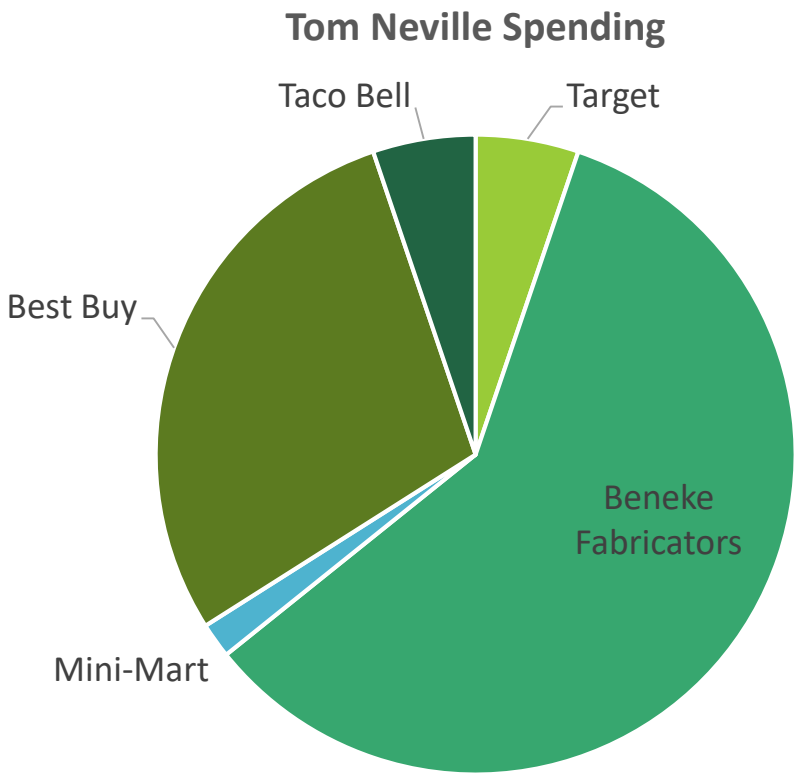
Date	Cust_ID	Vend_ID	Charge	Balance
12-01-2012	345-84-4475	2295	$22.13	$1,402.23
12-01-2012	345-84-4475	1245	$251.00	$1,151.23
12-01-2012	427-65-8847	1154	$8.85	$576.25
12-05-2012	345-84-4475	3011	$7.54	$1,143.69
12-06-2012	427-65-8847	9584	$100.01	$476.24
12-08-2012	427-65-8847	2295	$25.23	$451.01
12-12-2012	345-84-4475	9584	$122.45	$1,012.17
12-15-2012	345-84-4475	1154	$13.00	$999.17
12-19-2012	345-84-4475	1154	$9.07	$1,134.62
12-19-2012	427-65-8847	2224	$10.85	$440.16
12-19-2012	427-65-8847	1154	$9.75	$430.41
12-22-2012	427-65-8847	2224	$10.85	$419.56
12-23-1012	427-65-8847	3030	$1.22	$418.34
12-26-2012	427-65-8847	2295	$18.50	$399.84
12-27-2012	427-65-8847	1199	$19.25	$380.59
12-27-2012	427-65-8847	2224	$10.85	$369.74
12-29-2012	427-65-8847	9999	$20.00	$349.74

Vend_ID	Vendor
1154	Taco Bell
1199	Lowes
1245	Beneke Fabricators
2224	Los Pollos Hermanos
2295	Target
3011	Mini-Mart
3030	Quick Stop
9584	Best Buy
9999	ATM Cash Withdraw

Cust_ID	Customer
345-84-4475	Tom Neville
427-65-8847	Hal Wilkerson

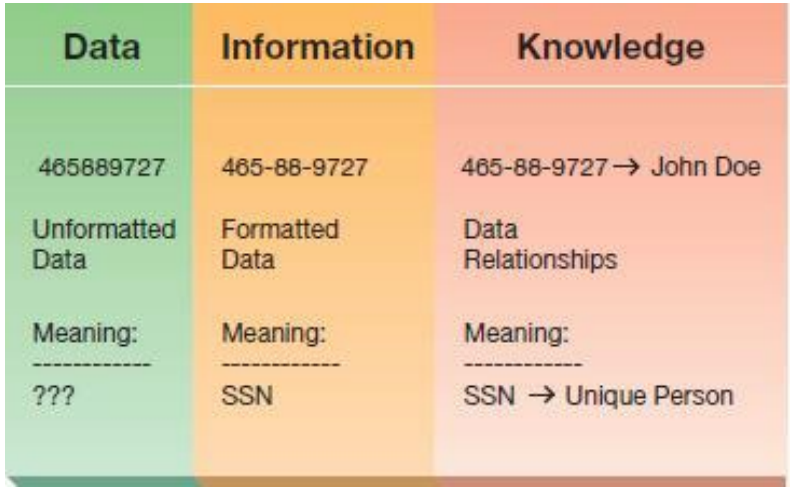
Information becomes knowledge

- Now we have a better understanding of our customers!
- Business Analytics



Data: The Root and Purpose of IS

- **Data**: Raw, unorganized facts
- **Information**: Data in context - transformed to have meaning
- **Knowledge**: Ability to understand information, form opinions, and make decisions or predictions



- We use information systems to convert data into information on which decisions can be based

Module 1.1
Data, Information, and Metadata

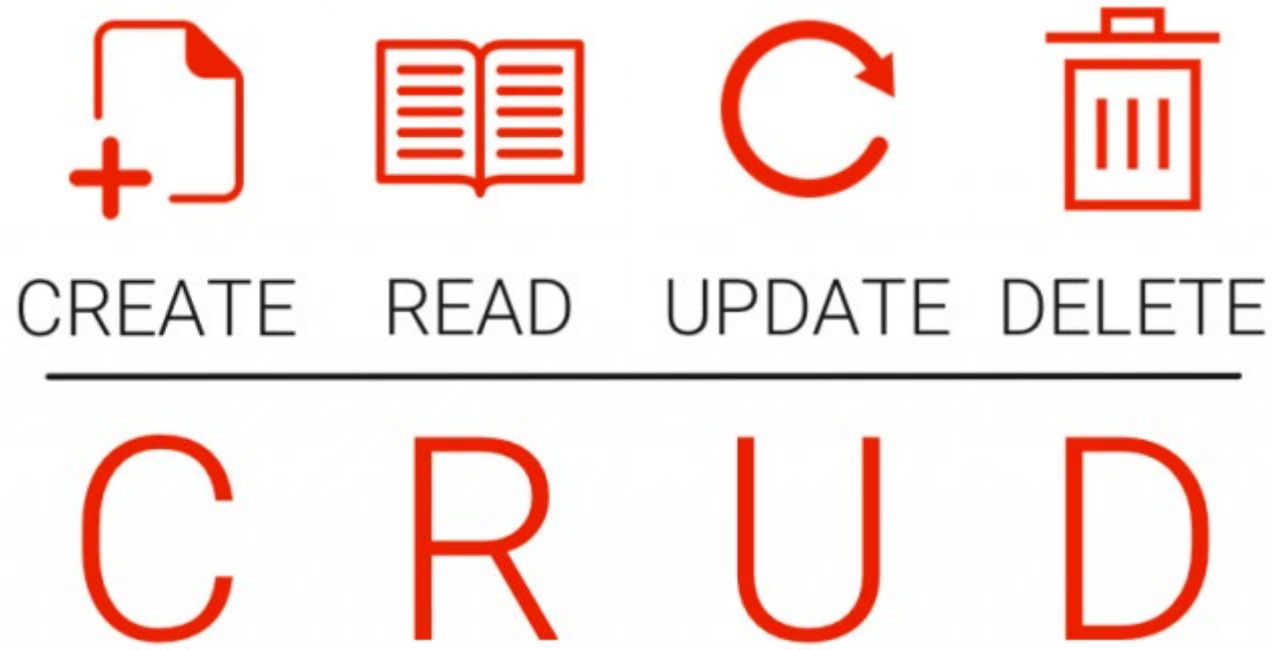
- What is the difference in data and information?
- What is metadata?

Module 1.2
Data Management

- What are the four actions of data management?
- What is a DBMS?

Four actions of data management

- An unfortunate abbreviation:



How do we manage a database?

- Database Management System (DBMS)
  - Oracle
  - Microsoft SQL Server
  - Microsoft Access
  - MySQL
  - PostgreSQL

ORACLE®

MySQL®

Microsoft® SQL Server®

Access

PostgreSQL

How do we manage a database?

- Database Management System (DBMS)
  - Provides a layer between the applications/users and the actual data

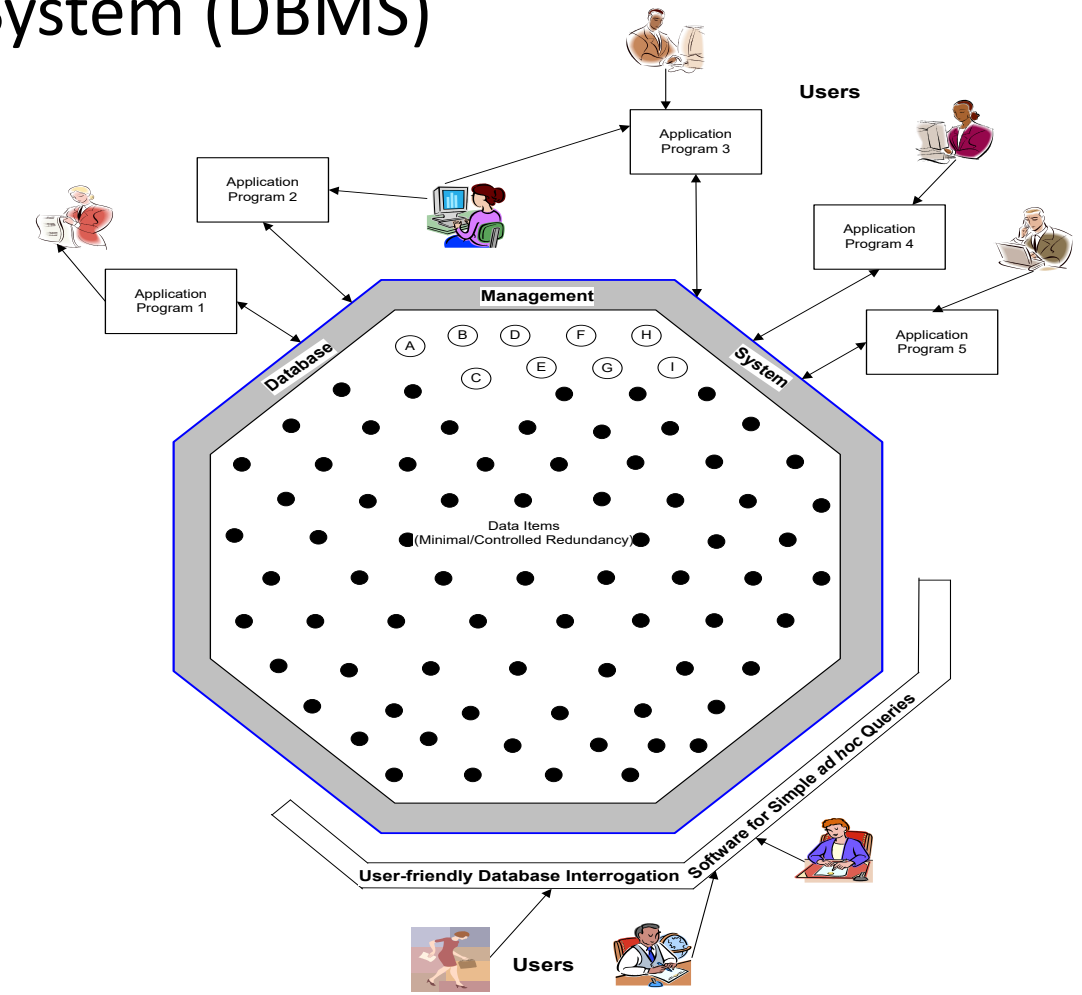
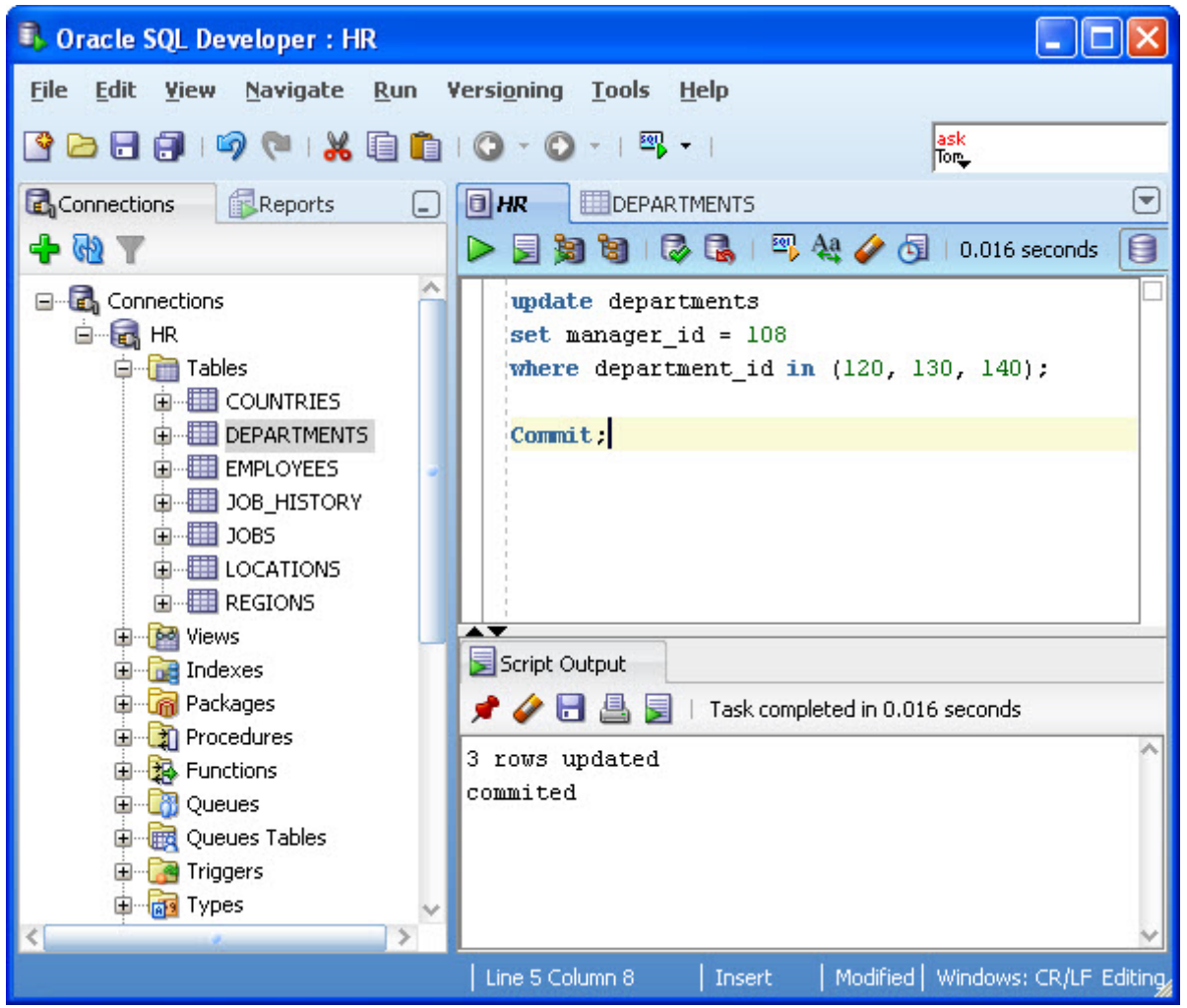


Figure 1.4 An early view of a database system*

*Adapted From Richard L. Nolan, "Computer Data Bases: The Future is Now," Harvard Business Review, (September-October, 1973)

How do we manage a database?

- Database Management System (DBMS)
  - Provides a layer between the applications/users and the actual data



Module 1.2
Data Management

- What are the four actions of data management?
- What is a DBMS?

Module 1.3
Limitations of File Systems

- What are three limitations of file-processing systems?
- What are the two fundamental problems that lead to these limitations?

The “old” way:
Data stored in files managed by the application

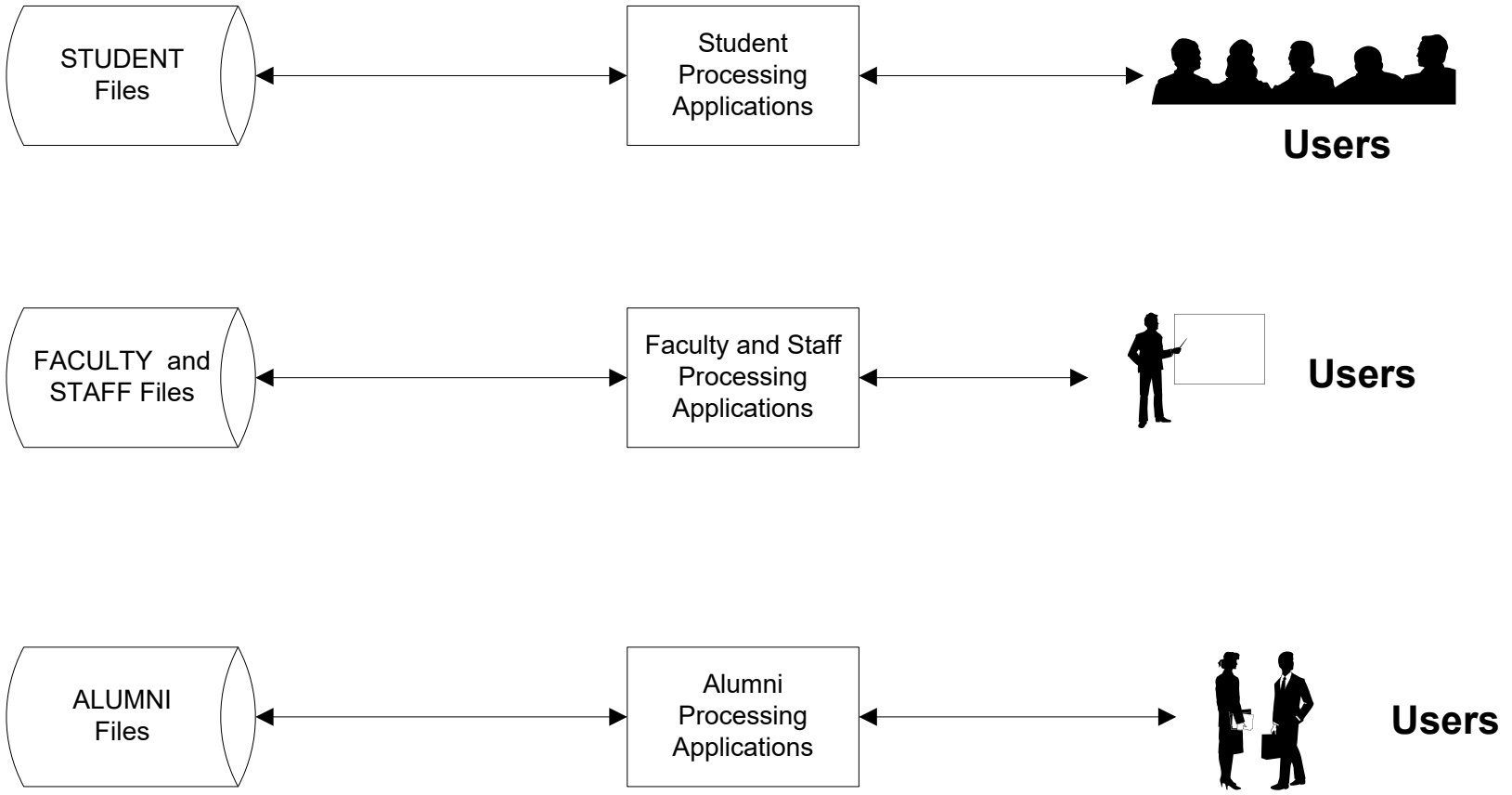


Figure 1.1 An example of a file processing environment

Limitations

- Lack of Data Integrity
  - Values may be incorrect, inconsistent, or out of date in isolated systems
- Lack of Standards
  - It may be difficult to keep data in the same format
  - Difficult to update data when standards change
- Lack of Flexibility/Maintainability
  - Dependent on a programmer to modify the data structures

Fundamental problems

- Lack of integration
  - Data are stored in separate, isolated files within the file system

- Lack of program-data independence
  - The structure of the data (i.e., metadata) is embedded in the application

Data Integration

- When we separate the DATA from the APPLICATION, great things happen:
  - Many applications can use the same data
  - Data can be updated independent of the application
  - Data can be modified without knowledge of programming

The “old” way

- Could a person be a Student, Alum, and Faculty all at the same time?

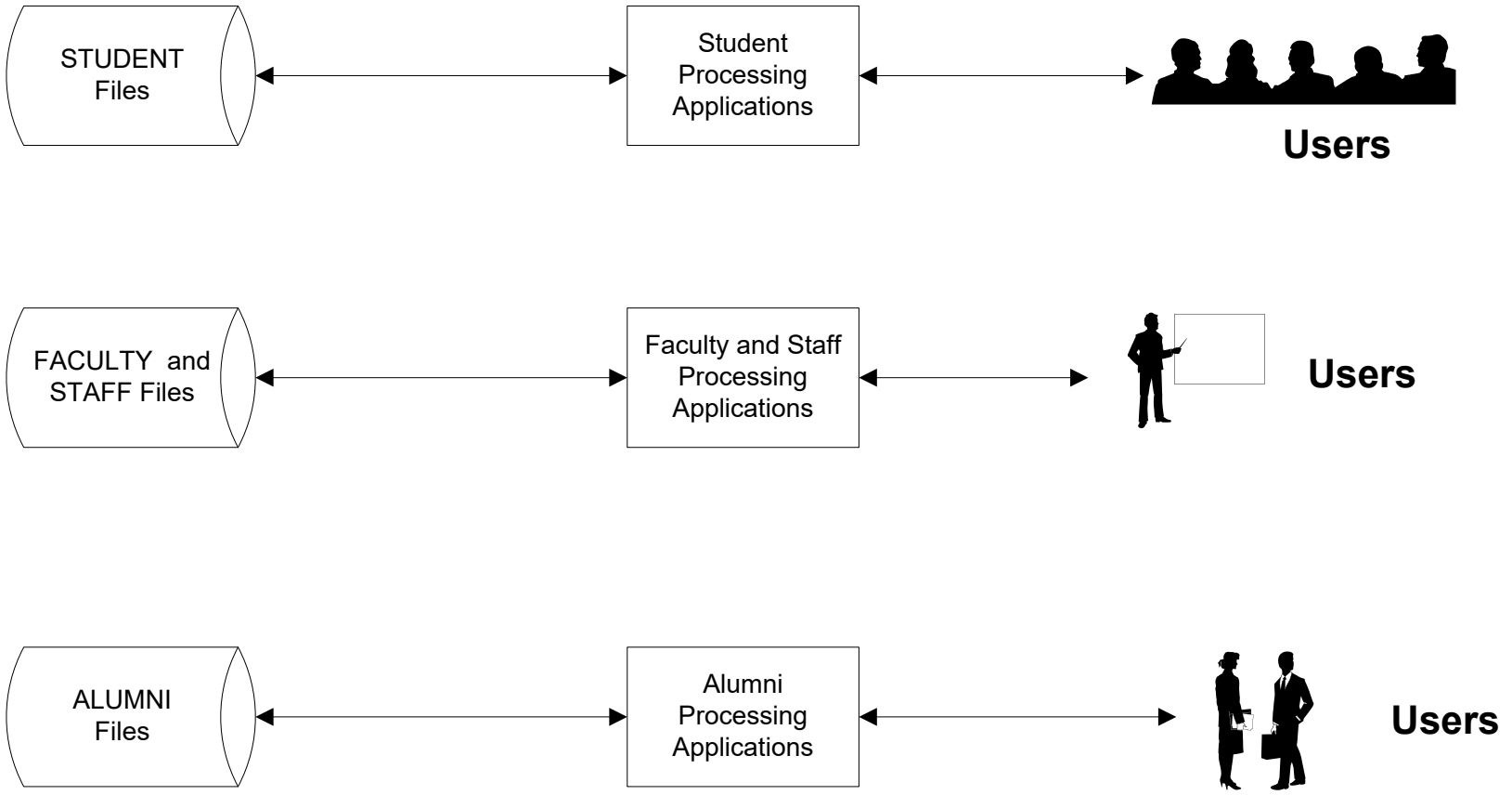


Figure 1.1 An example of a file processing environment

Integrated data

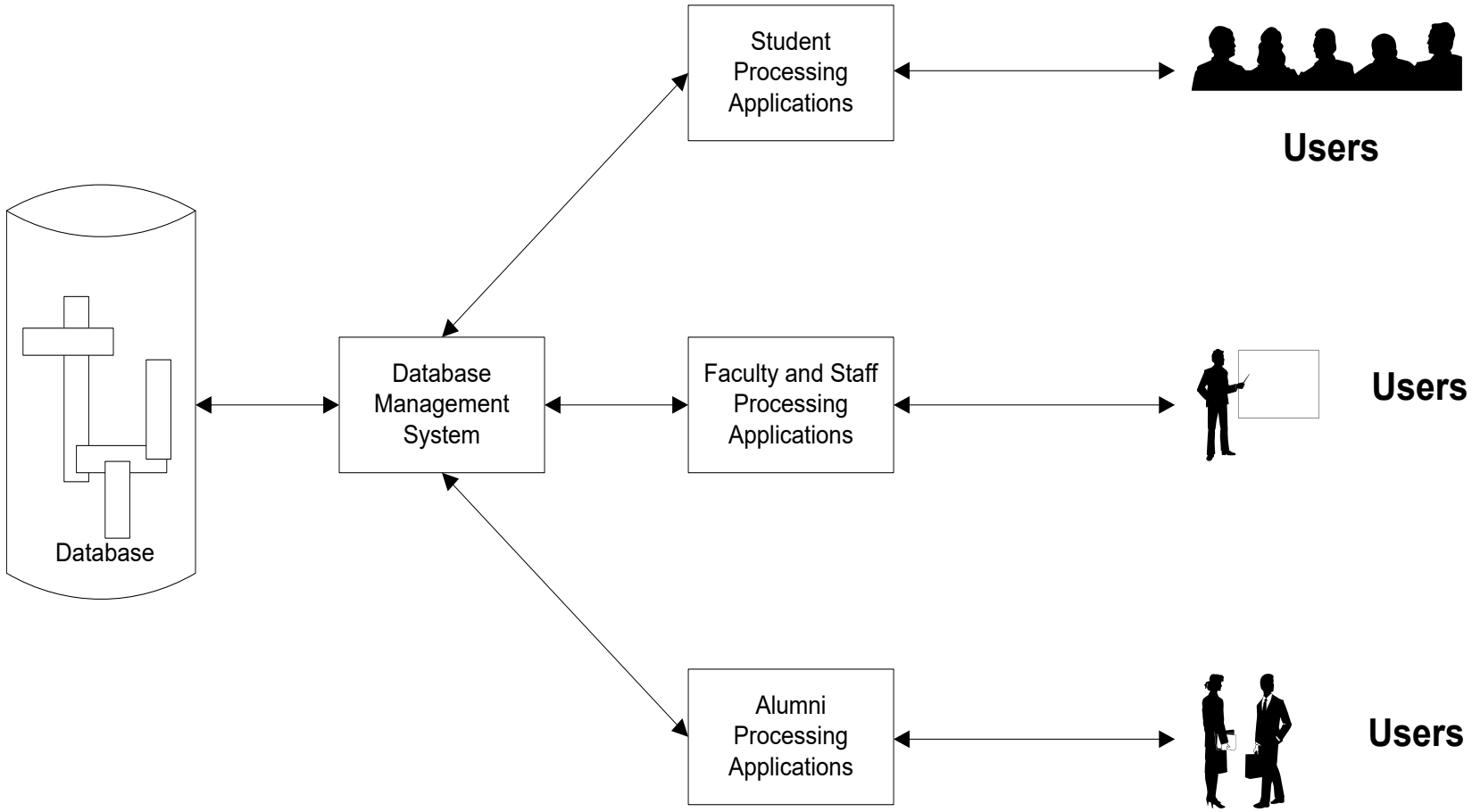
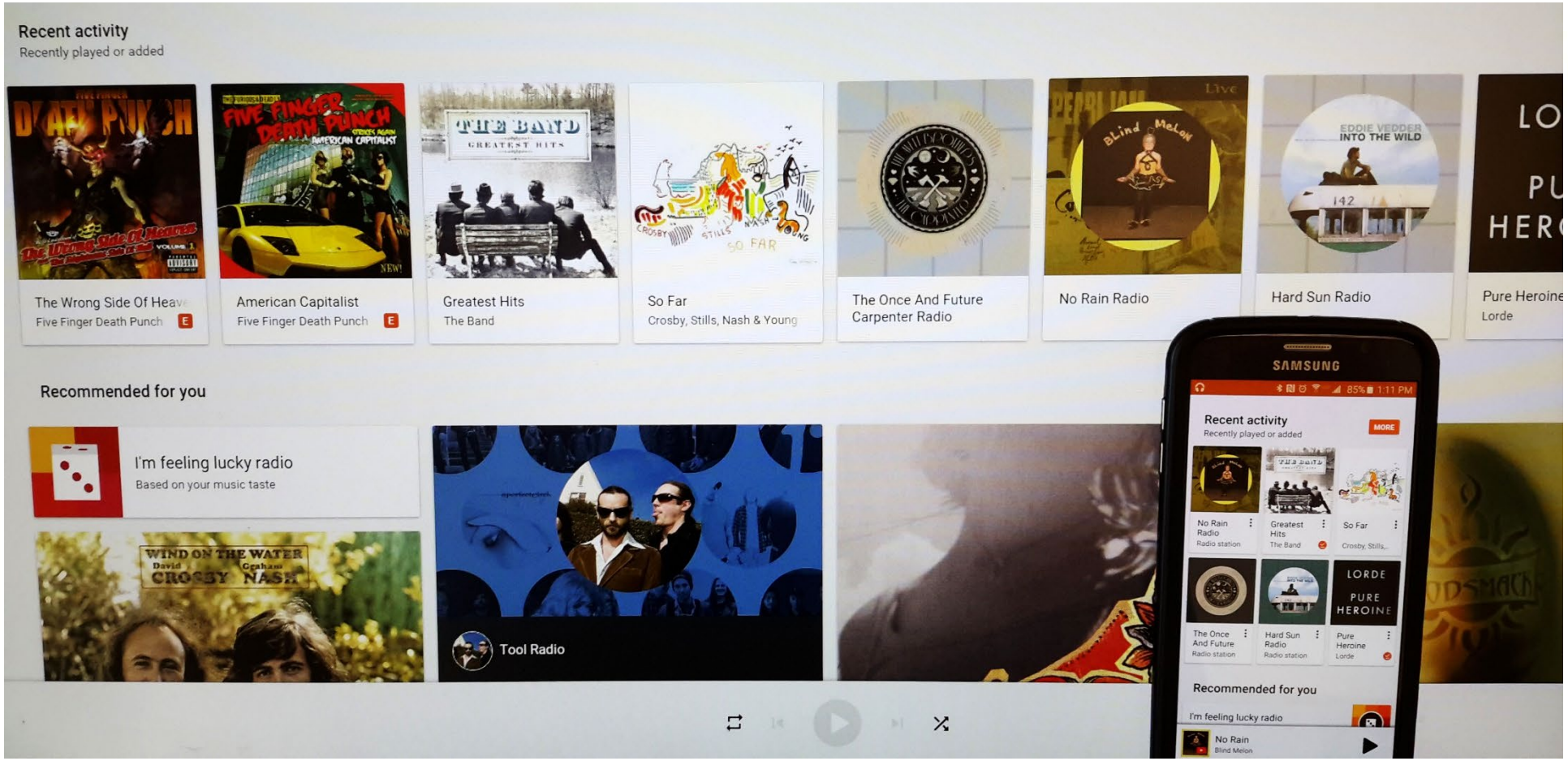


Figure 1.6 An example of a database system

The real world

- Applications used to be installed directly on your computer - now applications are often “in the cloud”



Module 1.3
Limitations of File Systems

- What are three limitations of file-processing systems?
- What are the two fundamental problems that lead to these limitations?

Module 1.4
Three Schema Architecture

- What is a schema?
- Explain the difference in internal, external, and conceptual schemas.

What do we really mean by these “limitations” & “Problems”?

- Limitations
  - Lack of data integrity
  - Lack of standards
  - Lack of Flexibility/maintainability

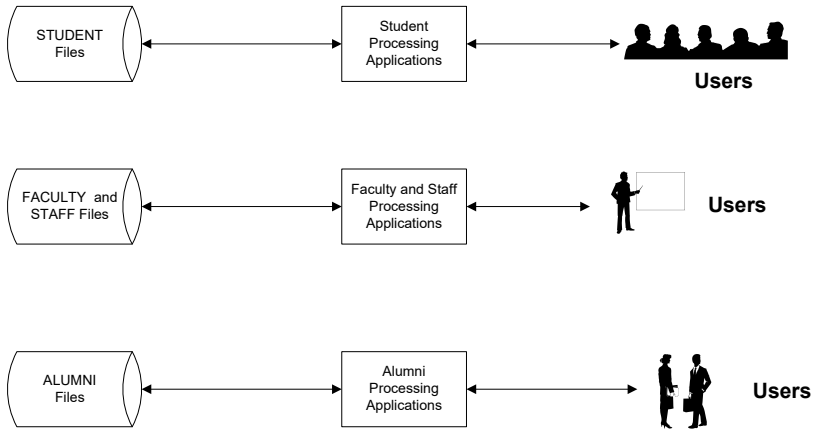


Figure 1.1 An example of a file processing environment

- Problems
  - Lack of integration
  - Lack of program/data independence

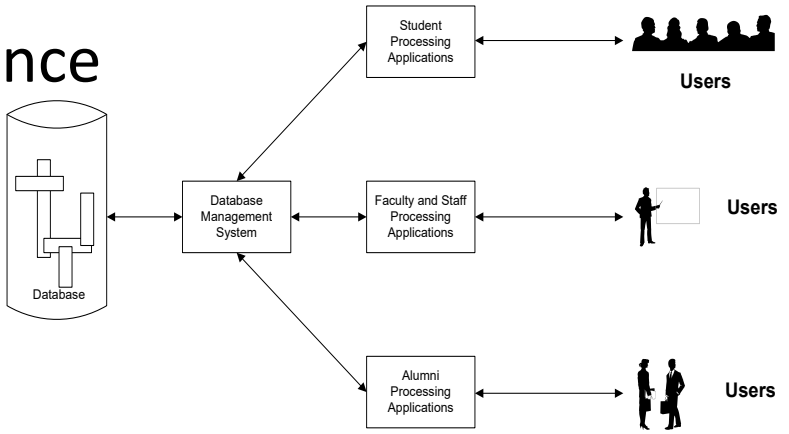


Figure 1.6 An example of a database system

What is desirable?

- Data that is integrated, not isolated

- Data that is independent of the application
  - Makes the app immune to changes in the data structure

  - Shows only what users need
    - Simpler and more secure

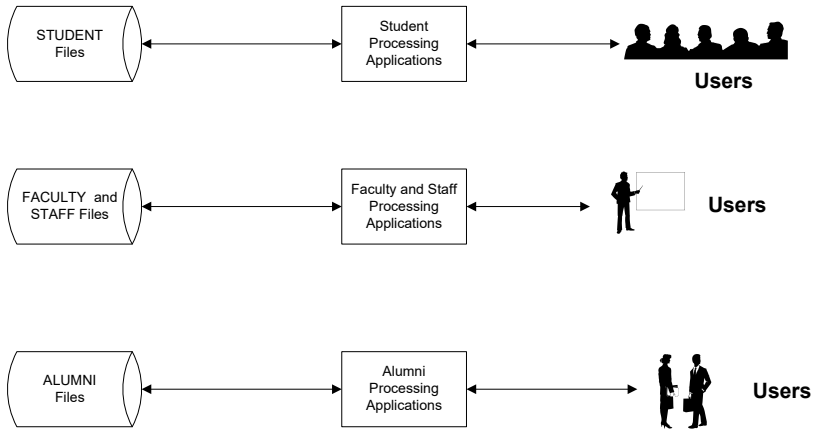


Figure 1.1 An example of a file processing environment

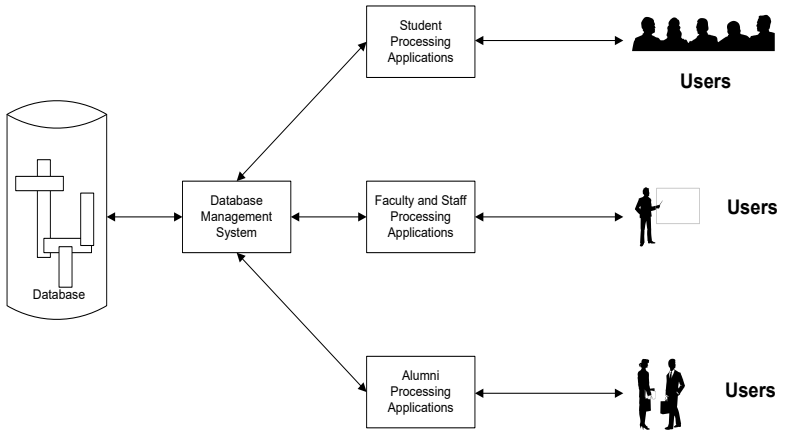


Figure 1.6 An example of a database system

Birth of data independence

- In the 1970s, the Standards Planning and Requirements Committee (SPARC) of the American National Standards Institute (ANSI) proposed what came to be known as:

- The ANSI/SPARC three-schema architecture:
  - External schema
  - Conceptual schema
  - Internal schema

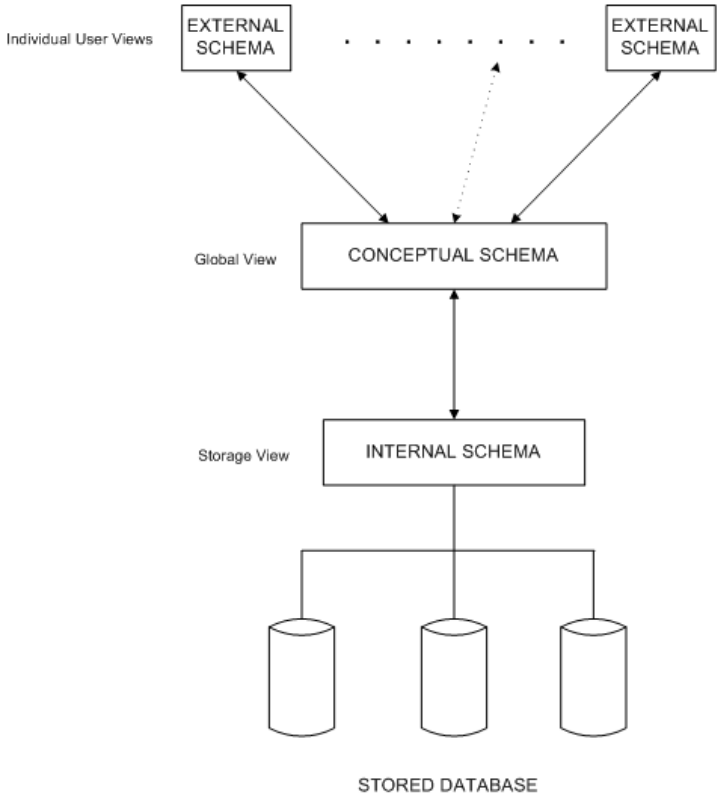


Figure 1.2 The ANSI/SPARC three-schema Architecture

What is a Schema?

- A description of your metadata
  - Data that describes the data that describes your data…

- The schema is a **map** that shows how things are related
  - What individual pieces of data make up a larger data element
  - How data is related to other data – for example, how students, courses, instructors and classrooms are related
  - Mapping where data is logically stored to where it is physically stored

Three perspectives of metadata

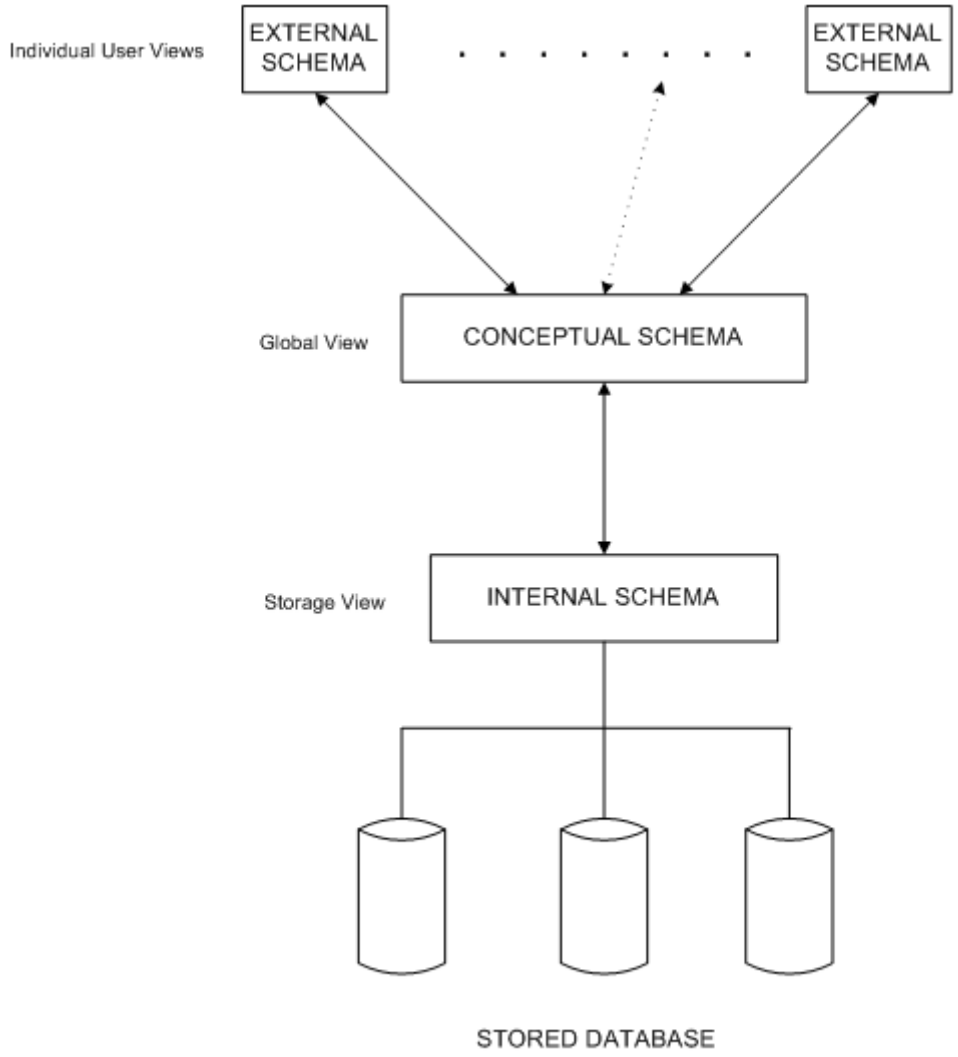
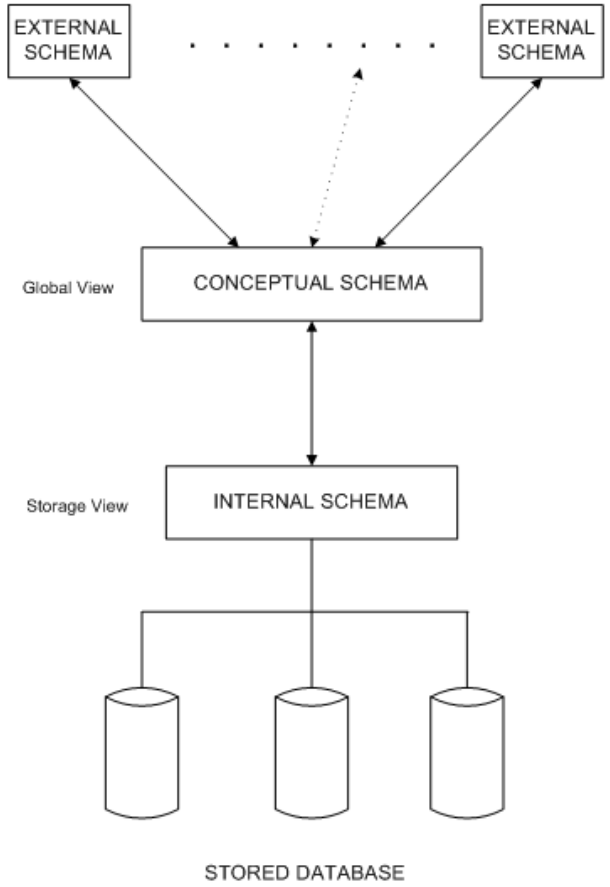


Figure 1.2 The ANSI/SPARC three-schema Architecture

Internal Schema

- The nuts and bolts of the database
- Describes the **physical organization** of the stored data (e.g., how the data is actually laid out on storage devices)
- Describes the mechanism used to implement access strategies (e.g., indexes, hashed addresses, etc.)
- Concerned with the efficiency of data storage and access mechanisms
- **Technology dependent**
  - Depends on the hardware/software you are using
  - Are you using Oracle, MSSQL, MySQL, etc?
  - What type of server/storage/network configuration do you have?



Internal Schema: Consider Amazon

- 119,928,851 products* as of April 2019
  - 44.2 Million Books
  - 10.1 Million Electronics
  - 6.6 Million Home Goods
  - 6.0 Million Digital Music
  - …about 53 million other things
- 310,000,000+ customers
  - 100,000,000 are Amazon Prime subscribers
- 3,000,000,000+ orders per year
  - That's over 8,000,000 orders per day!
  - 5% of all US retail sales (and 49% of all ecommerce sales)

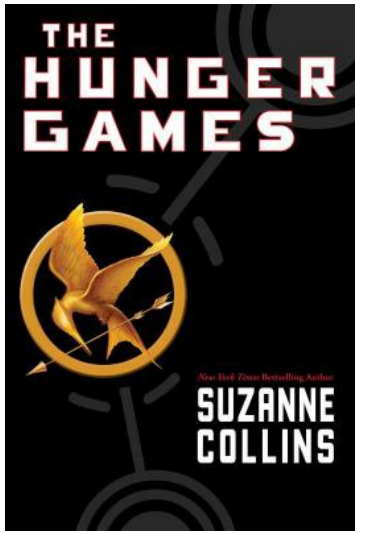
- Also keeps track of:
  - 3rd party sellers
  - Alexa devices and skills
  - Prime videos, music, etc.
  - LOTS of other stuff

- Too much for one poor server!

*Honestly these numbers vary wildly based on the source you are looking at. It's not public information as far as I know, but the point is the scale is insanely large. We could tell similar stories about Twitter, Facebook, Google, Tesla, Apple, and any number of other companies.

- Databases are spread across multiple servers, data centers, storage devices (HDDs, storage area networks, etc.)

Internal Schema: Consider Amazon



amazon

?

OK… Someone wants to buy The Hunger Games….
What server is that on? I don't know where to look…
This is going to be hard…

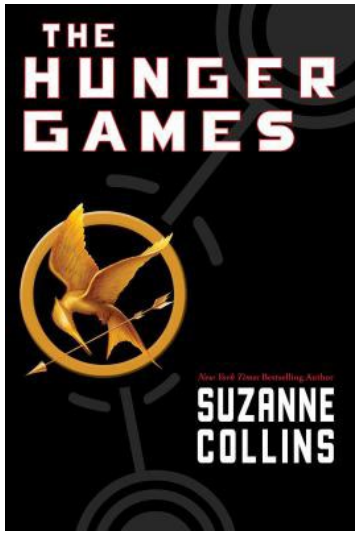
Keep in mind, these four "Books" boxes represent dozens, if not hundreds of database servers that could be located anywhere in the world!

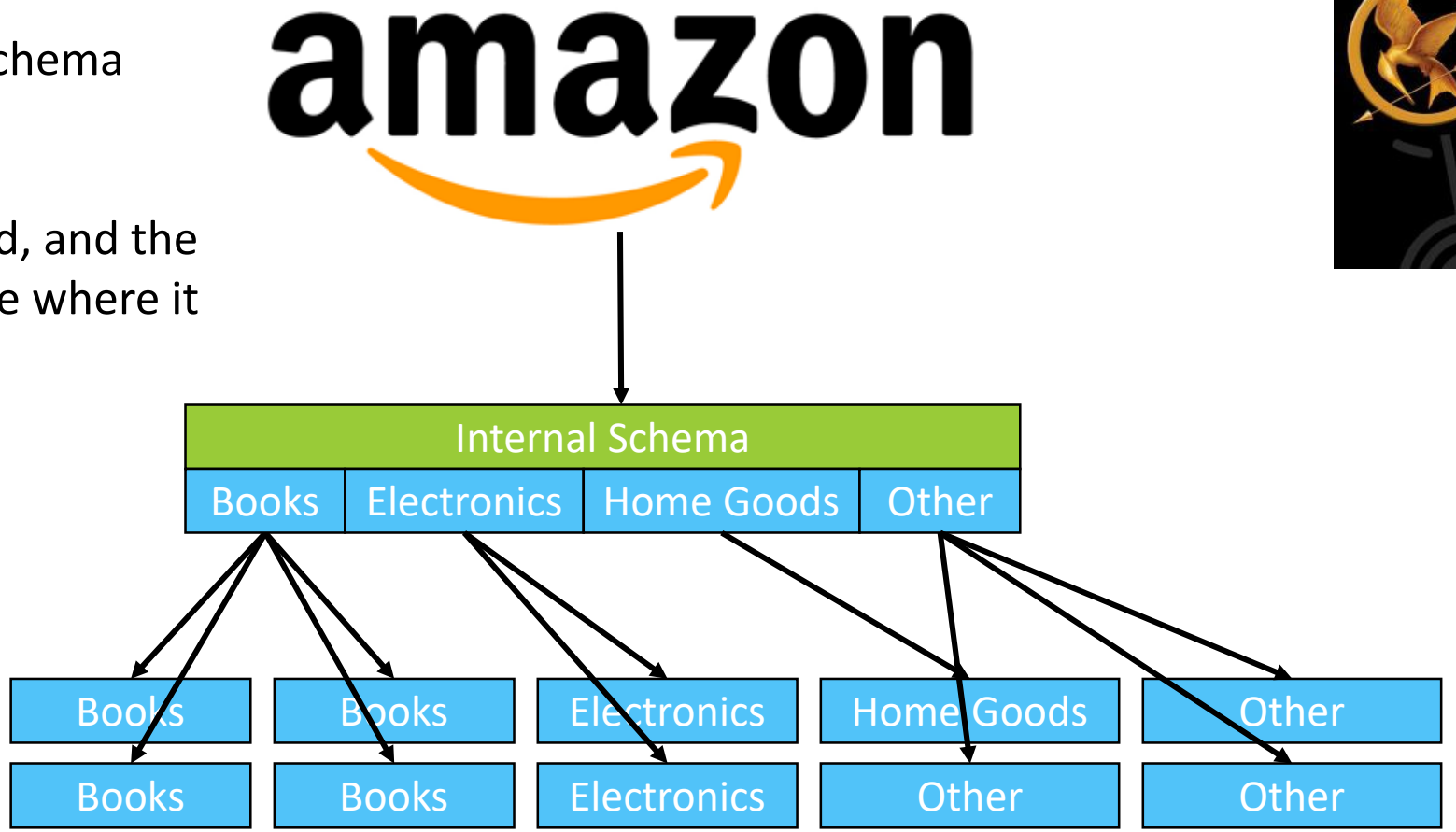
Books	Books	Electronics	Home Goods	Other
Books	Books	Electronics	Other	Other

Internal Schema: Consider Amazon

Oh good, the internal schema has organized all this!

I just ask for what I need, and the internal schema tells me where it Is - like map!



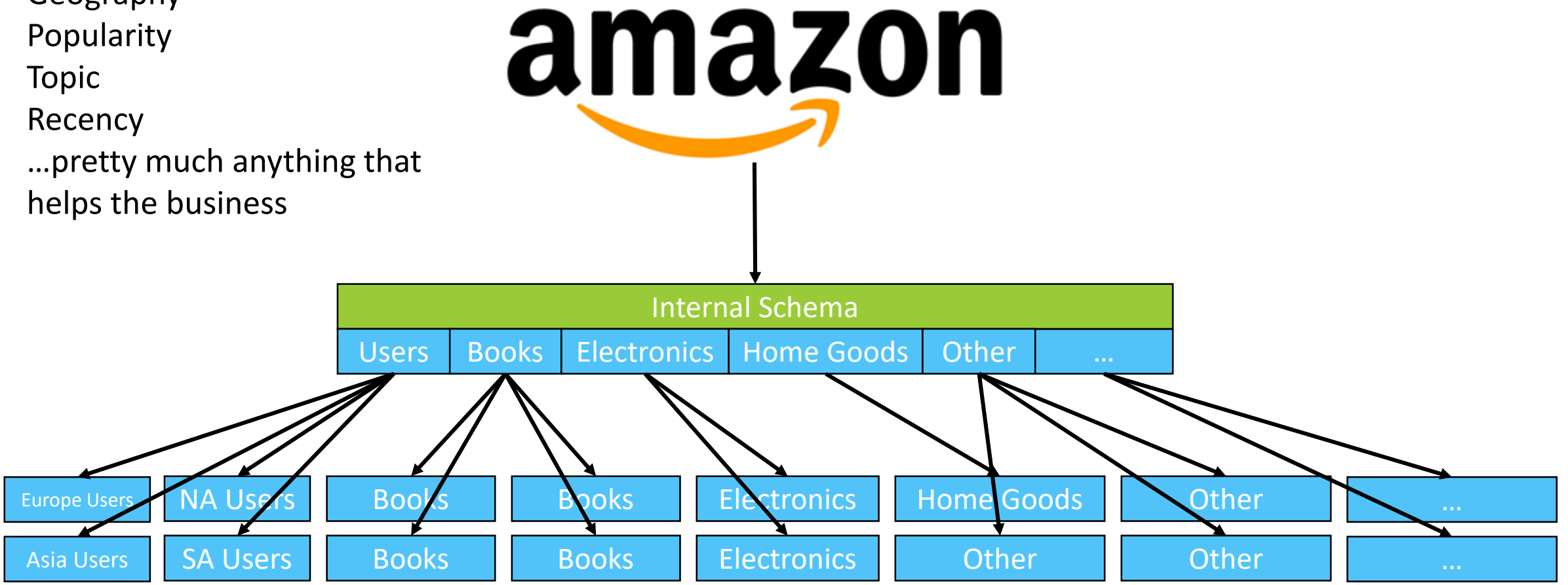


Internal Schema: Consider Amazon

Of course in reality this is MUCH bigger and more detailed than this…

Data may be segregated based on things like

- Geography
- Popularity
- Topic
- Recency
- …pretty much anything that helps the business



Conceptual Schema

- Core of the architecture - how is your data **logically organized**?
- Represents the **global view** of the structure of the **entire database** for a community of users
- Describes **all data items** and **relationships** between data together with integrity constraints
- Separates data from the program (or views from the physical storage structure)
- **Technology independent**
  - Just describes the data – will be the same regardless of the technology you use

Conceptual Schema: Consider Amazon

What is a user?

Just a collection of attributes

- Username
- First name
- Last name
- Address
- Credit Card Number, etc…

What is a product?

Just a collection of attributes

- ProductID
- Product Name
- Description
- Price, size, color, etc…

The conceptual schema is where these collections of attributes are defined and related to one another

Conceptual Schema: Consider Amazon

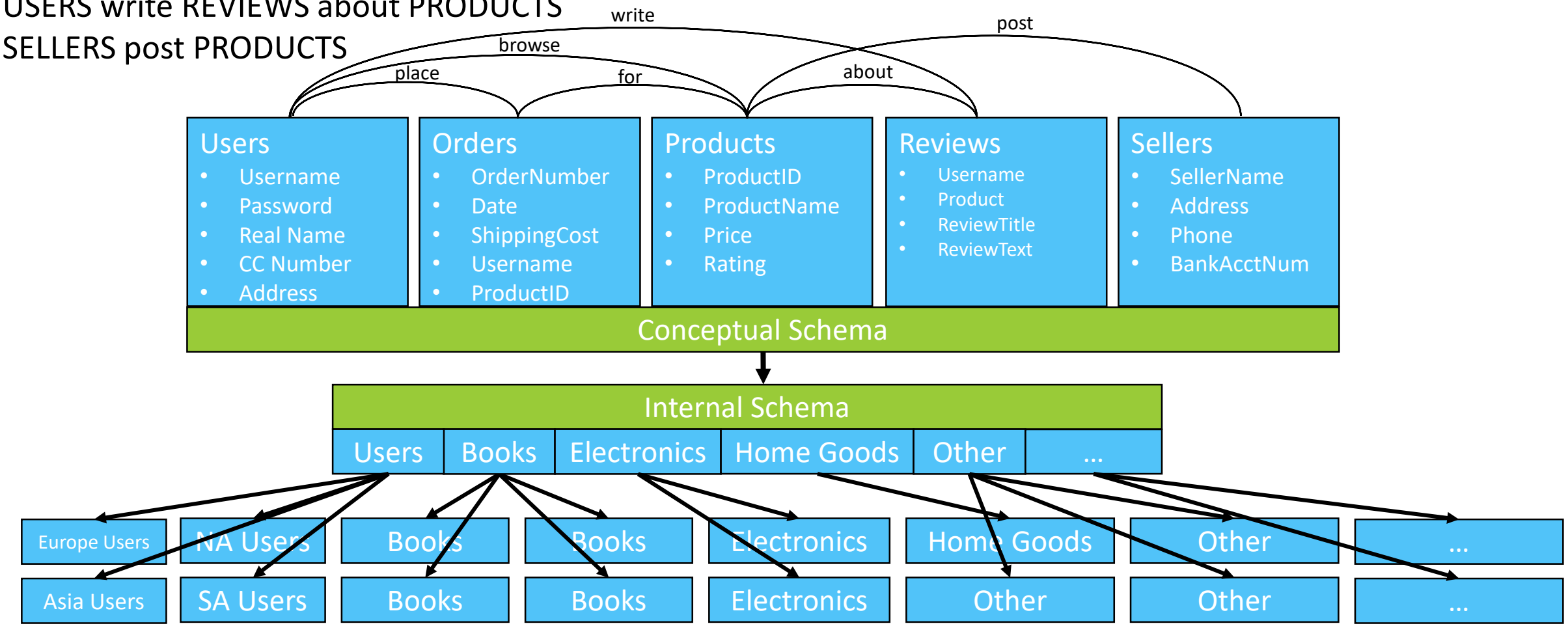
What happens on Amazon?

- Users login
- Users browse products
- Users place orders for products
- Users leave reviews about products
- Sellers post products for sale
- Etc.

- This could all be stored in one big nasty file (old school), but we group similar types of data and store them in **tables** in the database

Conceptual Schema: Consider Amazon

USERS login
USERS browse PRODUCTS
USERS place ORDERS for PRODUCTS
USERS write REVIEWS about PRODUCTS
SELLERS post PRODUCTS



External Schema

- Represents views for individual users and/or applications
  - Each view describes a portion of the database, which may include a single table, or many tables

- Each application (or user) has different data they care about
  - Other data should not be exposed for the sake of **simplicity & security**
  - A particular application's "view" of the data

- Views are generated by logical references to the **conceptual schema** elements

- **Technology independent**
  - Describes how users/applications will see the data – will be the same regardless of the technology you use

External Schema: Consider Amazon

External Schemas

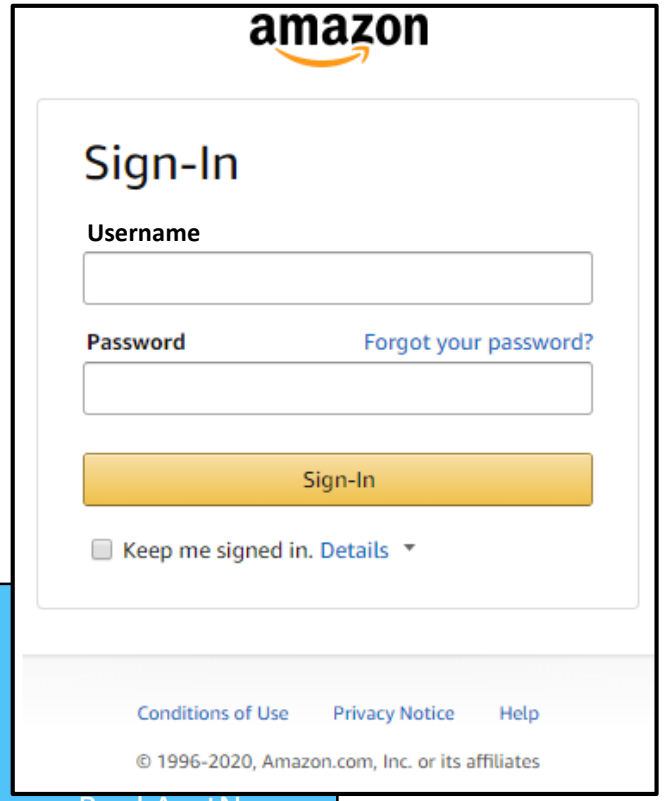
Sign-in Page

Users.username
Users.password

Is there any reason this Sign-in page would need access to your credit card number?

No!

The external schema makes your data more **secure** and more **simple** to work with



Users
- Username
- Password
- Real Name
- CC Number
- Address

Orders
- OrderNumber
- Date
- ShippingCost
- Username
- ProductID

Products
- ProductID
- ProductName
- Price
- Rating

Reviews
- Username
- Product
- ReviewTitle
- ReviewText

- BankAcctNum

Conceptual Schema

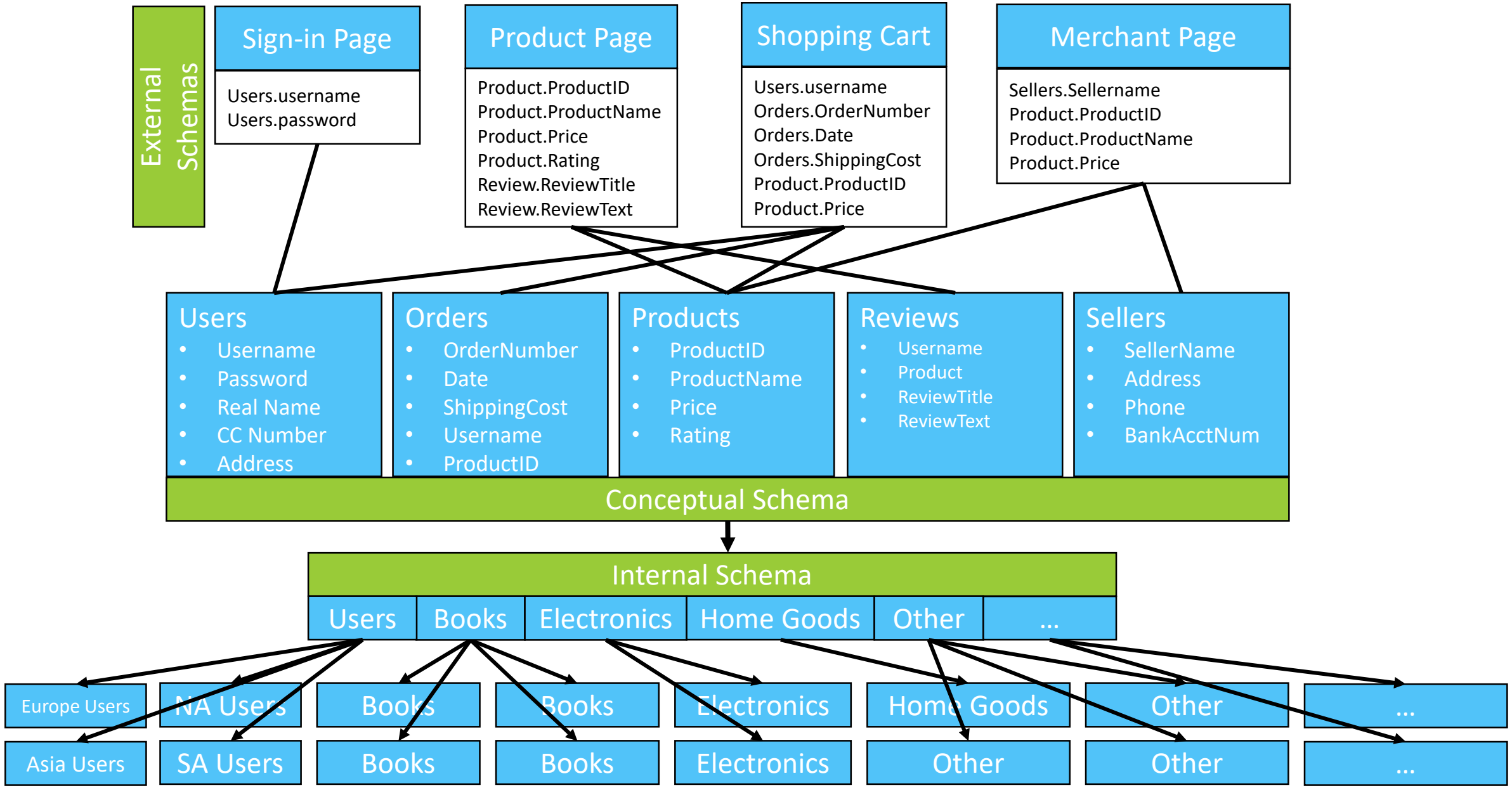
Internal Schema

Users | Books | Electronics | Home Goods | Other | …

Europe Users | NA Users | Books | Books | Electronics | Home Goods | Other | …

Asia Users | SA Users | Books | Books | Electronics | Other | Other | …

External Schema: Consider Amazon



External Schema

- Each application (or user) has different data they care about
  - Other data should not be exposed for the sake of **simplicity & security**
  - A particular application's "view" of the data

- For example:
  - The external schema for the part of an application that processes logins might contain only username and password (one table)

  - The external schema that displays reviews needs the text of the review, info about the product, and info about the user that left the review (data from three different tables)

- The external schema creates "views" of the data so the application can remain unaware of the underlying conceptual schema

Schemas

- Most of your time (at least in this course) will likely be spent thinking about conceptual and external schemas

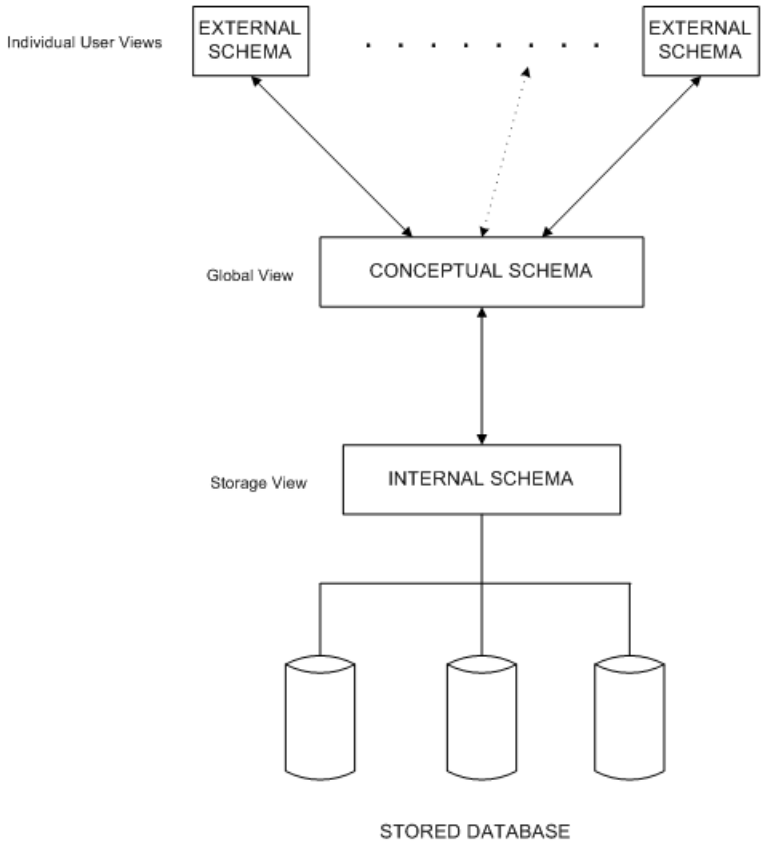


Figure 1.2 The ANSI/SPARC three-schema Architecture

Back to “data independence”

- When a schema at a lower level is changed, only the mapping information (managed by the DBMS) between this schema and higher-level schemas need to be changed

- Mapping: transforming requests & results between levels of schema

- The higher-level schemas themselves are unchanged.
  - Hence, the application programs need not be changed since they refer to the external schemas.

Data independence (continued)

- External views are unaffected by changes to the internal structure because of the conceptual schema between the external views and the internal schema

- External views are tailored for and accessible to application programs – the conceptual schema is not directly accessible by application program(s).

Physical Data Independence

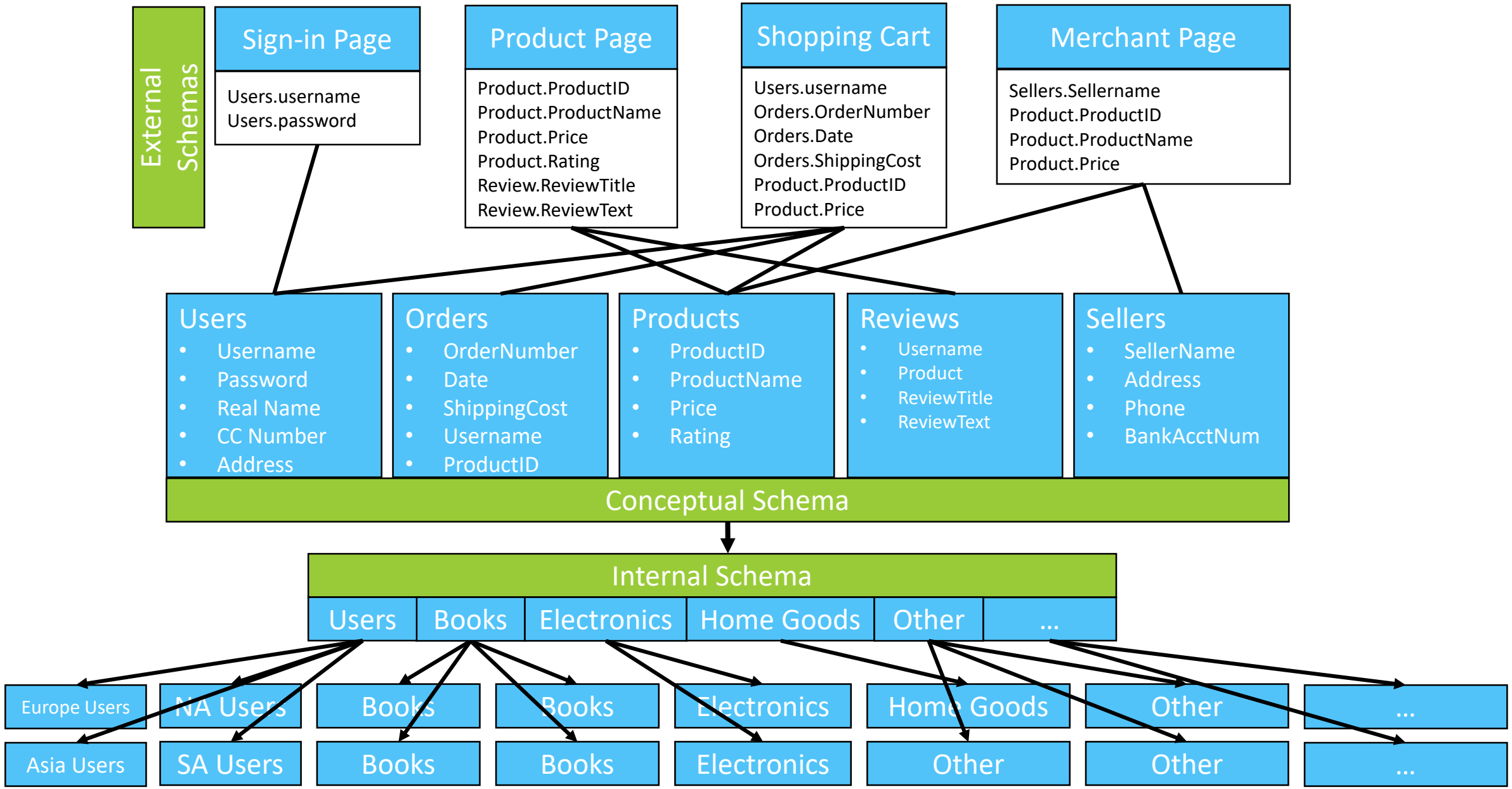
- The capacity to change the internal schema without having to change the conceptual or external schema.

- For example, the internal schema may be changed when certain file structures are reorganized or new indexes are created to improve database performance.

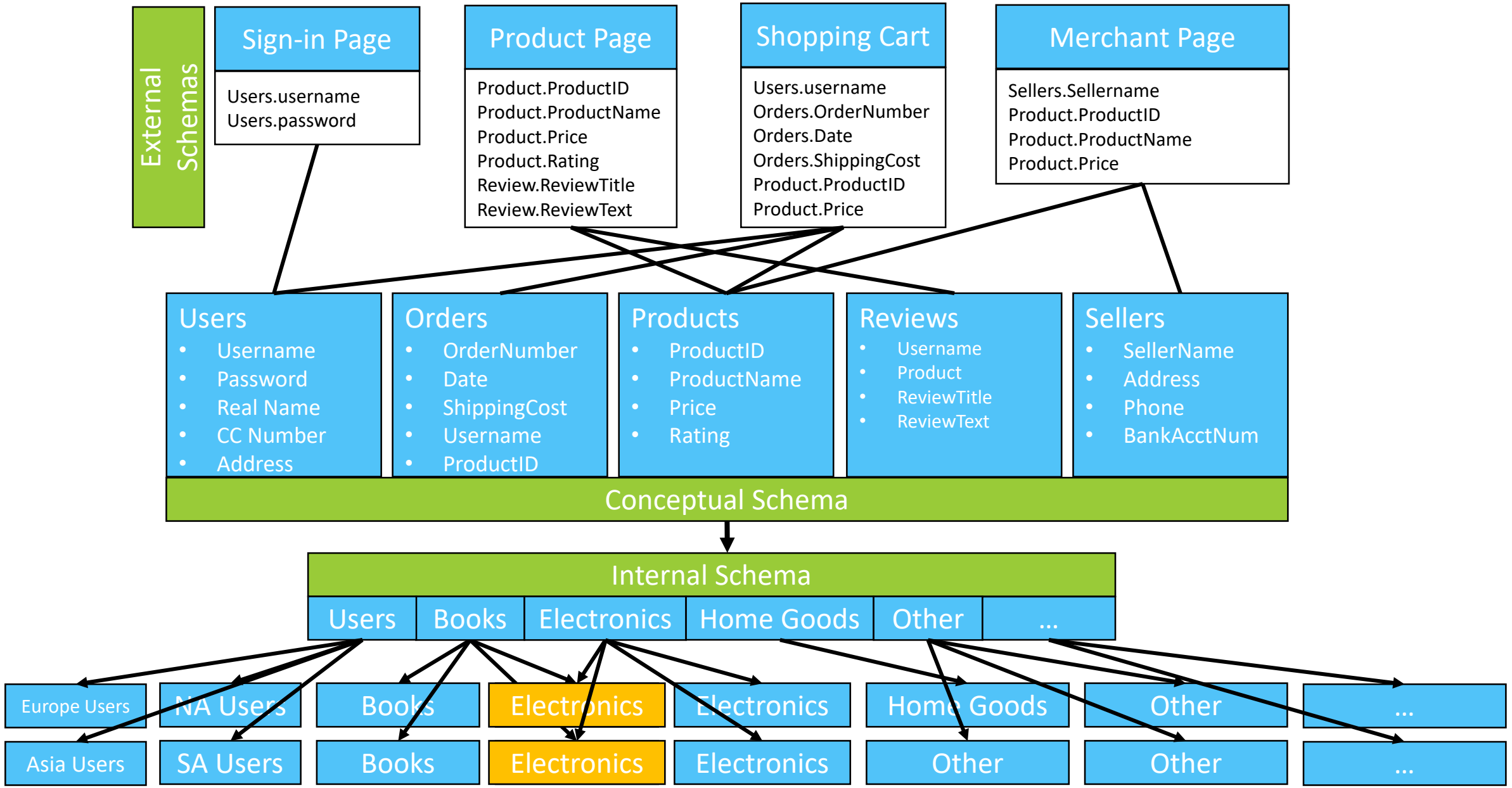
Logical Data Independence

- Definition:
  - External views unaffected by design changes (growth or restructuring) in conceptual schema
- How?
  - External views generated exclusively through logical reference to elements in the conceptual schema
- Consequence:
  - External views are unaffected by changes to other external views

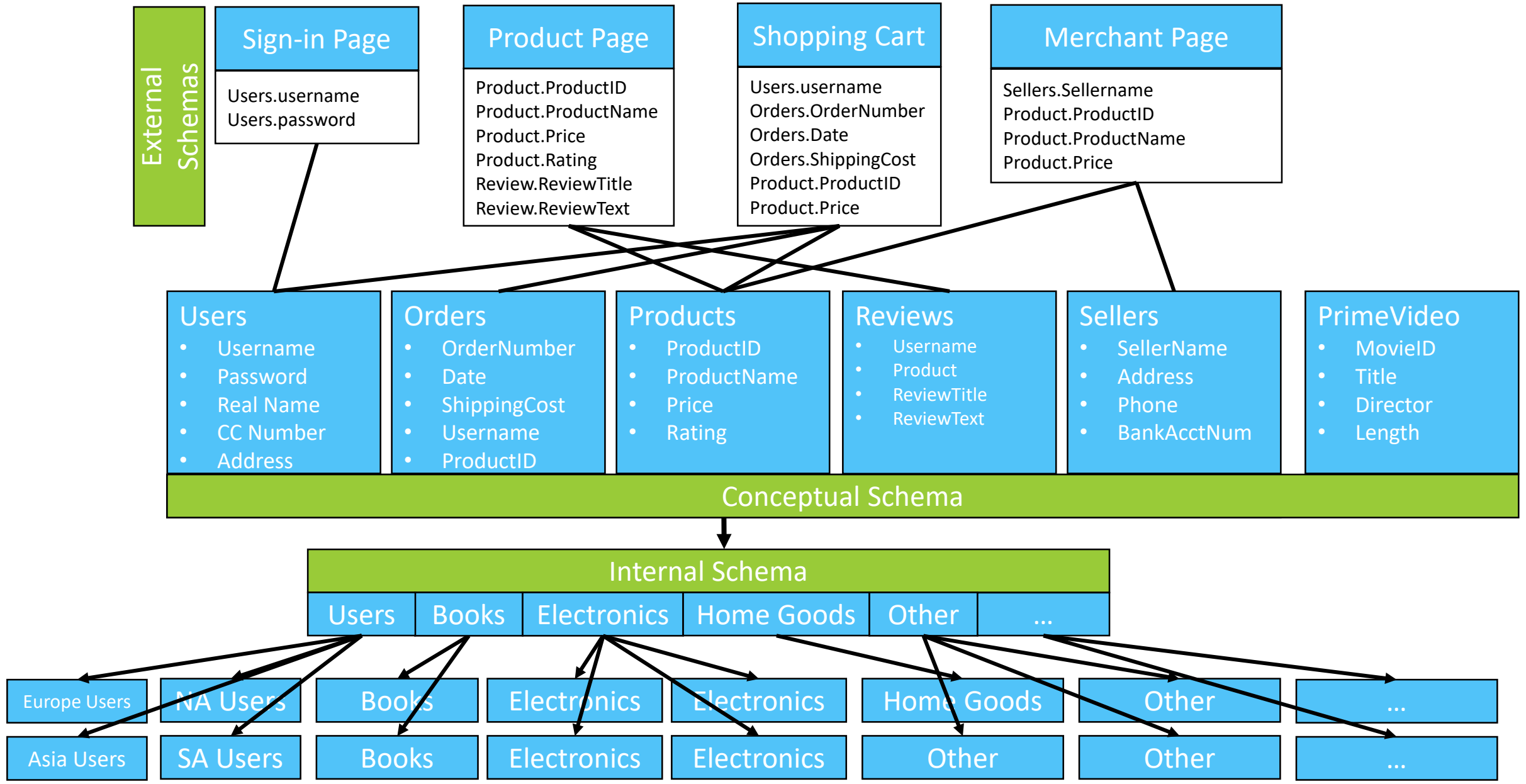
When lower level schemas change…



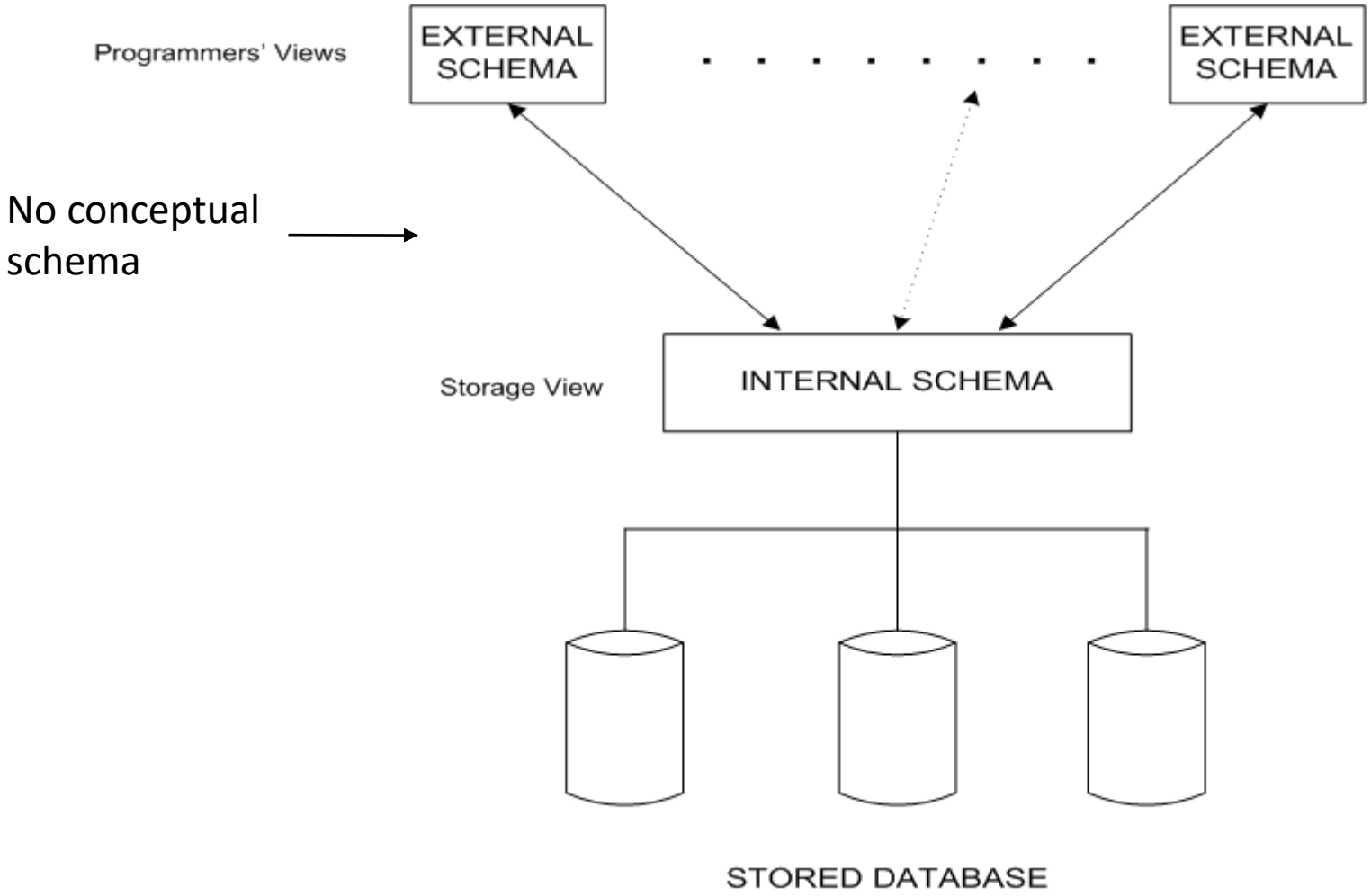
Imagine people start buying way more electronics than books…



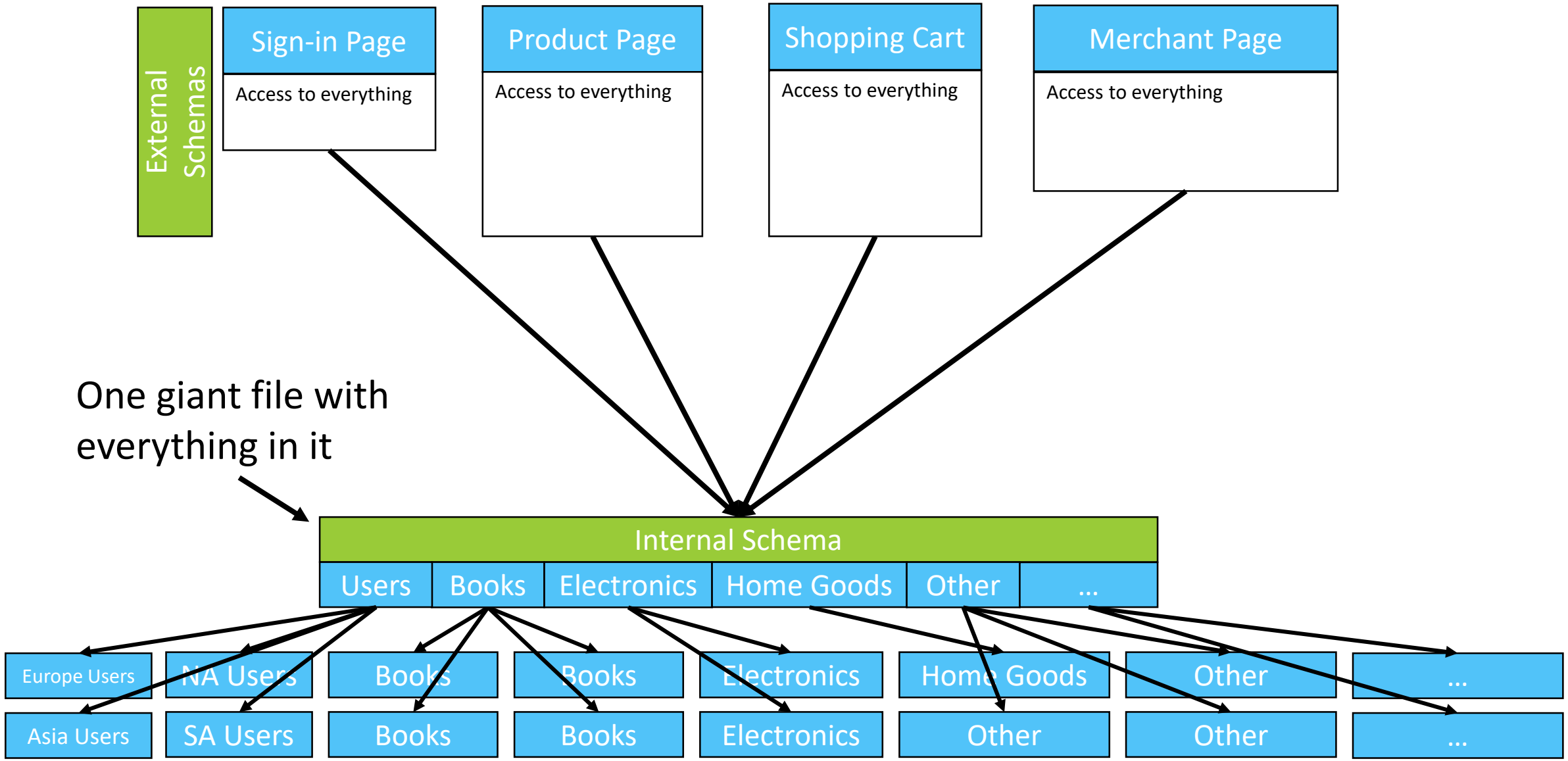
Let’s add a new feature to Amazon



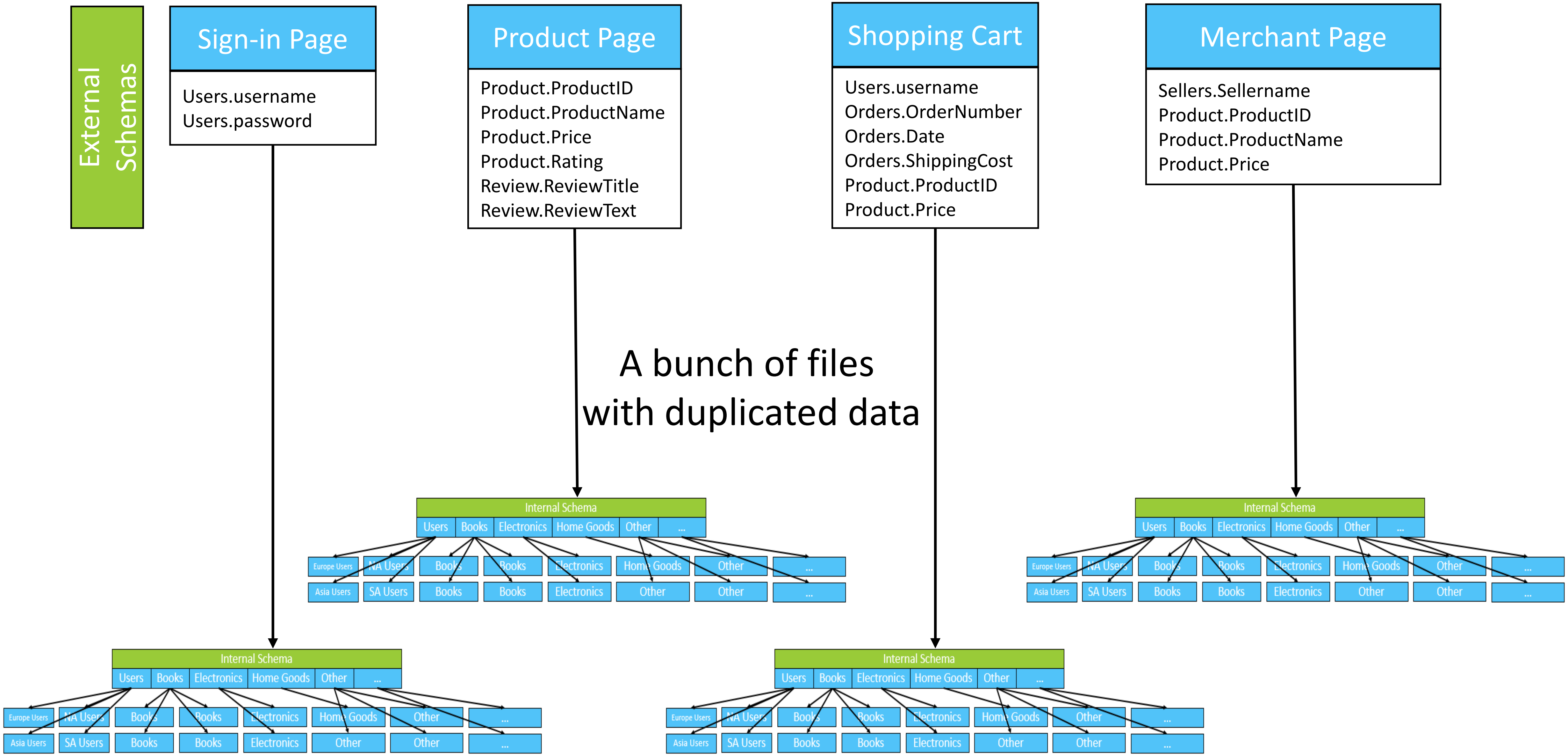
The old file system “Two schema” architecture



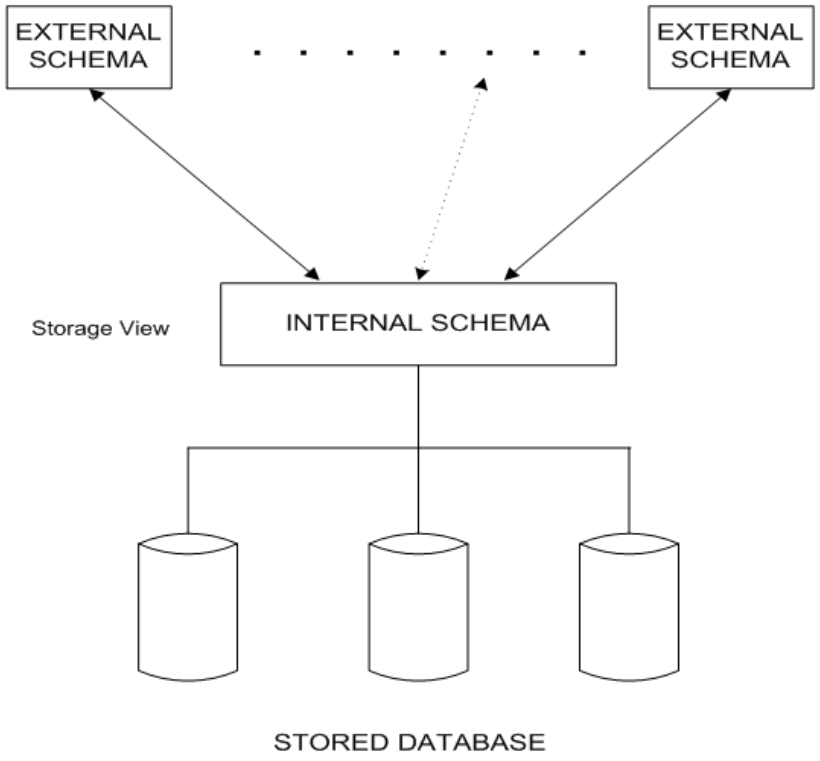
Two possibilities for a “two schema” Amazon



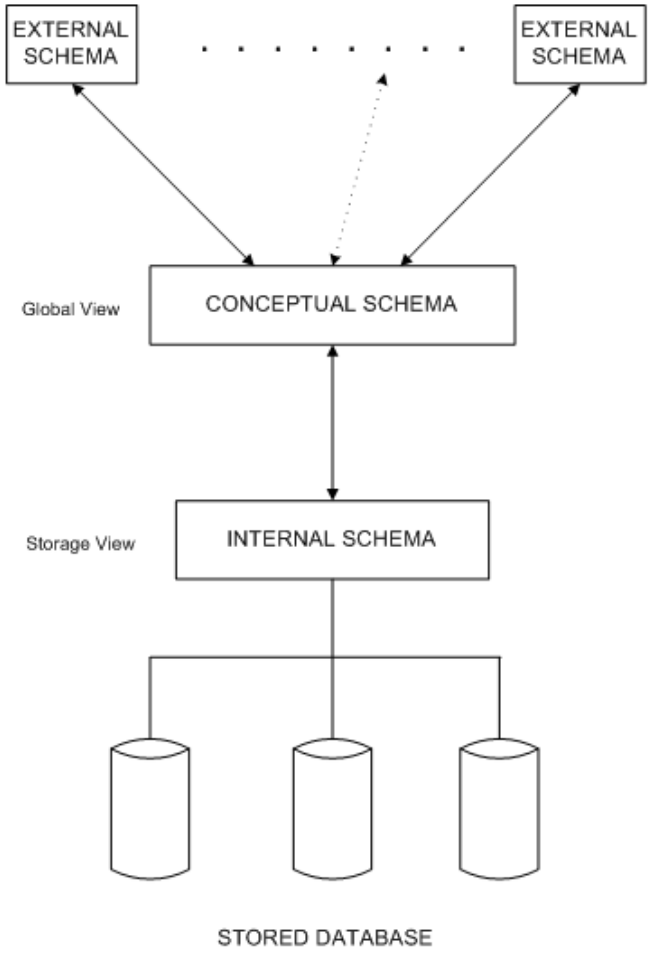
Two possibilities for a “two schema” Amazon



Two schema vs. Three schema

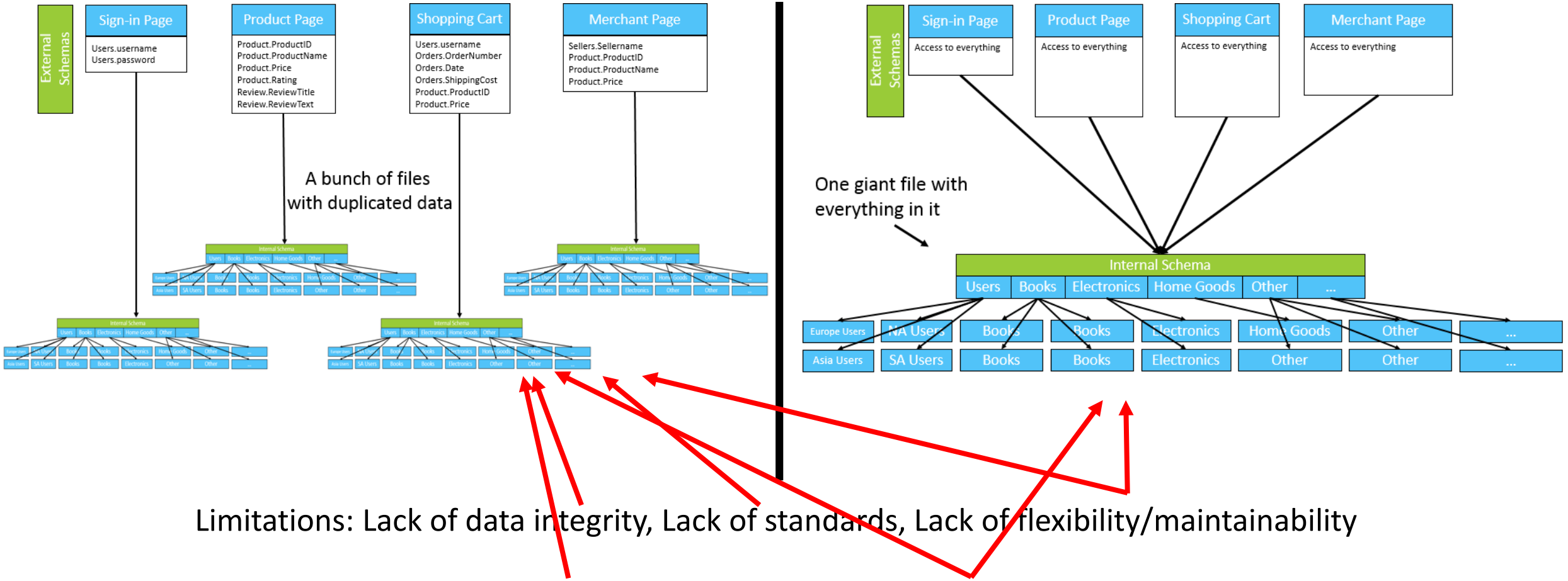


Old and Busted



Very Nice

Back to the limitations and problems

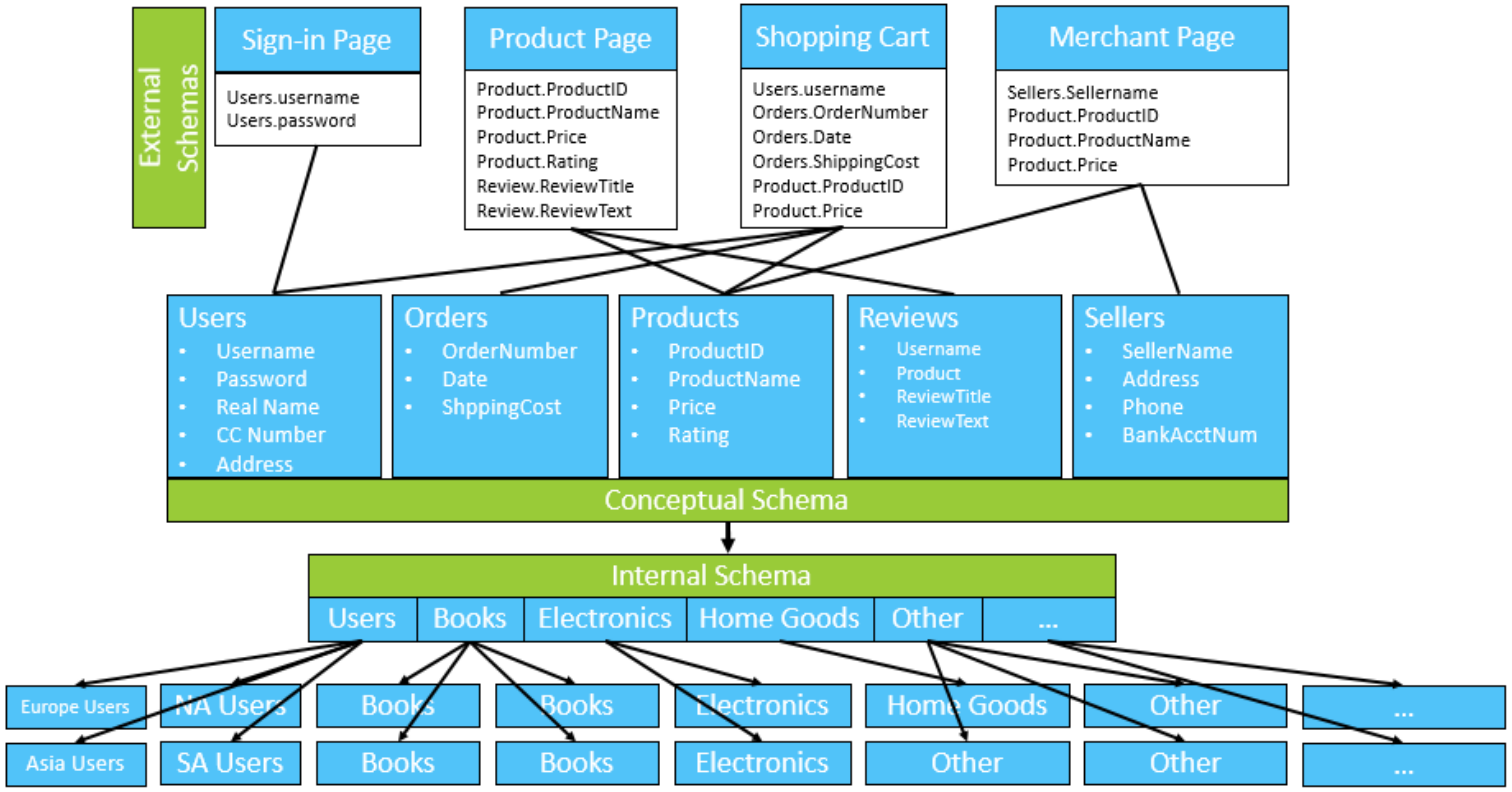


Limitations: Lack of data integrity, Lack of standards, Lack of flexibility/maintainability

Problems: Lack of integration, Lack of program-data independence

What is desirable?

- Data that is integrated, not isolated
- Data that is independent of the application
  - Immune to changes in the data structure
  - Shows just what applications/users need to see



Module 1.4
Three Schema Architecture

- What is a schema?
- Explain the difference in internal, external, and conceptual schemas.

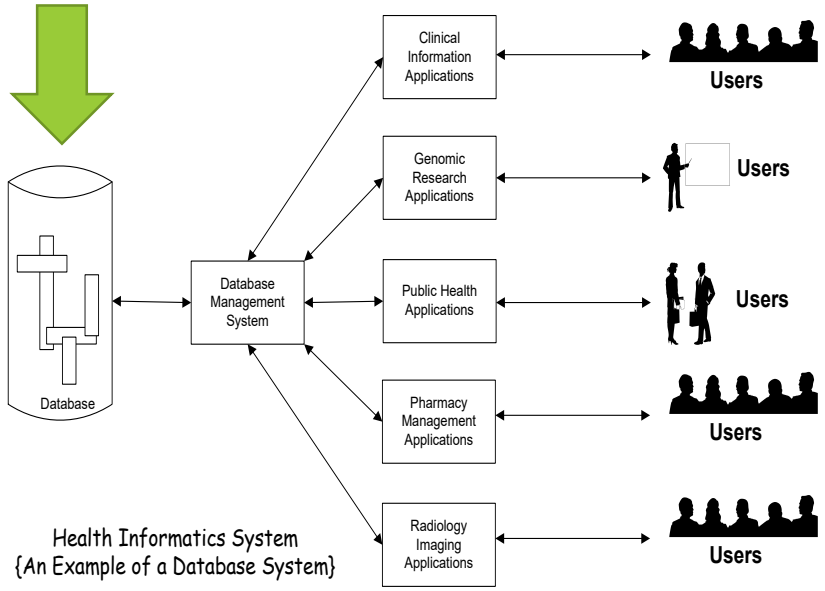
Module 1.5
Characteristics of Database Systems

- What is the difference in the “database” and the “DBMS”?
- Why is it important that a database is “self describing”?

What is a database system?

- Database systems were created to overcome the limitations of the old “file system” way of doing things

- Database: An integrated set of files
  - We still use files – but the DBMS is a system for managing the files and data contained within



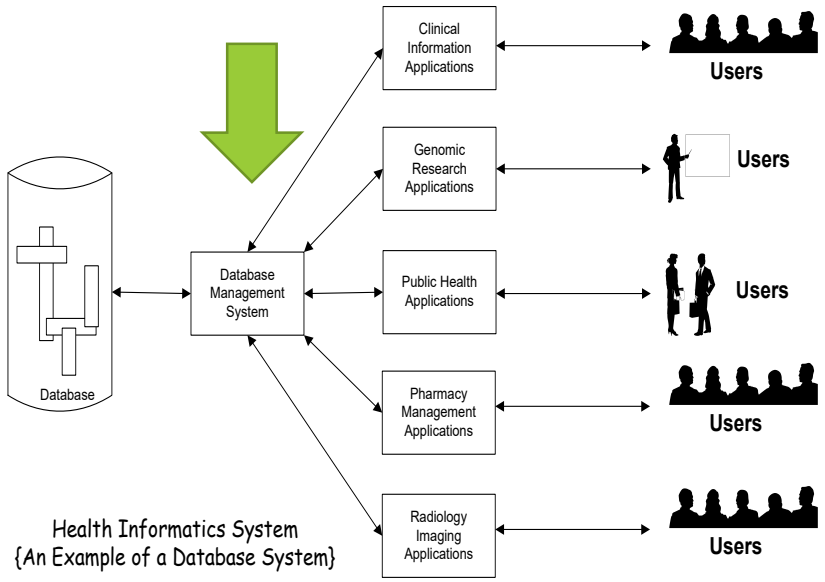
What is a database system?

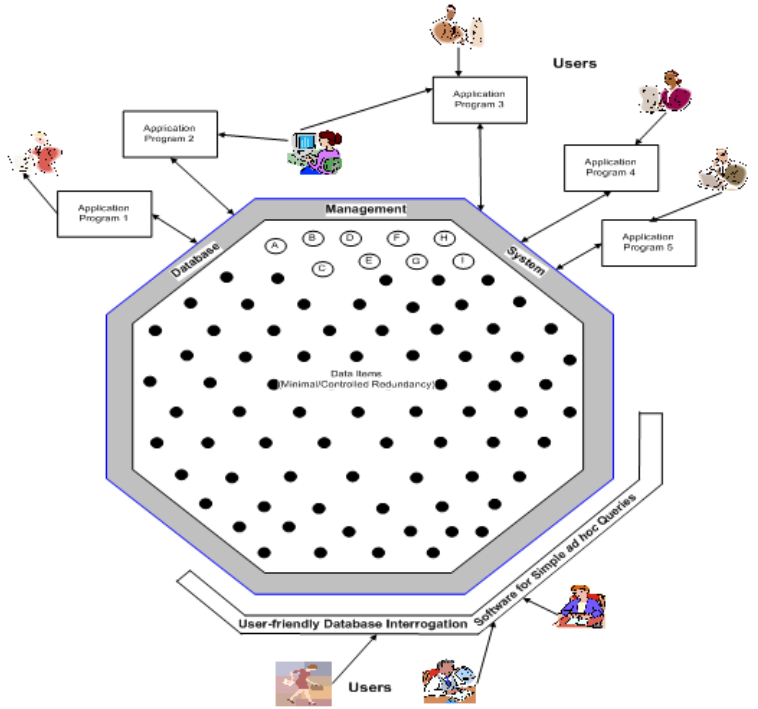
- A system is, generally, a set of interrelated components working together

- A **Database System** includes data and metadata
  - A database is **self-describing** in that the metadata is recorded within the database **(i.e., the schemas)**, not in application programs.
  - Data consists of **recorded facts** that have implicit meaning.
  - Viewed through the lens of **metadata**, the meaning of data becomes explicit.
  - A database is a collection of files whose records are logically related to one another. In contrast with that of a file-processing system
  - **Integration of data as needed is the responsibility of the DBMS software instead of the programmer.**

What is a database management system?

- DBMS: A collection of tools (software) that facilitate the process of defining, constructing, and manipulating data in a database.
- Rather than interacting with the data directly, the DMBS provides users and applications a method for “asking” for the data





Components of a DBMS

- A **data dictionary**
  - The metadata about your data
- One or more **query languages** (i.e., SQL)
- A **data manipulation language** (SQL, PL/SQL) for accessing the database
- A **data definition language** (SQL) to define the structure of data
- Tools for generating **reports**
- DBMS **utilities**
  - User security, importing data, data conversion, backup/restore, performance monitoring, reorganizing/indexing data,

Module 1.5
Characteristics of Database Systems

- What is the difference in the “database” and the “DBMS”?
- Why is it important that a database is “self describing”?

Module 1.6
Data Models

- What is a data model?
- What are the differences in conceptual, logical, and physical data models? Who is the intended audience for each

What is a model?

- All models are wrong, but some are useful
  - Box, George. E. P., and Draper, N. R., (1987), *Empirical Model Building and Response Surfaces*, John Wiley & Sons, New York, NY.

- If a model was perfectly correct, it would be the real thing!

What is a model?

- Simplified expression of observed or unobservable reality used to perceive relationships in the outside world.
  - A model is an approximation & entails assumptions

- Examples:
  - Model aircraft in wind tunnel testing
  - Mathematical models (econometric, optimization, etc.)
  - Computing models (e.g., analog models/directed graphs)
  - Data models, Process models, etc.

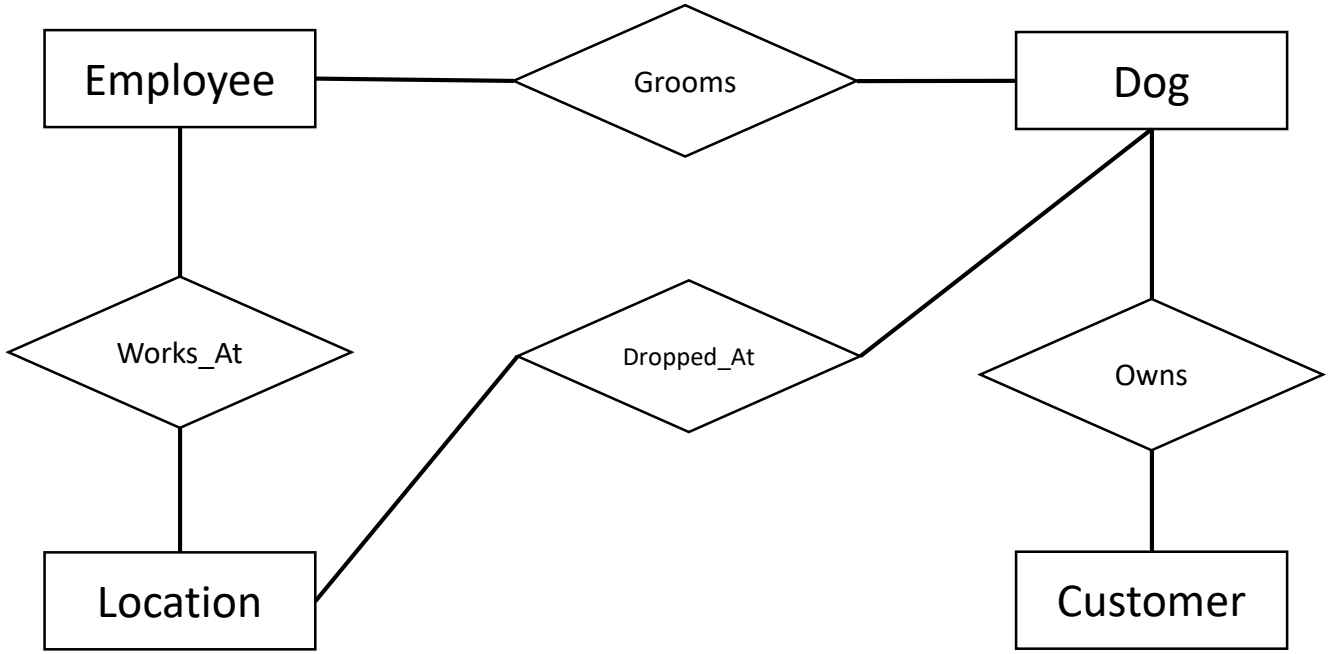
- A blue print for designing databases

Data modeling stages

- Conceptual modeling
  - Product: Conceptual schema
- Logical model/design
  - Product: Logical schema
- Physical design
  - Product: Physical/Internal schema

Conceptual modeling: Dog grooming service

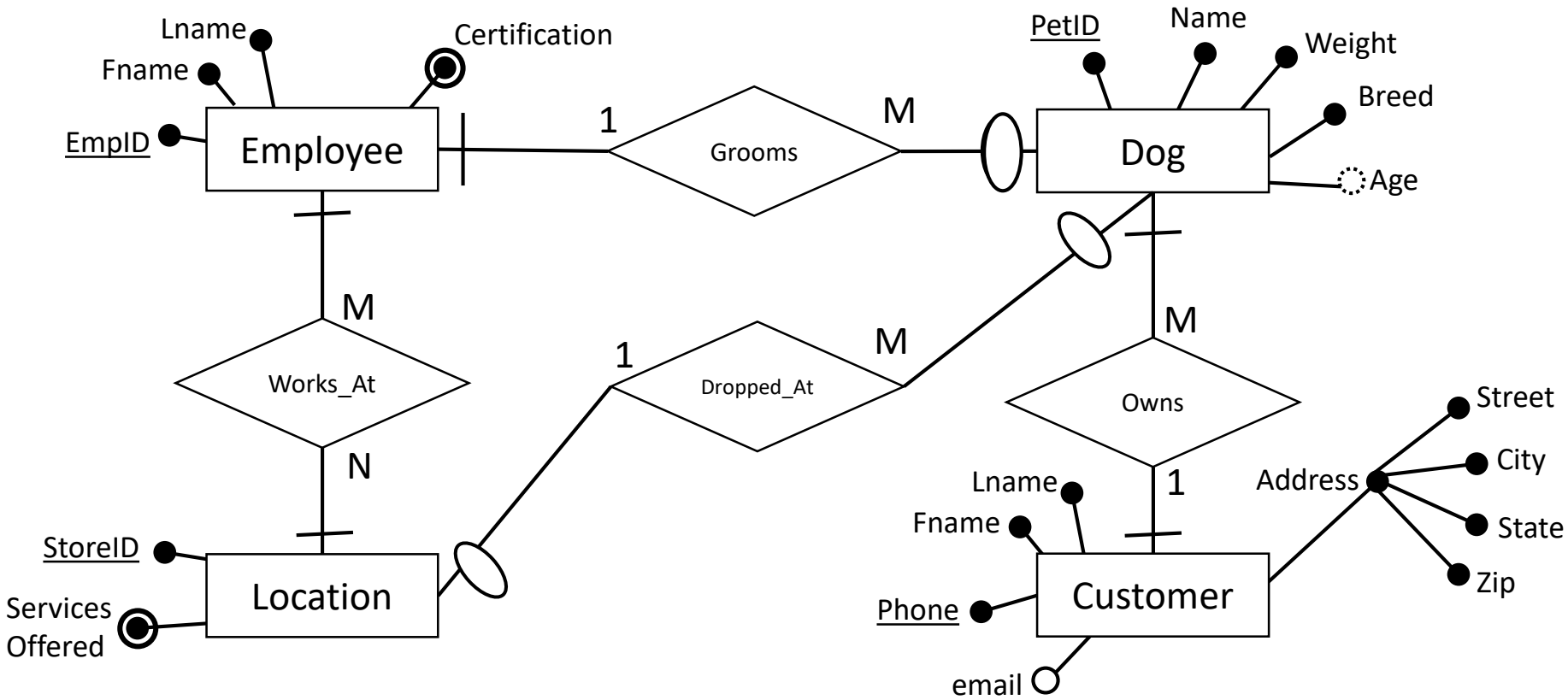
Customers of Dave's Dog Wash (DDW) own dogs. Employees of DDW groom dogs. There are multiple Locations of DDW that employees can work at. Dogs can be dropped off at any location.



Conceptual modeling: Dog grooming service

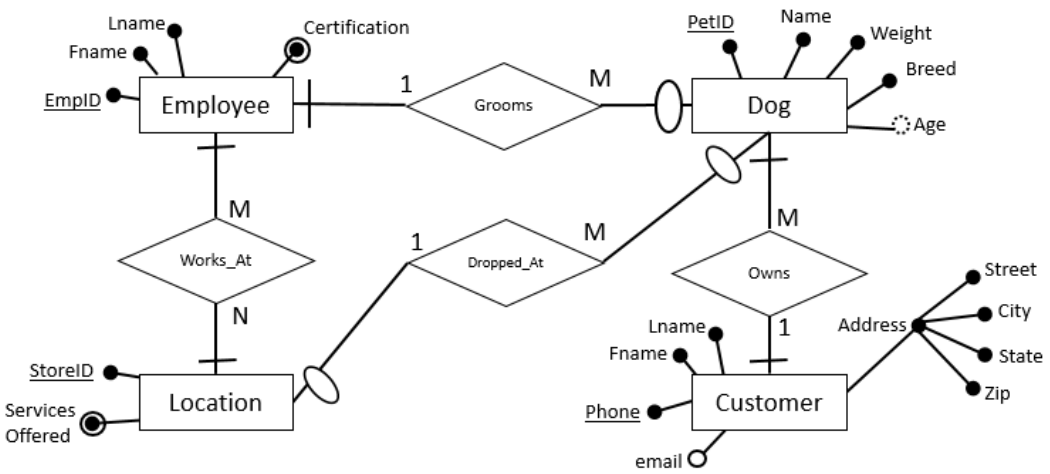
Customers of Dave's Dog Wash (DDW) own dogs. Employees of DDW groom dogs. There are multiple Locations of DDW that employees can work at. Dogs can be dropped off at any location.

Customers, Dogs, Employees, and Locations also have **attributes** that describe them!



Logical Schema

Conceptual schema



Logical schema

```
Employees(EmpID, Fname, lanme, certifications)

Dog(PetID, Name, Weight, Breed, DOB, FK_Cust, FK_Loc, FK_Emp)

Customer(Phone, Fname, Lname, email, Street, City, State, Zip)

Location(StoreID, ServicesOffered)

Emp_Loc(EmpID, StoreID)

Dog.FK_Cust ⊆ Customer.Phone
Dog.FK_Emp ⊆ Employee.EmpID
Dog.FK_Loc ⊆ Location.StoreID
Emp_Loc.EmpID ⊆ Employee.EmpID
Emp_Loc.StoreID ⊆ Location.StoreID
```

SQL Code to create the physical schema

```
CREATE TABLE employees (
  EmpID numeric(12,0) PRIMARY KEY,
  Fname varchar(50) NOT NULL,
  Lname varchar(50) NOT NULL,
  Certifications varchar(50)
);

CREATE TABLE customer (
  Phone varchar(14) PRIMARY KEY,
  Fname varchar(50) NOT NULL,
  Lname varchar(50) NOT NULL,
  email varchar(150),
  Street varchar(50) NOT NULL,
  City varchar(50) NOT NULL,
  State varchar(2) NOT NULL,
  Zip varchar(5) NOT NULL
);

CREATE TABLE location (
  StoreID numeric(12,0) PRIMARY KEY,
  ServicesOffered varchar(250) NOT NULL
);
```

```
CREATE TABLE Dog (
  PetID numeric(12,0) PRIMARY KEY,
  Name varchar(50) NOT NULL,
  Weight numeric(6,2) NOT NULL,
  Breed varchar(50) NOT NULL,
  DOB DATE NOT NULL,
  FK_Emp numeric(12,0),
  FK_Cust varchar(14),
  FK_Loc numeric(12,0),
  CONSTRAINT fk_groomedby FOREIGN KEY (FK_Emp) REFERENCES Employee (EmpID),
  CONSTRAINT fk_ownedby FOREIGN KEY (FK_Cust) REFERENCES Customer (Phone),
  CONSTRAINT fk_droppedat FOREIGN KEY (FK_Loc) REFERENCES Location (StoreID)
);

CREATE TABLE Emp_Loc (
  FK_EmpID numeric(12,0),
  FK_Loc numeric(12,0),
  CONSTRAINT pk_emploc PRIMARY KEY (FK_EmpID, FK_Loc),
  CONSTRAINT fk_emploc FOREIGN KEY (FK_EmpID) REFERENCES Employee (EmpID),
  CONSTRAINT fk_locemp FOREIGN KEY (FK_Loc) REFERENCES location (StoreID)
);
```

Module 1.6
Data Models

- What is a data model?
- What are the differences in conceptual, logical, and physical data models? Who is the intended audience for each

One last thing…

- Go here https://bit.ly/DB1-Spring24 and complete the introduction survey by Friday at 6:00 PM
  - Just go do it now! It will take less than five minutes.

Go forth and do great things!