Kaitlin Hoffmann

Office Hours:

SH 243

MR 11:00 AM - 1:45 PM and by appointment

Email: hoffmank4@newpaltz.edu

THE DATABASE ENVIRONMENT AND DEVELOPMENT PROCESS - PART 1

WEB DEVELOPMENT

OBJECTIVES

- Intro to Databases
- Traditional File Processing Systems
- The Database Approach

WHAT IS A DATABASE?

- A **database** is an organized collection of logically related data. It may be of any size and complexity. **For example:**
  - a salesperson may maintain a small database of customer contacts—consisting of a few **megabytes** of data—on her laptop computer.
  - A large corporation may build a large database consisting of several **terabytes** (trillion bytes) on a large mainframe computer that is used for decision support applications.

DATA

- **Data** is stored representations of *objects* and *events* that have meaning and importance in the user's environment.
  - **Example:** A social media site that uses a database to store **data** on a user – name, birthdate, email, profile picture, photos, posts and videos added, watch history, followers, etc.
  - However, the above example has two different types of data…

TYPES OF DATA – STRUCTURED VS UNSTRUCTURED

- **Structured Data** – typically categorized as *quantitative data* – is highly organized and easily decipherable by machine learning algorithms. The most important structured data types are **numeric**, **characters** and **dates**.
  - **Examples:** dates, names, addresses, and credit card numbers.
  - Relational (**SQL**) databases allow users to quickly input, search and manipulate structured data.

WHERE IS STRUCTURED DATA USED?

- **Customer relationship management (CRM):** CRM software runs structured data through analytical tools to create datasets that reveal customer behavior patterns and trends.
- **Online booking:** Hotel and ticket reservation data (e.g., dates, prices, destinations, etc.) fits the "rows and columns" format indicative of the pre-defined data model.
- **Accounting:** Accounting firms or departments use structured data to process and record financial transactions.

TYPES OF DATA – STRUCTURED VS UNSTRUCTURED

- **Unstructured data** – typically categorized as *qualitative data* – cannot be processed and analyzed via conventional data tools and methods.
  - **Examples:** documents, e-mails, tweets, Facebook posts, GPS information, maps, photographic images, sound, and video.
  - Best managed in non-relational (**NoSQL**) databases.

WHERE IS UNSTRUCTURED DATA USED?

- **Data mining:** Enables businesses to use unstructured data to identify consumer behavior, product sentiment, and purchasing patterns to better accommodate their customer base.
- **Predictive data analytics:** Alert businesses of important activity ahead of time so they can properly plan and accordingly adjust to significant market shifts.
- **Chatbots:** Perform text analysis to route customer questions to the appropriate answer sources.

STRUCTURED VS UNSTRUCTURED

- Structured and unstructured data are often **combined** in the same database to create a true multimedia environment.
  - For **example**, an automobile repair shop can combine structured data (describing customers and automobiles) with multimedia data (photo images of the damaged autos and scanned images of insurance claim forms).
  - Thus, it's okay to have both types in our relational database!

STRUCTURED VS UNSTRUCTURED

- From our example before, what is the type data for each:
- A social media site that uses a database to store data on a user –
  - **name**
  - **birthdate**
  - **email address**
  - **watch history**
  - **followers**
  - **profile picture**
  - **photos, posts and videos added**

Structured Data

Unstructured Data

DATA VS INFORMATION

- Data and information are often used interchangeably, however, they are different.
- **Information** is data that has been processed in such a way that the ***knowledge*** of the person who uses the data is increased.
- For **example**, consider the following list of facts. What can you decipher from this:

Baker, Kenneth D.	324917628
Doyle, Joan E.	476193248
Finkle, Clive R.	548429344
Lewis, John C.	551742186
McFerran, Debra R.	409723145

DATA VS INFORMATION

- This is essentially data that is useless in its present form. We need to convert this data into information!

- One way is with adding **context**:

Baker, Kenneth D.	324917628
Doyle, Joan E.	476193248
Finkle, Clive R.	548429344
Lewis, John C.	551742186
McFerran, Debra R.	409723145

Class Roster

Course: MGT 500 Business Policy

Semester: Spring 2015

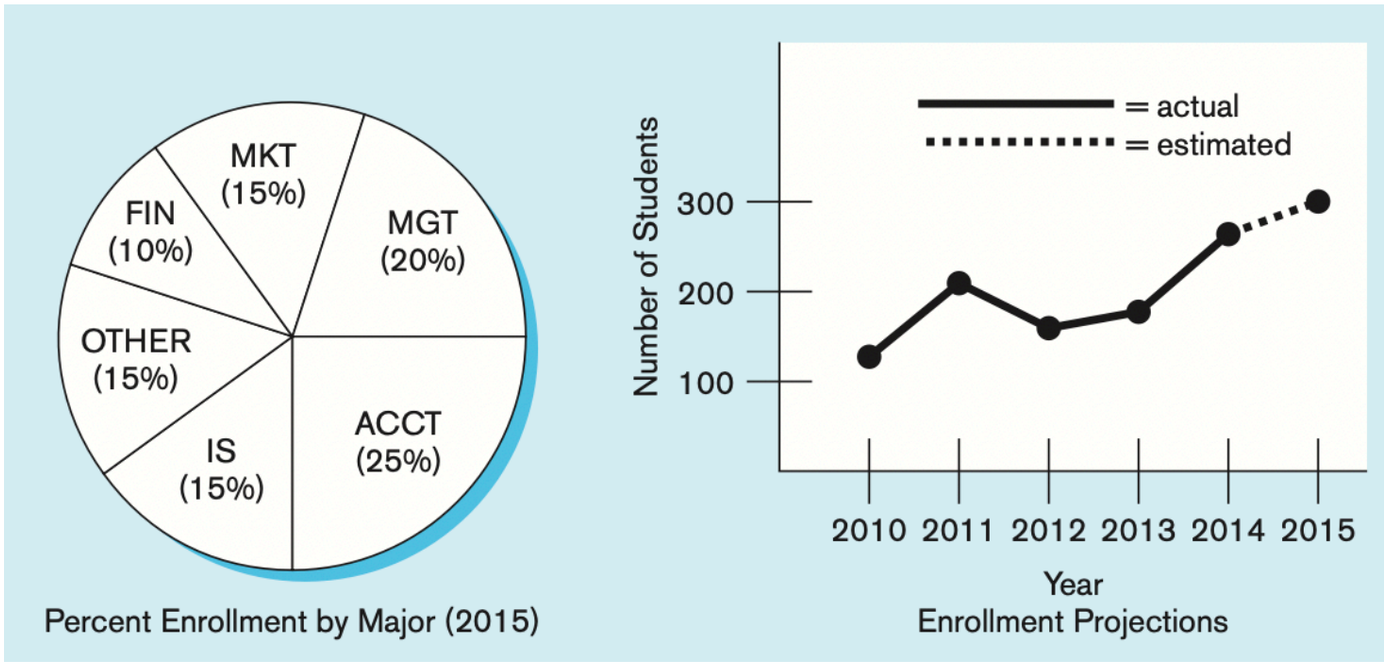
Section: 2

Name	ID	Major	GPA
Baker, Kenneth D.	324917628	MGT	2.9
Doyle, Joan E.	476193248	MKT	3.4
Finkle, Clive R.	548429344	PRM	2.8
Lewis, John C.	551742186	MGT	3.7
McFerran, Debra R.	409723145	IS	2.9
Sisneros, Michael	392416582	ACCT	3.3

DATA VS INFORMATION

- Another way to convert data into information is to **summarize** or process and present the data for human interpretation:

Baker, Kenneth D.	324917628
Doyle, Joan E.	476193248
Finkle, Clive R.	548429344
Lewis, John C.	551742186
McFerran, Debra R.	409723145



*This information could be used as a basis for deciding whether to add new courses or to hire new faculty members

DATABASES CAN CONTAIN BOTH DATA AND INFORMATION!

- Databases today may contain **either** data or information (or both).
- For **example:**
  - A database may contain an image of the **class roster** document shown before.
  - Data may often be **preprocessed** and stored in summarized form in databases that are used for decision support (Like whether or not to hire more faculty).

METADATA

- As we saw before, data becomes useful only when placed in some **context**. The primary mechanism for providing context for data is **metadata**.
- **Metadata** is data that describe the properties or characteristics of end-user data and the context of those data (Often times defined as **data about data**).
- Managing metadata is as crucial as managing the associated data because data without clear meaning can be confusing, misinterpreted, or erroneous.

METADATA EXAMPLE

Data Item - no sample data.

Metadata Notice there is **no** sample data. Metadata is once removed from data. This means metadata describes the data but are separate from the data

Data Item	Metadata					
Name	**Type**	**Length**	**Min**	**Max**	**Description**	**Source**
Course	Alphanumeric	30			Course ID and name	Academic Unit
Section	Integer	1	1	9	Section number	Registrar
Semester	Alphanumeric	10			Semester and year	Registrar
Name	Alphanumeric	30			Student name	Student IS
ID	Integer	9			Student ID (SSN)	Student IS
Major	Alphanumeric	4			Student major	Student IS
GPA	Decimal	3	0.0	4.0	Student grade point average	Academic Unit

METADATA

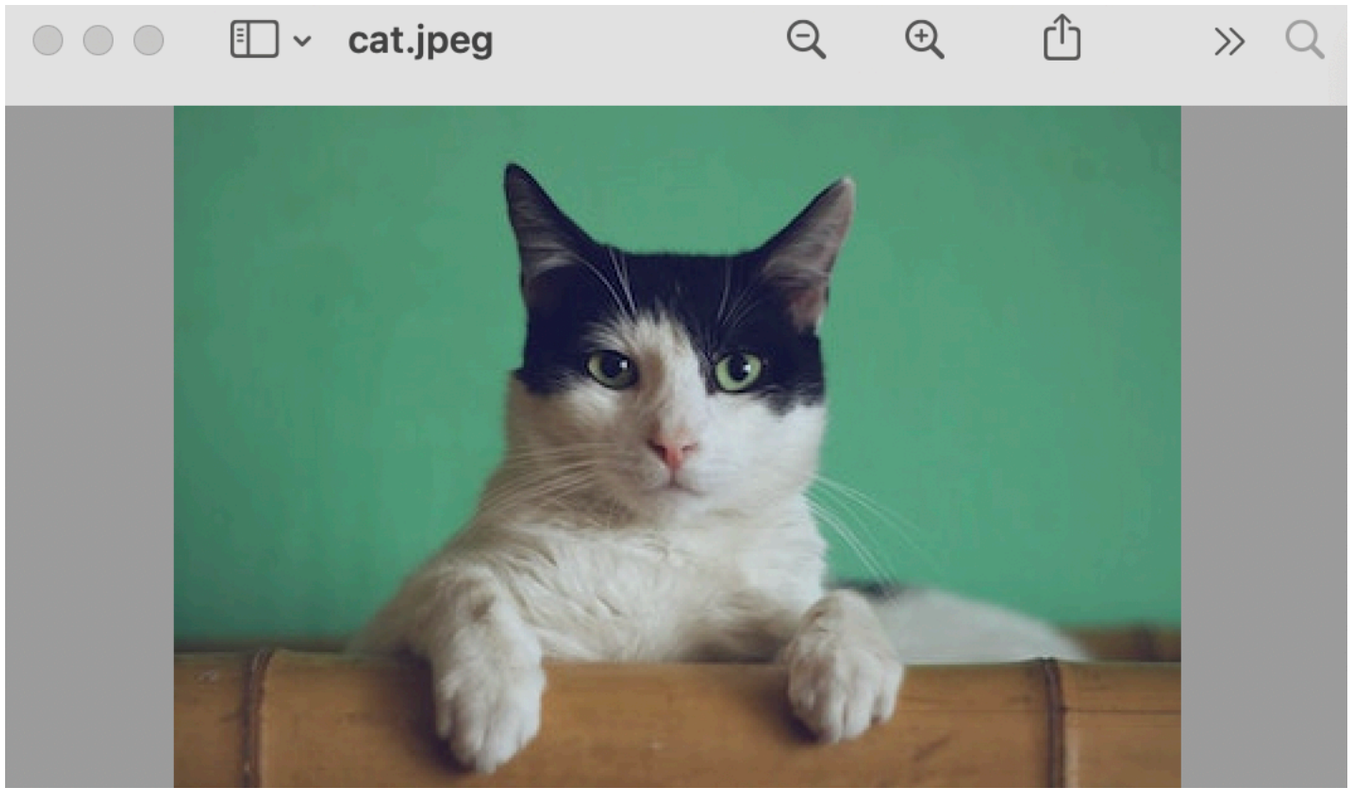
- Where else have we seen metadata??

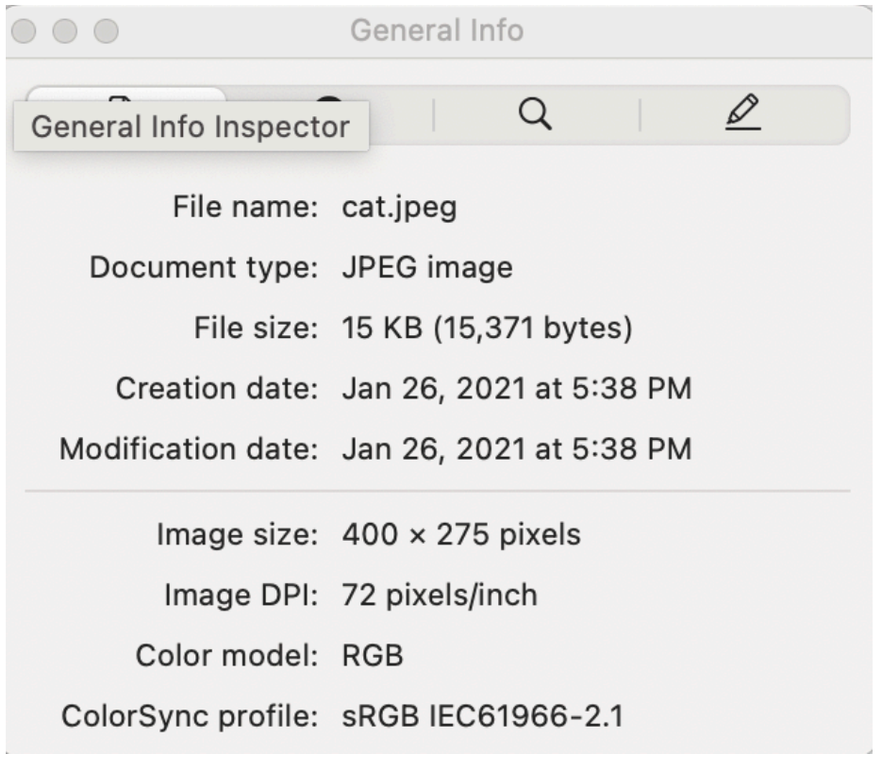
```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <meta name="keywords" content="empathy,the empathy project,empathy project,empathy,hu
  <meta name="description" content="The Empathy Project explores integrating empathy in
  <meta name="google-site-verification" content="S5NN5kkN4H7Tr2X1ZyMve9hsM_t59nLJiRb6Xf
  <link rel="stylesheet" type="text/css" href="/stylesheets/style.css">
  <link rel="shortcut icon" href="images/empathy_project_logo-01.png" type="image/x-ico
```

- In our **meta** tags inside our web pages! Metadata here is data about the data on the web page.

METADATA – ANOTHER EXAMPLE

- Images on your computer have metadata as well.





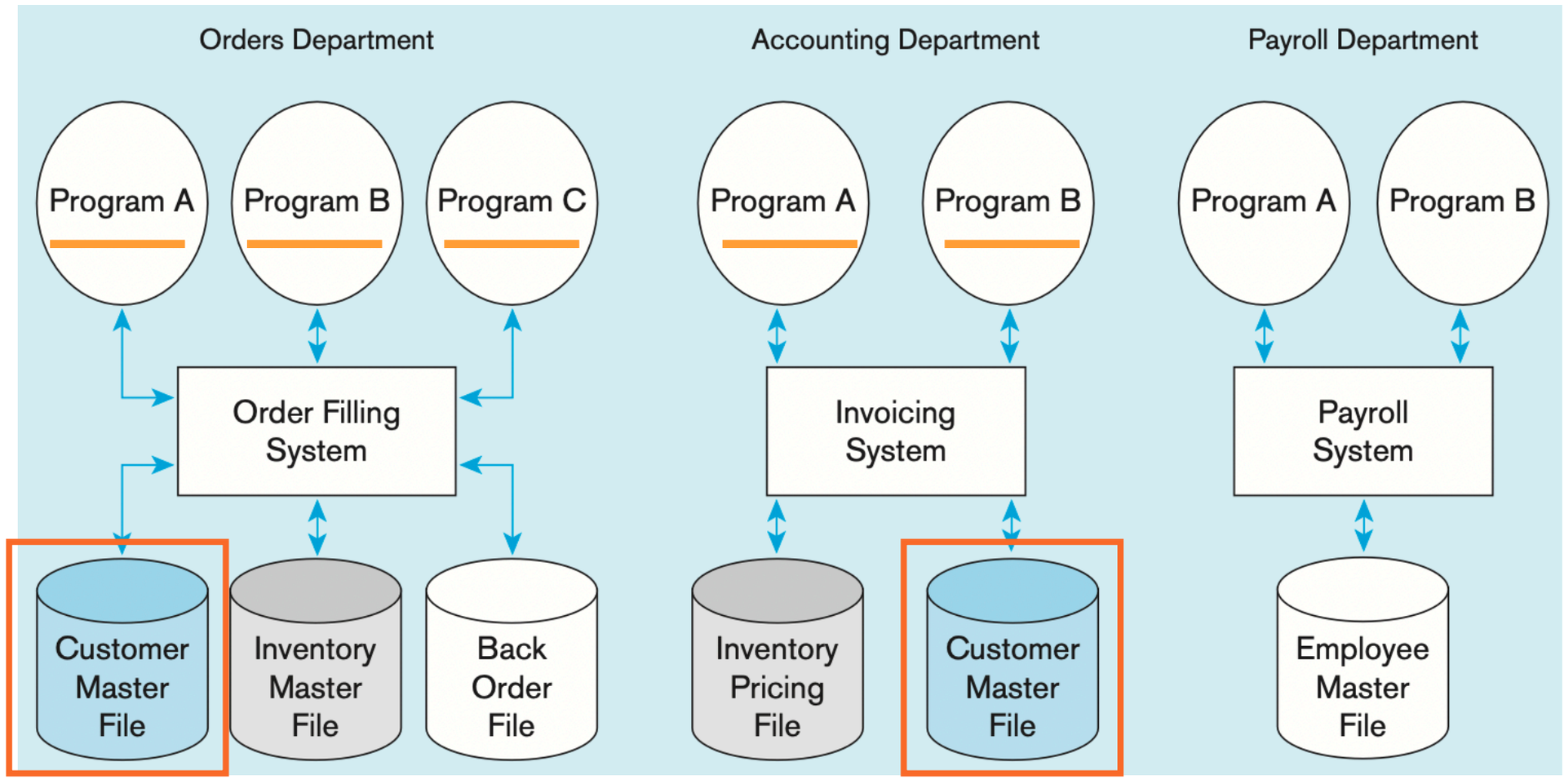
WHAT DID PEOPLE USE BEFORE DATABASES?

- Computer file processing systems were first developed for business applications to allow computers to store, manipulate and retrieve large files of data.
- As business applications became more complex, database processing systems began to replace file systems.
- **Excel file** generally fall into the same category as well!

DISADVANTAGES OF FILE PROCESSING SYSTEMS

- **Program-Data Dependence** - All programs maintain metadata for each file they use.
- Any changes made, the file descriptions in each program that is affected would have to be modified. It is often difficult to locate all programs affected. Errors are often introduced when making such changes.

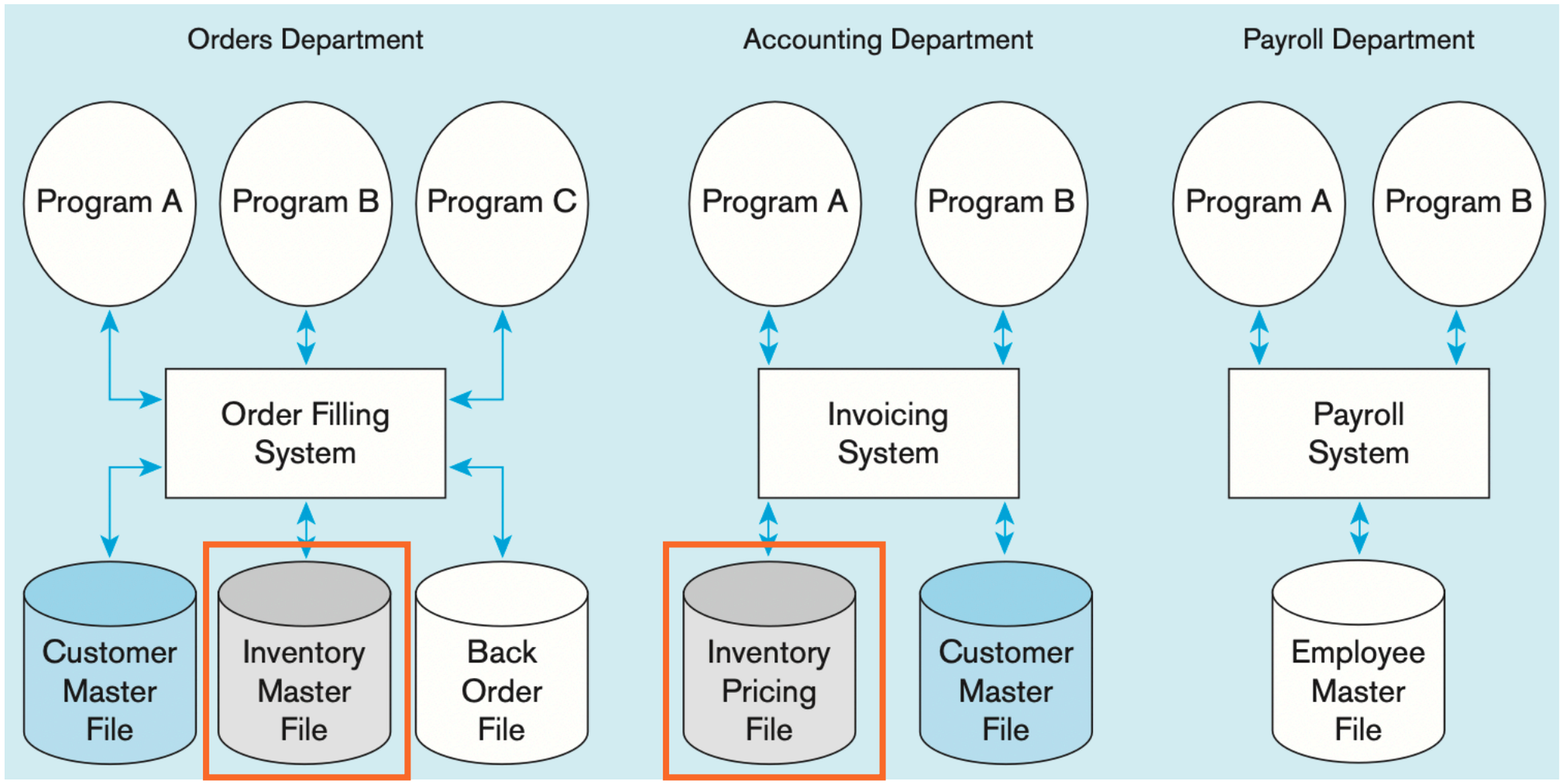
5 programs need to be modified!



DISADVANTAGES OF FILE PROCESSING SYSTEMS

- **Duplication of Data** - Different systems/programs have separate copies of the same data.
  - Data formats may be inconsistent and/or data values may not agree. Reliable metadata are very difficult to establish in file processing systems.

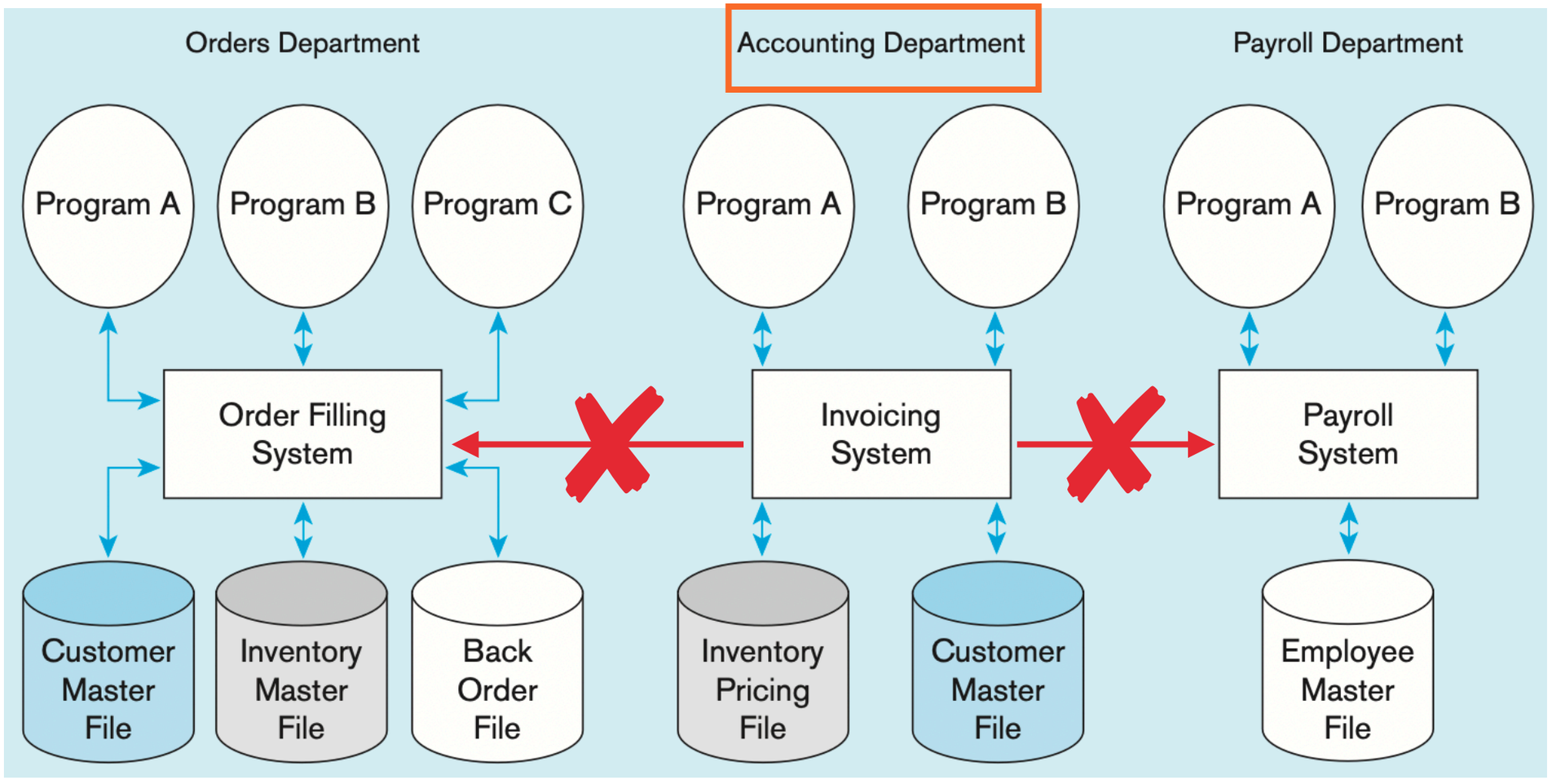
Example, the same data item may have different names in different files or, conversely, the same name may be used for different data items in different files.



DISADVANTAGES OF FILE PROCESSING SYSTEMS

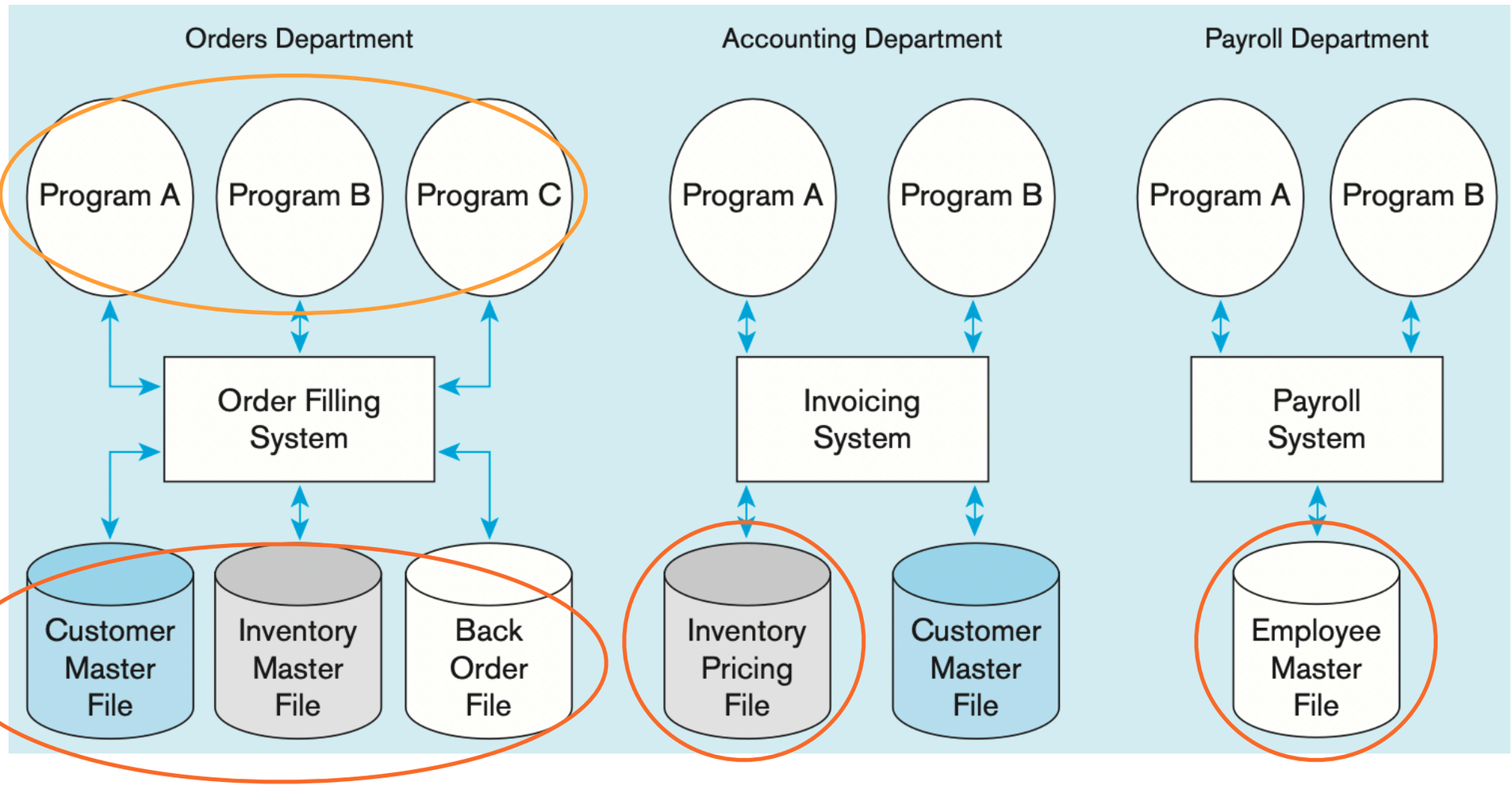
- **Limited Data Sharing** - No centralized control of data.
- With the traditional file processing approach, each application has its own private files, and users have little opportunity to share data outside their own applications.

Managers often find that a requested report requires a major programming effort because data must be drawn from several incompatible files in separate systems.



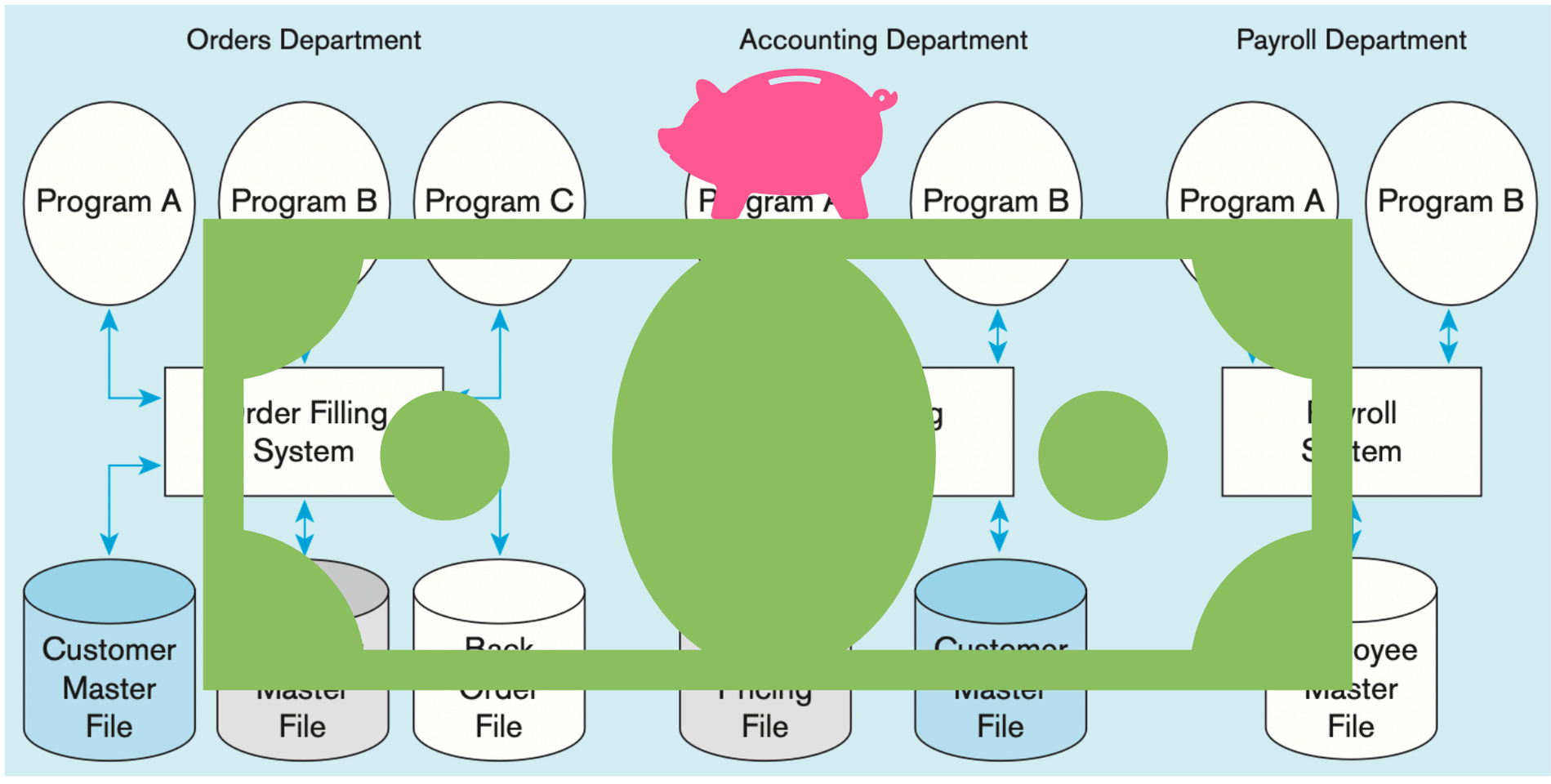
DISADVANTAGES OF FILE PROCESSING SYSTEMS

- **Lengthy Development Times** - Programmers must design their own file formats.
- Each new application requires that the developer essentially start from scratch by designing new file formats and descriptions and then writing the file access logic for each new program.



DISADVANTAGES OF FILE PROCESSING SYSTEMS

- **Excessive Program Maintenance** - 80% of information systems budget might be devoted to program maintenance.
- This in turn means that resources (time, people, and money) are not being spent on developing new applications.

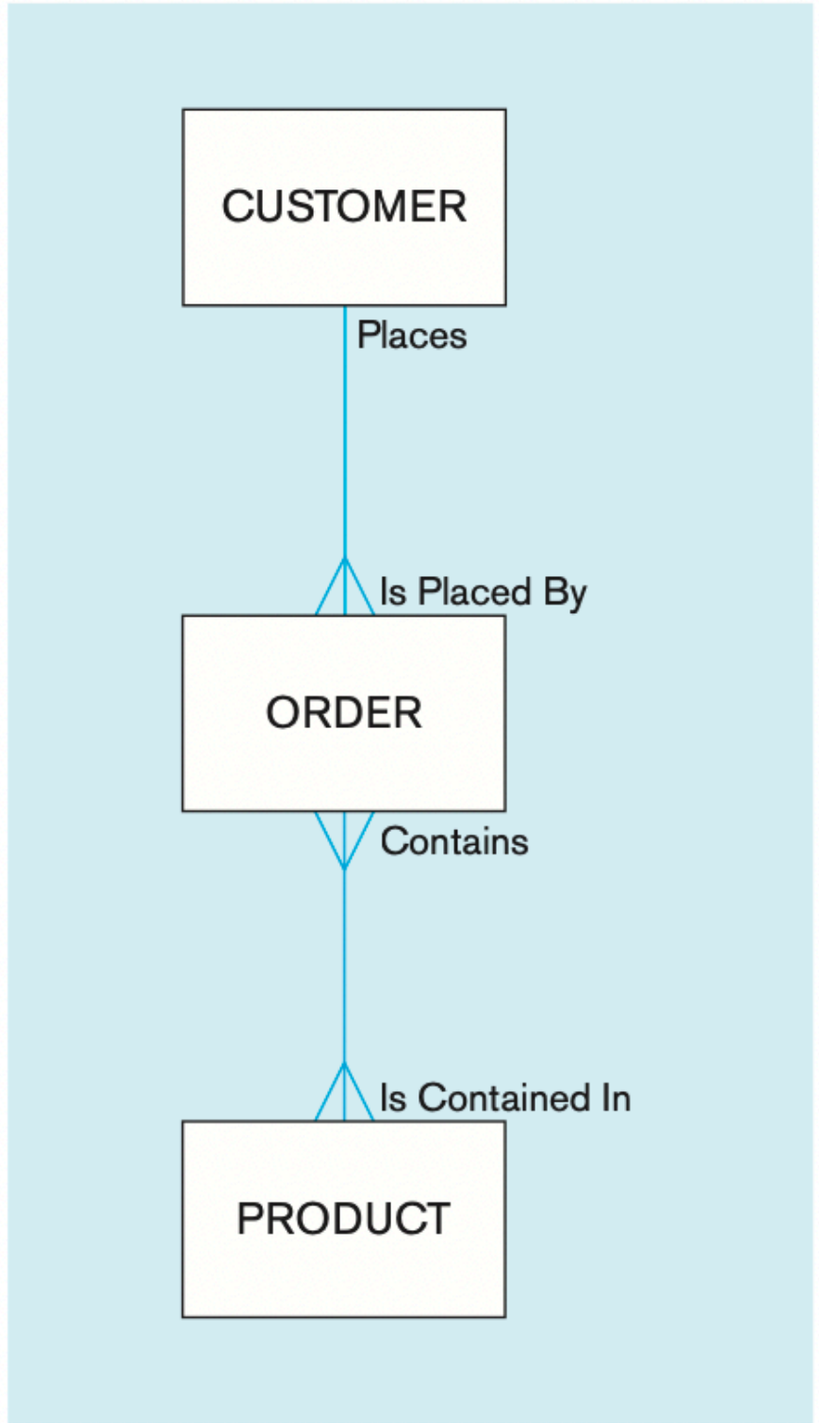


WHY SHOULD WE KNOW THE DISADVANTAGES TODAY?

- Many of the disadvantages of file processing we have mentioned can also be limitations of databases if an organization does not properly apply the database approach.
- **Example**, if an organization develops many separately managed databases (say, one for each division or business function) with little or no coordination of the metadata, then uncontrolled data duplication, limited data sharing, lengthy development time, and excessive program maintenance can occur.
- The **database approach** is a way to manage organizational data. It is also a set of technologies for **defining**, **creating**, **maintaining**, and **using** these data.

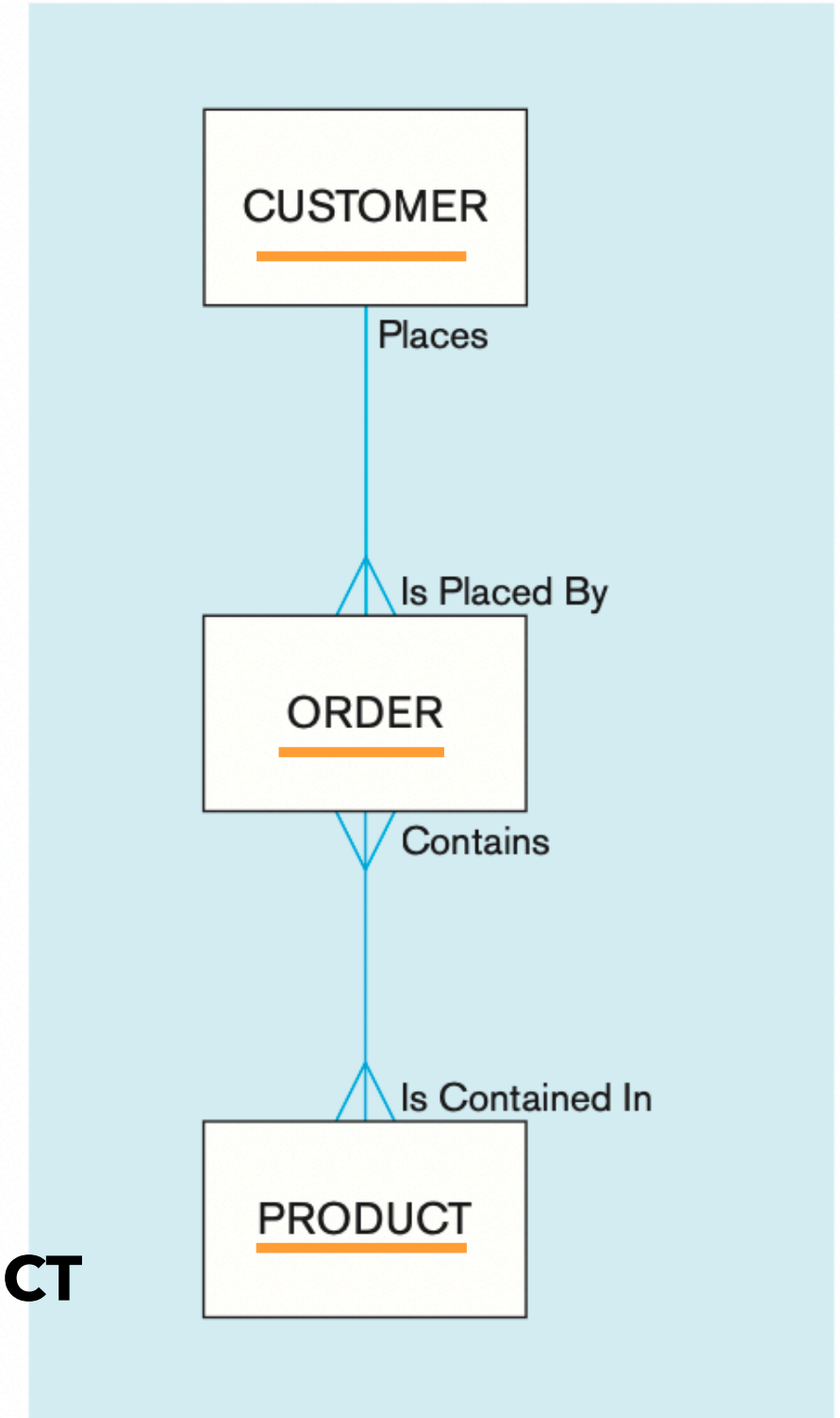
CORE CONCEPTS OF THE DATABASE APPROACH

- A **data model** is a graphical system used to capture the nature and relationships among data.
- The **effectiveness** and **efficiency** of a database is directly associated with the structure of the database.
- A typical data model is made up of **entities**, **attributes**, and **relationships**, and the most common data modeling representation is the **entity-relationship** model.



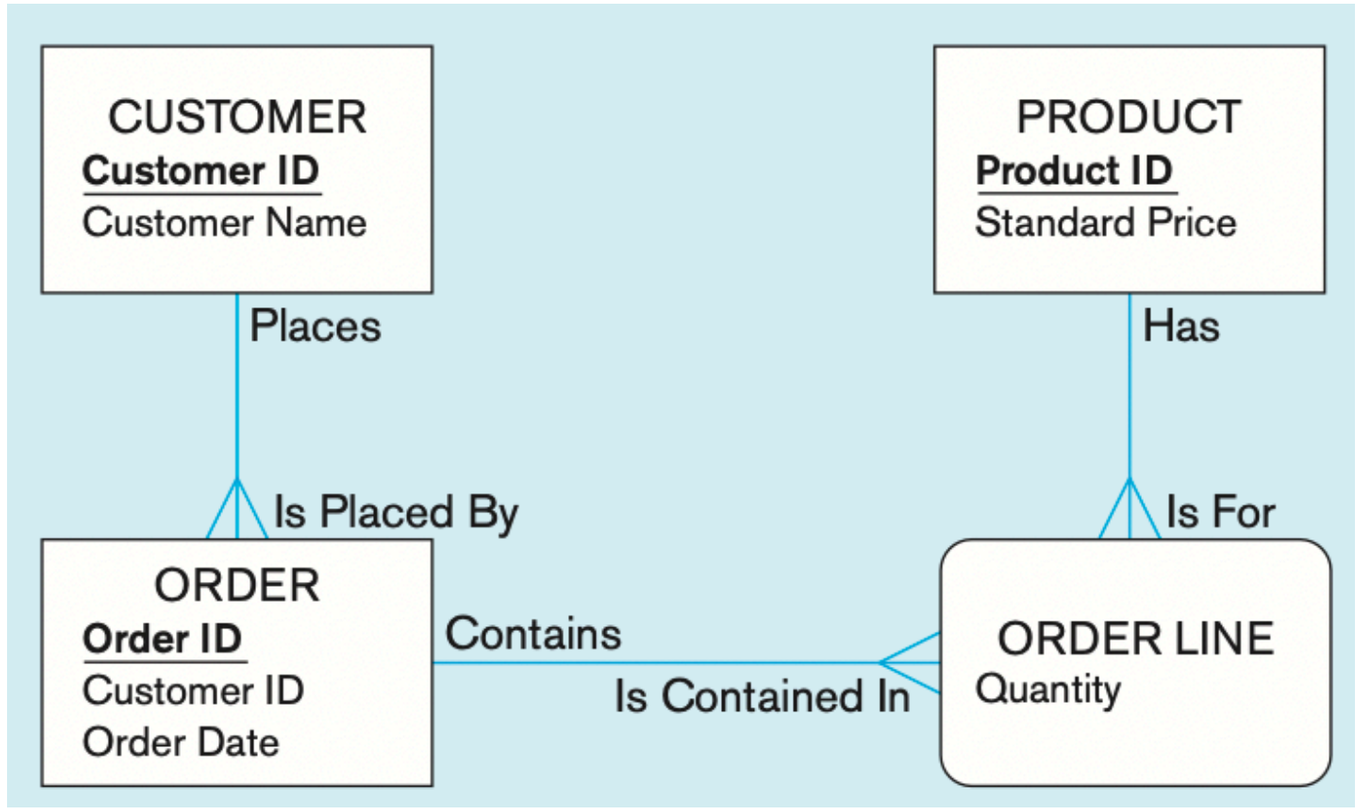
CORE CONCEPTS OF THE DATABASE APPROACH

- An **entity** is a person, a place, an object, an event, or a concept in the user environment about which the organization wishes to maintain data.
- Think of an entity as a **noun**! It **describes** a person, place, object, etc. in the business environment for which information must be recorded and retained.
- What are the entities in the adjacent data model? **CUSTOMER**, **ORDER**, and **PRODUCT**



ENTITIES CONTINUED

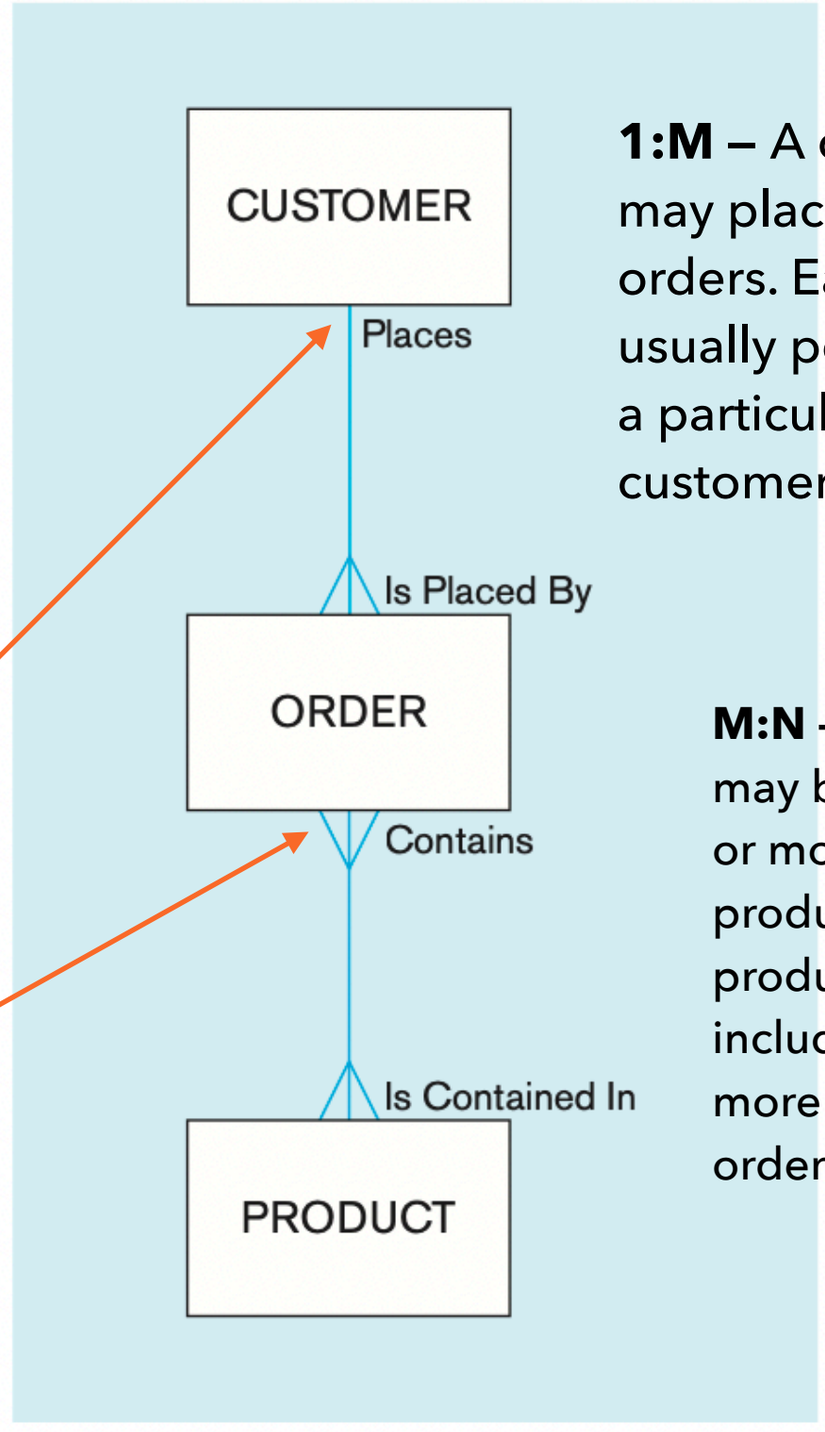
- The data you are interested in capturing about the entity (e.g., Customer Name) is called an ***attribute*** (like an attribute/instance variable when creating a Java class!)
- Data are recorded for **many** customers. Each customer's information is referred to as an ***instance*** of CUSTOMER (like when creating a Java object!)



RELATIONSHIPS

- A well-structured database establishes the relationships **between entities** that exist in organizational data so that desired information can be retrieved.

- Most relationships are one-to-many (**1:M**) or many-to-many (**M:N**).
  - The 1 nature of a relationship is marked by a **single** line.
  - The M nature of a relationship is marked by the **crow's foot** attached to the rectangle.

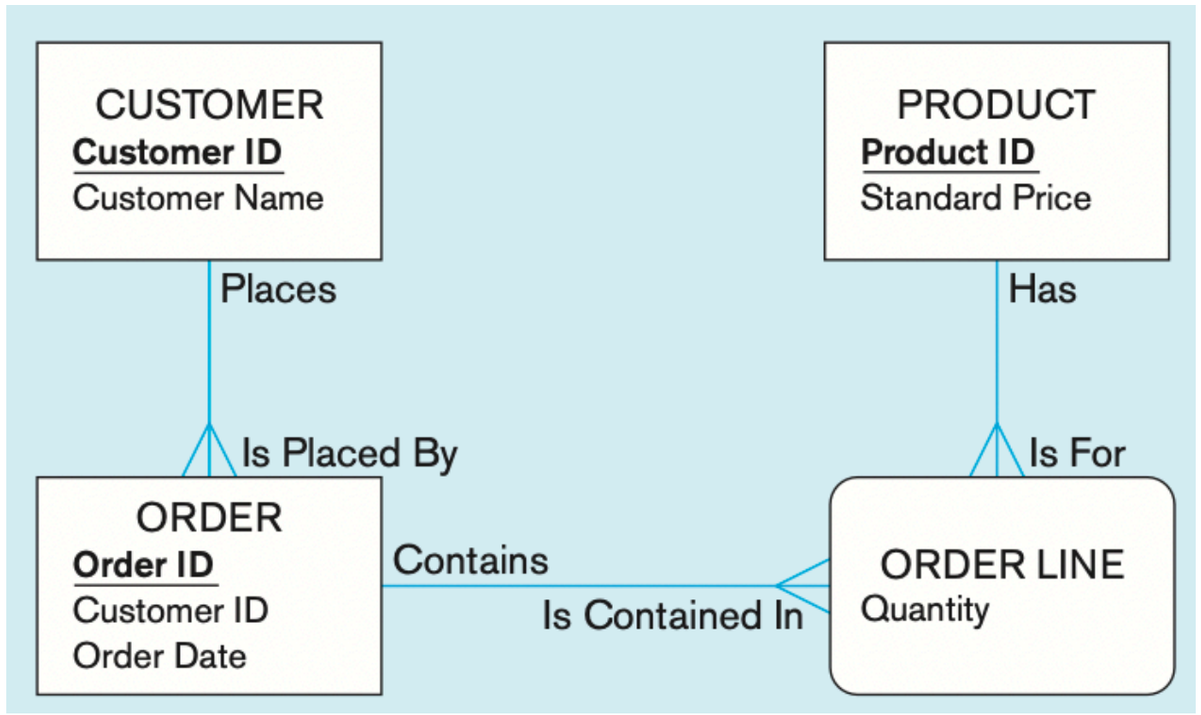


1:M – A customer may place many orders. Each order usually pertains to a particular customer.

M:N – An order may be for one or more products, and a product may be included on more than one order.

RELATIONAL DATABASES

- **Relational databases** establish the relationships between entities by means of common fields included in a file, called a **relation**.
- The relationship between a customer and the customer's order depicted in the data models in is established by including the **customer number** with the **customer's order**. Thus, a customer's identification number is included in the file (or relation) that holds customer information such as name, address, and so forth.
- Relational databases use the identification number to establish the relationship between customer and order.



DATABASE MANAGEMENT SYSTEM (DBMS)

- A **database management system** (DBMS) is a software system that enables the use of a database approach.
- The primary purpose of a DBMS is to provide a systematic method of **creating**, **updating**, **storing**, and **retrieving** the data stored in a database.
- It enables end users and application programmers to share data, and it enables data to be shared among multiple applications rather than propagated and stored in new files for every new application.
- A DBMS also provides facilities for controlling data access, enforcing data integrity, managing concurrency control, and restoring a database.