





#### Research & Statistics GROSSES - BROADWAY IN NYC



















 If you sell one or several books to a customer, what information would you include in the receipt?

**ISBN**, Book name, barcode, - books

**transaction number,** cashier name, Quantity, data of purchase, price, payment method, tax, discount, - transactions

**Store location,** store number, return policy. - stores

Shipping address, card info, **customer ID** - customers

## Thriftbooks.com receipt example



#### Order (#35092810)

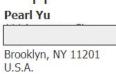
04/16/2022 09:22:12

1 Items		
My Own Words Acceptable condition (MD)		\$3.99
	Order Subtotal	\$3.99
	Sales Tax	\$0.47
	Shipping	\$1.29
	Order Total	\$5.75

#### **Payment**



#### Shipped To



**Print Order** 

#### This Order Has 1 Shipment

Shipment #168593280 From MD

Title Date

**My Own Words** 4/18/2022 12:00:00 AM



- A database is a collection of related data.
- Data are usually stored in the form of tables.
- There are different ways of storing data.

First Name	Last Name	Address	City	Age
	Name			



 If you sell one or several books to a customer, what information would you include in the receipt? What kind of database to store these info?

# Storing data



• Idea: Storing data in the form of a table.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15

## Storing data



The idea of storing data in the form of tables – "relational database" – was introduced by Edgar Frank Codd (1923 - 2003) in his seminal paper:

Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks". Communications of the ACM. 13 (6): 377–387.

- In the relational model, data is organized into table(s) (or "relations") of columns and rows, with a unique identifier for each row.
- The table(s) can capture entities (nouns) and relationships (verbs) between entities.

#### The 1981 ACM Turing Award Lecture

Delivered at ACM '81, Los Angeles, California, November 9, 1981



The 1981 ACM Turing Award was presented to Edgar F. Codd, an IBM Fellow of the San Jose Research Laboratory, by President Peter Denning on November 9, 1981 at the ACM Annual Conference in Los Angeles, California It is the Association's foremost award for technical contributions to the computing community.

Codd was selected by the ACM General Technical Achievement Award Committee for his "fundamental and continuing contributions to the theory and practice of database management systems." The originator of the relational model for databases, Codd has made further important contributions in the development of relational algebra, relational calculus, and normalization of

Edgar F. Codd joined IBM in 1949 to prepare programs for the Selective Sequence Electronic Calculator. Since then, his work in computing has encompassed logical design of computers (IBM 701 and Stretch), managing a computer center in Canada, heading the development of one of the first operating systems with a general multiprogramming capability, contributing to the logic of selfreproducing automata, developing high level techniques for software specifica-

tion, creating and extending the relational approach to database management, and developing an English analyzing and synthesizing subsystem for casual users of relational databases. He is also the author of Cellular Automata, an early volume in the ACM Monograph Series.

Codd received his B.A. and M.A. in Mathematics from Oxford University in England, and his M.Sc. and Ph.D. in Computer and Communication Sciences from the University of Michigan. He is a Member of the National Academy of Engineering (USA) and a Fellow of the British Computer Society.

The ACM Turing Award is presented each year in commemoration of A. M. Turing, the English mathematician who made major contributions to the computing sciences.

#### Relational Database: A Practical Foundation for Productivity

E. F. Codd IBM San Jose Research Laboratory

It is well known that the growth in demands from end users for new applications is outstripping the capability of data processing departments to implement the corresponding application programs. There are two complementary approaches to attacking this problem (and both approaches are needed); one is to put end users into direct touch with the information stored in computers; the other is to increase the productivity of data processing professionals in the development of application programs. It is less well known that a single technology,

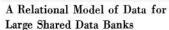
relational database management, provides a practical foundation for both approaches. It is explained why this

While developing this productivity theme, it is noted that the time has come to draw a very sharp line between relational and non-relational database systems, so that the label "relational" will not be used in misleading ways. The key to drawing this line is something called a "relational processing capability."

CR Categories and Subject Descriptors: H.2.0 [Database Management]: General; H.2.1 [Database Management]: Logical Design-data models, H.2.4 [Database Management]: Systems

General Terms: Human Factors, Languages

Additional Key Words and Phrases: database, relational database, relational model, data structure, data manipulation, data integrity, productivity



E. F. Copp

IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, jain, retrieval language, predicate calculus, security, data integrity

CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

#### I. Relational Model and Normal Form

#### 1.1. Introduction

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of data independence—the independence of application programs and terminal activities from growth in data types and changes in data representation-and certain kinds of data inconsistency which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for noninferential systems. It provides a means of describing data with its natural structure only-that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations-these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

#### 1.2. Data Dependencies in Present Systems

The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed without logically impairing some application programs is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. Ordering Dependence. Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

Author's Present Address: E. F. Codd, IBM Research Laboratory,

<sup>5600</sup> Cottle Road, San Jose, CA 95193. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission. © 1982 ACM 0001-0782/82/0200-0109 \$00.75



Consider the following database of the book seller.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15

NYU

STERN SCHOOL OF BUSINESS

- Scenario 1:
  - Add a book for:

Book id: EEE

Title: Business Administration

Author: K. Greene

What problems do you encounter?

Why is this not optimal?



Consider the following database of the book seller.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15



- Scenario 1:
  - Add a book for:

Book id: EEE

Title: Business Administration

Author: K. Greene

- What problems do you encounter?
  - Empty fields (4 fields have no data until purchased).
- Why is this not optimal?
  - Waste of computer space, especially if there is a large number (millions) of unsold books.



- Scenario 2:
  - Change the address of the customer M. Jones from

25 Sun

To

26 Moon

What problems do you encounter?



Consider the following database of the book seller.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15



- Scenario 2:
  - Change the address of the customer M. Jones from
     25 Sun

To

26 Moon

- What problems do you encounter?
  - Every record for M. Jones would have to be located and updated to ensure consistency throughout the database (waste of time, especially if the database is very large, with millions of records).

NYU

STERN SCHOOL OF BUSINESS

- Scenario 3:
  - Delete customer record for E. Rice.
- What problems do you encounter?



Consider the following database of the book seller.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15



- Scenario 3:
  - Delete customer record for E. Rice.
- What problems do you encounter?
  - Deletion of book information DDD, Spring, M. Mouse (loss of valuable information)



- Scenario 4:
  - The seller wants to know the most expensive book it carries.
- What problems do you encounter?



Consider the following database of the book seller.

Customer Name	Customer Address	Book ID	Book Title	Book Author	Purchase Date	Purchase Price (\$)
J. Smith	12 Elk	AAA	Peace	A. Bart	03/04/15	20
M. Jones	25 Sun	BBB	War	M. Hine	03/04/16	18
G. Hart	73 Sera	CCC	System	N. Vang	03/05/17	15
V. Hicks	22 Mann	AAA	Peace	A. Bart	03/19/18	20
E. Rice	69 Witt	DDD	How to fly	M. Mouse	03/06/19	25
M. Jones	25 Sun	CCC	System	N. Vang	03/15/19	15



- Scenario 4:
  - The seller wants to know the most expensive book it carries.
- What problems do you encounter?
  - Have to scan the entire "purchase" table.
  - Search the same information (for example, AAA Peace \$20) twice (redundancy and waste of time)
  - This can slow down the search if the database if large (thousands of books / millions of purchases).

## An Immediate Solution?



Splitting the long table into several smaller tables.

Custome r Name	Customer Address
J. Smith	12 Elk
M. Jones	25 Sun
G. Hart	73 Sera
V. Hicks	22 Mann
E. Rice	69 Witt

Book ID	Book Title	Book Author
AAA	Peace	A. Bart
BBB	War	M. Hine
CCC	System	N. Vang
DDD	How to fly	M. Mouse
CCC	System	N. Vang

Custome r Name	Book ID	Purchas e Date	Purchas e Price (\$)
J. Smith	AAA	03/04/15	20
M. Jones	BBB	03/04/16	18
G. Hart	CCC	03/05/17	15
V. Hicks	AAA	03/19/18	20
E. Rice	DDD	03/06/19	25
M. Jones	CCC	03/05/20	15

# A Systematic Solution - Data Modeling



- Entity-relationship model (ERM)
  - Most widely-used data modeling method
  - The method was developed from Peter Chen's (Harvard; MIT; UCLA; CMU; Louisiana State, Carnegie Mellon) foundational paper in 1976.
  - This method is intuitive but rigorous, and bridges business concepts with technical design.
  - ER modeling has been incorporated into many major database products.

# **ER Modeling**

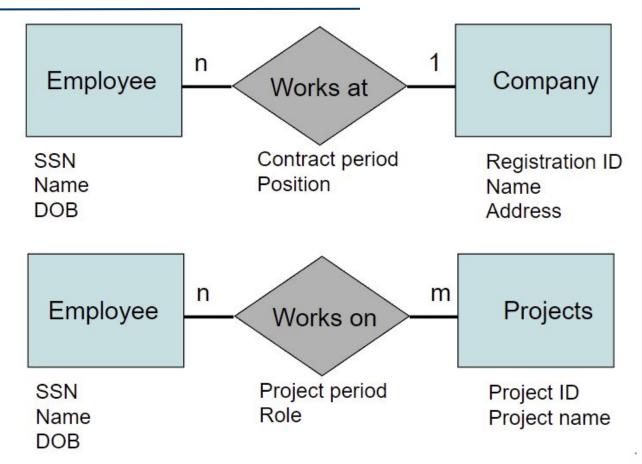


#### ER Model

- Abstract and conceptual representation of data in software engineering, database design, etc.
  - An entity: A thing which is recognized as being capable of an independent existence (a "Noun").
  - A relationship: Captures how two or more entities are related to one another (a "Verb")
  - Attributes: qualify the entity and relationship

# ER Modeling - Example





# Book purchasing

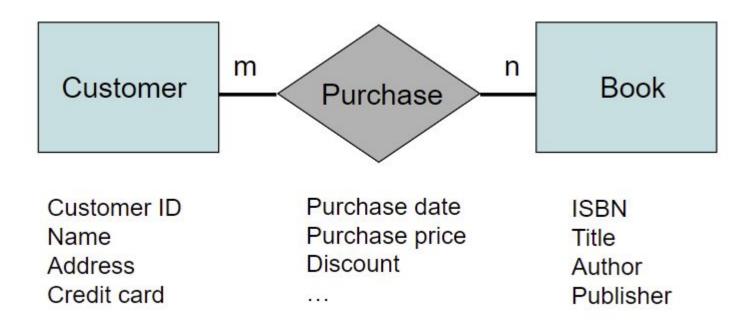


 How would you model purchasing activities (customers buy books) in a bookstore?

## Book purchasing

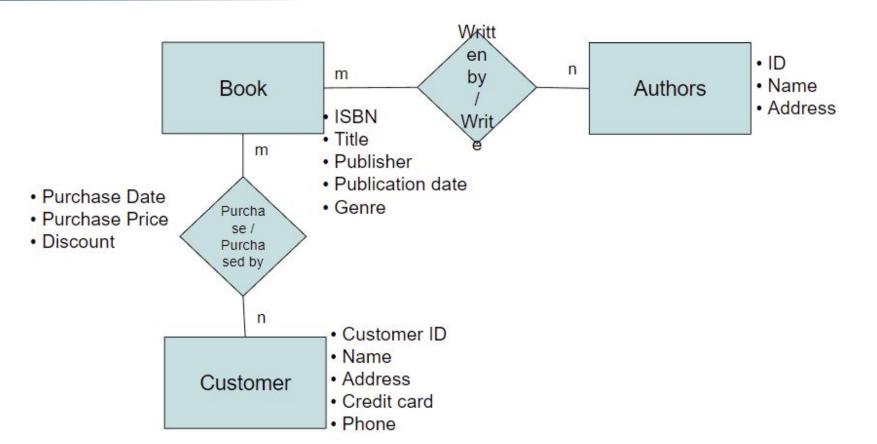


 How would you model purchasing activities (customers buy books) in a bookstore?



# Book purchasing

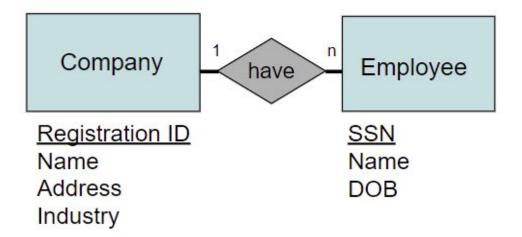








- The purpose of ER diagram:
  - Helping us understand business relationships and create effective and efficient tables.
- Conversion into tables:
  - How would you convert the following into tables?

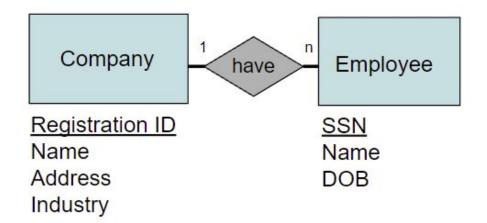




## Represent Entities

- Each regular entity is transformed into a table.
- The identifier of the entity type becomes the primary key of the corresponding table.
- The primary key must satisfy the following two conditions:
  - The value of the key must uniquely identify every row in the relation.
  - The key should be nonredundant.





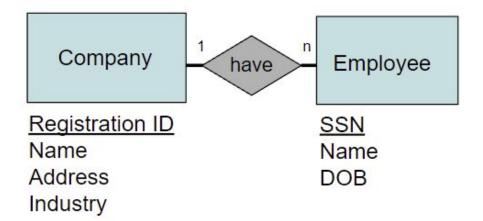
Registrati on ID	Name	Address	Industry
001	Nasdaq	1 K St	
002	Apple	3 B St	

<u>SSN</u>	Name	DOB	Registrati on ID
00123	J. Smith		001
00124	A. Rice		002



- Represent Relationships
- Basic rules to convert ER into tables:
  - For 1:1 cardinality relationships, the primary key of one side is placed as a foreign key in the other.
  - For 1:N cardinality relationship, post the primary key from the "one" side as an attribute to the "many" side.





Registrati on ID	Name	Address	Industry
001	Nasdaq	1 K St	
002	Apple	3 B St	

<u>SSN</u>	Name	DOB
00123	J. Smith	
00124	A. Rice	



#### **Assume:**

- -10k firms in NYC
- -200 employees per firm

Registr ation ID	Org. Name	Org. Addres s	Industr y	
001	Nasdaq	1 K St		
002	Apple	3 B St		
003	Amazon	1 V Ave.		
004	NYU	44 W4		

<u>SSN</u>	Employee Name	DOB	Registrati on ID
00123	J. Smith		001
10124	D. Duck		001
25666	M. Mouse		001
98765	Minnie		002

Size of Table 1: (assume 20 columns)

- 10k rows \* 20 columns = 200k cells

<u>Size of Table 2</u>: (assume 4 columns)

- 2 mil rows \* 4 columns = 8 mil cells

Total size: 200k + 8 mil = 8.2 million





#### **Assume:**

- -10k firms in NYC
- -200 employees per firm

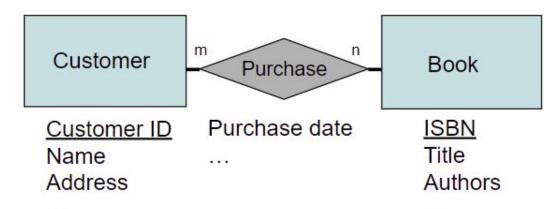
Registrati on ID	Org. Name	Org. Address	Industry	 SSN	Employe e Name	DOB
001	Nasdaq	1 K St		 00123	J. Smith	
001	Nasdaq	1 K St		 10124	D. Duck	
001	Nasdaq	1 K St		 25666	M. Mouse	
002	Apple	3 B St		 98765	Minnie	
003	Amazon	1 V Ave.		 19876	ABC	
003	Amazon	1 V Ave.		 08876	XYZ	

<u>Total size:</u> 2 mil rows \* (20 + 3) columns = 46 million cells



- Represent Relationships
- Three basic rules to convert ER into tables:
  - For M:N cardinality relationships: Create a new table and post the primary keys from each entity as attributes in the new table.
  - The primary key of the new table is the compound attributes.





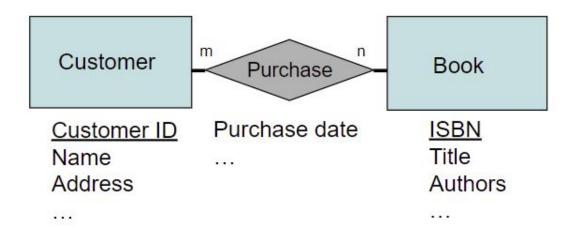
<u>Customer ID</u>	Name	Address	•••
001	J. Smith	1 K St	
002	M. Jones	3 B St	

<u>ISBN</u>	Title	Author	
00123	Outsourcing	M. Lacity	
00124	Computers	N.Cook	

Customer ID	<u>ISBN</u>	Purchase date
001	00123	09-20-2017
002	00124	08-12-2017

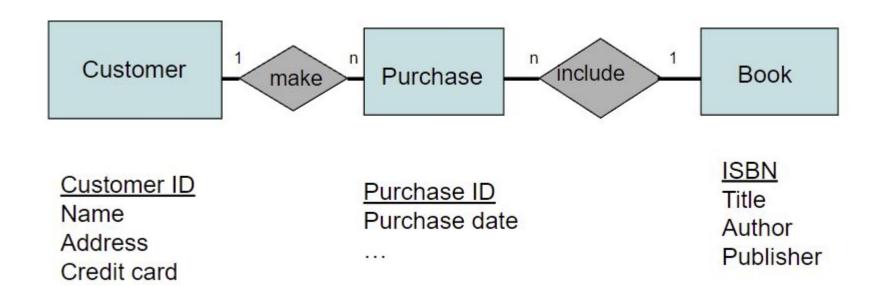


Can you convert a M:N relationship into multiple 1:N relationships?





Can you convert a M:N relationship into multiple 1:N relationships?



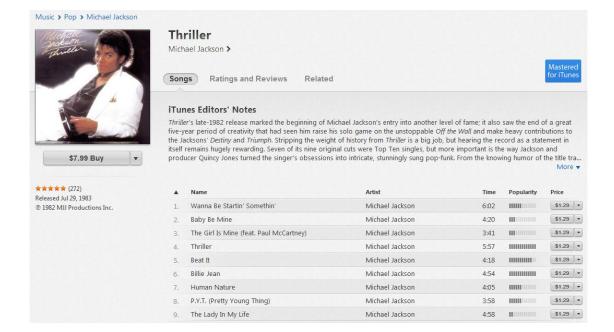
## **ER Modeling**



- A graphical description of entities, relationships, and attributes it reflects real world facts / business models, and is used as a blueprint for creating a database.
- It helps create a set of databases (tables) that are efficient (to retrieve information), small (in its consumption of storage space), and have little redundancy.
- The problems discussed previously can be avoided in these databases.

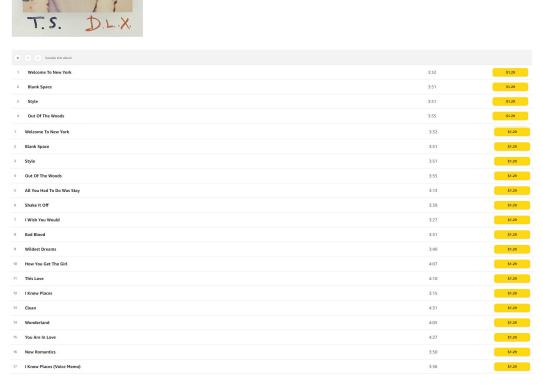
#### **ER Modeling**





#### Similar to







#### Similar to





K      H Listen to full-length playback of eligible Prime Music content	
1 Come Together (2019 Mix)	4.20 <b>                                    </b>
2 Something (2019 Mix)	3:02
3 Maxwell's Silver Hammer (2019 Mix)	3:28
4 Oh! Darling (2019 Mix)	3:27 ####################################
5 Octopus's Garden (2019 Mix)	2.50
6 I Want You (She's So Heavy) (2019 Mix)	7:47 // prime + Add

### **Exercise: ER Modeling**



#### Music > Pop > Michael Jackson



\$7.99 Buy

★★★★ (272) Released Jul 29, 1983 ® 1982 MJJ Productions Inc.

#### Thriller

Michael Jackson >

Songs

Ratings and Reviews Related



#### iTunes Editors' Notes

Thriller's late-1982 release marked the beginning of Michael Jackson's entry into another level of fame; it also saw the end of a great five-year period of creativity that had seen him raise his solo game on the unstoppable Off the Wall and make heavy contributions to the Jacksons' Destiny and Triumph. Stripping the weight of history from Thriller is a big job, but hearing the record as a statement in itself remains hugely rewarding. Seven of its nine original cuts were Top Ten singles, but more important is the way Jackson and producer Quincy Jones turned the singer's obsessions into intricate, stunningly sung pop-funk. From the knowing humor of the title tra...

More ▼

•	Name	Artist	Time	Popularity	Price
1.	Wanna Be Startin' Somethin'	Michael Jackson	6:02	IIIIII IIIIII	\$1.29
2.	Baby Be Mine	Michael Jackson	4:20	шании	\$1.29
3.	The Girl Is Mine (feat. Paul McCartney)	Michael Jackson	3:41	<b>III</b>	\$1.29
4.	Thriller	Michael Jackson	5:57	1111111111111	\$1.29
5.	Beat It	Michael Jackson	4:18	10000000	\$1.29
6.	Billie Jean	Michael Jackson	4:54		\$1.29
7.	Human Nature	Michael Jackson	4:05		\$1.29
8.	P.Y.T. (Pretty Young Thing)	Michael Jackson	3:58	HIHI WALL	\$1.29
9.	The Lady In My Life	Michael Jackson	4:58	Homon	\$1.29

#### Exercise: ER Modeling



- Identify the important entities, relationships, and attributes implied by this system output.
- Create an ER model showing the entities and attributes, and the relationships between these entities.