Principles of Database Systems (CS307)

程然

(Original Slides by 马昱欣 老师)

Department of Computer Science and Engineering
Southern University of Science and Technology

Instructor

- **Dr. Ran Cheng 程然** (Associate Professor in CSE)
  - Office: Room 316, South Tower, Engineering Building
  - Personal Website: https://chengran.tech/
  - Group Website: https://emi.sustech.edu.cn/

Lecture Notes, Sakai and QQ Group

- Slides and lab sheets will be on the Sakai site.
- QQ Group
  - ID: 949474976
  - Teaching assistants will be there to help you

Textbooks

- Reference book:
  - A. Silberschatz, H. Korth, and S. Sudarshan. **Database System Concepts**. McGraw-Hill, New York, 7th Edition, (2019).

Grading Policy

- Lecture and Lab Attendance (10%)
- Assignments 24%)
  - 4 Assignments (including 1 in-class quiz)
- Project (36%)
  - 2 Projects
- Final exam (30%)

Grading Policy

- Late Submission
  - We do not accept late submissions. All assigments, quizzes, and projects, etc. will receive a score of zero if you miss the deadline.
- Groups for Projects
  - Groups across different lab sessions are not allowed
  - Please try to find your teammate in the same lab session
- Grades
  - The teachers and TAs guarantee that your assignments and projects will be evaluated carefully and unbiasedly
  - We do not accept argueing with teachers over a certain grade once the decision has been made

Plagiarism

Zero tolerance!

Some Other Stuff

- Computing technologies advance very fast
  - Search online to learn more by yourself
    - Search engines (Google, Bing, Baidu, etc.), StackOverflow, GitHub.
  - The lecture notes can guide your self study

- You are encouraged to ask questions
  - At any time

- Practice makes perfect
  - No need to be afraid of trying new techniques/ideas/codes

Principles of Database Systems (CS307)

Lecture 1: Introduction to Databases

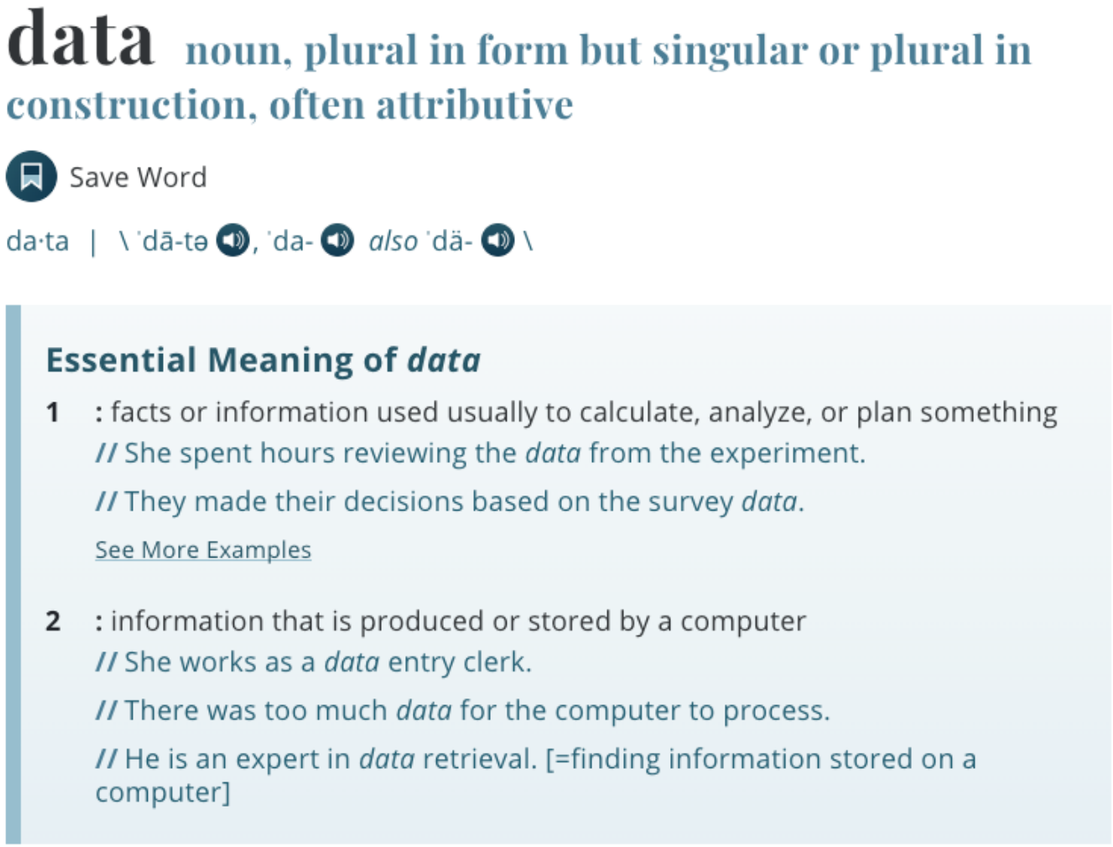
程然

Department of Computer Science and Engineering
Southern University of Science and Technology

- Most contents are from slides made by Stéphane Faroult and the authors of Database System Concepts (7th Edition).

What is a Database?

- Well, first, let's take a step back: What is data?



data noun, plural in form but singular or plural in construction, often attributive

Save Word

da·ta | \ ˈdā-tə, ˈda- *also* ˈdä- \

Essential Meaning of *data*

1 : facts or information used usually to calculate, analyze, or plan something
// She spent hours reviewing the *data* from the experiment.
// They made their decisions based on the survey *data*.
See More Examples

2 : information that is produced or stored by a computer
// She works as a *data* entry clerk.
// There was too much *data* for the computer to process.
// He is an expert in *data* retrieval. [=finding information stored on a computer]

What is a Database?

- A modern database system is a complex software system whose task is to manage a large, complex collection of data.
  - Collection of interrelated data
  - Set of programs to access the data
  - An environment that is both *convenient* and *efficient* to use
- Databases touch all aspects of our lives

Applications of Database

- Enterprise Information
  - **Sales**: customers, products, purchases
  - **Accounting**: payments, receipts, assets
  - **Human Resources**: Information about employees, salaries, payroll taxes.
- Manufacturing
  - Management of production, inventory, orders, supply chain.
- Universities
  - Registration, grades

Applications of Database

- Databases are everywhere today
  - … but the concept is old
  - The idea was to have one system doing once and for all the boring data storage/retrieval part
    - What boring parts?

Purpose of Database Systems

在 20 世纪 50 年代后期到 20 世纪 60 年代中期，计算机中的磁盘和磁鼓等直接存取设备开始普及。这时，可以将数据存储在计算机的磁盘上。这些数据都以文件的形式存储，然后通过文件系统来管理这些文件。

在 20 世纪 60 年代后期，随着网络技术的发展，计算机软/硬件的进步，出现了数据库技术，该阶段就是所谓的数据库系统阶段。
数据库系统阶段使用专门的数据库来管理数据，用户可以在数据库系统中建立数据库，然后在数据库中建立表，最后将数据存储在这些表中。用户可以直接通过数据库管理系统来查询表中的数据。

- In the early days ...
  - Database applications were built directly on top of file systems
    - (And we will have a lab session about it)
  - However, it suffers from many issues, including (but not limited to):
    - Data redundancy and inconsistency 数据冗余与不一致性
    - Data isolation 数据隔离
    - Integrity problems 完整性
    - Atomicity of operations 操作原子性
    - Concurrent access by multiple users 并发访问
    - Security problems 安全问题

Purpose of Database Systems

数据管理的3个阶段	人工管理（20 世纪 50 年代中期）	文件系统（50 年代末至 60 年代中期）	数据库系统（60 年代后期）
应用背景	科学计算	科学计算、管理	大规模数据、分布数据的管理
硬件背景	无直接存取存储设备	磁带、磁盘、磁鼓	大容量磁盘、可擦写光盘、按需增容磁带机等
软件背景	无专门管理的软件	利用操作系统的文件系统	由 DBMS 支撑
数据处理方式	批处理	联机实时处理、批处理	联机实时处理、批处理、分布处理
数据的管理者	用户/程序管理	文件系统代理	DBMS 管理
数据应用及其扩充	面向某一应用程序难以扩充	面向某一应用系统、不易扩充	面向多种应用系统、容易扩充
数据的共享性	无共享、冗余度极大	共享性差、冗余度大	共享性好、冗余度小
数据的独立性	数据的独立性差	物理独立性好、逻辑独立性差	具有高度的物理独立性、具有较好的逻辑独立性
数据的结构化	数据无结构	记录内有结构、整体无结构	统一数据模型、整体结构化
数据的安全性	应用程序保护	文件系统保护	由 DBMS 提供完善的安全保护

le systems

not limited to):

一致性

Let’s Show an Example

- Write a Java program to manage information of all students in CS307
  - We have classroom sessions and lab sessions
    - (potential redundancy of students)
  - A new Java class method for each task
    - (difficulty in accessing data)
  - Maybe you will split the students into files based on lab session times
    - (split into multiple files; hard to manage the files)
  - You need to check the validity of Student IDs and classes
    - (hard to maintain integrity constraints)
  - Atomicity? Concurrency? Security? All need to be handled by yourself

When there are redundancies, an organization tells some of its employees to leave because their jobs are no longer necessary or because the organization can no longer afford to pay them.

Let's Show an Example (Well, still remember how to read and write files in Java?)

- Write a Java program to manage information of all students in CS307
  - We have classroom sessions and lab sessions
    - (potential redundancy of students)
  - A new Java class method for each task
    - (difficulty in accessing data)
  - Maybe you will split the students into files based on lab session times
    - (split into multiple files; hard to manage the files)
  - You need to check the validity of Student IDs and classes
    - (hard to maintain integrity constraints)
  - Atomicity? Concurrency? Security? All need to be handled by yourself

Let’s Show an Example

- What is more,
  - Can you reuse the code you just wrote in a staff management system?

Purpose of Database Systems

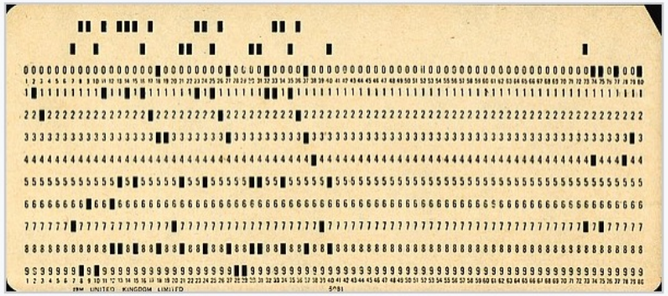
- In the early days …
  - Database applications were built directly on top of file systems
    - (And we will have a lab session about it)
  - However, it suffers from many issues, including (but not limited to):
    - Data redundancy and inconsistency 数据冗余与不一致性
    - Data isolation 数据隔离
    - Integrity problems 完整性
    - Atomicity of operations 操作原子性
    - Concurrent access by multiple users 并发访问
    - Security problems 安全问题

Database systems offer solutions to all the problems mentioned above

Database

A Bit of History

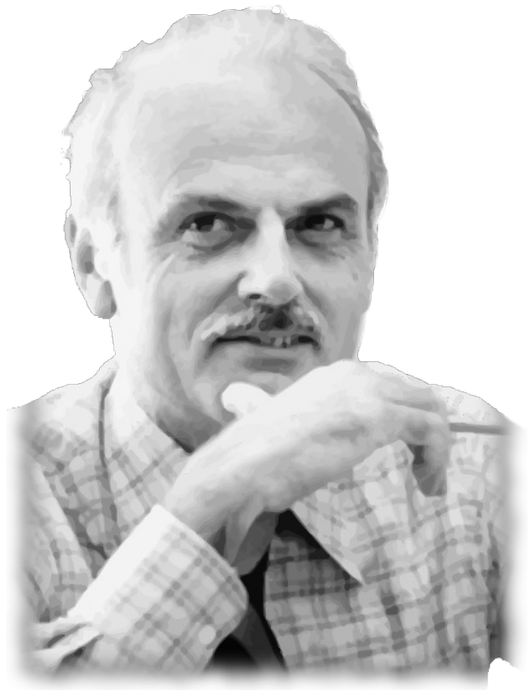
- 1950s and early 1960s: 打孔卡
  - Data processing using magnetic tapes for storage
    - Tapes provided only sequential access
  - Punched cards for input
- Late 1960s and 1970s: 硬盘
  - Hard disks allowed direct access to data
  - Network and hierarchical data models in widespread use
  - **Ted Codd** defines the relational data model
    - Would win the ACM Turing Award for this work
    - IBM Research begins System R prototype
    - UC Berkeley (Michael Stonebraker) begins Ingres prototype
    - Oracle releases first commercial relational database
  - High-performance (for the era) transaction processing



A 12-row/80-column IBM punched card from the mid-twentieth century

Internals of a 2.5-inch laptop hard disk drive

A Bit of History



Edgar F. “Ted” Codd
1923 – 2003
Turing Award 1981

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
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

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 non-inferential 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.

E. F. Codd, A Relational Model of Data for Large Shared Data Banks, Information Retrieval, June, 1970

A Bit of History

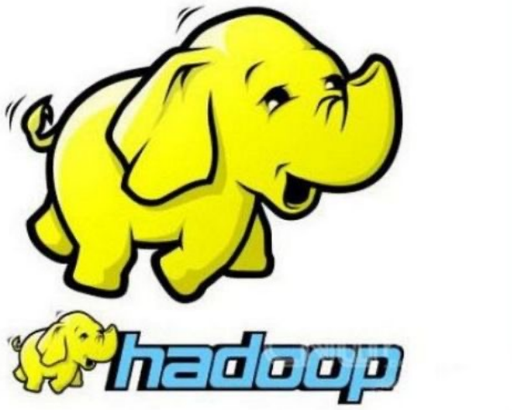
- 1980s: SQL工业应用
  - Research relational prototypes evolve into commercial systems
    - SQL becomes industrial standard
  - Parallel and distributed database systems
    - Wisconsin, IBM, Teradata
  - Object-oriented database systems
- 1990s: 互联网
  - Large decision support and data-mining applications
  - Large multi-terabyte data warehouses
  - Emergence of Web commerce

A Bit of History

- 2000s: 非SQL大规模数据库
  - Big data storage systems
    - Google BigTable, Yahoo PNuts, Amazon,
    - “NoSQL” systems.
  - Big data analysis: beyond SQL
    - MapReduce
- 2010s: SQL重出江湖
  - SQL reloaded
    - SQL front end to MapReduce systems
    - Massively parallel database systems
    - Multi-core main-memory databases

MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算。

概念"Map（映射）"和"Reduce（归约）"，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。



based on data stored in a tabular form

Relational Database

- Based on the relational model of data
  - Organizes data into one or more tables
  - Rows are also called records or tuples
  - Columns are also called attributes

In a theatre or cinema, or on a plane, each line of seats is called a row.
She was sitting in the front row.

a row of values in a relational database

tuple only has one meaning

An attribute is a quality or feature that someone or something has.

attributes

Columns (Attributes)

ID	*name*	*dept_name*	*salary*
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

Rows (Tuples)

(a) The *instructor* table

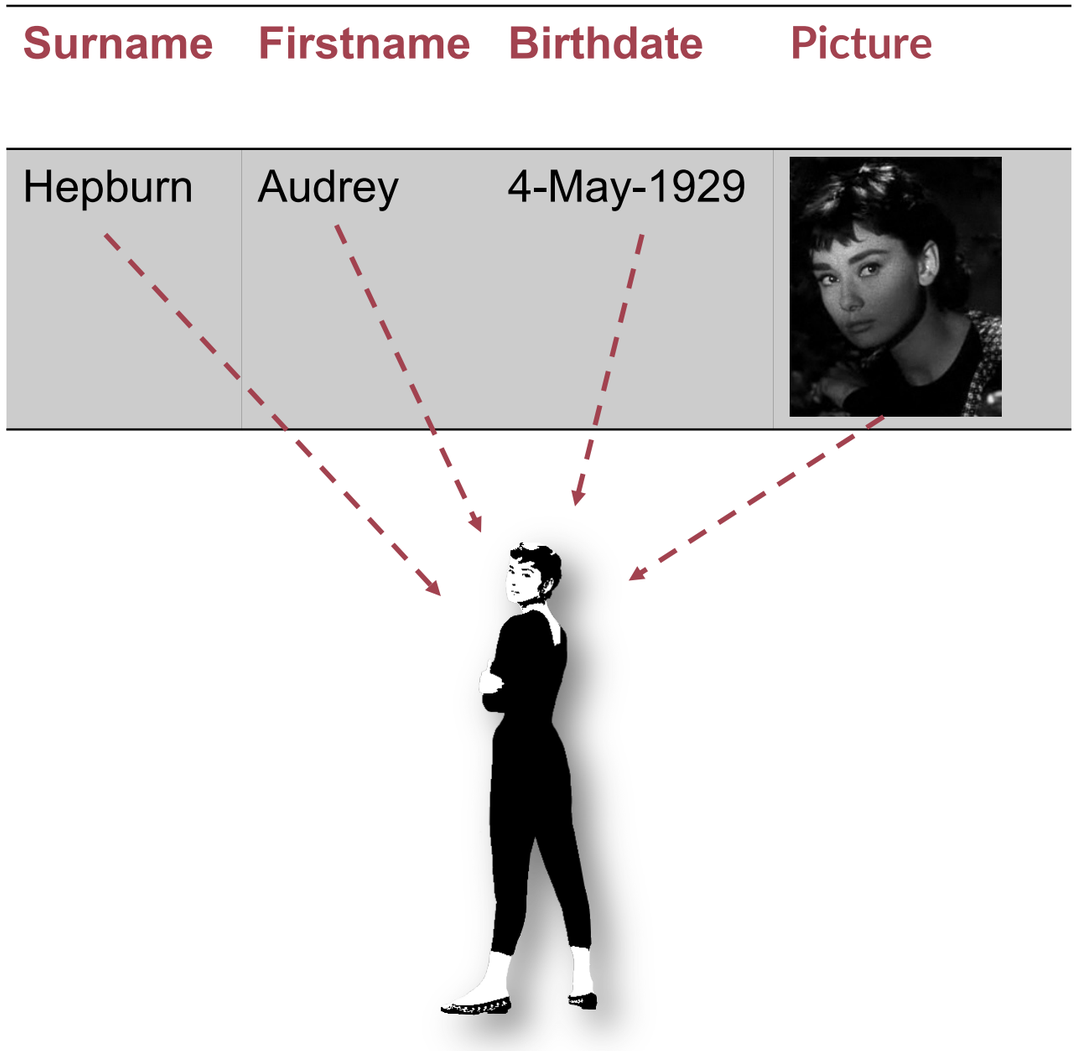
Relational Database

- Each column stores a piece of data
- One row represents a “known fact”:
  - “Audrey Hepburn was born on 1929/05/04 and looked like this.”

Surname	Firstname	Birthdate	Picture
Hepburn	Audrey	4-May-1929	

Relational Database

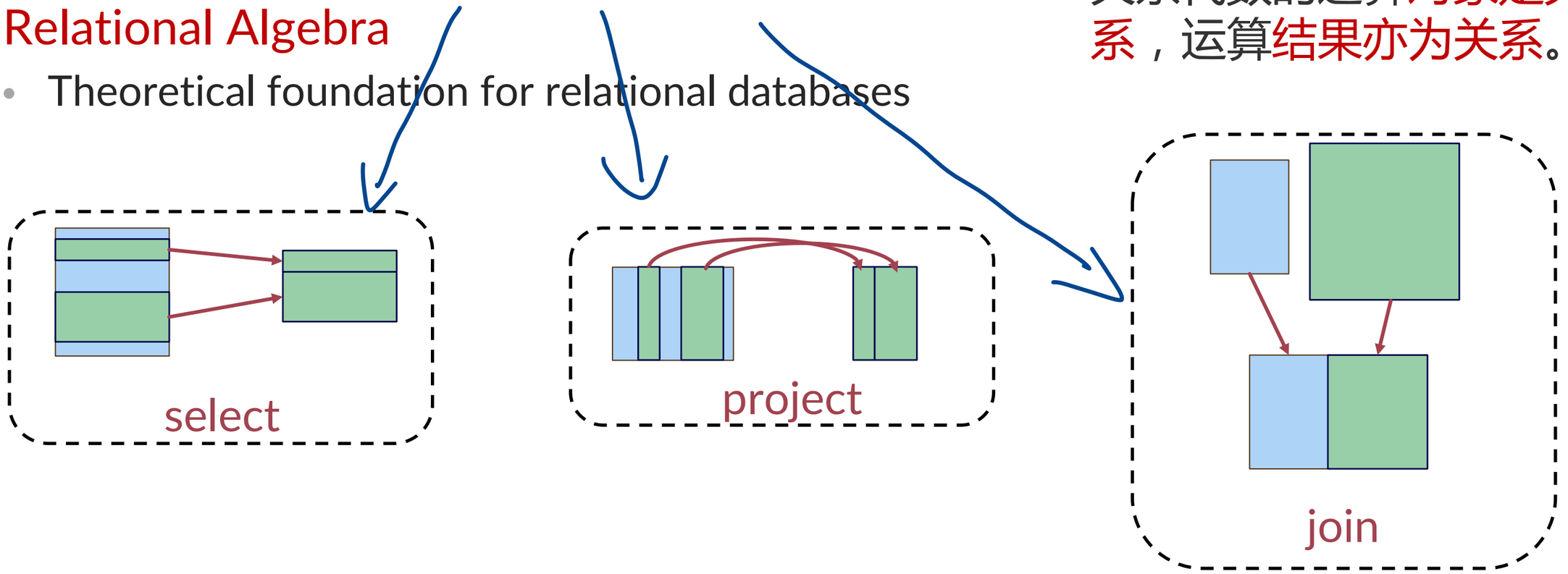
- Each column stores a piece of data
- One row represents a “known fact”:
  - “Audrey Hepburn was born on 1929/05/04 and looked like this.”
- All the pieces of data in a row are related, hence "**relational**".



Relational Database

- But Codd's "big idea"
  - You could operate on the relations and get new sets
- Relational Algebra
  - Theoretical foundation for relational databases

关系代数是一种抽象的查询语言，用对关系的运算来表达查询，作为研究关系数据语言的数学工具。关系代数的运算对象是关系，运算结果亦为关系。



Lecture 2 Pt 1 : key { primary key / unique

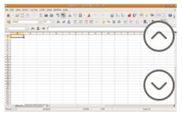
Key

- Example: A Film Database
  - Easy to find such as “The 100 greatest films ever”
  - Sometimes as a `.csv` file that you can load into a spreadsheet

	Movie Title	Country	Year	Director	Starring
2	Citizen Kane	US	1941	welles, o.	Orson Welles, joseph Cotten
3	La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
4	North by Northwest	US	1959	HITCHCOCK, A.	Cary GRANT, Eva Marie SAINT, James MASON
5	Singin' In the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
6	Rear Window	US	1954	HITCHCOCK, A.	James STEWART, Grace KELLY
7	City Lights	US	1931	CHAPLIN, C.	Charlie CHAPLIN, Virginia CHERRILL
8	The Third Man	GB	1949	Reed, C.	Joseph Cotten, Alida Valli, Orson Welles
9	The Searchers	US	1956	Ford, J.	John Wayne, Jeffrey Hunter, Natalie Wood
10	Ladri di biciclette	IT	1949	DeSica, V.	Lamberto Maggiorani, Enzo Staiola
11	Annie Hall	US	1977	Allen, W.	Woody Allen, Diane Keaton
12	On the Waterfront	US	1954	Kazan, E.	Marlon Brando, Eva Marie Saint, Karl Malden
13	All about Eve	US	1950	Mankiewicz, J.	Bette Davis, Anne Baxter, George Sanders
14	Casablanca	US	1942	Curtiz, M.	Humphrey Bogart,Ingrid Bergman, Claude Rains
15	The Treasure of the Sierra Madre	US	1948	HUSTON, J.	Humphrey BOGART, Walter HUSTON, Tim HOLT
16	High Noon	US	1952	Zinnemann, F.	Gary Cooper, Grace Kelly
17	Some Like It Hot	US	1959	Wilder, B.	Tony Curtis, Jack Lemmon, Marilyn Monroe
18					

电子试算表

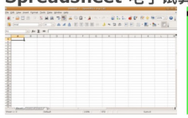
电子试算表（Spreadsheet），又称电子数据表，是一类模拟纸上计算表格的计算机进程。它会显示由一系列行与列构成的网格。单元格内可以存放数值、计算式、或文本。电子表格通常用于财务信息，因为它能够频繁的重新计算整个表格。

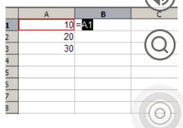


查看详情

英语百科参考

Spreadsheet 电子试算表





A spreadsheet is an interactive computer application for organization, analysis and storage of data in tabular form. Spreadsheets are developed as computerized simulations of paper accounting worksheets. The program operates on data entered in cells of an array, organized in rows and columns. Each cell of the array may contain either numeric or text data, or the results of formulas that automatically calculate and display a value based on the contents of other cells.

In SQL, keys are the set of attributes that used to identify the specific row in a table and to find or create the relation between two or more tables i.e keys identify the rows by combining one or more columns.

SQL provides super key, primary key, candidate key, alternate key, foreign key, compound key, composite key, and surrogate key.

SQL keys use constraints to uniquely identify rows from karger data.

Primary Key

Primary Key is a field that can be used to identify all the tuples uniquely in the database.

Only one of the columns can be declared as a primary key.

A Primary Key can not have a NULL value.

Foreign Key

A foreign key is a column which is known as Primary Key in the other table

i.e. A Primary Key in a table can be referred to as a Foreign Key in another table.

Foreign Key may have duplicate & NULL values if it is defined to accept NULL values.

Key

- Duplicates are forbidden in relational tables
  - Introduces potential errors such as when counting the number of movies

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason

Key

- How to identify “different rows” in a table?
  - A column (or a set of columns) to differentiate one row from another
    - In the film data … How about the movie titles?

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly

豆瓣电影 神雕侠侣

影讯&购票 选电影 电视剧 排行榜 分类 影评 2021年度榜单 2021书影音报告

搜索 神雕侠侣

神雕侠侣 神雕侠侣 (1995) [剧集] [可播放]

★★★★½ 9.2 (176564人评价)

中国香港 / 爱情 / 武侠 / 古装 / 新神雕侠侣 / Return Of The Condor Heroes / 45分钟

李添胜 / 古天乐 / 李若彤 / 白彪 / 魏秋桦 / 傅明宪 / 李绮虹 / 雪梨 / 简佩玛

话题 《神雕侠侣》哪个角色最让你惊喜?

10030人浏览 · 22篇文章

神雕侠侣 (2006) [剧集] [可播放]

★★★★☆ 7.5 (70883人评价)

中国大陆 / 武侠 / 古装 / The Return of the Condor Heroes / 40分钟

于敏 / 刘亦菲 / 黄晓明 / 陈紫函 / 杨幂 / 叮当 / 王洛勇 / 赵鸿飞 / 钱博

神雕侠侣 (2014) [剧集] [可播放]

★★½☆☆ 4.9 (26245人评价)

中国大陆 / 剧情 / 武侠 / 新神雕侠侣 / The Condor Heroes / 45分钟

李慧珠 / 邓伟恩 / 李达超 / 陈晓 / 陈妍希 / 张馨予 / 张雪迎 / 郑国霖 / 杨明娜 / 陈翔 / 毛晓彤

新神雕侠侣 (2022) [剧集]

☆☆☆☆☆ (尚未播出)

中国大陆 / 爱情 / 武侠 / 古装

林峰 / 佟梦实 / 毛晓慧 / 文淇 / 涂冰 / 邵兵 / 龚蓓苾 / 毛林林 / 宗峰岩

神雕侠侣 (2023)

☆☆☆☆☆ (尚未上映)

中国大陆 / 剧情 / 爱情 / 武侠 / 古装 / The Romance Of The Condor Heros

徐克

神雕侠侣 神雕侠侣 (1983) [剧集] [可播放]

★★★★☆ 7.9 (6328人评价)

中国香港 / 剧情 / 动作 / 爱情 / The Return of the Condor Heroes / 42分钟

范秀明 / 鞠觉亮 / 萧显辉 / 司徒立光 / 谭锐铭 / 刘德华 / 陈玉莲 / 梁家仁 / 欧阳佩珊 / 廖安丽 / 吕有慧 / 曾江

神雕侠侣 (1998) [剧集] [可播放]

★★★½☆ 7.1 (5767人评价)

新加坡 / 爱情 / 武侠 / 古装 / 神雕侠侣 98版 / Return of the Condor Heroes / 47分钟

马玉辉 / 谢敏洋 / 蔡晶盛 / 张龙敏 / 卢燕金 / 李铭顺 / 范文芳 / 朱厚任 / 何咏芳 / 林湘萍 / 丁岚 / 李南星

神雕侠侣 (1998) [剧集] [可播放]

★★½☆☆ 5.2 (10379人评价)

中国台湾 / 中国大陆 / 武侠 / 古装 / 杨过与小龙女 / 45分钟

李惠民 / 赖水清 / 任贤齐 / 吴倩莲 / 孙兴 / 季芹 / 李立群 / 夏文汐 / 蔡君茹 / 高捷

神雕侠侣 (2001) [剧集] [可播放]

★★★★☆ 7.5 (1772人评价)

日本 / 中国香港 / 动画 / 24分钟

案纳正美 / 高木淳 / 浪川大辅 / 园崎未惠 / 中田让治 / 唐泽润 / 木村亚希子 / 小村哲生 / 广田行生 / 高户靖广

神雕侠侣 (1984) [剧集]

★★★★☆ 7.5 (926人评价)

中国台湾 / 剧情 / 爱情 / 武侠 / 古装 / 95分钟

何东兴 / 孟飞 / 潘迎紫 / 傅娟

九一神雕侠侣 九一神鵰侠侶 (1991) [可播放]

★★★☆☆ 6.4 (8789人评价)

中国香港 / 动作 / 剧情 / 奇幻 / 爱情 / 科幻 / 神秘英豪 / 新神雕侠侣 / 92分钟

元奎 / 黎大炜 / 刘德华 / 梅艳芳 / 郭富城 / 叶蕴仪 / 刘嘉玲

神雕侠侣 神鵰侠侶 (1982) [可播放]

★★½☆☆ 5.2 (1402人评价)

中国香港 / 动作 / 爱情 / 武侠 / 古装 / 射雕英雄传4 / Brave Archer 4 / 100分钟

张彻 / 江生 / 郭追 / 傅声 / 龙天翔 / 黄淑仪

九二神雕之痴心情长剑 九二神鵰之痴心情長劍 (1992) [可播放]

★★★☆☆ 6.0 (6152人评价)

中国香港 / 喜剧 / 爱情 / 奇幻 / 武侠 / 神秘情侠 / 新神雕侠侣2(台) / 92分钟

元奎 / 黎大炜 / 刘德华 / 关之琳 / 吴耀汉 / 关淑怡

神雕侠侣 (1976) [剧集]

☆☆☆☆☆ (暂无评分)

中国香港 / 古装

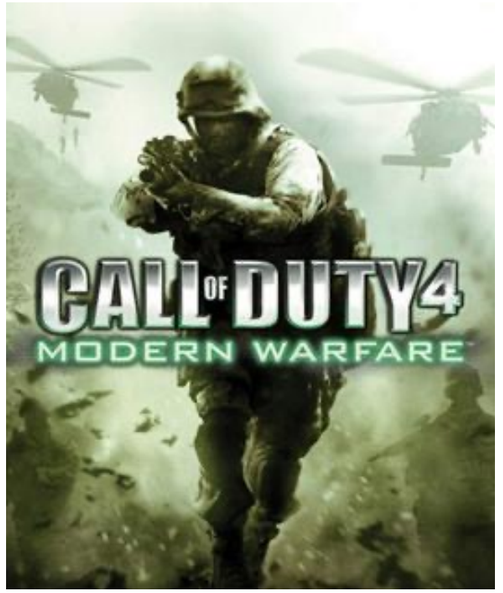
萧笙 / 罗乐林 / 李通明 / 白彪 / 米雪 / 曾江 Kenneth Tsang / 秦煌 / 郑裕玲 / 冯淬帆

Key

- How to identify “different rows” in a table?
  - A column (or a set of columns) to differentiate one row from another
    - In the film data … ~~How about the movie titles?~~ Title + Director?

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly

But, Hitchcock did It (So it was with the COD game)



Same developer (Infinite Ward)
Same publisher (Activision)
Two different years (2007, 2019)

Key

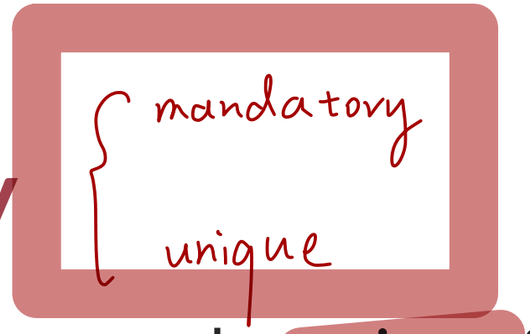
- How to identify “different rows” in a table?
  - A column (or a set of columns) to differentiate one row from another
    - In the film data … ~~How about the movie titles? Title + Director?~~
    - Title + Director + Year?

Movie Title	Country	Year	Director	Starring
Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly

Key

- Ok, you have made it this time, but –
  - The combination is too difficult for either remembering or computing
    - Need to compare multiple times on different columns
    - Think about deduplication (transitive) computing to remove (duplicated material) from a system
  - What if there are multiple items in a single column?
    - For example, more than one directors in a movie?

Primary Key



Additional material about creating a unique ID for each row:
`https://en.wikipedia.org/wiki/Universally_unique_identifier`

- Some of the keys may be unique for every row
  - Student ID, Email address, 18-digit ID number, etc.
- Usually, it is a good practice to choose the simplest one
  - (Or, create one)

Movie ID	Movie Title	Country	Year	Director	Starring
0	Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
1	La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
2	North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
3	Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
4	Rear Window	US	1954	HITCHCOCK, A.	James Stewart, Grace Kelly

Normalization

- A way of standardizing your data

Movie ID	Movie Title	Country	Year	Director	Starring
0	Citizen Kane	US	1941	welles, o.	Orson Welles, Joseph Cotten
1	La règle du jeu	FR	1939	Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
2	North By Northwest	US	1959	HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
3	Singin' in the Rain	US	1952	Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
4	Rear Window	US	1954	Alfred Hitchcock	James Stewart, Grace Kelly

Too many different ways of spelling

Normalization

Something that is normal is usual and ordinary, and is what people expect.

- "First Norm Rule" (1NF)
  - Each column should only contain ONE piece of information

First Normal Form (1NF):

If a relation contains a composite or multi-valued attribute, it violates the first normal form, or the relation is in first normal form if it does not contain any **composite** or **multi-valued attribute**. A relation is in first normal form if every attribute in that relation is singled valued attribute.

A table is in 1 NF iff:

1. There are only Single Valued Attributes.
2. Attribute Domain does not change.
3. There is a unique name for every Attribute/Column.
4. The order in which data is stored does not matter.

attribute ⇒ column

Director	Starring
welles, o.	Orson Welles, Joseph Cotten
Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
HITCHCOCK, A.	James Stewart, Grace Kelly

Director_Firstname	Director_Lastname
Alfred	Hitchcock
Orson	Welles

Starring_Firstname	Starring_Lastname
Orson	Welles
Joseph	Cotten

Normalization

- “First Norm Rule” (1NF)
  - Each column should only contain ONE piece of information

Director	Starring
welles, o.	Orson Welles, Joseph Cotten
Renoir, J.	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
HITCHCOCK, A.	Cary Grant, Eva Marie Saint, James Mason
Donen/Kelly	Gene Kelly, Debbie Reynolds, Donald O'Connor
HITCHCOCK, A.	James Stewart, Grace Kelly

Director_ Firstname	Director_ Lastname	Born	Died
Alfred	Hitchcock	1899	1980
Orson	Welles	1915	1985

Extend the table to represent all directors

use another table to represent different directors

Normalization

- “First Norm Rule” (1NF)
  - Each column should only contain ONE piece of information

Movie ID	Movie Title	Country	Year	Director ID	Starring
0	Citizen Kane	US	1941	2	Orson Welles, Joseph Cotten
1	La règle du jeu	FR	1939	5	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
2	North By Northwest	US	1959	1	Cary Grant, Eva Marie Saint, James Mason
3	Singin' in the Rain	US	1952	6	Gene Kelly, Debbie Reynolds, Donald O'Connor
4	Rear Window	US	1954	1	James Stewart, Grace Kelly

Director ID	Director_ Firstname	Director_L astname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	….	…	…	…

Link the director information between tables

Normalization

- Normal Form (NF)
  - 1NF: Simple attributes
  - 2NF: Attributes depend on the full key
  - 3NF: Non-key attributes do not depend on each other
  - And many others

	UNF (1970)	1NF (1970)	2NF (1971)	3NF (1971)	EKNF (1982)	BCNF (1974)	4NF (1977)	ETNF (2012)	5NF (1979)	DKNF (1981)	6NF (2003)
Primary key (no duplicate tuples)[4]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Atomic columns (cells cannot have tables as values)[5]	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Every non-trivial functional dependency either does not begin with a proper subset of a candidate key or ends with a prime attribute (no partial functional dependencies of non-prime attributes on candidate keys)[5]	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
Every non-trivial functional dependency either begins with a superkey or ends with a prime attribute (no transitive functional dependencies of non-prime attributes on candidate keys)[5]	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
Every non-trivial functional dependency either begins with a superkey or ends with an elementary prime attribute	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	N/A
Every non-trivial functional dependency begins with a superkey	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	N/A
Every non-trivial multivalued dependency begins with a superkey	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	N/A
Every join dependency has a superkey component[8]	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	N/A
Every join dependency has only superkey components	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	N/A
Every constraint is a consequence of domain constraints and key constraints	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Every join dependency is trivial	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

https://en.wikipedia.org/wiki/Database_normalization

Normalization

Every non key attribute must provide a fact about the key, the whole key, and nothing but the key.

William Kent (1936 – 2005)

William Kent. "A Simple Guide to Five Normal Forms in Relational Database Theory", Communications of the ACM 26 (2), Feb. 1983, pp. 120–125.

Normalization

Question:
What if there are multiple director?

- "First Norm Rule" (1NF)
  - Each column should only contain ONE piece of information

Movie ID	Movie Title	Country	Year	Director ID	Starring
0	Citizen Kane	US	1941	2	Orson Welles, Joseph Cotten
1	La règle du jeu	FR	1939	5	Roland Toutain, Nora Grégor, Marcel Dalio, Jean Renoir
2	North By Northwest	US	1959	1	Cary Grant, Eva Marie Saint, James Mason
3	Singin' in the Rain	US	1952	6	Gene Kelly, Debbie Reynolds, Donald O'Connor
4	Rear Window	US	1954	1	James Stewart, Grace Kelly

Director ID	Director_ Firstname	Director_Lastname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	....	...	...	...

Link the director information between tables

Normalization is a database design technique, which is used to design a relational database table up to higher normal form.[9] The process is progressive, and a higher level of database normalization cannot be achieved unless the previous levels have been satisfied.[10]

That means that, having data in unnormalized form (the least normalized) and aiming to achieve the highest level of normalization, the first step would be to ensure compliance to first normal form, the second step would be to ensure second normal form is satisfied, and so forth in order mentioned above, until the data conform to sixth normal form.

However, it is worth noting that normal forms beyond 4NF are mainly of academic interest, as the problems they exist to solve rarely appear in practice.[11]

The data in the following example were intentionally designed to contradict most of the normal forms. In real life, it is quite possible to be able to skip some of the normalization steps because the table doesn't contain anything contradicting the given normal form. It also commonly occurs that fixing a violation of one normal form also fixes a violation of a higher normal form in the process. Also one table has been chosen for normalization at each step, meaning that at the end of this example process, there might still be some tables not satisfying the highest normal form.

Entity and Relationship

- A bad idea: Add more columns in the Movie table

Movie ID	Movie Title	Country	Year	Director ID	Director 2 ID	Director 3 ID	Starring

  - Waste of space (not too many movies have 3 directors, let alone 6)

Entity and Relationship

- A bad idea: Add more columns in the Movie table

Movie ID	Movie Title	Country	Year	Director ID	Director 2 ID	Director 3 ID	Starring

  - Waste of space (not too many movies have 3 directors, let alone 6)
  - How about starring? 10+ more columns?

An entity is something that exists separately from other things and has a clear identity of its own.

Entity and Relationship

- Further refactoring of the tables …

Movie ID	Movie Title	Country	Year
0	Citizen Kane	US	1941
1	La règle du jeu	FR	1939
2	North By Northwest	US	1959
3	Singin' in the Rain	US	1952
4	Rear Window	US	1954

Movie Entities

Relationship?

Director ID	Director_ Firstname	Director_ Lastname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	….	…	…	…

Director Entities

Entity and Relationship

- Further refactoring of the tables …

Movie ID	Movie Title	Country	Year
0	Citizen Kane	US	1941
1	La règle du jeu	FR	1939
2	North By Northwest	US	1959
3	Singin' in the Rain	US	1952
4	Rear Window	US	1954

Movie Entities

Directed By

Movie ID	Director ID
0	2
1	5
2	1

Relationship!

Director ID	Director_ Firstname	Director_ Lastname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	….	…	…	…

Director Entities

Entity and Relationship

- Further refactoring of the tables … 把“列”转成“行”（矩阵相乘？）

Movie ID	Movie Title	Country	Year
0	Citizen Kane	US	1941
1	La règle du jeu	FR	1939
2	North By Northwest	US	1959
3	Singin' in the Rain	US	1952
4	Rear Window	US	1954

Movie Entities

Directed By

Movie ID	Director ID
0	2
1	5
2	1
…	…
16	8
16	9
16	10

Director ID	Director_ Firstname	Director_ Lastname	Born	Died
1	Alfred	Hitchcock	1899	1980
2	Orson	Welles	1915	1985
3	….	…	…	…

Director Entities

Question:
What if there are multiple directors?

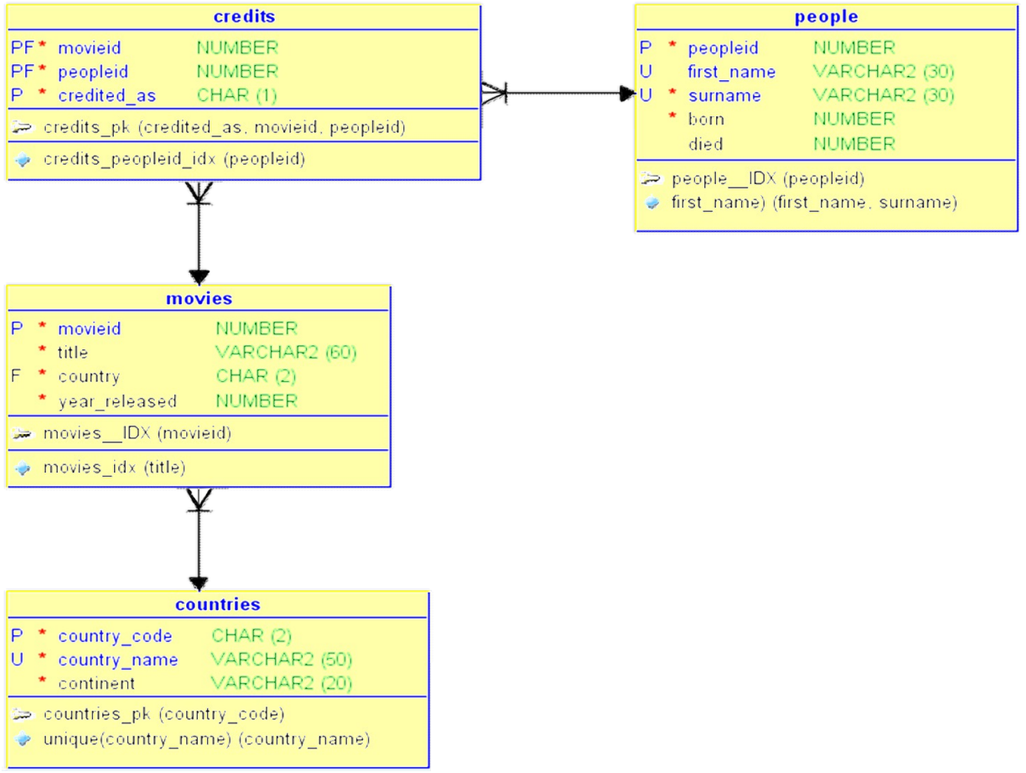
Answer:
Multiple rows in the relationship!

Entity and Relationship

- Starring -> Actor table
- Country -> Country and Region table
  - You can also link the movies with corresponding actors, countries/regions, etc.

- Entity Relationship Diagram (E/R Diagram, ER Diagram, ERD)
  - A way of representing entity tables and their relationships (relationship tables)

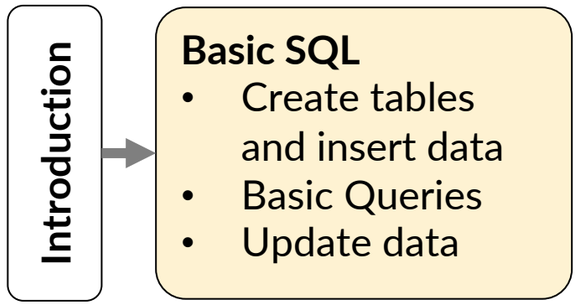
独立存在体的表格



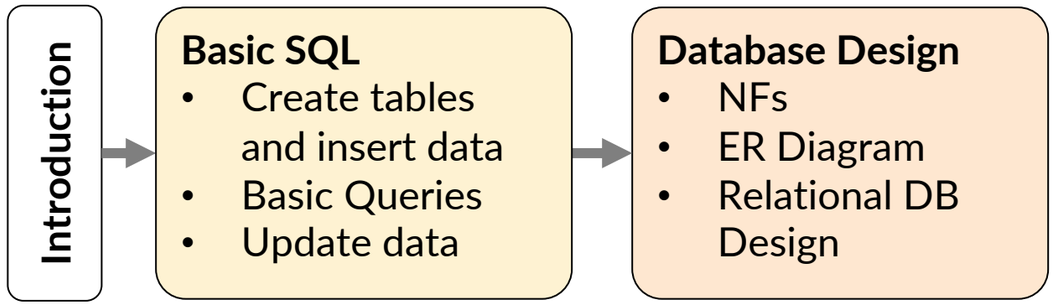
Outline

Introduction

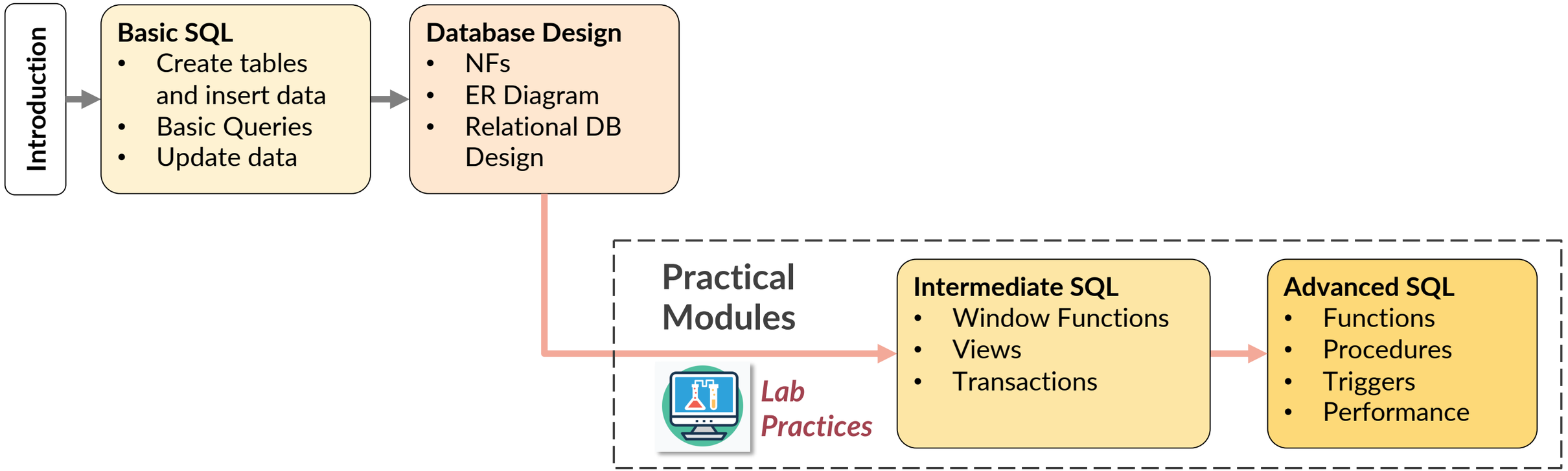
Outline


Introduction
Basic SQL
• Create tables and insert data
• Basic Queries
• Update data

Outline


Introduction
Basic SQL
• Create tables and insert data
• Basic Queries
• Update data
Database Design
• NFs
• ER Diagram
• Relational DB Design

Outline


Introduction
Basic SQL
• Create tables and insert data
• Basic Queries
• Update data
Database Design
• NFs
• ER Diagram
• Relational DB Design
Practical Modules
Lab Practices
Intermediate SQL
• Window Functions
• Views
• Transactions
Advanced SQL
• Functions
• Procedures
• Triggers
• Performance

Outline

Introduction

Basic SQL
- Create tables and insert data
- Basic Queries
- Update data

Database Design
- NFs
- ER Diagram
- Relational DB Design

Theoretical Modules

Big Data Analytics
- Big Data Storage, Querying, and Processing
- OLAP, Data Mining, Visualization

Storage Management
- Storage Structure
- Indexing

Application Design
- Access databases in programming
- Complex Data Types (JSON, XML)

Relational Algebra

Practical Modules

Lab Practices

Intermediate SQL
- Window Functions
- Views
- Transactions

Advanced SQL
- Functions
- Procedures
- Triggers
- Performance

Outline

- What we may NOT (fully) cover
  - Programming Language Support
    - We will focus on one or two languages
  - Parallel and Distributed Databases 分布式数据库
  - Object-based Databases 面向对象
  - Blockchain 区块链
  - Advanced Relational Algebra and Calculus 高级关系代数与微积分
  - Advanced Data Mining and Analytics 高级数据挖掘与分析
  - …（以及一些其他你也许听过的“新技术”）

Data Definition and Manipulation

- Data Definition Language (DDL)
  - DDL compiler generates a set of table templates stored in a data dictionary
    - Database schema
    - Integrity constraints (primary key, etc.)
    - Authorization (who can access it)

A schema is an outline of a plan or theory.

```
create table lab(
    id serial primary key,
    address varchar(20) not null,
    time varchar(20) not null,
    capacity int,
    teacher varchar(20),
    unique (address,time)
);
```

If you have integrity, you are honest and firm in your moral principles.

A constraint is something that limits or controls what you can do.

The integrity of something such as a group of people or a text is its state of being a united whole.

Data Definition and Manipulation

- Data Manipulation Language (DML)
  - **Language** for accessing and updating the data organized by the appropriate data model (also known as query language)
- SQL query language
  - Takes several tables as input (possibly only one) and always returns a single table

```
select * from lab;

select * from lab where time = '3-34';
```

A query is a question, especially one that you ask an organization, publication, or expert.

2. **VERB**

If you say that someone manipulates an event or situation, you disapprove of them because they use or control it for their own benefit, or cause it to develop in the way they want.

[*disapproval*]

She was unable, for once, to control and manipulate events. [*VERB noun*]

They felt he had been cowardly in manipulating the system to avoid the draft. [*VERB noun*]

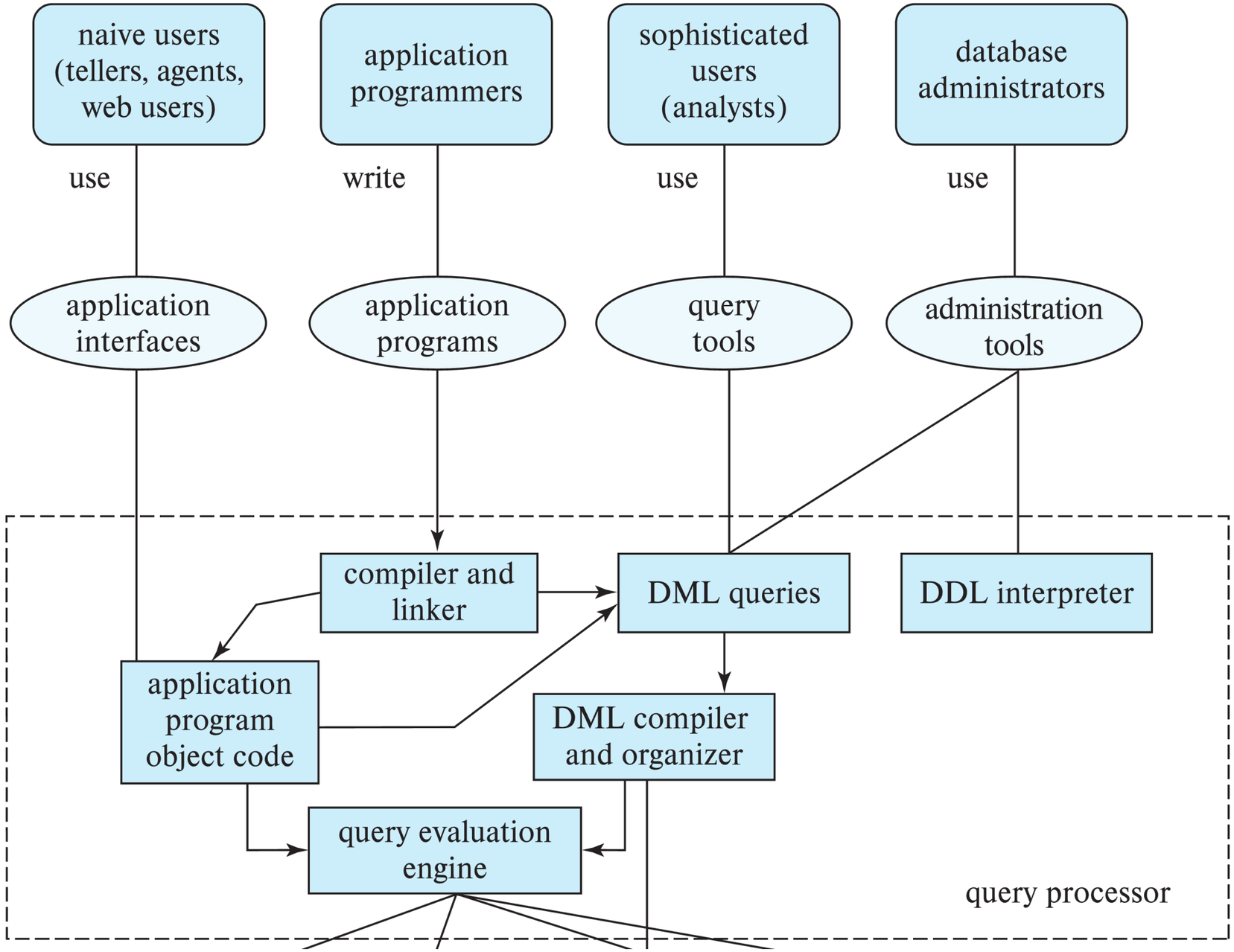
manipulation **VARIABLE NOUN**

...accusations of political manipulation.

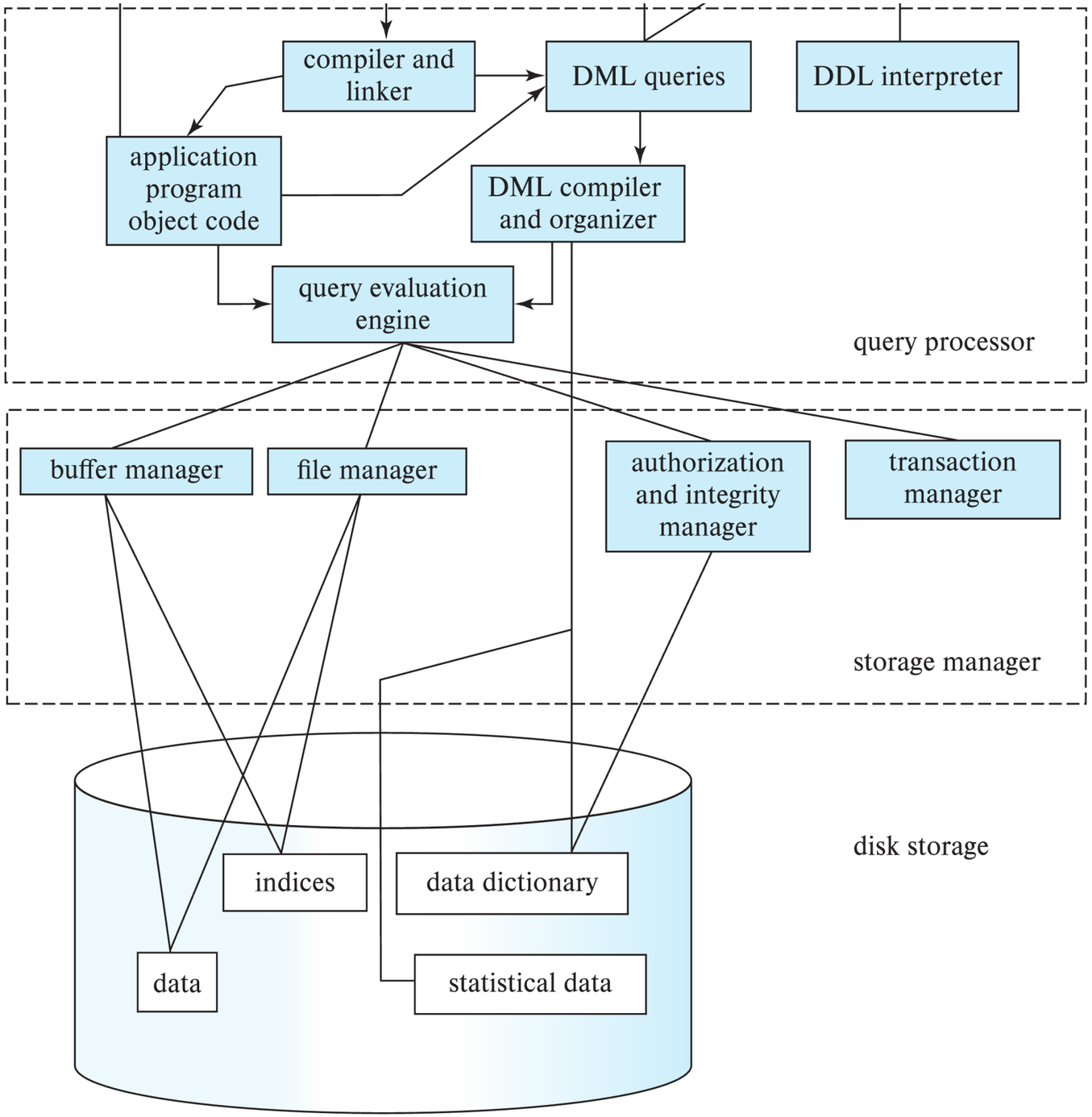
3. **VERB**

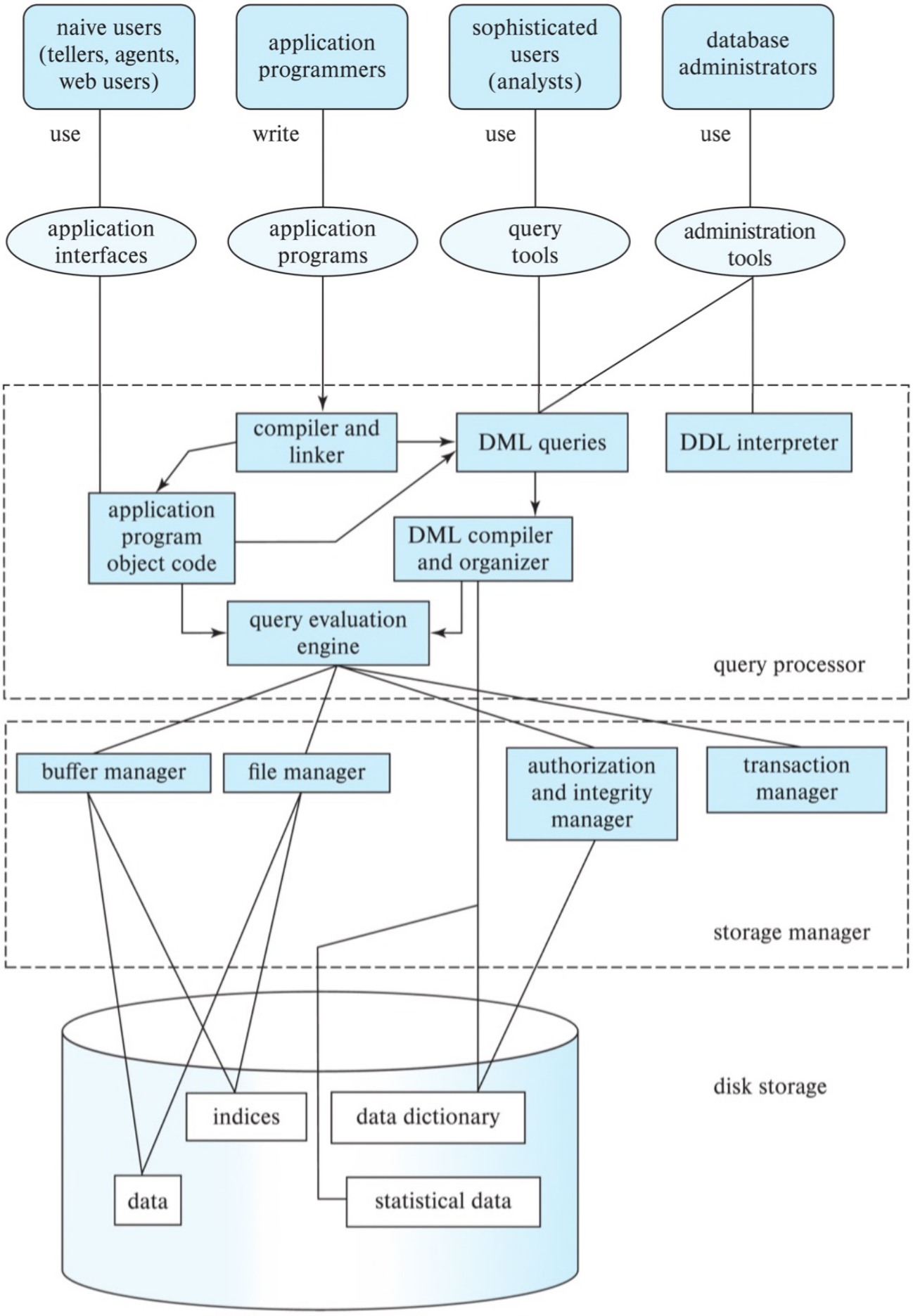
If you manipulate something that requires skill, such as a complicated piece of equipment or a difficult idea, you operate it or process it.

Database Users



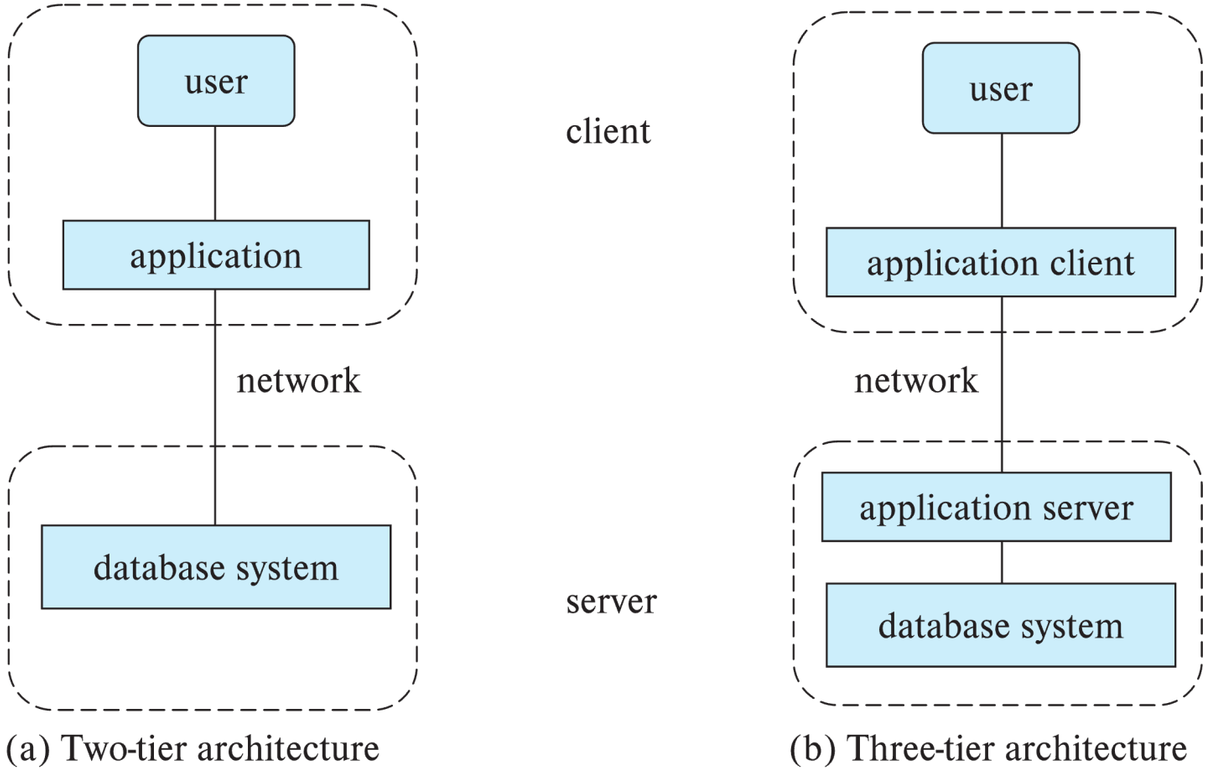
Database Architecture




naive users (tellers, agents, web users)
application programmers
sophisticated users (analysts)
database administrators
use
write
use
use
application interfaces
application programs
query tools
administration tools
compiler and linker
DML queries
DDL interpreter
application program object code
DML compiler and organizer
query evaluation engine
query processor
buffer manager
file manager
authorization and integrity manager
transaction manager
storage manager
indices
data dictionary
data
statistical data
disk storage

Database Applications

- Database applications are usually partitioned into two or three parts
- Application programs generally access databases through one of
  - Language extensions to allow embedded SQL
  - Application Program Interface (e.g., ODBC/JDBC) which allow SQL queries to be sent to a database



(a) Two-tier architecture

(b) Three-tier architecture

Thank you

- QQ Group: 949474976