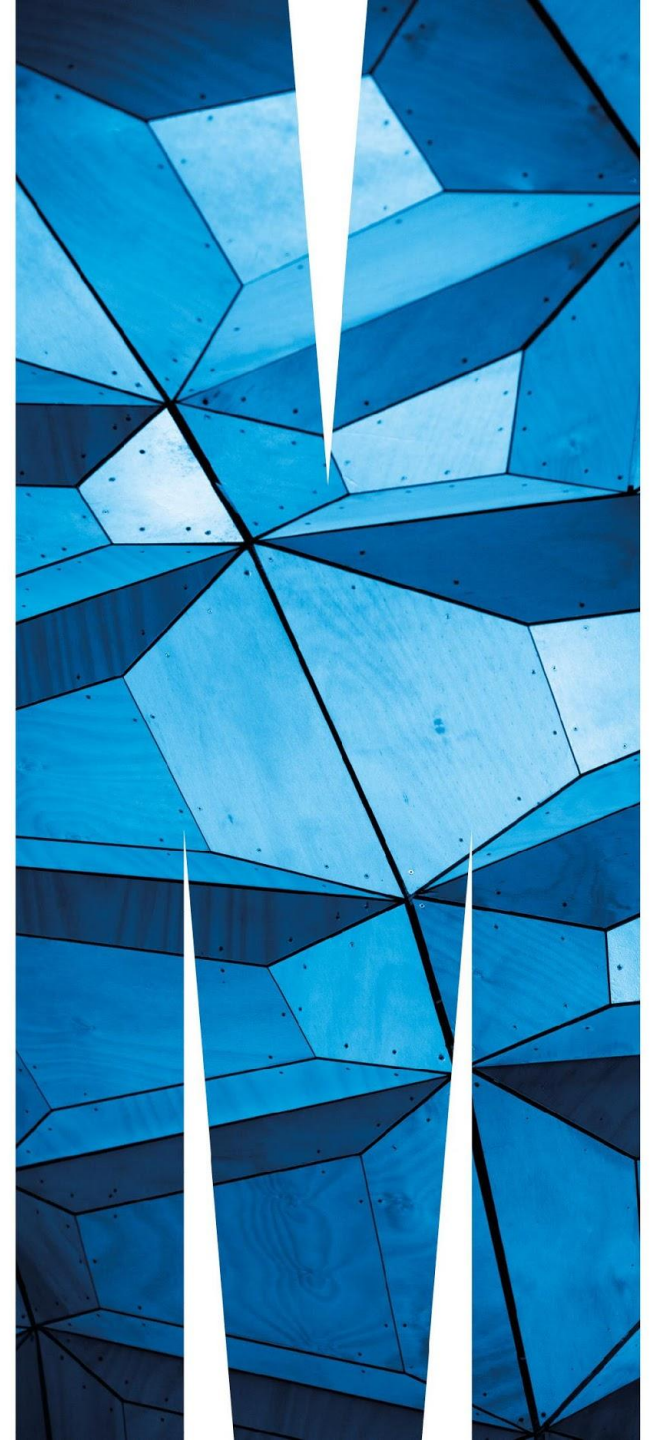


Week 2 – Relational Data Model

FIT2094 - FIT3171 Databases

Clayton Campus S1 2019.



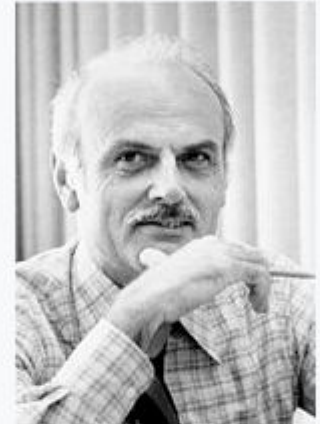
Overview

- Relational Model
- Relational Algebra

The Relational Model

- Introduced by Codd in 1970 - the fundamental basis for relational DBMS's
- Basic structure is the mathematical concept of a RELATION mapped to the 'concept' of a table (tabular representation of relation)
 - Relation - abstract object
 - Table - pictorial representation
 - Storage structure - "real thing"
 - e.g. ISAM file on disk.
 - “Indexed sequential access method” (ISAM) by IBM
 - “hierarchy of indexes” (Smith, 2006)
https://books.google.com.au/books?id=HJ9gds_zhVEC&pg=PA517&lpg=PA517&dq=isam+file+on+disk

Edgar "Ted" Codd



Born	Edgar Frank Codd 19 August 1923 ^{[1][2]} Fortuneswell, Dorset, England
Died	18 April 2003 (aged 79) Williams Island, Aventura, Florida, USA
Alma mater	Exeter College, Oxford University of Michigan
Known for	OLAP Relational model ^[3] Codd's cellular automaton Codd's 12 rules Boyce–Codd normal form
Awards	Turing Award (1981) ^[4]

Img src: Wikipedia.

The Relational Model

- Relational Model Terminology
 - DOMAIN - set of atomic (indivisible) values
 - ...specifies
 - **name**
 - data **type**
 - data **format**
- Examples:
 - **customer_number** domain - 5 character string of the form xxxdd
 - **name** domain - 20 character string
 - **address** domain - 30 character string containing street, town & postcode
 - **credit_limit** domain - money in the range \$1,000 to \$99,999

No Flux this slide.

Q: You are hired by Twitter Inc to create a new “Politics” version of Twitter. It’s identical to Twitter, with the only difference that it can only be used by politicians :-)

This web app requires an RDBMS to store data in the backend.

**You need to determine the DOMAIN for “message content”.
Chat with your neighbour.**

Remember: name, type, format.

TL;DR: “A domain describes the set of possible values for a given attribute, and can be considered a constraint on the value of the attribute”

(https://en.wikipedia.org/wiki/Relational_database#Domain)

A Relation

- A relation consists of two parts: **heading** and **body**
- Relation Heading
 - Also called **Relational Schema**.
 - Formally: it consists of a fixed set of attributes
 - **$R(A_1, A_2, \dots, A_n)$**
 - R = relation name, A_i = attribute i
 - Each attribute corresponds to one underlying domain:
 - Customer relation heading:
CUSTOMER (custno, custname, custadd, credlimit)
 - $\text{dom}(\text{custno}) = \text{customer_number}$
 - $\text{dom}(\text{custname}) = \text{name}$
 - $\text{dom}(\text{custadd}) = \text{address}$
 - $\text{dom}(\text{credlimit}) = \text{credit_limit}$

custno	custname	custadd	credlimit
--------	----------	---------	-----------

No Flux this slide.

Q: You are hired by Twitter Inc to create a new “Politics” version of Twitter. It’s identical to Twitter, with the only difference that it can only be used by politicians :-)

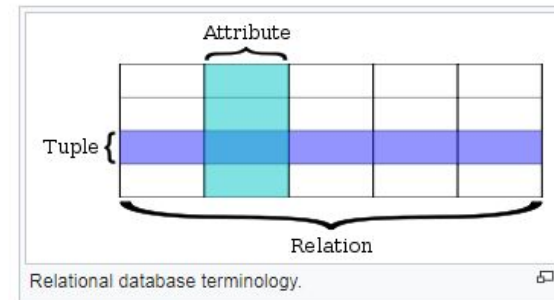
This web app requires an RDBMS to store data in the backend.

You need to determine some other ATTRIBUTES for your Twitter clone.

R(message_content, ...what else...)

Relation Body

- Relation Body
- Also called Relation Instance (state)
- Formally: $r(R) = \{t_1, t_2, t_3, \dots, t_m\}$
 - consists of a time-varying set of **n-tuples**
(we just call them tuples below for simplicity)
 - Relation R consists of tuples $t_1, t_2, t_3 \dots t_m$
 - m = number of tuples = **relation cardinality**
 - each **n-tuple** is an **ordered list** of n values
 - $t = \langle v_1, v_2, \dots, v_n \rangle$
 - n = number of values in tuple (no of attributes) = **relation degree**



Relation Body

- In the tabular representation:
 - Relation heading ⇨ column headings
 - Relation body ⇨ set of data rows

custno	custname	custadd	credlimit
SMI13	SMITH	Wide Rd, Clayton, 3168	2000
JON44	JONES	Narrow St, Clayton, 3168	10000
BRO23	BROWN	Here Rd, Clayton, 3168	10000

SQL term	Relational database term	Description
<i>Row</i>	<i>Tuple</i> or <i>record</i>	A data set representing a single item
<i>Column</i>	<i>Attribute</i> or <i>field</i>	A labeled element of a tuple, e.g. "Address" or "Date of birth"
<i>Table</i>	<i>Relation</i> or <i>Base relvar</i>	A set of tuples sharing the same attributes; a set of columns and rows

These slides with the blue background are Clayton FLUX slides!

[Q1] Clayton students: Trick question...

As discussed, if we have an relation body, consisting of a set of 'n-tuples', where m = number of tuples.

What is m , and what is n , in this example?

Refer to our **very simple** Twitter example below.

- A. $m=3, n=3$ C. $m=2, n=3$ E. $m=1, n=3$
B. $m=2, n=2$ D. $m=3, n=2$

message_content	user_name	date
Brace yourself... Elections are coming!	NedStark	05/03/2019
I am not a crook.	RealRichardNixon	17/11/1973

Relation Properties

- No duplicate tuples
 - **by definition** sets do not contain duplicate elements
 - hence tuples are unique
- Tuples are unordered within a relation
 - **by definition** sets are not ordered
 - hence tuples can only be accessed by content
- No ordering of attributes within a tuple
 - **by definition** sets are not ordered

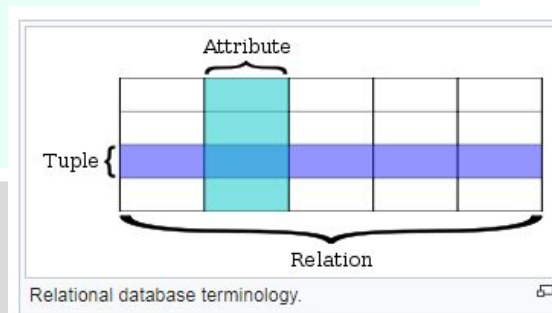
Relation Properties cont'd

- Tuple values are atomic - cannot be divided
 - **EMPLOYEE(eid, ename, departno, dependants)**
 - not allowed: dependants (depname, depage) multivalued
 - hence no multivalued (repeating) attributes allowed, called the **first normal form rule**
- COMPARE with tabular representation
 - say, an Excel shopping list?
 - normally nothing to prevent duplicate rows
 - **rows are ordered**
 - **columns are ordered**
 - tables and relations are not the same 'thing'

These slides with the blue background are Clayton FLUX slides!

Q2. Which of the following statements is TRUE according the characteristics of relational table?

- A. All values in a column need to be from the same domain.
- B. Each column needs to have a distinct name.
- C. The order of attributes (column) and tuples (row) matters.
- D. Each intersection of a column and a row represent a single value.
- E. More than one statement is TRUE.
- F. All (A-E above) FALSE - trick question.



surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Green	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Green	Maria	BITS	31-08-1995

▪ Functional Dependency:

- A set of attributes X functionally determines an attribute Y
... **if, and only if (IIF)** for each X value, there is exactly one Y value in the relation.
- It is denoted as $X \rightarrow Y$.

▪ For example, given the data above:

- firstname, surname \rightarrow degree (hint: no dupl. full names)
 - *but*
- firstname \rightarrow degree does not hold (hint: BBIS, BCS)
- What about: degree \rightarrow firstname, surname? (hint: BITS)

These slides with the blue background are Clayton FLUX slides!

Q3. Which of the following statement is TRUE when the concept of **functional dependency** is applied to the data shown here?

Assume this data represents the entire table and is not going to change in any way.

- A. surname \rightarrow firstname.
- B. surname, firstname \rightarrow degree
- C. DOB \rightarrow surname
- D. degree \rightarrow surname
- E. firstname, degree \rightarrow DOB
- F. surname, degree \rightarrow DOB
- G. options B, C, and E are correct
- H. options B, C, E, and F are correct
- I. (Help, this Q too difficult!)

surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Green	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Green	Maria	BITS	31-08-1995

Relational Keys

- A **superkey** K is an attribute or set of attributes which only exhibits the uniqueness property
 - No two tuples of R have the same value for K
(Uniqueness property)
- A **candidate key** K of a relation R is an attribute or set of attributes which exhibits the following properties:
 - Uniqueness property (as above), *and*
 - No proper subset of K has the uniqueness property
(Minimality or Irreducibility property)
- One candidate key is chosen to be the **primary key** of a relation. Remaining candidate keys are termed alternate keys.

These slides
with the blue
background
are Clayton
FLUX slides!

surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Green	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Green	Maria	BITS	31-08-1995

Q4. Superkey: A set of attributes X that uniquely identifies each row in a relation R. Which of the following is **NOT** a superkey based on the above data?

- A. DOB
- B. DOB, degree
- C. surname, firstname
- D. surname, firstname, DOB
- E. surname, degree
- F. All of the above (A,B,C,D,E) are superkeys
- G. None of the above (A,B,C,D,E) is a superkey

These slides
with the blue
background
are Clayton
FLUX slides!

surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Green	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Green	Maria	BITS	31-08-1995

Q5. Candidate Key: A **minimal** set of attributes X that uniquely identifies each row in a relation R. i.e. $X \rightarrow \{R\}$ **AND** for any subset Y of X, $Y \rightarrow \{R\}$ does not hold.

Which of the following is a candidate key based on the above data?

- A. DOB
- B. DOB, degree
- C. surname, firstname
- D. surname, firstname, DOB
- E. Options A and C are candidate keys
- F. Options A, B, and C are candidate keys
- G. All of the above (A, B, C, D) are candidate keys

These slides
with the blue
background
are Clayton
FLUX slides!

surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Green	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Green	Maria	BITS	31-08-1995

Q6. Candidate Key: A **minimal** set of attributes X that uniquely identifies each row in a relation R. i.e. $X \rightarrow \{R\}$ **AND** for any subset Y of X, $Y \rightarrow \{R\}$ does not hold.

How many candidate keys are there in the table based on **the above data**?

- A. 0
- B. 1
- C. 2
- D. 3
- E. 4

These slides
with the blue
background
are Clayton
FLUX slides!

surname	firstname	degree	DOB
Black	Sam	BBIS	02-02-1996
Brown	Jane	BITS	01-01-1995
Chen	Chan	BITS	09-02-1996
Grey	Maria	BCS	15-12-1995
Indigo	Jose	BITS	28-10-1995
Black	Jet	BCS	13-05-1996
Chen	Maria	BITS	31-08-1995

Q7. How many primary keys are there in the table based on the the above data?

- a. 0
- b. 1
- c. 2
- d. 3
- e. 4

Selection of a Primary key

- A primary key must be chosen considering the data that *may be added to the table in the future*
 - Names, dates of birth etc are rarely unique and as such are not a good option
 - PK should be free of 'extra' semantic meaning, preferably single attribute, preferably numeric (see Table 5.3 Coronel & Morris, PTO)
 - Natural vs Surrogate - natural PK is part of the data, surrogate is “generated and then stored with the rest of the columns in a record... no meaning associated with the value” (Larsen, 2011)
 - <https://www.databasejournal.com/features/mssql/article.php/3922066/SQL-Server-Natural-Key-Verses-Surrogate-Key.htm>

stu_no	surname	firstname	degree	DOB
1111	Black	Sam	BBIS	02-02-1996
1112	Brown	Jane	BITS	01-01-1995
1113	Chen	Chan	BITS	09-02-1996
1114	Grey	Maria	BCS	15-12-1995
1115	Indigo	Jose	BITS	28-10-1995
1116	Black	Jet	BCS	13-05-1996
1117	Chen	Maria	BBIS	31-08-1995

TABLE 5.3**DESIRABLE PRIMARY KEY CHARACTERISTICS**

PK CHARACTERISTIC	RATIONALE
Unique values	The PK must uniquely identify each entity instance. A primary key must be able to guarantee unique values. It cannot contain nulls.
Nonintelligent	The PK should not have embedded semantic meaning other than to uniquely identify each entity instance. An attribute with embedded semantic meaning is probably better used as a descriptive characteristic of the entity than as an identifier. For example, a student ID of 650973 would be preferred over Smith, Martha L. as a primary key identifier.
No change over time	If an attribute has semantic meaning, it might be subject to updates, which is why names do not make good primary keys. If Vickie Smith is the primary key, what happens if she changes her name when she gets married? If a primary key is subject to change, the foreign key values must be updated, thus adding to the database work load. Furthermore, changing a primary key value means that you are basically changing the identity of an entity. In short, the PK should be permanent and unchangeable.
Preferably single-attribute	A primary key should have the minimum number of attributes possible (irreducible). Single-attribute primary keys are desirable but not required. Single-attribute primary keys simplify the implementation of foreign keys. Having multiple-attribute primary keys can cause primary keys of related entities to grow through the possible addition of many attributes, thus adding to the database workload and making (application) coding more cumbersome.
Preferably numeric	Unique values can be better managed when they are numeric, because the database can use internal routines to implement a counter-style attribute that automatically increments values with the addition of each new row. In fact, most database systems include the ability to use special constructs, such as Autonumber in Microsoft Access, sequence in Oracle, or uniqueidentifier in MS SQL Server to support self-incrementing primary key attributes.
Security-compliant	The selected primary key must not be composed of any attribute(s) that might be considered a security risk or violation. For example, using a Social Security number as a PK in an EMPLOYEE table is not a good idea.

Writing Relations

- Relations may be represented using the following notation:
 - **relation_name(attribute1, attribute2,...)**
- The primary key is underlined.
- Example:
 - **staff(staffid, surname, initials, address, phone)**

Relational Database

- A relational database is a **collection of normalised relations**.
- **Normalisation** is part of the design phase of the database and will be discussed in a later lecture.

Example relational database:

order (order_id, orderdate,)

order-line (order_id, product_id, quantity)

product (product_id, description, unit_price)

Foreign Key (FK)

- An attribute/s in a table that exists in the same, or another table as a Primary Key.
- **Referential Integrity**
 - **A Foreign Key value must either match the *primary key* in another table or be NULL.**
- The pairing of PK and FK creates relationships (logical connections) between tables. Hence the abstraction away from the underlying storage model.

These slides with the blue background are Clayton FLUX slides!

MANAGER

	PROJECT_MANAGER	MANAGER_PHONE	MANAGER_ADDRESS
▶	Holly B. Parker	904-338-3416	3334 Lee Rd., Gainesville, FL 37123
	Jane D. Grant	615-898-9909	218 Clark Blvd., Nashville, TN 36362
	George F. Dorts	615-227-1245	124 River Dr., Franklin, TN 29185
	William K. Moor	904-445-2719	216 Morton Rd., Stetson, FL 30155

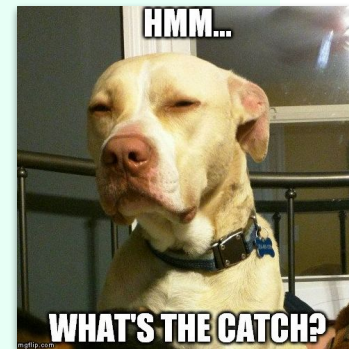
PROJECT

	PROJECT_CODE	PROJECT_BID_PRICE
▶	21-5Z	\$16,833,460.00
	25-2D	\$12,500,000.00
	25-5A	\$32,512,420.00
	25-9T	\$21,563,234.00
	27-4Q	\$10,314,545.00
	29-2D	\$25,559,999.00
	31-7P	\$56,850,000.00

Q7. If the above two tables are to be created in a relational database, in which table would you assign the FK (and using which attribute) to create the logical link?

A manager may manage many projects.

- A. MANAGER table using project_manager attribute.
- B. PROJECT table using project_code attribute.
- C. MANAGER table using manager_phone attribute.
- D. PROJECT table using project_manager attribute
- E. None of the above, a relationship is not needed.



Q8. Where are the foreign keys in these two relations?

Note: a supervisor is a staff member

**STAFF (staff_id, surname, initials, address, phone, dept_id,
supervisor_id)**

DEPARTMENT (dept_id, deptname)

- A. dept_id in staff relation.
- B. dept_id in department relation.
- C. staff_id in staff relation.
- D. supervisor_id in staff relation.
- E. More than one answer is correct

Data Integrity

- Entity integrity ('PK unique, non-null')
 - Primary key values must be unique
 - Primary key value must not be NULL.
- Referential integrity ('FK is another's PK or NULL')
 - The values of FK must either match a value of the PK in the related relation or be NULL.
- Column/Domain integrity
 - All values in a given column must come from the same domain (the same data type and range).

MANAGER

	PROJECT_MANAGER	MANAGER_PHONE	MANAGER_ADDRESS
▶	Holly B. Parker	904-338-3416	3334 Lee Rd., Gainesville, FL 37123
	Jane D. Grant	615-898-9909	218 Clark Blvd., Nashville, TN 36362
	George F. Dorts	615-227-1245	124 River Dr., Franklin, TN 29185
	William K. Moor	904-527-19	216 Morton St., Stetson, FL 30155

PROJECT

	PROJECT_CODE	PROJECT_MANAGER	PROJECT_BID_PRICE
▶	21-5Z	Holly B. Parker	\$16,833,460.00
	25-2D	Jane D. Grant	\$12,500,000.00
	25-5A	George F. Dorts	\$32,512,420.00
	25-9T	Holly B. Parker	\$21,563,234.00
	27-4Q	George F. Dorts	\$10,314,545.00
	29-2D	Holly B. Parker	\$25,559,999.00
	31-7P	William K. Moor	\$56,850,000.00

Q9. Suppose that the manager William K. Moor leaves the company and we delete his record from the manager table. Which of the following actions will satisfy the data integrity constraints?

- A. The last row in PROJECT table must be deleted
- B. The PROJECT_MANAGER value in the last row of PROJECT table must be set to NULL (empty)
- C. The PROJECT_MANAGER value in the last row of PROJECT table must be set to any string (e.g., "XYZ")
- D. The options A and B
- E. All of the above



Coffee break - see you in 10 minutes.

Relational DMLs

- Relational Calculus
- Relational Algebra
- Transform Oriented Languages (e.g. SQL)
- Graphical Languages
- Exhibit the “closure” property - queries on relations produce relations

Relational Calculus

- Based on mathematical logic.
- Non-procedural.
- Primarily of **theoretical importance**.
- May be used as a yardstick for measuring the power of other relational languages (“relational completeness”).
- Operators may be applied to any number of relations.
- NB: Can be combined
 - Analogy: programming/Excel
 - $\text{sum}(\text{sum}(1,2,3), 4,5,6) \rightarrow \text{sum}(6,4,5,6) \rightarrow 21$

RELATIONAL ALGEBRA

Manipulation of relational data

Relational Algebra

- Relationally complete.
- Procedural.
- Operators only apply to **at most two relations at a time**.
- 8 basic operations:
 - single relation: selection, projection
 - Cartesian product, join
 - union
 - intersection
 - difference
 - division

Relational Operation PROJECT

π

ID	Show Number	Air Date	Round	Category	Value	Question	Answer
5623	4680	31/12/2004	Jeopardy!	HISTORY	\$200	For the last 8 years of his life, Galileo	Copernicus
5624	4680	31/12/2004	Jeopardy!	ESPN's TOP 10 ALL-TIME ATHLETES	\$200	No. 2: 1912 Olympian; football star	Jim Thorpe
5625	4680	31/12/2004	Jeopardy!	EVERYBODY TALKS ABOUT IT...	\$200	The city of Yuma in this state has a	Arizona
6622	4680	31/12/2004	Jeopardy!	THE COMPANY LINE	\$200	In 1963, live on "The Art Linkletter	McDonald's
6623	4680	31/12/2004	Jeopardy!	EPITAPHS & TRIBUTES	\$200	Signer of the Dec. of Indep., framed	John Adams
6624	4680	31/12/2004	Jeopardy!	3-LETTER WORDS	\$200	In the title of an Aesop fable, this	the ant
6625	4680	31/12/2004	Jeopardy!	HISTORY	\$400	Built in 312 B.C. to link Rome & the	the Appian Way
6626	4680	31/12/2004	Double Jeopardy!	DR. SEUSS AT THE MULTIPLEX	\$400	<a href="http://www.j-archive.co	Horton
7898	4680	31/12/2004	Double Jeopardy!	PRESIDENTIAL STATES OF BIRTH	\$400	California	Nixon
7899	4680	31/12/2004	Double Jeopardy!	AIRLINE TRAVEL	\$400	It can be a place to leave your pupa	a kennel
7900	4680	31/12/2004	Double Jeopardy!	THAT OLD-TIME RELIGION	\$400	He's considered the author of the	Moses
7901	4680	31/12/2004	Double Jeopardy!	MUSICAL TRAINS	\$400	Steven Tyler of this band lent his s	Aerosmith
7902	5957	6/07/2010	Jeopardy!	GEOGRAPHY "E"	\$200	It's the largest kingdom in the Uni	England
7903	5957	6/07/2010	Jeopardy!	RADIO DISNEY	\$200	"Party In The U.S.A." is by this sin	Miley Cyrus

Memory aid: **pi** starts with 'P'. Project starts with 'P'.

Dataset by J-Archive / https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/

Relational Operation SELECT

σ

ID	Show Number	Air Date	Round	Category	Value	Question	Answer
5623	4680	31/12/2004	Jeopardy!	HISTORY	\$200	For the last 8 years of his life, Galileo	Copernicus
5624	4680	31/12/2004	Jeopardy!	ESPN's TOP 10 ALL-TIME ATHLETES	\$200	No. 2: 1912 Olympian; football star	Jim Thorpe
5625	4680	31/12/2004	Jeopardy!	EVERYBODY TALKS ABOUT IT...	\$200	The city of Yuma in this state has a	Arizona
6622	4680	31/12/2004	Jeopardy!	THE COMPANY LINE	\$200	In 1963, live on "The Art Linkletter	McDonald's
6623	4680	31/12/2004	Jeopardy!	EPITAPHS & TRIBUTES	\$200	Signer of the Dec. of Indep., framed	John Adams
6624	4680	31/12/2004	Jeopardy!	3-LETTER WORDS	\$200	In the title of an Aesop fable, this	the ant
6625	4680	31/12/2004	Jeopardy!	HISTORY	\$400	Built in 312 B.C. to link Rome & the	the Appian Way
6626	4680	31/12/2004	Double Jeopardy!	DR. SEUSS AT THE MULTIPLEX	\$400	<a href="http://www.j-archive.co	Horton
7898	4680	31/12/2004	Double Jeopardy!	PRESIDENTIAL STATES OF BIRTH	\$400	California	Nixon
7899	4680	31/12/2004	Double Jeopardy!	AIRLINE TRAVEL	\$400	It can be a place to leave your pupa	a kennel
7900	4680	31/12/2004	Double Jeopardy!	THAT OLD-TIME RELIGION	\$400	He's considered the author of the	Moses
7901	4680	31/12/2004	Double Jeopardy!	MUSICAL TRAINS	\$400	Steven Tyler of this band lent his s	Aerosmith
7902	5957	6/07/2010	Jeopardy!	GEOGRAPHY "E"	\$200	It's the largest kingdom in the Uni	England
7903	5957	6/07/2010	Jeopardy!	RADIO DISNEY	\$200	"Party In The U.S.A." is by this sin	Miley Cyrus

Memory aid: **sigma** starts with 'S'. Select starts with 'S'.

Dataset by J-Archive / https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/

Relational Operation Multiple Actions

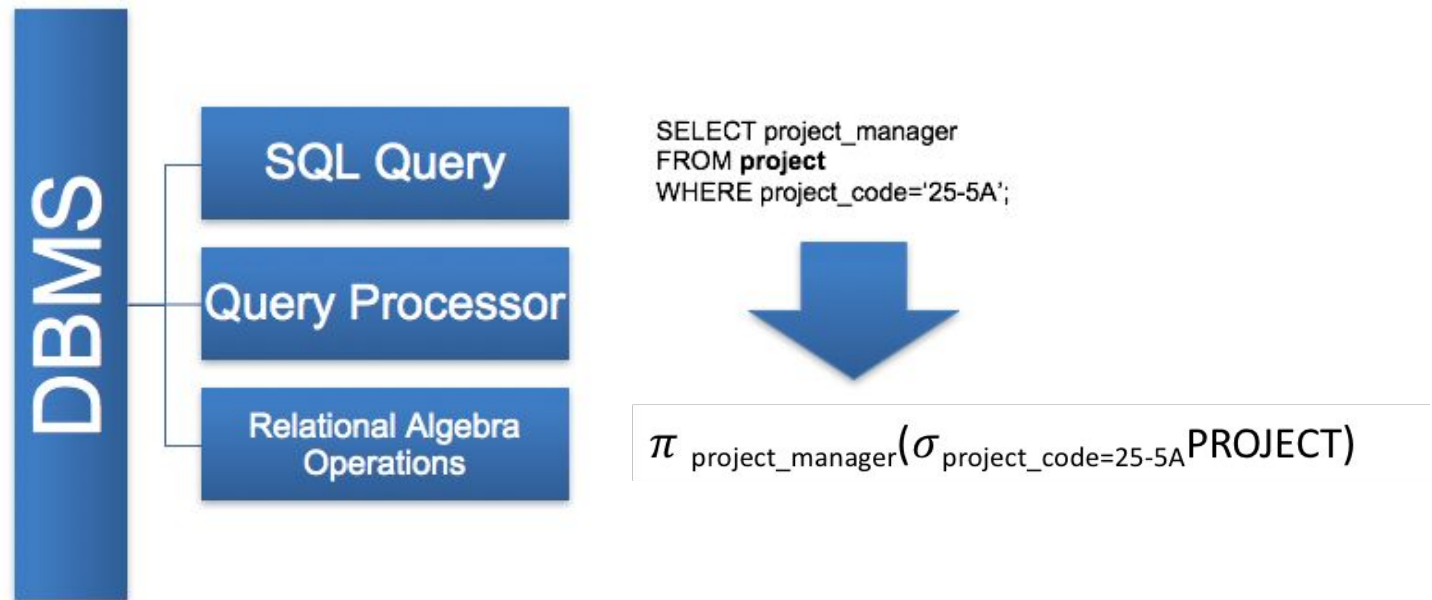
2

ID	Show Number	Air Date	Round	Category	Value	Question	Answer
5623	4680	31/12/2004	Jeopardy!	HISTORY	\$200	For the last 8 years of his life, Galileo	Copernicus
5624	4680	31/12/2004	Jeopardy!	ESPN's TOP 10 ALL-TIME ATHLETES	\$200	No. 2: 1912 Olympian; football star	Jim Thorpe
5625	4680	31/12/2004	Jeopardy!	EVERYBODY TALKS ABOUT IT...	\$200	The city of Yuma in this state has a	Arizona
6622	4680	31/12/2004	Jeopardy!	THE COMPANY LINE	\$200	In 1963, live on "The Art Linkletter	McDonald's
6623	4680	31/12/2004	Jeopardy!	EPITAPHS & TRIBUTES	\$200	Signer of the Dec. of Indep., framed	John Adams
6624	4680	31/12/2004	Jeopardy!	3-LETTER WORDS	\$200	In the title of an Aesop fable, this	the ant
6625	4680	31/12/2004	Jeopardy!	HISTORY	\$400	Built in 312 B.C. to link Rome & the	the Appian Way
6626	4680	31/12/2004	Double Jeopardy!	DR. SEUSS AT THE MULTIPLEX	\$400	<a href="http://www.j-archive.co	Horton
7898	4680	31/12/2004	Double Jeopardy!	PRESIDENTIAL STATES OF BIRTH	\$400	California	Nixon
7899	4680	31/12/2004	Double Jeopardy!	AIRLINE TRAVEL	\$400	It can be a place to leave your pupa	a kennel
7900	4680	31/12/2004	Double Jeopardy!	THAT OLD-TIME RELIGION	\$400	He's considered the author of the	Moses
7901	4680	31/12/2004	Double Jeopardy!	MUSICAL TRAINS	\$400	Steven Tyler of this band lent his s	Aerosmith
7902	5957	6/07/2010	Jeopardy!	GEOGRAPHY "E"	\$200	It's the largest kingdom in the Uni	England
1 7903	5957	6/07/2010	Jeopardy!	RADIO DISNEY	\$200	"Party In The U.S.A." is by this sin	Miley Cyrus

Result = $\pi_{\text{category}}(\sigma_{\text{id}=7903} \text{JEOPARDY})$

SQL vs Relational Algebra in the Database

Case study: 'Project Management'



NB: in this example there is a table called PROJECT which is only a coincidence to the **pi** operator!

STUDENT

course	name	sid
BE	Anne	21333
BE	Dave	21876
BSc	John	21531
BSc	Tim	21623

MARK

stude	subj	mark
21333	1011	74
21333	1021	70
21333	2011	68
21531	1011	94
21531	1021	90
21623	1011	50

Q10. Which of the following statements returns the student ids of the students who got more than 70 marks in the subject 1011.

- A. $\sigma_{\text{mark} > 70} (\pi_{\text{stude}} (\text{MARK}))$
- B. $\sigma_{\text{mark} > 70} (\text{MARK})$
- C. $\sigma_{\text{mark} > 70 \text{ AND } \text{subj} = 1011} (\pi_{\text{stude}} (\text{MARK}))$
- D. $\sigma_{\text{mark} > 70 \text{ AND } \text{subj} = 1011} (\text{MARK})$
- E. $\pi_{\text{stude}} (\sigma_{\text{mark} > 70 \text{ AND } \text{subj} = 1011} (\text{MARK}))$

JOIN

- Join operator used to combine data from two or more relations, based on a common attribute or attributes.
- Different types:
 - theta-join
 - equi-join
 - natural join
 - outer join

THETA JOIN (Generalised join)

$(\text{Relation_1}) \bowtie_F (\text{Relation_2})$

- F is a predicate (i.e. truth-valued function) which is of the form $\text{Relation_1}.a_i \theta \text{Relation_2}.b_i$
- θ is one of the standard arithmetic comparison operators, i.e. $<, \leq, =, \geq, >$
- Most commonly, θ is equals ($=$)
- \bowtie just means ‘natural join’ per Codd
 - see [https://en.wikipedia.org/wiki/Relational_algebra#Natural_join_\(%E2%8B%88\)](https://en.wikipedia.org/wiki/Relational_algebra#Natural_join_(%E2%8B%88))

STUDENT	course	name	sid
	BE	Anne	21333
	BE	Dave	21876
	BSc	John	21531
	BSc	Tim	21623

MARK	stude	subj	mark
	21333	1011	74
	21333	1021	70
	21333	2011	68
	21531	1011	94
	21531	1021	90
	21623	1011	50

Q11. How many rows are generated when the product (Cartesian Product) of the STUDENT and MARK relations is taken? i.e. the number of rows in STUDENT X MARK.

- A. 24
- B. 6
- C. 18
- D. 7
- E. none of the above

STUDENT	course	name	sid
	BE	Anne	21333
	BE	Dave	21876
	BSc	John	21531
	BSc	Tim	21623

MARK	stude	subj	mark
	21333	1011	74
	21333	1021	70
	21333	2011	68
	21531	1011	94
	21531	1021	90
	21623	1011	50

Q12. How many columns are generated when the product (Cartesian Product) of the STUDENT and MARK relations is taken? i.e. the number of columns in STUDENT X MARK.

- A. 9
- B. 6
- C. 5
- D. 7
- E. none of the above

NATURAL JOIN

STUDENT		MARK		
ID	Name	ID	Subj	Marks
1	Alice	1	1004	95
2	Bob	2	1045	55
		1	1045	90

Step 1: STUDENT X MARK

Step 2: delete rows where IDs do not match (select =)

STUDENT. ID	Name	MARK.ID	Subj	Marks
1	Alice	1	1004	95
1	Alice	2	1045	55
1	Alice	1	1045	90
2	Bob	1	1004	95
2	Bob	2	1045	55
2	Bob	1	1045	90

NATURAL JOIN

STUDENT		MARK		
ID	Name	ID	Subj	Marks
1	Alice	1	1004	95
2	Bob	2	1045	55
		1	1045	90

Step 1: STUDENT X MARK

Step 2: delete rows where IDs do not match (select =)

Step 3: delete duplicate columns (project away)

STUDENT.ID	Name	MARK.ID	Subj	Marks
1	Alice	1	1004	95
1	Alice	1	1045	90
2	Bob	2	1045	55

NATURAL JOIN

STUDENT		⋈	MARK		
ID	Name		ID	Subj	Marks
1	Alice	⋈	1	1004	95
2	Bob		2	1045	55
			1	1045	90

Step 1: STUDENT X MARK

Step 2: delete rows where IDs do not match (select =)

Step 3: delete duplicate columns (project away)

ID	Name	Subj	Marks
1	Alice	1004	95
1	Alice	1045	90
2	Bob	1045	55

A natural join of STUDENT and MARK

STUDENT	course	name	sid
	BE	Anne	21333
	BE	Dave	21876
	BSc	John	21531
	BSc	Tim	21623

MARK	stude	subj	mark
	21333	1011	74
	21333	1021	70
	21333	2011	68
	21531	1011	94
	21531	1021	90
	21623	1011	50

Q13. Which of the following statements returns a natural join of the two relations on the student ids (sid and stude)?



- A. $\sigma_{\text{sid} = \text{stude}} (\text{STUDENT} \times \text{MARK})$
- B. $\pi_{\text{course, name, sid, subj, mark}} (\sigma_{\text{sid} = \text{stude}} (\text{STUDENT} \times \text{MARK}))$
- C. $\sigma_{\text{sid} = \text{stude}} (\pi_{\text{course, name, sid, subj, mark}} (\text{STUDENT} \times \text{MARK}))$
- D. All of the above
- E. None of the above

STUDENT		MARK		
ID	Name	ID	Subj	Marks
1	Alice	1	1004	95
2	Bob	2	1045	55
3	Chris	1	1045	90
		4	1004	100

Q14. Which of the following statements returns names and subject codes for which the students got more than 70 marks in the subject.

- A. $\sigma_{\text{Marks} > 70} (\text{STUDENT} \bowtie \text{MARK})$
- B. $\pi_{\text{Name, subj}} (\sigma_{\text{Marks} > 70} (\text{STUDENT} \bowtie \text{MARK}))$
- C. $\pi_{\text{Name, subj}} (\text{STUDENT} \bowtie \text{MARK})$
- D. None of the above

OUTER JOIN

STUDENT			MARK		
ID	Name		ID	Subj	Marks
1	Alice	⋈	1	1004	95
2	Bob		2	1045	55
3	Chris 		1	1045	90
			4 	1004	100

No information for Chris (no mark, e.g. just enrolled) and the student with ID 4 (no student, e.g. quit uni)

ID	Name	Subj	Marks
1	Alice	1004	95
1	Alice	1045	90
2	Bob	1045	55

A natural join of STUDENT and MARK

FULL OUTER JOIN

STUDENT			MARK		
ID	Name		ID	Subj	Marks
1	Alice	⋈	1	1004	95
2	Bob		2	1045	55
3	Chris		1	1045	90
			4	1004	100

Get (incomplete) information of both Chris and student with ID 4

ID	Name	Subj	Marks
1	Alice	1004	95
1	Alice	1045	90
2	Bob	1045	55
3	Chris	Null	Null
4	Null	1004	100

A full outer join of STUDENT and MARK

LEFT OUTER JOIN

STUDENT			MARK		
ID	Name		ID	Subj	Marks
1	Alice	⌋	1	1004	95
2	Bob		2	1045	55
3	Chris		1	1045	90
			4	1004	100

← *Get (incomplete) information of only Chris*

ID	Name	Subj	Marks
1	Alice	1004	95
1	Alice	1045	90
2	Bob	1045	55
3	Chris	Null	Null

A left outer join of STUDENT and MARK
Memory aid: Chris is on the ← LEFT of the nulls.

RIGHT OUTER JOIN

STUDENT			MARK		
ID	Name		ID	Subj	Marks
1	Alice	⋈	1	1004	95
2	Bob		2	1045	55
3	Chris		1	1045	90
			4	1004	100

Get (incomplete) information of the student with ID 4 →

ID	Name	Subj	Marks
1	Alice	1004	95
1	Alice	1045	90
2	Bob	1045	55
4	Null	1004	100

A right outer join of STUDENT and MARK.
Memory aid: the marks data is on the RIGHT → of the null.

No Flux this slide.
Live demo with free chocolate.

Live demo!

We need 3 volunteers from the audience to come forward

We need 4 pieces of candy.

STUDENT

course	name	sid
BE	Anne	21333
BE	Dave	21876
BSc	John	21531
BSc	Tim	21623

MARK

stude	subj	mark
21333	1011	74
21333	1021	70
21333	2011	68
21531	1011	94
21531	1021	90
21623	1011	50

Q15. Consider the above relations.

Assume that we want to join the them to obtain the information of all students (Anne, Dave, John and Tim). Which of the following is **WRONG? (Hint: Dave just enrolled!)**

- A. Left outer join on STUDENT and MARK
- B. Right outer join on MARK and STUDENT
- C. Right outer join on STUDENT and MARK
- D. Full outer join on STUDENT and MARK
- E. Select if (B and C are wrong)
- F. Select if (B, C and D are wrong)