CPE 365 Introduction to Databases Winter 2023

Homework 1

Due: Tuesday, February 21

The homework is due the day of our midterm exam.

Problem 1 *Consider the following relations:*

R:

A	B	C
b	c	3
c	c	3
c	a	3
b	b	1
c	a	4
b	a	2

S:

A	B	C
c	c	2
c	a	3
b	b	1
a	b	3
b	c	3
a	a	1
c	c	3

T:

B	D	E	F
a	a	1	2
c	d	2	4
b	b	3	2
d	b	3	2
a	a	2	3
b	c	4	1
d	a	1	4

W:

C	D
1	a
2	b
3	c
4	d

For each of the following queries: (1) compute the answer, (2) draw the query tree.

a. $R \cap S$

b. $S - R$

c. $R - S$

d. $\pi_{A,B}(S)$

e. $\pi_{B,C}(R) \cup \pi_{B,C}(S)$

f. $\pi_{A,B}(R) - \pi_{A,B}(S)$

g. $\pi_C(W) \times \pi_A(S) \times \pi_B(T)$

h. $\sigma_{E>F}(T)$

i. $\sigma_{A \neq B}(R)$

j. $\pi_{B,F}(\sigma_{F \geq E}(T))$

k. $\sigma_{A=D \vee B=D}(\pi_{A,B}(R) \times W)$

l. $\sigma_{A=b \wedge C>1}(R) \cup \sigma_{B=b \vee C \neq 3}(S)$

m. $\sigma_{\neg(B=d)}(T)$

n. $\pi_{A,B,R.C,D}(\sigma_{R.C=W.C}(R \times W))$

o. $W \bowtie R$

p. $W \bowtie_{R.C=W.C} R$

q. $T \bowtie_{F>C} W$

r. $R \bowtie S$

s. $R \bowtie_{R.B=S.A} S$

t. $(R \bowtie T) \bowtie \pi_{A,C,D}(S \bowtie W)$

u. $\pi_{T1.D,T2.B}(\rho_{T1}(T) \bowtie_{T1.D=T2.B} \rho_{T2}(T))$

v. $\pi_{B,D,E}(\sigma_{F \leq C}(T \bowtie W))$

w. $\pi_{R.A,R.B}(R \bowtie_{R.C \neq S.C} S) \bowtie \sigma_{D=a}(T)$

x. $\pi_A(\pi_B(\pi_C(R \cup S)))$

y. $\sigma_{A \neq a}(S) \bowtie \sigma_{D \neq c}(W)$

z. $\sigma_{C=1}(R) \bowtie \sigma_{C=2}(S)$

Problem 2 *Consider the following database:*

bands:

Id	Name	Formed_in	Country
1	'Pink Floyd'	1965	'UK'
2	'King Crimson'	1969	'UK'
3	'Can'	1968	'Germany'
4	'Doors'	1967	'USA'
5	'Velvet Underground'	1967	'USA'
6	'Gong'	1969	'France'
7	'Yes'	1968	'UK'

albums:

AId	Title	BandId	Year
1	'Meddle'	1	1970
2	'Animals'	1	1977
3	'Red'	2	1974
4	'Landed'	3	1975
5	'Soundtracks'	3	1970
6	'Beat'	2	1982
7	'The Doors'	4	1967
8	'Strange Days'	4	1967
9	'Loaded'	5	1970
10	'You'	6	1974
11	'Shamal'	6	1976
12	'Fragile'	7	1971
13	'Close to the Edge'	7	1972

musicians:

MId	Name	BandId	From	To
1	'Roger Waters'	1	1966	1983
2	'Syd Barrett'	1	1966	1968
3	'Robert Fripp'	2	1969	2002
4	'Adrian Belew'	2	1981	2002
5	'Irmin Schmidt'	3	1968	1978
6	'Michael Karoli'	3	1968	1978
7	'Jim Morrison'	4	1967	1971
8	'Lou Reed'	5	1967	1973
9	'Daevid Allen'	6	1969	1974
10	'Pierre Moerlen'	6	1975	1978
11	'Jon Anderson'	7	1968	1981
12	'Bill Bruford'	7	1968	1972
13	'Bill Bruford'	2	1972	2010

Notes: (a) BandID is a *foreign key* on bands in both relations where this attribute appears. (b) Consider all attributes indicating years to be integers with appropriate comparison operators applicable. (c) In what follows we assume that if a musician is listed as being in the band during a year in which an album was recorded, the musician played on it. (d) First attribute of each relation is its primary key.

If you need to refer to relation names in queries you can use A for albums, B for bands and M for musicians.

Translate the following queries into Relational Algebra and compute their answers based on the given database. As an extra exercise, draw the query trees for the relational algebra expressions you construct.

1. Find all musicians who played with their band in year 1970. Output the names of the musicians.
2. Find all musicians who played with their band in year 1970. Output the names of the musicians and the name of the band they played in.
3. Find all musicians who played for `'Gong'`. Output the names of the musicians and the years they played in the band.
4. Find all musicians who played in the band `'King Crimson'` in the year 1974. Output the names of the musicians.
5. Find all musicians who played in their band in the band's year of inception. Output the names of the musicians and the names of the bands.
6. Find the band that released the album `'Loaded'`. Report the name of the band.
7. Find the band in which `'Jim Morrison'` played. Report the name of the band.
8. Find all albums recoreded by bands from UK. For each record output its name, year and the name of the band that recorded it.
9. Find all musicians who participated in recording of the album `'Fragile'`. Output the names of the musicians.
10. Find all `'Pink Floyd'` band members who did NOT participate in the recording of the album `'Meddle'`. (note: a musician did not participate in a recording of an album if he did not play in the band that year).
11. Find all bands that recorded two albums in the same year. For each band, output its name, titles of both albums and the year of their release.
12. Find all bands in which `'Lou Reed'` DID NOT play. Output the names of the bands.
13. Find all albums recorded by the bands that had at least one musician leave befire 1972. Output the name of the album and the year of release.
14. Find all albums recorded by `'King Crimson'` before `'Adrian Belew'` joined the band. Report the album titles and years.
15. Find all musicians who played in at least two different bands. Report their names.

Problem 3

You have a CSV file describing the performance of NBA basketball players, which contains the following columns:

```
FirstName: string  -- first name of the player
LastName: string   -- last name of the player
TeamCity: string   -- city of the player's team (e.g., Orlando, New York)
TeamName: string   -- name of the player's team (e.g., Lakers, Jazz)
Height  : int      -- player height in centimeters
Position: string   -- main position of the player on the field (e.g., PG, SG, PF - abbreviated)
Season  : int      -- year of the NBA season (year the NBA final game for the season is played)
Games   : int      -- games played in the given season for the given team
Starts  : int      -- games started in the given season for the given team
Shots   : int      -- total number of shots taken while playing for the team in the season
Points  : int      -- total number of points scored while playing for the team in the season
Assists : int      -- total number of assists made while playing for the team in the season
Rebounds: int      -- total number of rebounds made while playing for the team in the season
Steals  : int      -- total number of steals made while playing for the team in the season
```

For the purposes of this task, you can assume that no two players on the same team have the same first name - last name combination. Also, assume that this table covers current NBA teams. You want to create a relational table `PlayerStats` containing all these columns.

1. Find one candidate key for this table. Explain why it is a candidate key.
2. Explain why the combination of attributes `(TeamName, Shots, Rebounds)` is not a candidate key.
3. Explain why the combination of attributes `(LastName, Season)` is not a candidate key.
4. Explain why the combination of attributes `(FirstName, LastName, TeamCity, TeamName, Height, Position, Season, Games)` is not a candidate key.

Problem 4

You have a CSV file describing a passenger train schedule in a European country. All trains are operated by a single train operator (national railroad company). The CSV file contains the following columns.

```
RouteNumber: string   -- unique number corresponding to a train route
Origin: string        -- City (railway station) of origin
Departure: time       -- time of departure
Destination: string   -- City (railway station) of final destination of the train
Arrival: time         -- time of arrival
NDays: int            -- number of days of travel (0 if train arrives same day, 1 if next day, and so on).
TrainType: string     -- "regular" or "express"
NStops                -- number of stops along the route
FirstClass: int       -- number of 1st class rail cars
SecondClass: int      -- number of 2nd class rail cars
SleepingCars: int     -- number of sleeping rail cars
FirstClassFare: float -- first class fare for full trip
SecondClassFare:float -- second class fare for full trip
SleepingCarFare:float -- sleeping care fare for full trip
```

Basically, this CSV file describes each route with a point of origin and destination inside the country, and specifies for each route it's origin, destination, departure and arrival times, number of stops, type of train, and information about availability and the cost of different kinds of accommodations on the train.

You want to create a relational table `Schedule` that contains all the columns above.

1. Make some reasonable assumptions about train schedules, and based on those assumptions, identify all candidate keys for this table. (specify your assumptions along with the candidate keys).
2. Out of the candidate keys identified select your primary key.
3. Explain why `(RouteNumber, Orign, NDays, NStops)` is not a candidate key.
4. Explain why `(Origin, Destination)` is not a candidate key.
5. Explain why `(Origin, TrainType, NStops)` is not a candidate key.

Problem 5

You have a CSV file containing historic information about all people who were elected[1] as a State Legislator in any of the US State Legislatures. One row of this table describes one term of service for a given person (that is, if someone was elected multiple times, we will see a record of them for each term of their service). A term has a start year and and end year (e.g., 2018-2019). Most US state legislatures are bicameral (i.e., they have a Senate and an Assembly). Only Nebraska is an exception, it has unacameral legislature.

The CSV file has the following columns.

```
StartYear: int     -- start year of the term
EndYear: int       -- end  year of the term
State: string      -- two-letter abbreviation of the state (e.g., CA, IA, KY)
Chamber: string    -- the legislative chamber (Senate, Assembly)
District: int      -- legislative chamber district
FirstName: string  -- first name of the legislator
MiddleName: string -- middle name of the legislator
LastName: string   -- last name of the legislator
party: string      -- political party affiliation of the legislator (Democrat, Republican, Green, Independent,
Birthday: date     -- legislator's date of birth
Biography:text     -- legislator's biography
Email: string      -- legislator's email address
Phone: string      -- phone number of the legislator's office
Web: string        -- URL of legislator's web page
```

You want to create a relational table `StateLegislators` with all the columns listed above. You are worried that `FirstName, MiddleName, LastName` by itself may not be enough to uniquely identify a person (note that each row of the table describes not a person, but a single term served by them, however your concern is still a valid one).

1. Suggest at least five candidate keys for this table. If you had to make any assumptions along the way, state them explicitly.
2. Select a primary key for this table.
3. Explain why `(Email, Phone, Web)` is not a valid candidate key.
4. Explain why `(StartYear, EndYear, State)` is not a valid candidate key.
5. Explain why `(FirstName, MiddleName, LastName, StartYear)` is not a valid candidate key.
6. Explain why `(StartYear, EndYear, State, Chamber, Phone)` is not a valid candidate key.

[1] In some situations an elected lawmaker leaves the Legislature mid-term and is replaced. This table only stores information about people who actually were elected to their term, not their replacements.