Final Assignment

Introduction to Data Analytics for Business

Giuliano Sposito - gsposito@gmail.com

The final assessment for course 1 is peer-graded. While it won't incorporate every idea, we've covered in the course, it will include key elements from each of Modules 1 through 4 in one integrated exercise.

The assessment has four parts:

- **Conceptual business model**. You'll construct a conceptual business model similar to the one we discussed in Module 1.
- **Relational data model**. You'll then design a simple relational data model to represent some of the ideas from your conceptual model (like in Module 2), and describe what types of systems you think the data might come from (Module 1).
- **SQL queries**. You'll write two SQL queries to extract an interesting data set from your data model (Module 3)
- **Sensitive data and data quality issues**. Finally, you'll identify whether your model contains certain types of sensitive data, and assess where your model might be susceptible to data quality issues (Module 4).

Part 1: Conceptual Model

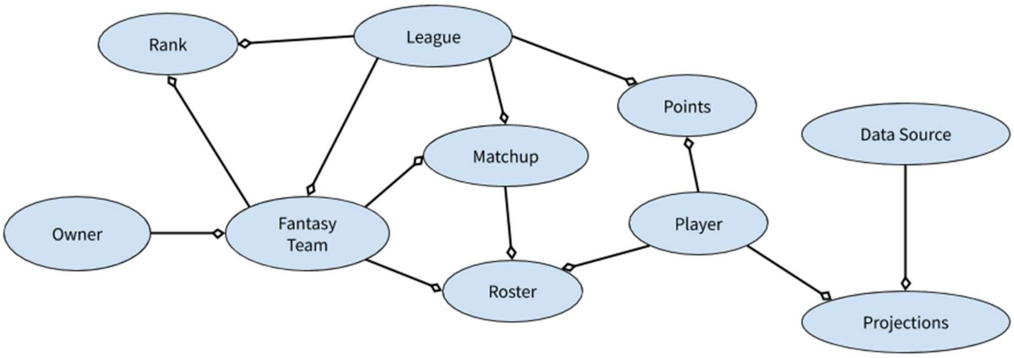
Instruction

Conceptual business model. Construct a conceptual business model for an industry or business that you are familiar with or have interest in. Visually it should be similar to the one we illustrated in Module 1, Video 2.

- Your model should represent at least 10 ideas
- It should visually represent one to one, many to one, or many to many relationships among ideas

Conceptual Model

For this exercises I used the NFL Fantasy as conceptual and data models. Fantasy football is a game in which participants assemble an imaginary team of real life footballers and score points based on those players' actual statistical performance or their perceived contribution on the field of play. Usually players are selected from one specific division in a particular country, although there are many variations. The original game was created in England by Bernie Donnelly on Saturday 14 August 1971 and is still going strong 45 years later. Check this reference to know more about it.



NFL Fantasy Conceptual Model

Explaining the entities

- **League**: it's a set of fantasy players that join to build teams and play a championship between them.
- **Owner**: it's the fantasy player member of a league that owns one team
- **Fantasy Team**: it's the team in the league, owned by a fantasy player
- **Matchup**: It's the set of matchs between teams in a league, each team plays once a week.
- **Roster**: It's the lineup of football players for each team in a week.
- **Player**: It's the set of football players.
- **Points**: It's the weekly point record of each player of a league
- **Projection**: It's the "prediction" of each weekly player score.
- **Data Source**: It's the web site that make weekly "prediction" about players score.

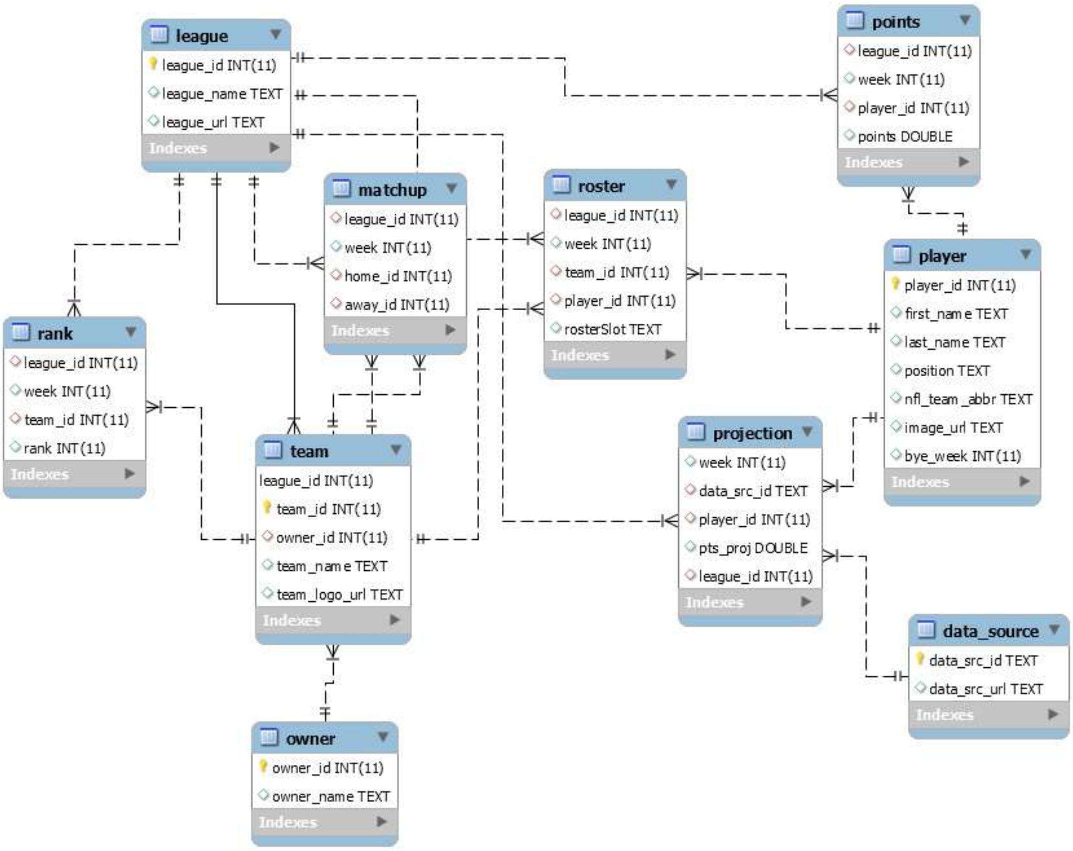
Part 2: Relational data model.

Instruction

Take a subset of the ideas from the conceptual model you constructed in Part 1 and design a simple relationship model similar to the ones we discussed in Module 2, Video 4

- Your model should have at least 5 tables
- You should include at least 20 attributes, or fields, in your model (20 total across all tables, not per table)
- Your model should be normalized
- Identify the primary key in each table, and state whether it is a natural or surrogate key
- For each relationship between tables, identify any foreign keys needed to define the relationship
- For each table, identify what type of system or systems you think the data might come from, like those we discussed in Module 1, Video 6.

Relational Data Model



Fantasy Datamodel

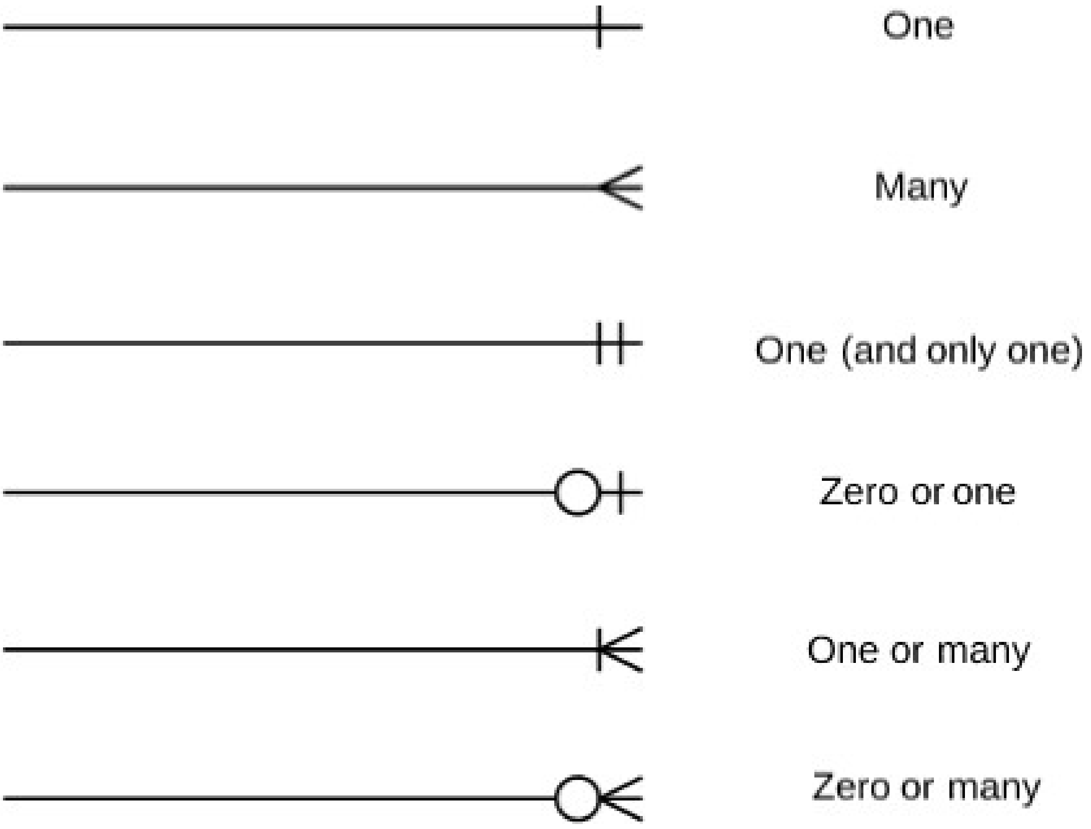
Keys

The relational data model is pretty straightforward with respect the conceptual model, e create 10 tables with 42 fields, including the key. The model is normalized, for each table the key are:

Table	Key(s)	PK/FK	Type
League	league_id	PK	surrogate
Team	team_id	PK	surrogate
Owner	owner_id	PK	surrogate
Player	player_id	PK	surrogate
Datasource	data_src_id	PK	surrogate
Rank	league_id, team_id, week	FK	composite
Matchup	league_id, home_id, away_id, week	FK	composite
Roster	league_id, team_id, week, player_id	FK	composite
Points	league_id, player_id, week	FK	composite
Projection	data_src_id, league_id, player_id, week	FK	composite

Relationships

The relationship are explicit in the model, that the follows the standard Entity-Relationship Diagram Notation



Cardinality and ordinality Notation

Worth to note that the in the **matchup** table, the teams that will play in this match are named as *home_id* and *away_id* for home and visitors team, this fields are the *team_id* in the **team** table, so one match always has two teams.

Sources

We can imagine 3 diferente type os “sources” from this info, the Fantasy Game itself, the actual NFL game stats and the internet for poinst predictions, with each table coming from:

Table	System Source
League	Fantasy Game
Team	Fantasy Game
Owner	Fantasy Game
Player	NFL Stats
Rank	Fantasy Game
Matchup	Fantasy Game
Roster	Fantasy Game
Points	NFL Stats
Datasource	Internet
Projection	Internet

Understanding data_source and projection tables

These tables makes part of a “prediction system”, the table **data_source** tells witch internet web site has NFL Players Prediction that can be used by a Fantasy Player choose and lineup Football Players. Let’s see the table content.

```
select *
from data_source
```

Displaying records 1 - 10

data_src_id	data_src_url
CBS	https://www.cbssports.com/fantasy/football/stats/
ESPN	http://games.espn.com/ffl/tools/projections
FantasyData	https://fantasydata.com/nfl-stats/fantasy-football-weekly-projections.aspx
FantasyFootballNerd	http://www.fantasyfootballnerd.com/service/
FantasyPros	https://www.fantasypros.com/nfl/projections/
FantasySharks	https://www.fantasysharks.com/apps/bert/forecasts/projections.php
FFToday	http://www.fftoday.com/rankings/
FleaFlicker	https://www.fleaflicker.com/nfl/leaders
NFL	http://api.fantasy.nfl.com/v1/players/stats
NumberFire	https://www.numberfire.com/nfl/fantasy/

As we see, each record of the table **data_source** is a diferent website projection provider. Now let’s look part of the **projection* table, who stores the weekly projected score for each player made by each website provider.

```sql
select pr.league_id, pr.week, pr.data_src_id, pr.pts_proj, pl.first_name,
pl.last_name, pl.position, pl.nfl_team_abbr
from projection pr
inner join player pl
on pl.player_id=pr.player_id
order by pl.player_id desc, week asc, data_src_id asc
limit 15
```

Displaying records 1 - 10

league_id	week	data_src_id	pts_proj	first_name	last_name	position	nfl_team_abbr
3940933	5	CBS	1.970000	D’Ernest	Johnson	RB	CLE
3940933	5	FantasySharks	1.015662	D’Ernest	Johnson	RB	CLE
3940933	5	NFL	1.608000	D’Ernest	Johnson	RB	CLE
3940933	5	Yahoo	0.640000	D’Ernest	Johnson	RB	CLE
3940933	5	CBS	6.400000	Joey	Slye	K	CAR
3940933	5	FantasySharks	9.000000	Joey	Slye	K	CAR
3940933	5	NumberFire	9.130000	Joey	Slye	K	CAR
3940933	5	Yahoo	8.500000	Joey	Slye	K	CAR
3940933	5	FantasyPros	1.540000	Darrius	Shepherd	WR	GB
3940933	5	FantasySharks	0.180000	Darrius	Shepherd	WR	GB

Part 3: SQL queries.

Using the data model you constructed in Part 2, come up **with two data extracts** you think would be interesting, then write SQL queries to provide each one.

- For each query, state what data you are trying to get and why it would be interesting
- Provide the SQL query using the commands and syntax we learned in Module 3
- For maximum credit, at least one of your queries should involve a join across two or more tables.

Queries: Week Matchup Result

As first query, let's find which team won and what score they have in a specific week (5 for example) in the fantasy league. To do so, we'll have to join the **roster** of each **team** in that week with the individual **player'a points**, sum then up to find the team's score and compare with the opponent team score.

For this exercise we'll use, data extract from my **league**, named "It's football, dudes", which I extract the data from Fantasy website.

```sql
-- what are the league_id?
select *
from league
```

1 records

league_id	league_name	league_url
3940933	Its Football Dudes	https://dudesfootball.netlify.com/

There is only one league in the data set, we'll use it's code: 3940933. Let's make this in two steps, first let's find which teams plays against each other in the week 5.

Query 1 - Games Schedule

```sql
-- from the matchup
select
    m.league_id,
    m.week,
  m.home_id,
  hmt.team_name as home_team_name,
  m.away_id,
  awt.team_name as away_team_name
from matchup m
-- join the home team with team table
inner join team hmt on hmt.team_id = m.home_id
-- joint the visitor team with team table
inner join team awt on awt.team_id = m.away_id
-- week 5 on the league of interest
where m.week=5 and m.league_id=3940933;
```

5 records

league_id	week	home_id	home_team_name	away_id	away_team_name
3940933	5	6	Rio Claro Pfeiferians	1	Paulinia Robots
3940933	5	5	Sorocaba Steelers	7	London Knights
3940933	5	4	Amparo Bikers	8	Jersey Boys
3940933	5	3	Indaiatuba Riders	9	Indaiatuba Blues
3940933	5	2	Sorocaba Wild Mules	11	Campinas Giants

Now we know what are the matchups for week 5, let's calculate each team's score in that week, to do this we join the roster of each team with the individual players points and sum up in a team score.

Query 2 - Roster Score

```
-- week team total score
select league_id, week, team_id, sum(points)
from (
    -- join roster with points to get individual player's points in the r
osters
    select r.league_id, r.week, r.team_id, r.player_id, r.rosterSlot, p.p
oints
    from roster r inner join points p
    on r.league_id=p.league_id and r.week=p.week and r.player_id=p.player
_id
) roster_points
-- exclude players in the bench, their pontuation don't count as team poi
nts
where rosterSlot != "BN" and week=5 and league_id=3940933
group by league_id, week, team_id;
```

Displaying records 1 - 10

league_id	week	team_id	sum(points)
3940933	5	1	166.36
3940933	5	2	130.26
3940933	5	3	121.52
3940933	5	4	134.20
3940933	5	5	133.34
3940933	5	6	109.04
3940933	5	7	156.48
3940933	5	8	133.44
3940933	5	9	87.32
3940933	5	11	184.34

Now, let’s up all toghether.

Query 3 - Game Score

```sql
-- round results query
select
    hmt.team_name as home_team_name,
    htpts.total_points as home_team_points,
    awt.team_name as away_team_name,
    awpts.total_points as away_team_points
from matchup m
-- to get tha names of home and visitor teams
inner join team hmt on hmt.team_id = m.home_id
inner join team awt on awt.team_id = m.away_id
-- score for home team
inner join (
    select league_id, week, team_id, round(sum(points),1) as total_points
    from (
        -- join roster with points to get individual player's points in the rosters
        select r.league_id, r.week, r.team_id, r.player_id, r.rosterSlot, p.points
        from roster r inner join points p
        on r.league_id=p.league_id and r.week=p.week and r.player_id=p.player_id
    ) roster_points
    -- exclude players in the bench, their pontuation don't count as team points
    where rosterSlot != "BN"
    group by league_id, week, team_id
) htpts on m.league_id=htpts.league_id and m.week=htpts.week and hmt.team_id=htpts.team_id
-- score for visitor team
inner join (
    select league_id, week, team_id, round(sum(points),1) total_points
    from (
        -- join roster with points to get individual player's points in the rosters
        select r.league_id, r.week, r.team_id, r.player_id, r.rosterSlot, p.points
        from roster r inner join points p
        on r.league_id=p.league_id and r.week=p.week and r.player_id=p.player_id
    ) roster_points
    -- exclude players in the bench, their pontuation don't count as team points
    where rosterSlot != "BN"
    group by league_id, week, team_id
) awpts on m.league_id=awpts.league_id and m.week=awpts.week and awt.team_id=awpts.team_id
-- just from week and league of interest
where m.week=5 and m.league_id=3940933
order by home_team_points desc
```

5 records

home_team_name	home_team_points	away_team_name	away_team_points
Amparo Bikers	134.2	Jersey Boys	133.4
Sorocaba Steelers	133.3	London Knights	156.5
Sorocaba Wild Mules	130.3	Campinas Giants	184.3
Indaiatuba Riders	121.5	Indaiatuba Blues	87.3
Rio Claro Pfeiferians	109.0	Paulinia Robots	166.4

Finally we get the round *5* of league *3940933*, with the names of the teams and the score marked for each team.

Part 4: Sensitive data and data quality issues.

Consider the data privacy and data quality implications of the data model you constructed in Part 2.

1) Identify any fields you think might be PII, CFI, CPNI, or PHI as we defined in Module 4, Video 4

There is no real implications of CFI, CPNI or PHI, once the dataset has no information about network activities, financial or helth informations. For personal information there is two table candidates, the **Owners Table** identifies the game player, there is a field *name* on it, but is used by players as *alias*, so the risk is minimal.

```
select *
from owner
```

Displaying records 1 - 10

owner_id	owner_name
11286409	Roander
11660207	Leonel
11663146	Langas
11665758	Giuliano
12023425	Marcos
12023441	Vinicius
12102270	Thiago
13963520	Fernando
17609775	Hilton
21990516	Mota Bonitao

The other PII risk is in the **Players Table**, that identifies real football player, but I think in this case is *by design*, you play Fantay with real football players performances and results.

```
select player_id, first_name, last_name, position, nfl_team_abbr
from player
```

Displaying records 1 - 10

player_id	first_name	last_name	position	nfl_team_abbr
252	Chad	Henne	QB	KC
264	Josh	Johnson	QB	
310	Matt	Ryan	QB	ATL
382	Joe	Flacco	QB	DEN
949	Jonathan	Stewart	RB	
1581	DeSean	Jackson	WR	PHI
2346	Pierre	Garcon	WR	
2649	Danny	Amendola	WR	DET
4487	Matthew	Slater	WR	NE
71265	Jared	Cook	TE	NO

What data elements in your model will present the most significant data quality challenges? Explain your reasoning.

In this case there only three data source provider:

1. Fantasy Game
2. NFL Stats
3. Projections Websites

Firsts two, Fantasy Game and NFL Stats, could have **Provenance** or **Timeliness** problems, but the probability is low because both are provided by NFL itself. But in the last,

Projections Websites are, in fact, multple sources and could have several problems as **Conformance** (each site provides the information with differents formats), **Provenance** (each sites has your owmn quality standard) and also **Accuracy and Completeness** problems, because they haven't to comply necessarily with a Fantasy Game.