Maxime Theokritoff

CSCI - 6622

3 – Project Assessment

a – How does this design fulfills the requirements of the job?

Storing information from LinkedIn:

The Database can store data on people, skills, companies and jobs. This is the main information available on LinkedIn

Returning the value of a skill:

-Based on salary
-Based on number of job openings
-Based on number of job openings
- Based on number of jobs needing it
- Based on people with that skill
-Flexibility to add new tables: We are not sure this has been achieved.

Data Accuracy:

Reference tables are present for the JOBS, SKILLS and COMPANIES AND TIMES tables. They limit the data that can be added to each of these tables. A database administrator will need to update them

b – What issues did you faced in creating your design?

Solved issues

The first issue was to lock the problem and settle on a definition for skill value as many perspectives are valid and a few had to be settled for.

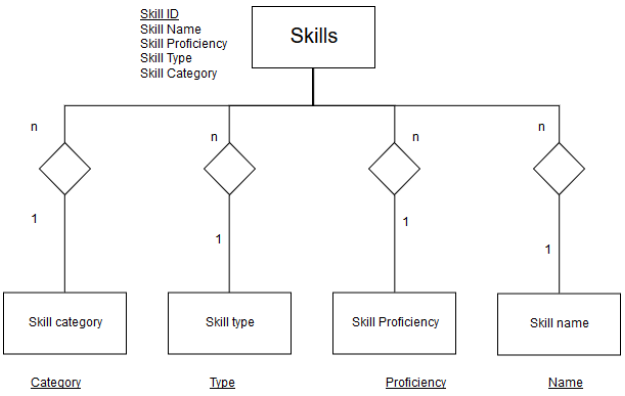
During the design phase We showed great enthusiasm for reference tables. Some of them had to be dropped as they didn't really have any utility and would have added maintenance.

When working on the SQL queries, we noticed that some of the data created with red gate did not make sense.

Existing Issues

Due to lack of time, little thought was given to the time dimension. Jobs, people, companies, skills will change with time, not paying attention to this factor could cause data inaccuracy in a couple of years.

When creating data with Red Gate, an issue appeared linked to reference tables. For example, data in the SKILLS tables made little sense. The SKILLS tables have multiple fields related to skills. Four of these fields are linked to reference tables. Skill Name, Skill Proficiency, Skill type and Skill Category. These four fields will be identical over multiple records. On the screen shot we can see 16 skills with the name Writing and identical Skill Proficiency, Skill type and Skill Category. This should not happen. As a skill name should have 5 proficiencies, 1 type and 1 Category. An answer might be the use of Unique keys. Due to time constraints this possibility has not been investigated.



SkillID	SkillName	SkillProficiency	SkillType	SkillCategory
988	Ability to work in a team	1	Hard	Business
760	Ability to work in a team	1	Hard	Business
686	Ability to work in a team	1	Hard	Business
658	Ability to work in a team	1	Hard	Business
467	Ability to work in a team	1	Hard	Business
395	Ability to work in a team	1	Hard	Business
289	Ability to work in a team	1	Hard	Business
163	Ability to work in a team	1	Hard	Business

With actual data, we doubt that this issue will happen, but it is a possibility and should be fixed.

The skill name consistency issue has not been addressed. Project management and managing projects would give separate results. This is an issue as it is the same skill. We would like users to be able to search for both project management and managing projects and get the same result.

Other issues certainly exist but have not been found out yet. We do not know what we do not know and are ready for surprises.

c – What areas would you improve if you had more time?

We would get actual data from LinkedIn, Glass Door and possibly other sources.

We would add a table or a column with information on the data source.

Add tables for Hobbies, Volunteering and recommendations.

Rethink the relationship between jobs and companies.

Other potential information linked to jobs, skills and resumes would have to be investigated further.

The existing issues would also need to be fixed.

d – Is your ER diagram in 3rd Normal Form? If so, why? If not, why not?

First Normal Form (NF)

We have the 1NF as all our columns are singled valued. A column value is linked to only one piece of information. E.g. Address information is split between address line 1, address line 2, city, zip, state and country. Each column contains a specific type of information.

The values stored in a column are of the same domain. First Name for the PERSONS tables only contains first name information.

All column names of the tables are unique.

The order in which the data is stored is not relevant.

Second and Third Normal Form

We have the 2NF and 3NF as we do not have partial dependencies. All columns of a table are linked to the primary key. There are no transitive dependencies. No column depends on a column depending on the primary key.

e – What grade do you think you deserve for this project?

95/100

We paid attention to details creating small python scripts to make sure the data would make as much sense as possible.

There is significant improvement between the first and last ER diagrams.

Many improvements could be made, but time constraints got the better of us.

The objective for the database has been reached.

Quality of your ER Design 30/30

The Final ER Diagram is clear and has all the relationships noted properly. 30/30

"Professionalness" & Cleanliness of your project 20/20

All documents are labeled appropriately, annexes where created, the project is easy to navigate

Quality of your write-ups & assessments 17/20

All documents are clear and concise. There can always be improvements, more graphs and images could be added, a summary should be created, the documents should be grouped in one single pdf file.

Accuracy of your SQL & quality of your reports 15/15

The SQL statements work as they are expected to, and the code is commented on needed.

Relevance of your project to the problem 9/10

The project enables answers for skill valuation based on different perspective, the database design enables some flexibility and already is set up to respond to a few potential future needs. The design puts constraints on data entries to maintain data accuracy.

Robustness of your data 4/5

The data was created with Red Gate. We wrote many Python scripts (see annex), to make sure the data made as much sense as possible. Further research was done to gather actual skill list, job titles …

There are still issues with the data as it was partially randomly generated. None the less, the database design holds the data perfectly and we can run queries.