INST 327 - Database Design and Modeling, Section 0203
Project Report - 5/12/24
Team 2
Samiha Naser, Maya Patel, Illia Polishchuk, Adrien Rozario, Andy Yang

Project Report on Polish IT Job Postings

Introduction

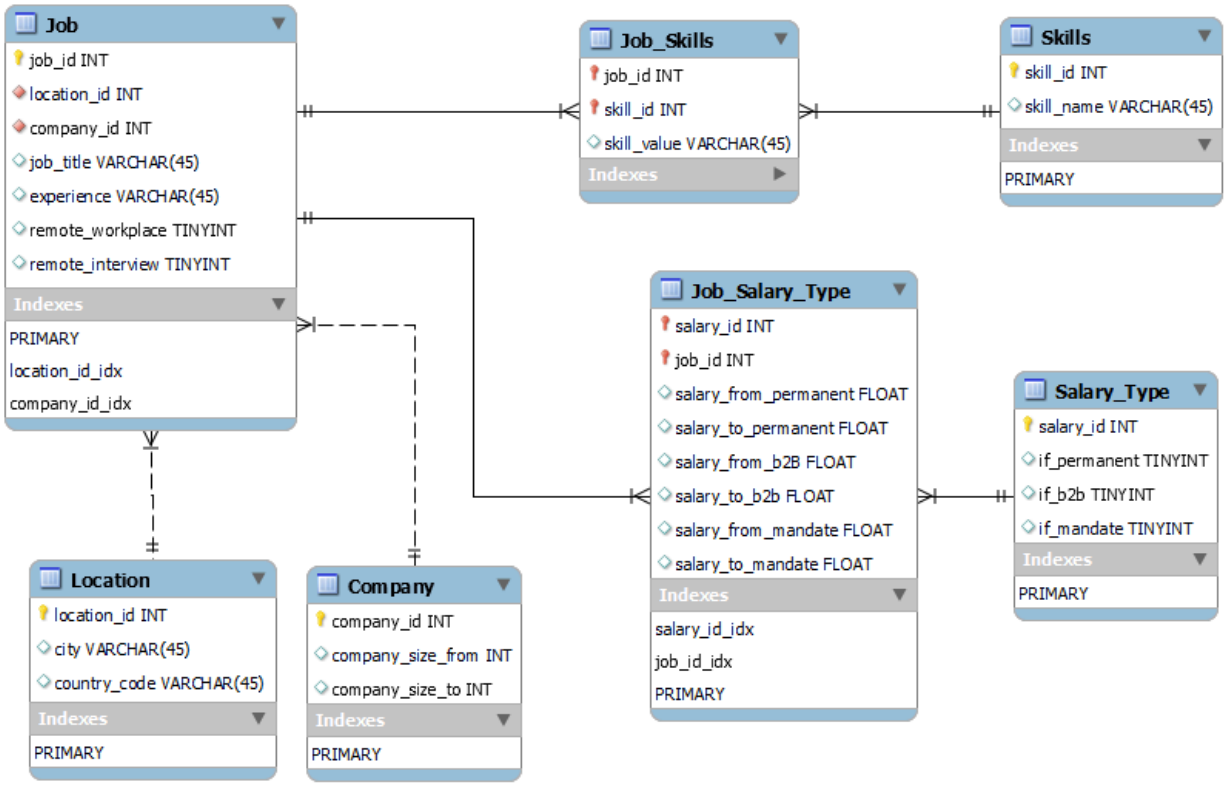
A rising issue that Information Technology (IT) careers face is the constantly changing job market, which can be hard to keep up with. Rife with competition, it is essential to identify current trends in striving towards being a successful job applicant. Ranging from company selection to job qualifications, many factors can contribute to a competitive portfolio. As such, our team aims to research which qualities are the most preferable for landing an IT job. In exploring these relationships, we seek to provide a better understanding for those who may be seeking work in the IT field.

Deriving our data from a Kaggle dataset, Polish IT Job Postings, this dataset was created by a Polish startup company and lists IT job postings for different European countries from February of 2022 to November of 2022. With 35 columns, some of the specific ones of interest are job title, workplace, salary, and experience. To focus our scope of interest, some repetitive and discriminatory columns are excluded. In representing our chosen columns, we developed a database in MySQLWorkbench using Structured Query Language (SQL) that showcases the specific relationships these entities and attributes have.

Database Description

This database is meant to serve as a tool for job applicants to find jobs that suit their qualifications and interests, consisting of 7 tables and a sample data size of 15 job postings. Each job posting retains information organized in such a way best represented by our entity relationship diagram below.

Logical Design



This database design organizes the significant details about the job, company, location, skills, a particular job's required skills, salary type, and a particular job's provided salary type into different tables. Looking at the table relations, location-to-job and company-to-job are one-to-many relationships, job-to-qualifications and job-to-salary_type are many-to-many relationships. Job and skills are connected by the job_skills linking table, and job and salary_type are connected by the job_salary_type table. As a job has both skills and a salary type, the job ID is a foreign key in both linking tables, alongside the skill ID and salary ID in their respective linking tables. In the job table, location ID and company ID are foreign keys referring to the values in the location and company tables respectively. Because all foreign keys are incremental ID values, they are non-null. This structure was finalized after 3rd normalization of the data, allowing for the most straightforward interpretation of the information; all information connects to the main job table. Any large set of job posting data can be utilized to an applicant's advantage in this format.

Physical Database

In order to use a sample of job posting data in this database, we had to import it based on its structure, which stems from the main table, job, and branches out to the location and company associated with it. The location and company table have no foreign keys so they were the first to be extracted, transformed, and loaded into the database using the MySQL Workbench import wizard. Similarly, the skill and salary table have no foreign keys so they were next to be imported. The only table with foreign keys that exist as primary keys in the already imported tables is the job table, thus it was next to be imported. Then, we looked at the linking tables, job_skills and job_salary_type, because they both contain a foreign key that exists as a primary key in the already imported job table. After all the tables were imported and the database was functional, we wrote queries and exported the file as a backup. Other team members were able to restore the database, view the tables, and run the queries, which will be discussed later.

Sample Data

Throughout the duration of the project, the scope of our data focus greatly narrowed due to the immense size of the Polish IT Job Postings dataset. The original dataset had over 37,000 rows of data, mostly with job listings in Poland and England. Between 1 and 2 percent of the rows corresponded to postings based in the United States, so in narrowing down our rows to under 500 rows, we chose only listings based in the United States which narrowed the dataset down to roughly 600 rows. We got rid of job postings with currencies other than the US dollar and those with extensive null values to populate our database with the strongest rows. Of these rows, 15 were chosen that best demonstrated a variety of desired job information. Our final sample data consists of these 15 rows and all 25 columns shown in the entity relationship diagram.

Views and Queries

Query Name	**JOIN (X4)**	**FILTER (X3)**	**AGGREGATE (X2)**	**LINKING (X1)**	**SUB-QUERY (X1)**
Q1: valuable_jobs	**X**				

Q2: marketable_skills	**X**	**X**	**X**		
Q3: big_company_cities	**X**	**X**			
Q4: average_highest_paying_cities	**X**		**X**	**X**	**X**
Q5: desired_experience_levels	**X**	**X**			

The following list describes what each query displays:

Query 1: Creates a view that shows the highest paying B2B jobs, ranked from highest to lowest paying jobs.

Query 2: Creates a view that displays cities with the most remote job opportunities, ranked from highest to lowest remote position count.

Query 3: Creates a view that lists companies from highest to lowest company size.

Query 4: Creates a view that depicts the average salaries for B2B jobs, ranking from highest to lowest average salary.

Query 5: Creates a view of the most valued skills and experience level for each job position, ordered from highest to lowest value.

Changes from Original Design

After months of reviewing and filtering our database, there are several changes we made to the original design to strengthen the clarity, focus, and effectiveness of our research. In regards to the database design, we refined the entities and tables, focusing on the most significant attributes to ensure it is relevant to any IT job applicant. While the original database included over 37,000 rows, we narrowed them down to around 300 in our original design, and 15 in our final. To achieve this, we excluded countries outside of the US and salaries that were not listed in US dollars. Additionally, we expanded on the diversity, equity, and inclusion considerations, emphasizing the importance of avoiding biases and promoting equal opportunities within the IT

job market. For instance, the “Open_to_Ukranians” column was removed due to our focus on creating a diverse and welcoming database. Other columns we excluded are the “if_other” salary type and currency because every row contains null values (false, 0, or unknown). The “marker icons” and the “currency exchange rate” columns were excluded because they are redundant. We also cleaned out rows with missing or irrelevant values in order to provide the highest quality of information.

Database Ethics Considerations

To ensure our database is inclusive within the IT job market, we've refined our criteria to prioritize diversity and prevent biases. Our approach involves collecting data to adapt to a wider variety of social, historical, and demographic factors. We're still focused on avoiding biases related to gender, ethnicity, and other demographic variables, ensuring that our database doesn't strengthen stereotypes or discrimination. We use inclusive language in our database, specifically in job qualifications, to avoid gendered or ethnically specific terms. None of the columns separate males or females, ensuring transparency in the hiring process. There is a column for companies who are open to hiring Ukranians, which can exclude potential applicants from Ukraine. This is a section we will attempt to improve and increase inclusivity in if possible. We still ensure that skill proficiency levels are defined objectively, along with regularly checking the database for biases in job postings, skill categorizations, and salary ranges to ensure equal representation. We also ensure that no candidates are excluded due to their socio-economic or educational backgrounds, since companies focus on practical skills and over degrees and areas of study. For this reason, we are not considering fields or tables for levels of education, such as college degrees. By implementing these measures, the database can promote fairness, inclusivity, and equal opportunities for individuals from diverse backgrounds in the IT job market.

After reviewing the database design our team came up with there does not seem to be any privacy, copyright, or ethical concerns. The database we are creating does not contain any sensitive data that can cause any harm in case of a privacy breach. Since the only information found in our database are job postings, company and position information, which is generally accessible by the public through resources like LinkedIn or Handshake, there are no legal or ethical problems with us using this data. Company names in this database are also excluded to avoid potential disputes. Sensitive or confidential information like salary for a certain position

could be a problem if it was disclosed by one of the current employees, however, since this data only contains information that companies were willing to disclose themselves when looking for new employees, using this data for our database is both ethical and legal. Our database is meant to be used by potential job seekers and people trying to analyze the job market for research reasons, hence it cannot cause harm.

Lessons Learned

Throughout the process of creating and setting up this database using the database management system, MySQL Workbench, our team has implemented numerous concepts from this course. This gave us an opportunity to practice normalization, create an entity relationship diagram, forward engineer, import data, write functional queries, backup and restore data. Along the way we ran into issues with normalization as it is an abstract concept, so we had to seek guidance on what the best normalized form looks like. We learned that normalization is the most simplest interpretation of the data and we did not have to include everything. We also had issues making sure the entity relationship diagram is an accurate representation of the data's relationships in order to successfully forward engineer. We learned that all tables include one primary key, linking tables don't include a primary key but instead include foreign keys, and ID values are non-null. At last, we resolved these issues and imported the data in the correct order thanks to the availability of resources provided in class. We learned that data is imported in a particular order, it can only be imported once, and it must not include the same primary keys if it is imported to an existing table. Through persistent efforts, consistent collaboration, and effective communication, our team divided and conquered the work. As a result, we produced an optimal database design and ensured the database runs smoothly.

Future Work

In evaluating considerations for the future of the project, there are multiple possible directions to take the database. Some potential future additions to this project include expanding the database to include all countries and currencies. In doing so, our scope will broaden to include IT job postings abroad for any applicants looking to apply to such jobs. That means creating more views with information relevant to job applicants abroad, and more table relationships that can be drawn out of different country data. For example, viewing the top skills

and experience valued in each country (if not all, then any select few). Another future implementation could be adding the currency exchange rate alongside the job salary and currency to help job applicants assess their options based on their desired pay and the current job market. Our team also discussed the possibility of utilizing this database as the framework for a job finding website, such as the website the original creator of the database used. The website would have a very simplified version of job information, but users would be able to see a variety of data relationships in order to satisfy their information seeking needs. In other words, the views we create would provide specific job posting statistics that the user can search through.

Citations

Just join it. Job Offers IT - Just Join IT. (n.d.-b). https://justjoin.it/

RSKriegs. "Polish It Job Board Data from 2022." *Kaggle*, 20 Nov. 2022, www.kaggle.com/datasets/kriegsmaschine/polish-it-job-board-data-from-2022/data.