INST 327
Section 0202
Team 10
Ashleigh Taylor, Shishir Poreddy, Timothy Chung, Yunlong Ou, Zoe Cheng

Team 10 Project Final Report

Introduction:

In order to create our database, we first had to understand all the information inside of the dataset that we chose. As a group, we collectively decided on the Polish IT Job Postings. This dataset contains information regarding a job opening all over Europe, such as workplace types, experience levels, and company size. After realizing that the dataset covered countries all over Europe, we wanted to create a more concise database that only covered one of the countries, which is why we chose to cover Poland's IT job listings. However, filtering the dataset to just jobs in one country won't fill the information need at hand. The information need that we wanted to fulfill was to see whether the position is remote, above a minimum salary, is for experienced workers, and for certain coding languages. We eventually fulfilled these needs by declaring that we will focus on postings in August 2022 that are either remote and hybrid, for senior level positions, and use any language other than .net.

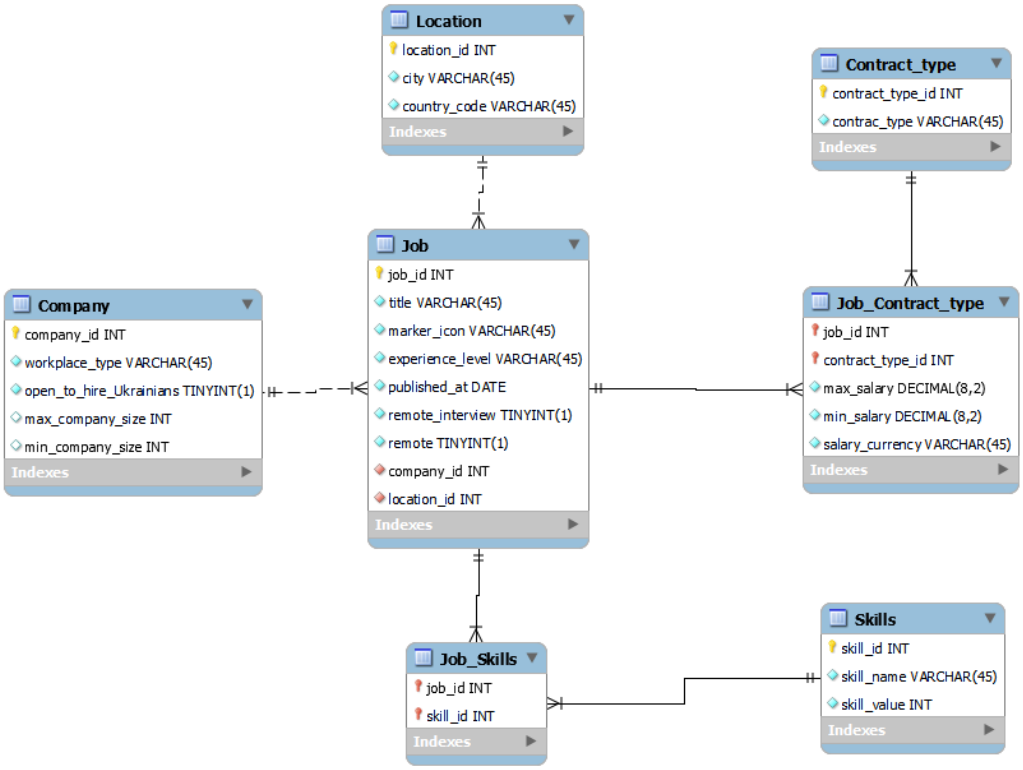
Database Description:

Logical Design:

The original scope of our project as stated in our proposal was to filter the dataset for only the IT job listings in Poland. After reading the reviews and feedback from our peers, we have decided to keep this as our project scope. More specifically, the categories we will store in our database are similar to those of the original data set, although they will be curated to fit our project's scope. Our database will provide all the listings in August 2022 that are remote and hybrid. Our database will also provide the job listings that are above a certain minimum salary amount and are only for senior level positions. Along with these characteristics of our database, the size of the company will be displayed. We want to provide our users with companies that are well established and have a substantial amount of staff already. Finally, the last characteristic of our project's scope is to provide our users with job listings that use more current and popular coding languages. We will remove any job listings that specialize in .net because of how outdated it is.

INST 327
Section 0202
Team 10
Ashleigh Taylor, Shishir Poreddy, Timothy Chung, Yunlong Ou, Zoe Cheng

ERD:



Physical Design:

Our database contains 7 tables, with the main one being the Job table. Starting from the Job table, which is the main table due to it having the most data that we will obtain from the database, we then created connections based on descriptions of the Job itself, resulting in the connections to the Company table, Location table, Job_Skills table, and the Job_Contract_Type table. While there are four connections to the main table, only two of them are connected to a linking table, Job_Skills and Job_Contract_Type. The Job_Contract_Type table is connected to the Contract_Type table via a many-to-one relationship, also containing information regarding salaries based on the type of contract and job. The Job_Skills table is connected to the Skills table using a many–to-one relationship, which contains information regarding the various skills that will be required and how valued those skills are.

Sample Data:

The sample data was from an online community platform called Kaggle which is used by people in the data analytics and machine learning field. We aim to create a customized, smaller, and more manageable dataset derived from the Polish IT Posting database for our sample data plan, focusing on all level job postings in Poland, during August 2022. To get the data from the data set imported into our database a team member went through and cleaned the data. This means removing null's, unknowns, and duplicate values from the data. We then imported this data going table by table starting with tables without foreign keys and then moving on to tables

with foreign keys from the initial tables. We continued this process until all the data was moved into the database below is a sample of the data in our Job Table.

Job ID	Title	Marker Icon	Experience Level	Published At	Remote Interview	Remote
1	Senior Data Engineer	data	senior	2022-08-11	1	1
2	Fullstack Software Engineer	javascript	senior	2022-08-11	1	1
3	Fullstack Software Engineer	javascript	senior	2022-08-02	1	1
4	Middle/Senior .NET Developer	net	mid	2022-08-01	1	0
5	Senior QA Engineer	testing	senior	2022-08-11	0	1
6	Full Stack Developer	javascript	senior	2022-08-02	1	1
7	Net developer	net	senior	2022-08-01	1	1

Views + Queries:

Query list:

1. **JobsOpentoUkrainians**: Query to find all jobs that are open to hire Ukrainians including the job title, company name, and contract type. The query includes job, company, job_contract_type, contract_type tables
2. **AvgSalariesByContract**: Query to find the average minimum and maximum salaries for contract type: b2b and permanent including contract_type, job_contract_type tables
3. **CitiesWithAboveOneJobs**: This query counts the distinct job titles for each city, then filters out cities that have more than one job including job and location tables
4. **HighPayingJobsForUkrainians:** This query will show the jobs that are open to hire Ukrainians and filter out jobs that have salaries greater than the average range of all jobs including job, company, job_contract_type, and contract_type tables

5. **LargerThanAvgMinCompaniesSize:** This query is to find all companies that have a minimum size greater than the average minimum company size including job and company tables.

View	Req.A	Req.B	Req.C	Req.D	Req.E
JobsOpentoUkrainians	X	X		X	
AvgSalariesByContract	X	X	X		
CitiesWithAboveOneJobs	X	X	X		
HighPayingJobsForUkrainians	X	X			X
LargerThanAvgMinCompaniesSize	X	X			

Changes From Original Design:

There has not been much change from the initial proposal. First of all, despite some feedback given by peers and TAs, we have decided to stick with our initial database plan as well as keep our scope limited to Poland as this allows us to keep the assumption of one location with our data. While this is a potential lack of diversity, equity, and inclusion, this can be avoided through the types of jobs and the other characteristics regarding the Polish IT jobs. We have also decided to focus on a specific type of job due to the excessive nature of the dataset as the original dataset is extremely large. With this, we have decided to focus on junior-level Java jobs that are also remote in Poland from the dataset.

Database Ethics Considerations:

Diversity, Equity, and Inclusion Considerations:

From our proposed database design and sample data plan, our envisioned database is going to be an inclusive information resource in its domain. As stated earlier, we are going to keep the scope of our database to IT job listings in Poland. We feel that our database will still be able to be diverse without including data from all over Europe. In order to accomplish this, we will be providing characteristics regarding the various Polish

IT job listings, such as level of experience, salary range, company size, and type of workplace. Because of this, our database design will be very diverse.

We will do our best to make our database as inclusive as possible in its design. Based on our feedback, some suggestions that we are planning on including are including demographics in the job listings. Including characteristics such as gender or ethnicity could provide a more accurate representation of different people who are working at the specific job listings. Some examples of biases that we may encounter are having data that is more directed to a specific country or city. Not all of our countries/cities are the same size and there may be more data found for a particular country or city as opposed to other countries/cities. This is something that we are going to be keeping in mind as we create our database. We will make sure to keep our database focused on Polish IT job listings as best as possible.

Data Privacy, Fair Use, Other Ethical Considerations:

Through the finalization of the database we have been able to look at the potential issues that arose in our proposal and now have a better understanding of if these are true issues or not. The first issue that was considered was two fold. Simply put we were worried that any data regarding users of the database might be used to an advantage of either other users or companies. After further development of the database we do not see this as an issue as we will not be keeping track of people who use the database, which negates the issue of their data being used. Regarding users of the database we were also worried about identifiable information involving the user being leaked, but again since we will not be storing any user data this is not an issue. The last issue discussed was that of in-house job postings. We have decided not to move forward with trying to obtain or use any information that is not publicly available so there will not be any fair use or privacy issues regarding that data. Throughout the process we have kept the issues of Data Privacy and Fair/Ethical Use at the forefront of our minds which has allowed us to create a database that should not harbor any issues regarding the subject.

Lessons Learned:

INST 327
Section 0202
Team 10
Ashleigh Taylor, Shishir Poreddy, Timothy Chung, Yunlong Ou, Zoe Cheng

Throughout the course of working on creating our database, we have faced many challenges that as a group have overcome. Facing these challenges is what allowed us to learn and problem solve as a group. Many of us have busy schedules which makes it difficult to meet up as a group and collectively come to decisions. However, through consistent communication and trust in one another we were able to communicate effectively and meet every deadline. In order to plan meeting times we would communicate our availability at the beginning of the week, which allowed us to plan future meetings. Throughout the semester we became a lot more organized, which made managing everyone's schedules a lot easier. Ultimately, we learned a lot of practical skills that come with working in teams that will translate to our future jobs.

Along with an improvement in practical skills, we learned the importance of diversity and inclusion when designing databases. During the review of our peer groups, we understood that it can be easy to forget about certain factors, which could lead to certain types of bias. After the review, we had to make sure we didn't accidentally implement a bias. After thorough review we were able to conclude that we didn't accidentally implement a bias. In conclusion, we learned that acknowledging diversity and inclusion on a grand scheme is important because of how easy it is to accidentally implement bias or accidentally disinclude specific groups.

Potential Future Work:

After the completion of the database, potential future implementations include incorporating further diversity. For our project we made our scope specifically focused in Poland, however by expanding it to other countries and regions outside of Poland it would increase diversity and inclusion. An increase in diversity and inclusion would allow us to properly represent a larger portion of Europe, instead of minimizing it to just one country. Another future implementation would be to create visualizations that represent the data that we filtered into our database. Implementing visualizations would allow the user to better understand the data we filtered into our database. Having a visual representation of any kind of work is always beneficial in increasing anyones understanding on a topic. Finally, after potentially incorporating a larger scope and data visualizations, analyzing the data for trends could be another potential implementation. If we are able to analyze the data and eventually recognize trends it would be beneficial to any companies hiring, allowing them to change anything in their process they see isn't working.

INST 327
Section 0202
Team 10
Ashleigh Taylor, Shishir Poreddy, Timothy Chung, Yunlong Ou, Zoe Cheng