INST 327, Section WB21 – Database Design and Modeling

Final Report

Team 10

Joanne Jeong, David Villeda, Tarun Gunaseelan, Jessica Fayer

8/19/23

Introduction

The Information Technology (IT) sector is dynamic, constantly presenting both employers and job seekers with challenges related to finding appropriate matches. As the reliance on digital platforms grows, the need for an organized, accessible, and comprehensive database for job postings becomes paramount. Recognizing this, our team initiated the development of a specialized database focusing exclusively on permanent IT job positions.

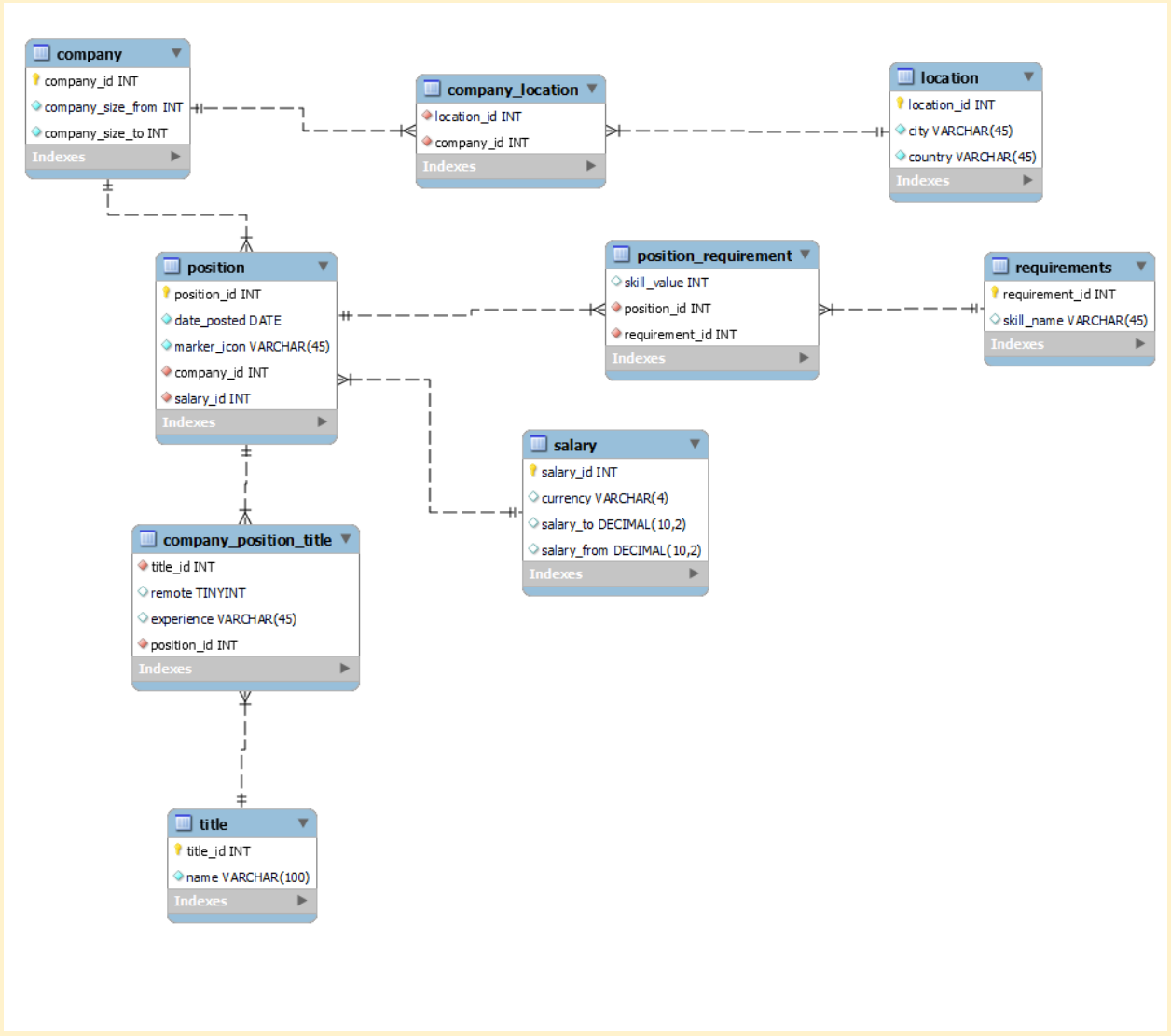
The primary data source for this endeavor was the "Polish IT Job Postings" dataset. Our objective was not limited to the Polish IT market but extended to relevant job postings from other countries within our project's scope. Early in the project's lifecycle, the importance of data accuracy and relevance became evident, prompting a stringent data refinement process.

The resulting database comprises nine tables, each meticulously designed to capture specific facets of job postings, from company details to job requirements. Throughout the development process, continuous evaluations and consultations ensured the integrity and utility of the database design.

This report provides a detailed account of the project's progression, from inception to completion, emphasizing challenges encountered, methodologies employed, and the rationale for critical decisions.

Database Description

Logical Design



Sample Data

The data in our database was obtained via an online community platform named Kaggle, this dataset in particular contains data extracted from JustJoin.IT, a Polish startup that lists IT job postings from across Europe. This dataset was developed by Kaggle user *rskriegs* with BigQuery. Each row of the data represents a specific job posting found by the user, The job postings date from February of 2022 to November of 2022. For the sake of accurately selecting data that pertains to our targeted audience, we reduce the data of 37,789 rows to a sample data of

230 rows containing job postings of a more current dating (August 2022 - November 2022) We did not however include job posting publishing dates due to a formatting issue and deemed it unnecessary for anything queries we expected to run.

position_id	name	company_id	title_id	remote
230	Database Admin	72	196	1
229	Project Manager	72	195	1
228	Fullstack Developer (.NET + Angular)	72	194	1
227	Full-stack engineer	71	193	1
226	Frontend Tech Lead (Typescript, React)	71	192	1
225	Backend Tech Lead (Javascript/ Python)	71	191	1

Image 2. A snippet of company_position_title joining the title table to display the total number of job postings along with their designated IDs and names.

Views/Queries

Query Name	**JOIN (x3)**	**FILTER (x3)**	**Aggregate (x2)**	**LINKING (x1)**	**SUB-QUERY (x1)**
company_average_salary	**X**		**X**		
jobs_per_country	**X**			**X**	
polish_min_max_mean		**X**	**X**		
posting_average_salary	**X**		**X**		
remote_jobs_number		**X**			
valued_skills					
companies_above_avg_salary	**X**	**X**			**X**
Total	**4**	**3**	**3**		**1**

The following list describe what each query we wrote for our database displays

- company_average_salary
    - Creates a view which displays the average salary per company
- jobs_per_country
    - Creates a view which displays number of IT jobs per country
- polish_min_max_mean

    - Creates a view which displays the minimum, mean, and maximum values of salaries for Polish IT positions
- posting_average_salary
    - Creates a view which displays
- remote_jobs_number
    - Creates a view which displays number of remote jobs
- valued_skills
    - Creates a view which displays most valued/sought after coding languages/skills
- companies_above_avg_salary
    - Creates a view which displays

Changes from Original Design

Given that our selected target audience in regards to the “Polish IT Job Postings” would include possible job applicants who are looking for a job, we have reduced the data to look at only current job postings limited only to the calendar year 2023. While importing CSV files into the database we ran into a few complications/errors which resulted in changes in table data types, including the deletion of the “dated_posted” value which was discarded due to an importing error and deeming it as unnecessary information that had no relation to any other tables. The name column in the title table which originally had a datatype of VARCHAR(45) was increased to VARCHAR(100), salary_from and salary_to columns from the salary table were changed from INT data types to DECIMAL(10,2) as some of the salaries contained decimal values which were being altered when imported. We also discovered some salary fields and company size fields to be empty and realistically speaking not all job postings have a posted salary range or even publicly displayed company size we changed salary_from and salary_to to allow NULL values as well as company_from and company_to.

Database Ethics Considerations

The “Polish IT Job Postings" in regards to ethics had some original values such as availability for Remote interviews and Remote work which we believe would be very inclusive to those who are not only in a location that may be of large distance from the worksite but also beneficial to those who may have any disabilities or other hardships which would make it

difficult to interview or work onsite. Our target audience did not have any specificity in regards to gender but we were able to find a couple of job titles that had gender pronouns He/She/They which we believe was to signify gender inclusivity, gender however was not a value found within the CSV.

Another value we noticed was a value for hiring Ukrainians which we assume was for Ukrainian refugees however the CSV didn't state whether or not they'd be sponsoring visas. In general database ethics considerations of diversity, equity, inclusion, data privacy, or fair use were not relevant to the project data as none of these considerations were relevant to the scope of our project, things such as gender considerations within the technology field or pay discrepancies would have been good data which could have been included in relation to our scope but the CSV file didn't contain any. If one wanted to, they'd likely be able to make a comparison between our database as a baseline and other datasets containing gender values to see if the average male or females (or non-binary gendered persons) have a different annual rate than those recorded in ours.

Lessons Learned

During the development of our database, our team encountered several challenges that offered valuable learning experiences. One of the most prominent issues surfaced during the logical design phase, specifically regarding the normalization of our tables. Our initial database structures, while comprehensive, showed redundancies that had the potential to compromise data integrity in future interactions. The importance of refining our design to align with normalization principles became evident, as this would guarantee efficient data storage and retrieval.

However, our efforts were initially hindered by limitations in the dataset. The dataset, comprehensive as it was, did not provide distinct identifiers for companies, such as their names. We had access only to generic data like company size. This lack of specific identifiers posed a challenge in effectively differentiating and categorizing companies in our design.

Our interactions with Professor Duffy proved to be instrumental at this juncture. Through detailed consultations, we gained a more lucid understanding of normalization techniques. Guided by Professor Duffy's expertise, we restructured our design to adhere closely to normalization standards, ensuring a well-organized and efficient table structure.

The physical design phase brought forth its own set of challenges. We adopted an iterative approach, consistently evaluating and refining our design based on emerging insights. One pivotal discussion centered on the relevance of certain data columns, especially those related to part-time positions and mandate work. After thorough deliberation, keeping in mind our primary focus and the potential requirements of end-users, we opted to exclude certain columns. This decision was rooted in our aim to maintain the database's focus squarely on permanent IT job positions.

Reflecting upon this journey, these challenges, while demanding, deepened our grasp of database design. They underscored the significance of adaptability, effective mentorship, and maintaining clarity of purpose throughout the development process.

Potential Future Work

Given more time and resources, we would like to increase the size and add more columns to our dataset. The given dataset did not provide information such as the company names, website, contact information, education requirements, or specific addresses. To incorporate this new data, we would most likely have to retrieve the data ourselves from various job sites. The company name alone would have been a great aid in the development of this database, as we had little to differentiate between the companies besides their sizes.

In addition to adding more columns to our data site, we would also like to advance the company's profiles. We would like to expand beyond incorporating companies' profiles and go deeper into building inclusive company profiles. Some examples of the data we can collect will be company values, company history, employees' experiences, and consumers' experiences.

With this feature, it will offer users a more detailed insight. With future enhancements, and more time and resources, we were able to recognize the potential of our database to make it more inclusive for our users. In the future, we envision further enhancement of our specialized IT database. The database can expand on a global surface and be accessible to different countries and regions. To bring traction to the database and make it more global expansion. To make this feature accessible, the database would have to adapt for the users, such as different job markets, new regions and locations, and language requirements.

We strive to make the data site ethically correct. For future enhancements, we want the database to gain prominence. To exceed this, we can explore different issues that are present in

the IT job marketplace. The database can explore different ethical issues such as equality and diversity. to examine data on job salaries, gender roles, and salary discrepancies. This can provide useful information on improving and encouraging diversity in the IT industry.

Conclusion

In our endeavor to develop a specialized database for IT job postings, our primary objective was to create a platform that would be easily accessible to job seekers. This platform would serve as a comprehensive resource for researching potential employers and understanding various components of the IT job market. The initial phase of our project illuminated the vast intricacies involved in creating a robust database. Throughout this journey, we faced multiple challenges, especially in refining the database design to ensure its reliability and efficiency.

These obstacles, however, proved invaluable. They drove us to adapt, understand the core needs of our project, and seek effective solutions. Reflecting upon our journey, several key lessons stand out. The essence of collaboration, the importance of seeking guidance, the value of time, and the virtue of patience were all reinforced. These learnings played a pivotal role in shaping our database to its current form. As we look forward, we are optimistic about the potential of our project. We aspire to continually enhance our database to serve the IT community better. Our ultimate goal is to offer job seekers a platform where they can find promising employment opportunities and connect with companies that uphold ethical and sustainable practices.