Blue text on a black background

Description automatically generated2023/2024

## **Base de Dados Avançadas**

**Project Report**

Group 2

Artur Costa nº 62740

Enzo Souza nº 62744

Ivo Estrela nº 51051

## **Description of the Dataset**

This dataset encompasses an extensive compilation of over 33,000 job listings. Each specific posting is equipped with 27 significant characteristics, encompassing details like job title, job description, salary, location, application URL, and work types. Additionally, there are distinct files for benefits, skills, and industries related to each job listing. Most positions are correlated with a corresponding company, the details of which are provided in a separate CSV file. This file includes attributes such as company description, headquarters location, number of employees, and follower count. To streamline and to not bloat our project we decided to not use some of the separate csv files mentioned above, with the remaining files being the following:

### **job\_postings.csv**

* **job\_id**: The job ID as defined by LinkedIn (<https://www.linkedin.com/jobs/view/> job\_id )
* **company\_id**: Identifier for the company associated with the job posting (maps to companies.csv)
* **title**: Job title.
* **description**: Job description. - **REMOVED**
* **max\_salary**: Maximum salary
* **med\_salary**: Median salary
* **min\_salary**: Minimum salary
* **pay\_period**: Pay period for salary (Hourly, Monthly, Yearly)
* **formatted\_work\_type**: Type of work
* **location:** Job location
* **applies**: Number of applications that have been submitted
* **original\_listed\_time**: Original time the job was listed
* **remote\_allowed**: Whether job permits remote work
* **views**: Number of times the job posting has been viewed
* **job\_posting\_url**: URL to the job posting on a platform
* **application\_url**: URL where applications can be submitted
* **application\_type**: Type of application process
* **expiry**: Expiration date or time for the job listing
* **closed\_time**: Time to close job listing
* **formatted\_experience\_level**: Job experience level
* **skills\_desc**: Description detailing required skills for job
* **listed\_time**: Time when the job was listed
* **posting\_domain**: Domain of the website with application
* **sponsored**: Whether the job listing is sponsored or promoted.
* **work\_type**: Type of work associated with the job
* **currency**: Currency in which the salary is provided.
* **compensation\_type**: Type of compensation for the job.

### **companies.csv**

* **company\_id**: The company ID as defined by LinkedIn
* **name**: Company name
* **description**: Company description
* **company\_size**: Company grouping based on number of employees (0 Smallest - 7 Largest)
* **country**: Country of company headquarters.
* **state**: State of company headquarters.
* **city**: City of company headquarters.
* **zip\_code**: ZIP code of company's headquarters.
* **address**: Address of company's headquarters
* **url**: Link to company's LinkedIn page

### **employee\_counts.csv**

* **company\_id**: The company ID
* **employee\_count**: Number of employees at company
* **follower\_count**: Number of company followers on LinkedIn
* **time\_recorded**: Unix time of data collection

### **benefits.csv**

* **job\_id**: The job ID
* **type**: Type of benefit provided (401K, Medical Insurance, etc)
* **inferred**: Whether the benefit was explicitly tagged or inferred through text by LinkedIn

### **salaries.csv**

* **job\_id**: The job ID
* **salary\_id**: The salary id
* **max\_salary:** Maximum salary
* **med\_salary:** Median salary
* **min\_salary:** Minimum salary
* **pay\_period:** Pay period for salary (Hourly, Monthly, Yearly)
* **currency: Type of currency**
* **compensation\_type: Method of compensation**

## **Scheme for both Databases**

MySQL

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generatedMongoDB

## **Discussion**

Queries Simples:

3a - 1

A black background with white text

Description automatically generatedEsta consulta busca os salários que são pagos anualmente, com um valor superior a 10000 e com um salário máximo superior a 50000.

**MySQL** - Results with no optimization:

A screenshot of a computer

Description automatically generatedFor optimization of this query, we decided to use a composed index in pay\_period, max\_salary, min\_salary to help the query Here are the following results:

A black background with white numbers and symbols

Description automatically generatedSince the table salaries is small, the difference in execution time is negligible. Nevertheless, the number of rows analyzed dropped from 259 (All rows) to 119.

**Mongo** -Results with no optimization:

A black background with white text

Description automatically generatedWe used the same optimization strategy to produce a slight improvement in execution time:

3a - 2

A próxima utiliza uma expressão de tabela comum e a função de janela ROW\_NUMBER() para classificar as empresas com base no número de ofertas de emprego em cada localização. O resultado inclui apenas as empresas com o maior número de ofertas de emprego em cada local.

Queries Complexas

3b - 1

Esta consulta recupera informações sobre empresas, incluindo o seu ID, nome, número de empregados e o número de ofertas de emprego em que o título contém "er". Também filtra as empresas com mais de 5 ofertas de emprego e ordena os resultados pela contagem de ofertas por ordem decrescente. A utilização de LEFT JOINs garante que as empresas sem entradas correspondentes nas tabelas benefits ou employee\_counts continuam a ser incluídas nos resultados.

3b - 2

Essa consulta calcula os valores médio, mínimo e máximo da coluna max\_salary da tabela salaries, considerando apenas as linhas em que os anúncios de emprego correspondentes têm um max\_salary maior que 5000. O RIGHT JOIN garante que todas as linhas da tabela job\_postings sejam incluídas, e as linhas correspondentes da tabela salaries sejam incluídas com valores NULL se não houver correspondência.

UPDATE benefits SET type = 'test' WHERE type = 'Medical insurance' LIMIT 10;

Esta consulta atualiza até 10 linhas na tabela benefits, alterando o valor da coluna type de 'Medical insurance' para 'test'. O LIMIT 10 garante que apenas um máximo de 10 linhas sejam atualizadas, mesmo que haja mais linhas que satisfaçam a condição.

INSERT INTO companies

(company\_id, name, description, company\_size, state, country, city, zip\_code, address, url)

VALUES (1, 'Empresas Empresas', 'Fazemos tudo e mais alguma coisa', 15, 'CA', 'USA', 'Los Angeles', '2625-136', 'Rua 29 de Fevereiro', 'https://www.example.com');

#INSERT into multiple tables. ???

Trade offs:  
Keep in mind that adding indexes comes with some trade-offs, such as increased storage space and potential performance overhead during write operations. It's essential to analyze the specific requirements of your application and database workload to determine the optimal indexing strategy. Additionally, monitor the performance after creating indexes to ensure they have the desired impact.