Pontíficia Universidade Católica do Rio de Janeiro

Mariane Macedo Viola

Pós-graduação em Ciência de Dados e Analytics

Análise das habilidades solicitadas em vagas de emprego na Holanda.

Rotterdam, 2025

Trabalho de conclusão da sprint de Engenharia de dados que consiste em um MVP para avaliação das habilidades mais solicitadas no mercado de trabalho da Holanda, para a instituição PUC Rio e acompanhamento do orientador Victor Almeida.

SUMÁRIO

1. Objetivo

Com uma recente mudança para a Holanda, desejo entender a dinâmica do mercado de trabalho local. O objetivo é identificar quais habilidades (skills) são mais demandadas no país e qual o peso das principais habilidades no conjunto total de requisitos.

O pipeline de dados deverá ser capaz de responder as seguintes perguntas:

a. Demanda por habilidades: Quais são as 10 habilidades mais citadas nos anúncios de emprego no conjunto de dados?
b. Categorias: Qual a distribuição de vagas por categoria?
c. Foco em software: Qual a porcentagem de todas as habilidades exigidas que são classificadas como relacionadas a software?

2. Coleta

a. Fonte de dados e procedência

O conjunto de dados selecionado para o projeto é derivado da base pública lightcast_lightcast_job_postings_global_sample.global_sample, que é disponibilizada através do catálogo de Public Datasets no ambiente Databricks Free Edition.

b. Persistência na nuvem

A base de dados original é sujeita a atualizações diárias. Para garantir um escopo de análise estático e reprodutível a base foi extraída em formato CSV nomeada como *NL_jobs* e subsequentemente carregada para o catálogo *jobs* no ambiente do Databricks, sob o nome de tabela *nl_jobs*.

Este procedimento forçou a parada da atualização da base de origem para os propósitos deste MVP.

c. Armazenamento técnico

Para documentar a persistência no pipeline de ELT, o passo de Coleta foi formalizado através da criação da camada Bronze no catálogo *jobs*. Esta camada é uma cópia exata do CSV persistido, garantindo que o dado original esteja armazenado e acessível para as próximas etapas de transformação:

Processo: nl_jobs (CSV) → jobs.bronze.dados_brutos (Data Lake Persistido)

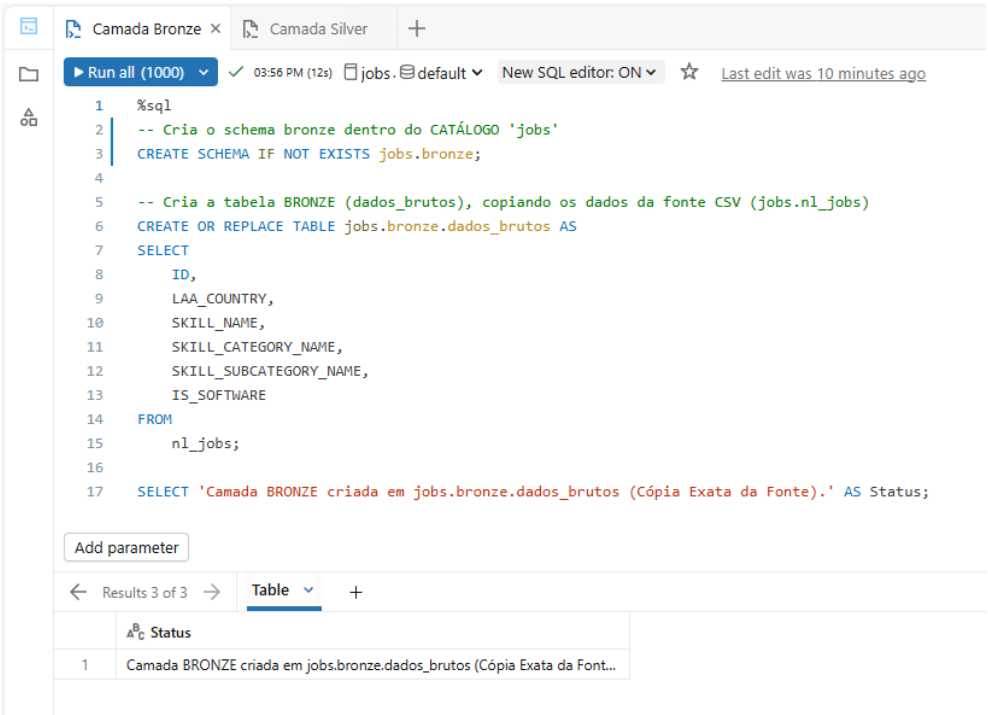

Imagem 1 – evidência query criação da camada Bronze.

3. Modelagem

A modelagem de dados foi realizada para estruturar os dados da camada Silver e otimizar as consultas analíticas, resultando na camada Gold *(jobs.gold)*. O modelo foi construído utilizando o esquema estrela.

a. Estrutura do esquema estrela

Composto por uma tabela fato central *(jobs.gold.fato_demanda)* que mede a exigência de habilidades por vaga, e está ligada a uma tabela de dimensão de suporte *(jobs.gold.dim_habilidade)* que descreve as características das habilidades.

A relação de modelagem é de 1 (dimensão) para N (fato), com a lógica de que cada habilidade única *(dim_habilidade)* pode ser exigida em muitas vagas *(fato_demanda)*.

b. Criação de chaves

A chave primária *(ID_HABILIDADE_HASH)* foi gerada utilizando a técnica de Hashing *(SHA256)* a partir da concatenação de todos os atributos descritivos da habilidade.

c. Documentação do catálogo de dados

Coluna	Descrição Detalhada	Domínio e Tipo	Detalhes Específicos / Valores Esperados
ID	Chave da Vaga.	String	Chave Estrangeira (FK) que compõe a Chave Composta do Fato.
ID_HABILIDADE_HASH	Chave Primária (PK). Identificador único e técnico da habilidade.	String (SHA256)	Gerada via Hashing (SHA256) na Camada GOLD.
ID_HABILIDADE_HASH	Chave da Habilidade.	String (SHA256)	Chave Estrangeira (FK) que se conecta à DIM_HABILIDADE.
LAA_COUNTRY	Código ISO do País. Atributo utilizado para filtragem.	String Categórica	Valor Mínimo/Máximo Esperado: No Catálogo SILVER, o domínio é fixado em apenas um valor: 'NL' (Holanda).
SKILL_NAME	Nome da Habilidade.	String Categórica	Categorias Possíveis: Qualquer termo, padronizado para *Uppercase*.
SKILL_CATEGORY_NAME	Categoria principal de Negócio.	String Categórica	Categorias Possíveis: Os 30 valores distintos da origem.
IS_SOFTWARE	Indicador se a habilidade é classificada como Software.	BOOLEAN	Domínio de Valores: TRUE ou FALSE.

Imagem 2 – Catálogo de dados das tabelas Dimensão e Fato e Atributo de filtro.

d. Linhagem dos dados

Camada Bronze *(jobs.bronze.dados_brutos)*

- Origem: Arquivo CSV *(nl_jobs)*, extraído da base pública.
- Técnica para compor o conjunto de dados: Cópia direta do dataset (Extração) e persistência no Data Lake.

Camada Silver (jobs.silver.vagas_skills_limpas)

- Origem: *jobs.bronze.dados_brutos*
- Técnica para compor o conjunto de dados:
  Filtro de Escopo *(LAA_COUNTRY)* - O campo de país foi utilizado no WHERE para restringir o escopo do projeto à Holanda *('NL')*.
  Qualidade - Padronização *(upper/trim)* e remoção de nulos *(is not null).*

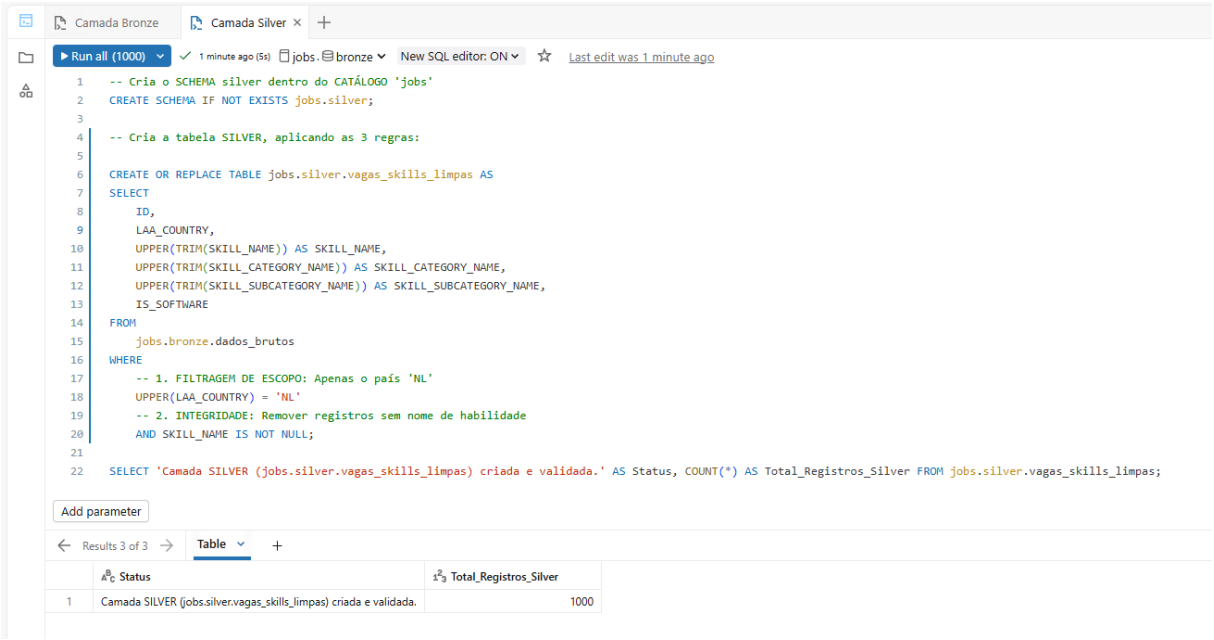

Imagem 3 - evidência query criação da camada Silver

Camada Gold *(jobs.gold.dim_habilidade / fato_demanda)*

- Origem: *jobs.silver.vagas_skills_limpas*
- Técnica para compor o conjunto de dados: Modelagem dimensional - Implementação do esquema estrela.

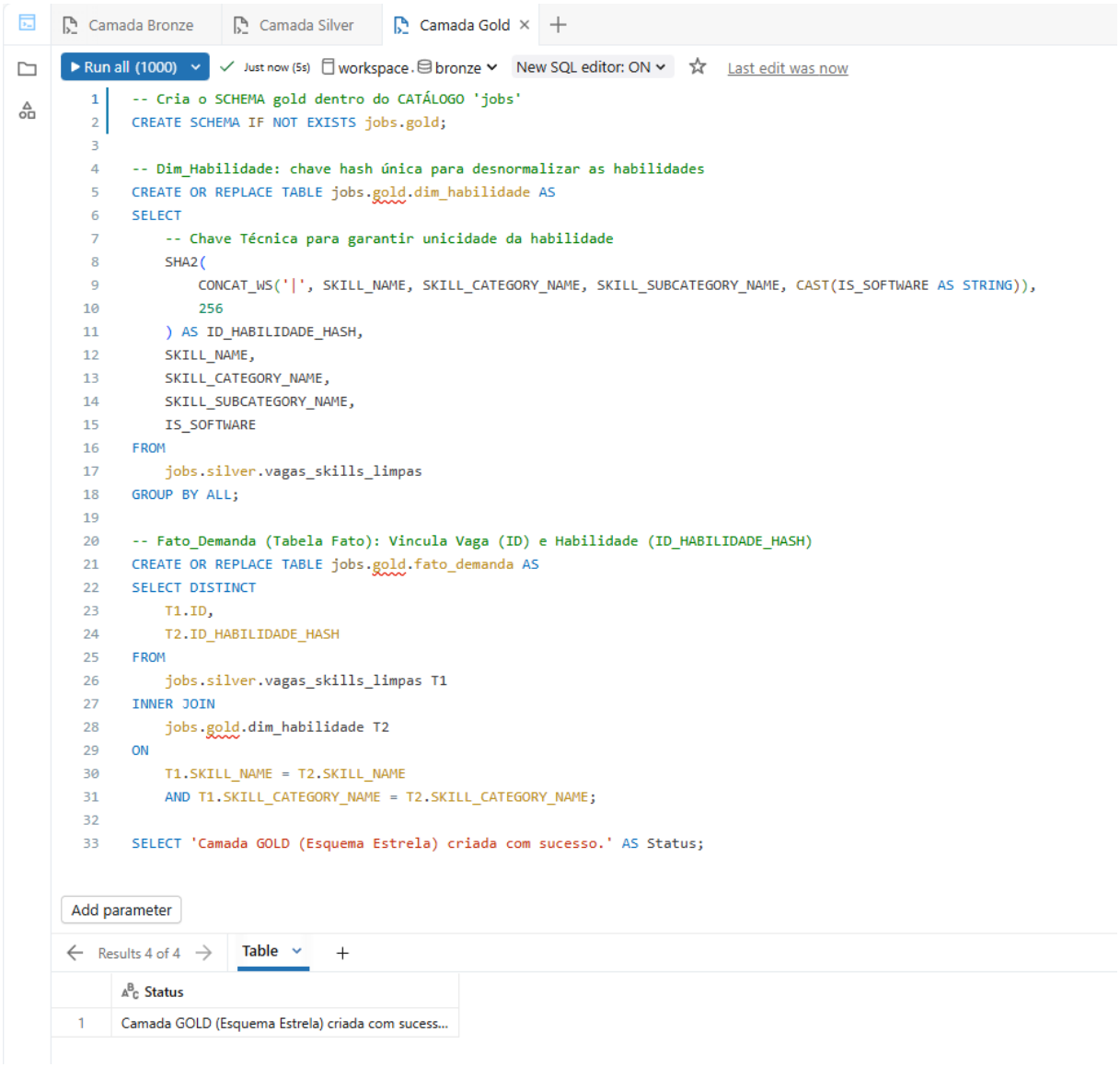

Imagem 4 - evidência query criação da camada Gold

4. Carga

a. Pipeline ETL

O pipeline de ETL (Extração, Transformação e Carga) foi implementado na plataforma Databricks, realizando a transformação dos dados brutos da camada Bronze para a camada Silver, onde foram aplicadas as regras de qualidade e de negócio.

b. Persistência dos dados

A persistência dos dados é garantida pelo uso do comando *CREATE OR REPLACE TABLE.*

A tabela *jobs.silver.vagas_skills_limpas* foi criada e persistida no Data Lake, garantindo que os dados transformados estejam estáveis e acessíveis para a próxima etapa (Camada Gold).

c. Documentação de processo de transformação

- Filtro de escopo: *LAA_COUNTRY*
  Detalhe: Aplicação de filtro *WHERE* para manter apenas o país *'NL'* (Holanda).
- Integridade: *SKILL_NAME*
  Detalhe: Remoção de registros onde o nome da habilidade era nulo *(SKILL_NAME IS NOT NULL).*
- Padronizaçao: *SKILL_NAME; SKILL_CATEGORY_NAME; SKILL_SUBCATEGORY_NAME*
  Detalhe: Uso das funções *UPPER/TRIM* para converter todo o texto para letras maiúsculas e remover espaços indesejados.

d. Código SQL executado **(evidência vide imagem 3)**

```
CREATE SCHEMA IF NOT EXISTS jobs.silver;                                    Criar o SCHEMA silver

CREATE OR REPLACE TABLE jobs.silver.vagas_skills_limpas AS      Cria a tabela SILVER, lendo da BRONZE
SELECT
  ID,
  LAA_COUNTRY,
  UPPER(TRIM(SKILL_NAME)) AS SKILL_NAME,                                    Padronização
  UPPER(TRIM(SKILL_CATEGORY_NAME)) AS SKILL_CATEGORY_NAME,                  Padronização
  UPPER(TRIM(SKILL_SUBCATEGORY_NAME)) AS SKILL_SUBCATEGORY_NAME,            Padronização
  IS_SOFTWARE
FROM
  jobs.bronze.dados_brutos                                          Fonte na Camada BRONZE
WHERE
  UPPER(LAA_COUNTRY) = 'NL'                                                 Filtro de Escopo
  AND SKILL_NAME IS NOT NULL;                                               Integridade
```

5. Análise

a. Qualidade dos dados

A análise da qualidade de dados foi realizada de forma proativa durante as etapas do pipeline ETL, onde foram identificados e resolvidos três problemas principais no conjunto de dados:

- Atualização: o *dataset* tem atualização diária, então para garantir a imutabilidade da base de dados foi necessária a extração em formato *CSV*.
- Escopo: o *dataset* é global, mas como o objetivo era a análise apenas das vagas na Holanda foi necessário o filtro de *WHERE (LAA_COUNTRY) = 'NL'*
- Habilidades vazias: o conjunto de dados contava com muitos campos indispensáveis em branco, portanto foi necessário a remoção dessas tuplas com *SKILL_NAME IS NULL*.
- Escritas múltiplas: tínhamos campos de habilidades com case sensistivity, o que identificava maiúsculas e minúsculas de maneira diferente e, portanto, diferenciava as classificações. Corrigido com a função *UPPER*.

b. Solução do problema

As perguntas de negócio foram respondidas através de queries na linguagem SQL, utilizando o modelo dimensional (Esquema Estrela) da Camada Gold.

Query 1 - Quais são as 10 habilidades mais citadas nos anúncios de emprego no conjunto de dados?

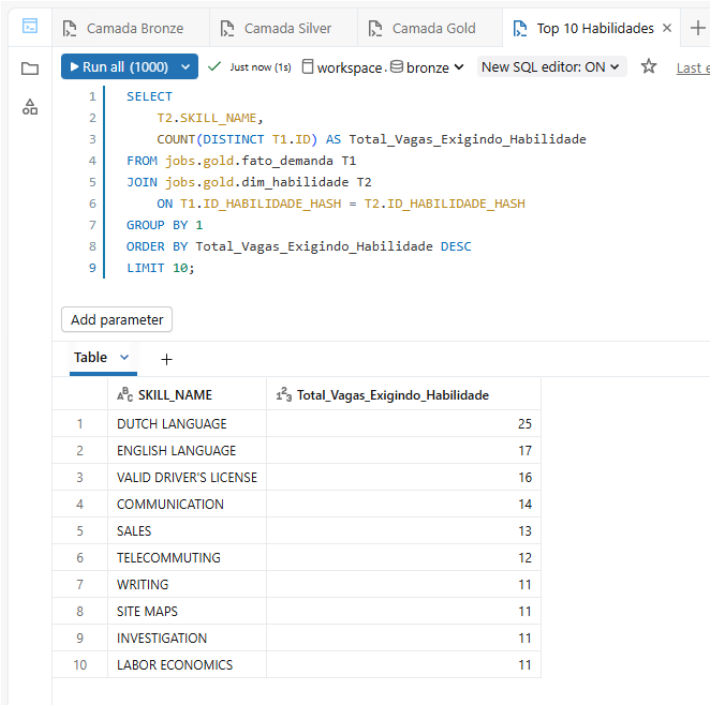

Imagem 5 – evidência query para top 10 habilidades

O resultado surpreendente indica que as habilidades mais demandadas no mercado holandês neste conjunto de dados são linguísticas e comuns, não técnicas. A alta exigência de *Dutch language* (idioma Holandês) e *English Language* (idioma Inglês) sugere que a fluência em ambos os idiomas é um pré-requisito básico e recorrente para a maioria das vagas.

Também fica evidente a pulverização de habilidades solicitadas já que em um conjunto de 1.000 habilidades solicitadas, a skill que mais aparece esta presente em apenas 25 delas.

Query 2 - Qual a distribuição de vagas por categoria?

Camada Bronze | Camada Silver | Camada Gold | Top 10 Habilidades | Vagas por Categoria

Run all (1000) · Just now (1s) · workspace.bronze · New SQL editor: ON · Last edit was now

```
WITH TotalVagas AS (
    SELECT COUNT(DISTINCT ID) AS total_vagas FROM jobs.silver.vagas_skills_limpas
)
SELECT
    T2.SKILL_CATEGORY_NAME,
    COUNT(DISTINCT T1.ID) AS Total_Vagas_Impactadas,
    ROUND(COUNT(DISTINCT T1.ID) * 100.0 / (SELECT total_vagas FROM TotalVagas), 2) AS Porcentagem_Vaga
FROM jobs.gold.fato_demanda T1
JOIN jobs.gold.dim_habilidade T2
    ON T1.ID_HABILIDADE_HASH = T2.ID_HABILIDADE_HASH
GROUP BY 1
ORDER BY Total_Vagas_Impactadas DESC;
```

Add parameter

Table

	SKILL_CATEGORY_NAME	Total_Vagas_Impactadas	Porcentagem_Vagas
1	MEDIA AND COMMUNICATIONS	66	40.49
2	PHYSICAL AND INHERENT ABILITIES	63	38.65
3	INFORMATION TECHNOLOGY	53	32.52
4	LAW, REGULATION, AND COMPLIANCE	51	31.29
5	BUSINESS	45	27.61
6	ENGINEERING	38	23.31
7	HEALTH CARE	34	20.86
8	ADMINISTRATION	33	20.25
9	MANUFACTURING AND PRODUCTION	32	19.63
10	TRANSPORTATION, SUPPLY CHAIN, AND LOGISTICS	28	17.18
11	HUMAN RESOURCES	25	15.34
12	SALES	24	14.72
13	ECONOMICS, POLICY, AND SOCIAL STUDIES	20	12.27
14	FINANCE	20	12.27
15	SCIENCE AND RESEARCH	18	11.04
16	MARKETING AND PUBLIC RELATIONS	17	10.43
17	EDUCATION AND TRAINING	16	9.82
18	DESIGN	16	9.82
19	ANALYSIS	15	9.20
20	MAINTENANCE, REPAIR, AND FACILITY SERVICES	13	7.98
21	HOSPITALITY AND FOOD SERVICES	12	7.36
22	CUSTOMER AND CLIENT SUPPORT	10	6.13
23	PERSONAL CARE AND SERVICES	8	4.91
24	ARCHITECTURE AND CONSTRUCTION	6	3.68
25	ENVIRONMENT	5	3.07
26	ENERGY AND UTILITIES	4	2.45
27	PROPERTY AND REAL ESTATE	3	1.84
28	PUBLIC SAFETY AND NATIONAL SECURITY	3	1.84
29	SOCIAL AND HUMAN SERVICES	1	0.61
30	AGRICULTURE, HORTICULTURE, AND LANDSCAPING	1	0.61

30 rows | 1.47s runtime

Imagem 6 – evidência query para vagas por categoria

A análise da distribuição revela que 40,5% *Media and Communications* (Mídia e Comunicação) e 38,6% *Physical and Inherent Abilities* (Habilidades Físicas e Inerentes) impactam a maior porcentagem de vagas, onde se encaixam comunicação, conhecimento linguístico e habilidades gerais. Isso reforça a conclusão da *Query 1* de que o mercado valoriza majoritariamente as habilidades de comunicação e competências transversais.

Query 3 – Qual a porcentagem de todas as habilidades exigidas que são classificadas como relacionadas a software?

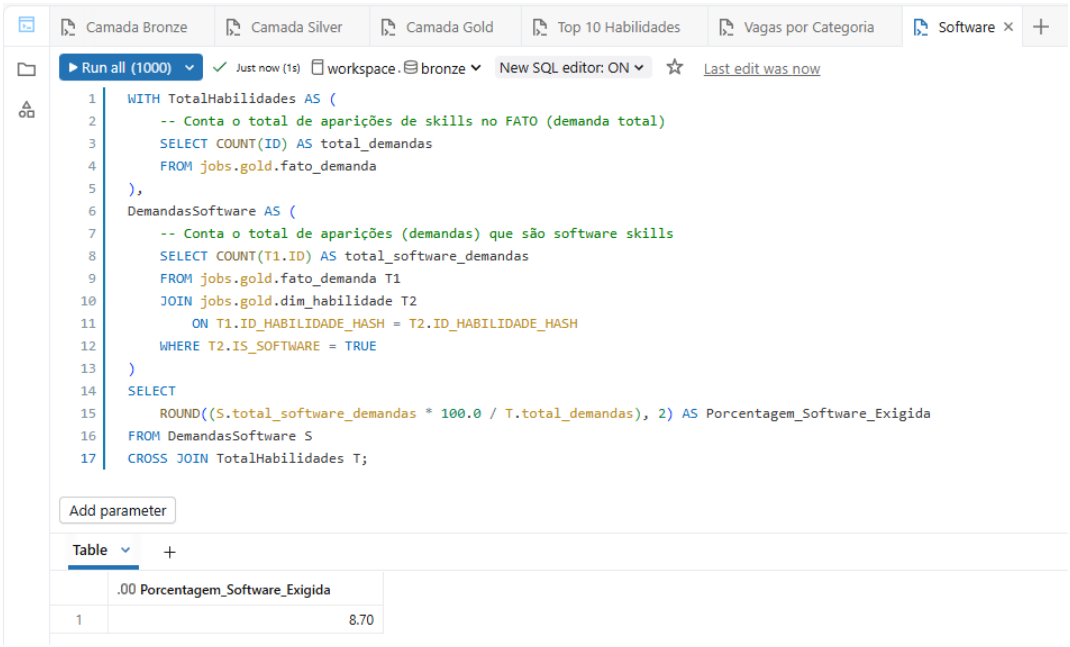

Imagem 7 – evidência query para conhecimento de software

A baixa porcentagem (8.70%) de todas as demandas que são *software skills* é um resultado chave. Mesmo que *Information Technology* (Tecnologia da Informação) seja a terceira categoria mais citada, o volume de habilidades de *software* é baixo, indicando que mesmo em habilidades consideradas majoritariamente técnicas, o conhecimento de *softwares* específicos representa a minoria das oportunidades.

c. Discussão geral

A solução do problema central (entender a demanda do mercado de trabalho) foi bem-sucedida, demonstrando que, para o mercado holandês neste *dataset*, o domínio de Línguas e Comunicação (habilidades comuns) é um filtro mais rígido do que as habilidades técnicas especializadas.

Idioma local: apesar de aparecer como principal habilidade listada no conjunto de dados, o conhecimento do Holandês ainda sim está presente em apenas 25 vagas.

Requisitos pulverizados: não existem habilidades específicas e diretas que estejam presentes de forma massiva e repetida nas vagas.

Prioridade em habilidades comuns: A maior barreira de entrada está nas habilidades Linguísticas e de Comunicação, que impactam cerca de 40% das vagas.

Tecnologia não é o foco principal: Embora presente, a demanda por *software skills* (8.70%) é baixa em relação à demanda geral, sugerindo que as empresas esperam as competências básicas antes de exigir a especialização técnica.

6. Autoavaliação

Com uma trajetória profissional consolidada de 10 anos em áreas de Marketing, Conteúdo e Mídia, o meu papel com dados sempre foi o de demandante e consumidor final das informações geradas pelas áreas de Analytics. A transição para a implementação técnica e engenharia do dado foi o maior desafio deste projeto.

A execução do pipeline de ELT (Bronze, Silver, Gold), a aplicação de SQL avançado para modelagem dimensional e a geração de chaves técnicas exigiram um esforço significativo, incluindo muitas horas de estudo e consulta a tutoriais externos (como o *YouTube*), além do material fornecido em aula.

Apesar da curva de aprendizado íngreme na parte técnica, os objetivos principais do projeto foram integralmente atingidos:

- Implementação do pipeline ETL
- Construção do modelo dimensional de esquema estrela
- Resposta às perguntas de negócio através de *querys* em SQL

Para enriquecer este projeto e expandir a base de conhecimento técnica, os seguintes trabalhos futuros são possibilidades para o portfólio:

- Integração de dimensão temporal para permitir análises de tendências sazonais ou anuais na demanda por habilidades.
- Visualização de Dados: Criar um dashboard interativo (utilizando Power BI ou outra ferramenta) para transformar os resultados das queries em uma narrativa visual.