**CASE JOIN**

**TESTE LÍDER TÉCNICO - ENGENHEIRO DE DADOS**

Olá, candidato!

Abaixo passamos as principais instruções para realização deste teste.

1. Tranquilidade 😁!
2. Não pesquise as respostas na internet. Utilize apenas seus conhecimentos atuais. É importante ser sincero consigo mesmo. 😉
3. Caso avance nas etapas do processo, iremos conversar a respeito de suas respostas. O bate-papo será o mais importante. Porém, queremos avaliar o seu conhecimento. 🤓

BOA SORTE!

**VAMOS COMEÇAR REFORÇANDO A CABEÇA**

**Case Prático**

Utilize o Database Type Ecommerce disponível em:

***Host:*** psql-mock-database-cloud.postgres.database.azure.com

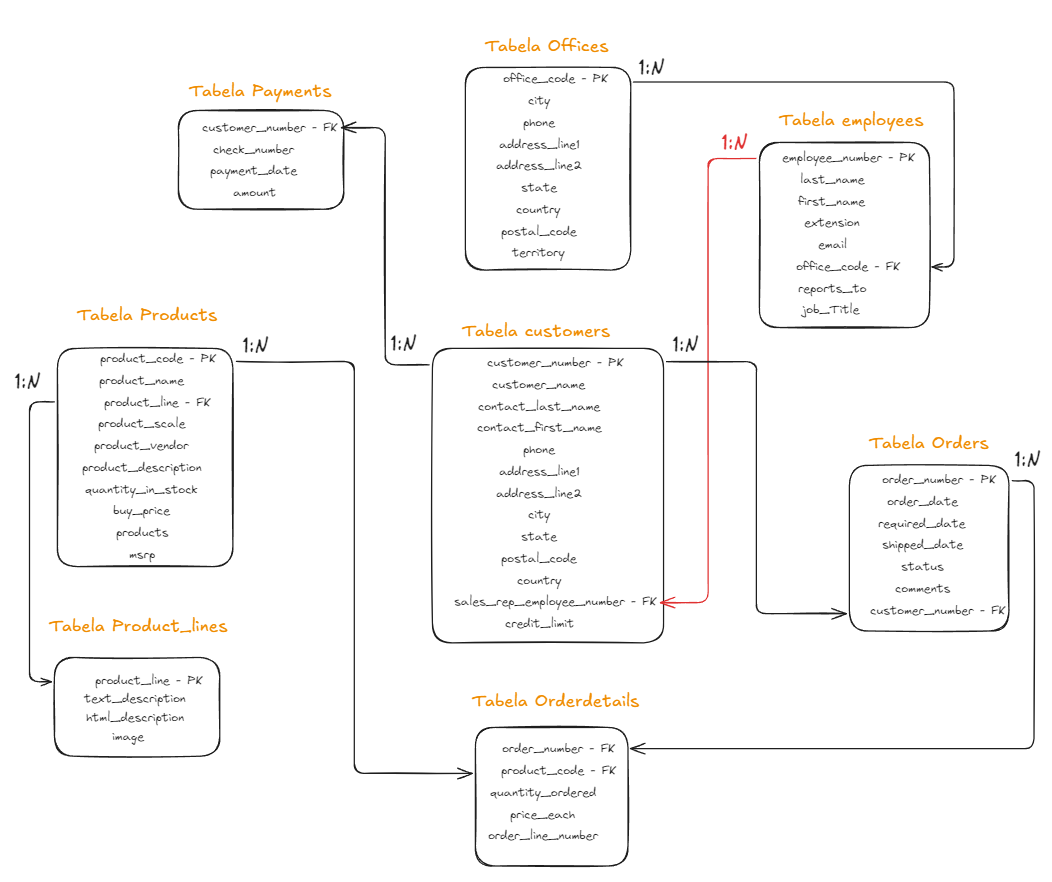
***Banco de Dados:*** ecom1692155331663giqokzaqmuqlogbu

***Port:*** 5432

***Username:*** eolowynayhvayxbhluzaqxfp@psql-mock-database-cloud

***Password:*** hdzvzutlssuozdonhflhwyjm

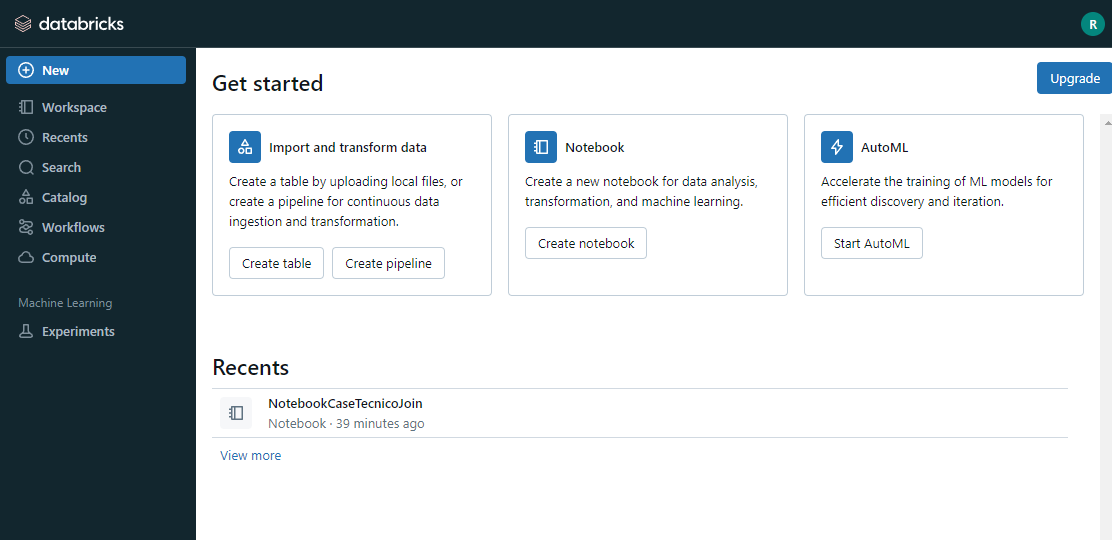
1. Desenhe o diagrama de ER do banco. Imagem, DBML ou diagrama são aceitáveis.



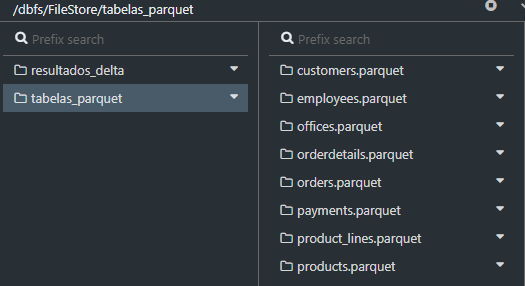
Obs: Ao utilizar o DBeaver para gerar o modelo ER, ou mesmo ao realizar consultas nas tabelas do information\_schema, as chaves primárias (PK) e estrangeiras (FK) não estão declaradas. Portanto, os relacionamentos apresentados são suposições baseadas nos nomes das colunas. O ideal seria que essas informações estivessem presentes nos metadados.

2. Crie uma conta no Databricks Community

(<https://community.cloud.databricks.com>)

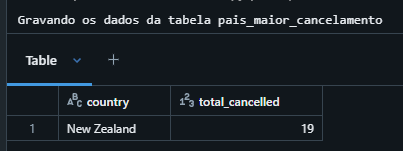


3. Realizar a ingestão das tabelas do Database Type Ecommerce no formato .parquet (1 arquivo por tabela) usando PySpark

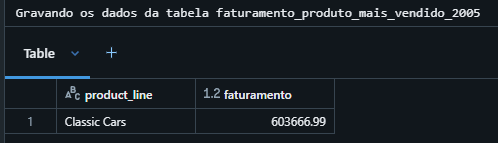


4. Criar as querys ou código utilizando a linguagem de sua preferência que respondam às seguintes perguntas:

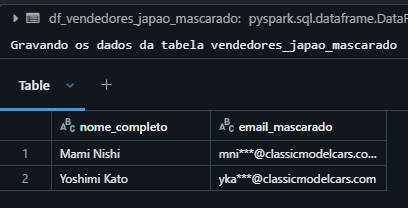
* 1. Qual país possui a maior quantidade de itens cancelados?



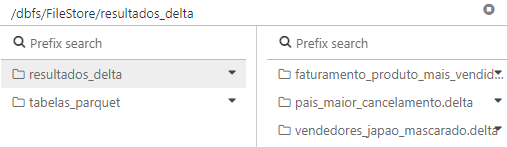
* 1. Qual o faturamento da linha de produto mais vendido, considere os pedidos com status 'Shipped', cujo pedido foi realizado no ano de 2005?

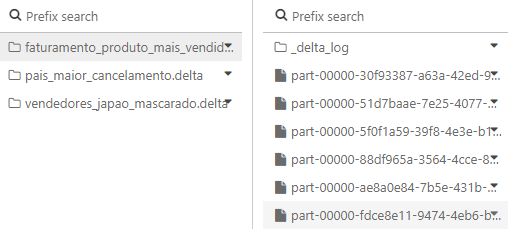


* 1. Traga na consulta o Nome, sobrenome e e-mail dos vendedores do Japão, lembrando que o local-part do e-mail deve estar mascarado.



5. Salve os resultados em formato delta.





6. Crie um README.md documentando a solução e o processo usado para chegar no resultado.



7. Disponibilize os artefatos (Notebooks, README.md, Diagrama de ER) num repositório público no GitHub para consulta.

<https://github.com/ronnanlimadataeng/CaseTecnicoJoin>

**Perguntas Teóricas**

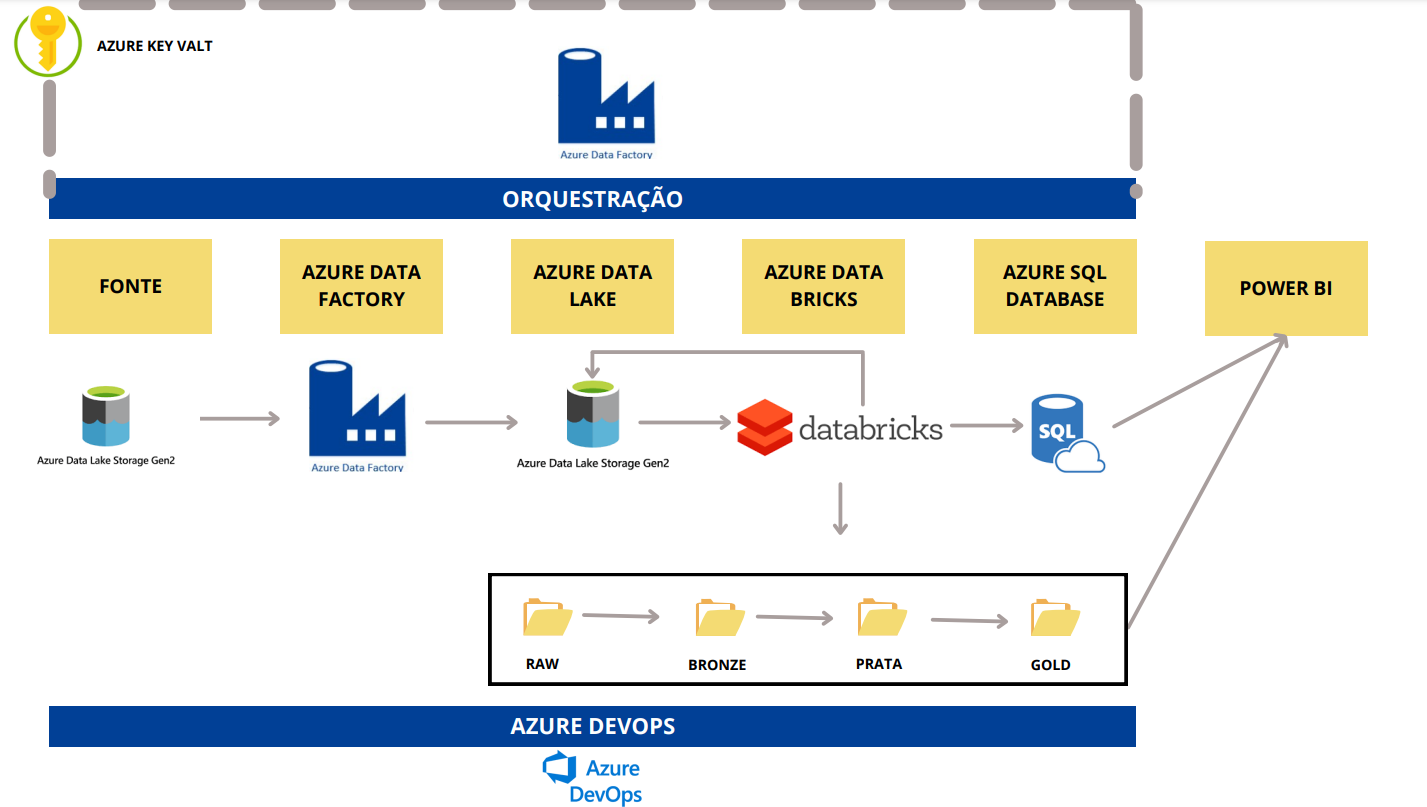
1. Como você utiliza o Delta Lake no Azure Databricks para garantir a integridade dos dados? Uso Delta Lake para versionamento e controle de transações ACID, garantindo consistência e recuperação de falhas, um exemplo bem claro disso pode ser

2. Quais são as vantagens do uso do Spark em comparação com outras tecnologias de processamento de dados? Escalabilidade, processamento distribuído em memória, e suporte a grandes volumes de dados com performance superior. Por muito tempo foi utilizado hadoop mas o processamento em disco tende a ser muito mais lento.

3. Descreva um caso em que você precisou sincronizar dados entre diferentes sistemas. Sincronizei dados migrando sistemas legados para Azure, diria que um grande exemplo no passo foi a migração de bancos Sybase e Oracle, onde construímos visões de layers e organizando em camadas (Raw, Bronze, Silver, Gold).

4. Desenhe uma arquitetura de dados comentada para uma empresa que utiliza Azure e Databricks, incluindo armazenamento, processamento e análise.

Armazenamento no Data Lake, processamento com Databricks/Spark, orquestração com Data Factory, CI/CD no DevOps, e análise no Power BI.



De forma bem resumida, está seria uma boa arquitetura para implementação.

Inclusive tenho outro projeto de case, executado a algum tempo atrás com ela implementada.

<https://github.com/ronnanlimadataeng/Projeto_criacao_arquitetura_de_dados_Azure>

5. Como você garante a escalabilidade e a robustez da arquitetura de dados? Utilizando arquiteturas modulares com escalabilidade elástica em nuvem, pipelines distribuídos, e monitoramento contínuo via Azure Monitor, Alertas no Databricks e etc.

6. Como você implementa a criptografia de dados em repouso e em trânsito? utilizo criptografia automática nas camadas de armazenamento (como Azure Data Lake) com gerenciamento de chaves via Azure Key Vault.

7. Como você gerencia a qualidade dos dados em um pipeline de dados? Implemento checagens de integridade em cada etapa, com regras de validação e alertas automáticos em pipelines.

8. Qual a importância do FinOps para a engenharia de dados? Diria que é essencial para otimizar custos em ambientes de dados em nuvem e inclusive tenho um bom exemplo que recentemente consegui implementar uma redução de custos mensais de R$44 mil para R$24 mil em um projeto de otimização de plataformas de dados, onde eliminei execuções desnecessárias e removi tabelas não utilizadas, gerando uma economia anual de R$235 mil, sem afetar a performance.

9. Como o DevOps ajuda o engenheiro de dados?

Automatiza deploys e versionamento, reduzindo erros, tempo de entrega e claro também garantindo que os produtos que sobem para produção possam passar por um review de codigo.

10. Como iniciamos um projeto de pipeline de dados? Definimos requisitos de negócio, mapeamos fontes de dados, projetamos arquitetura e garantimos conformidade com segurança e escalabilidade.

11. Como realizar CI/CD em um pipeline de dados? Com Azure DevOps, podemos automatizar testes, validações e deploys em ambientes DEV/HML/PRD e com isso garantir a rastreabilidade.

12. Quais ferramentas de orquestração você já trabalhou? Azure Data Factory, Airflow, e Databricks Jobs.

13. Quais suas motivações para ser um engenheiro de dados? é usar dados para transformar negócios. Gosto de atuar com projetos de dados que possuem valor para as estratégias de negócio então desde otimizações, construções, apoio com código, reviews, gestão de times, é parte das motivações, e claro também gosto de liderar e compartilhar conhecimento.