Descrição do Case

Este case é composto por 2 etapas.
Na primeira etapa, propomos uma reflexão sobre a gestão, acompanhamento e otimização do consumo e dos pipelines de dados. O objetivo é avaliar sua capacidade de identificar ineficiências e sugerir melhorias nos processos de tratamento e fluxo de informações.
Na segunda etapa, apresentamos uma situação analítica, e queremos entender a capacidade para solucionar problemas de negócio.

Etapa 1. Gerenciamento de dados

A empresa precisa de uma solução de monitoramento e otimização de seus pipelines de dados, garantindo que os processos rodem sem interrupções e com consumo de recursos otimizado. O desafio é criar um pipeline generalista que possa ser monitorado e ajustado para diferentes tipos de dados e volumes, garantindo que os custos sejam controlados e que alertas sejam configurados para possíveis falhas ou problemas de performance.

Informações gerais

A empresa possui infraestrutura de dados em GCP e apresenta uso massivo de dashboards em PowerBI. Também é utilizado o Metabase para consultas ao banco e produção de relatórios e dashboards.

Case 1:

1. Gestão de Consumos e Otimização de Custos:

- **Monitorar o consumo de recursos**: Você deve propor uma solução para monitorar o uso de recursos (armazenamento e processamento), considerando diferentes ferramentas e provedores de serviços de nuvem.
- **Propor a implementação e KPIs a serem avaliados** em dashboards e relatórios que forneçam insights sobre o consumo de recursos e gastos relacionados, com sugestões para otimização de custos.
- **Propor automações que possam ser aplicadas para escalar o uso de recursos com base na demanda**, otimizando o custo de operação.

- **Propor monitoramento específico dos pipes de PowerBi, identificando problemas de atualização, propondo formas de auto-atualização** e otimização de consumo no fluxo de dados.

2. Documentação e Boas Práticas:

- Documentar como a solução de monitoramento seria implementada, explicando as abordagens de otimização de custos e de processos.

Entregáveis:

Apresentação (PPT) contendo os seguintes pontos:

Etapa 1 - Gerenciamento de dados

1. **Solução de monitoramento de recursos**: Proposta de pipeline de monitoramento configurável com dashboards e alertas.
2. **Plano de otimização de custos**: Estratégias para identificar e mitigar pontos de ineficiência no consumo de recursos, com sugestões de escalonamento automático.
3. **Solução de orquestração de processos**: Proposta de orquestração de pipelines de dados com monitoramento contínuo e alertas configurados para diferentes cenários de falhas.
4. **Diagrama arquitetural:** Representação visual do pipeline de dados, desde a ingestão até o monitoramento e os alertas.

Etapa 2. Analítico

Informações gerais

Temos uma parte do time, dedicada à Monetização, onde é feito negociações com grandes marcas nacionais. Esta área é fortemente ligada às áreas comerciais da empresa visto a sinergia entre ambas.

Dito isso, é comum termos processos em que, respeitando a LGPD, há **data sharing** para os parceiros entenderem sua dimensão dentro do nosso ecossistema. Indo mais além, é possível também termos produtos de dados, **Data Products**, onde o nosso parceiro solicita informações mais específicas e detalhadas sobre seu segmento com determinada periodicidade, 1 x por mês etc. Esse segundo ponto é feito através de acordos comerciais.

Ex. a empresa Y vende o produto Z e gostaria de saber como está as vendas do produto dentro do nosso ecossistema, além da representatividade do produto sob seu respectivo segmento.

O case é para demonstrar a sua expertise como profissional de dados, em usar seu conhecimento ao seu favor e expor as boas práticas adotadas.

Normalmente, os casos de Data Products seguem o seguinte fluxo.

1. Análise da demanda feita pelo parceiro e possibilidades de aperfeiçoamento;
2. Extração, refinamento, anonimização e adequação dos dados solicitados;
3. Automação do processo e sistemas de governança sobre o mesmo.
    a. Garantir a fidelidade dos dados dentro do combinado,
    b. Criar estruturas que sirvam como alerta em caso algo saia do padrão
    c. Seja um sistema seguro, visto serem dados sensíveis
4. **Geração insights e análises podem ser obtidos em cima das informações compartilhadas, a fim de ampliar o valor agregado, com a empresa parceira?**
5. Envio para a empresa parceira através do formato solicitado, de forma automatizada.

De modo que, construa um case genérico, respeitando as etapas acima e ***MELHORANDO*** o que achar necessário

Case 2:

Imaginando que fechamos o contrato com a empresa fictícia *OuchHungry,* líder no segmento de refrigerantes e salgados, com abrangência nacional. A empresa quer saber a posição do seu refrigerante, de nome ‘Hungry’ e de seu salgado ‘Ouch’. O envio será feito de forma diária, a nível analítico e mensal de forma agrupada por cidades em que há vendas. As variáveis significantes (preço, unidades vendidas, clientes atingidos).

A empresa gostaria que enviasse os relatórios para um sistema de SFTP, além de seu próprio sistema via API. Adicionalmente, a empresa parceira gostaria que enviássemos e-mails automatizados após cada atualização de bases.

Entregável: Empresa Fictícia - Desafio Analítico

Objetivo: O candidato deverá desenvolver uma solução que simule a extração, manipulação, automação e monitoramento de dados, utilizando Python e princípios de engenharia de dados.

O foco não é a extração real de dados ou a conexão com sistemas de produção, mas sim demonstrar a capacidade de estruturar um pipeline robusto e flexível, aplicando boas práticas de código, clareza e segurança.

1. Criação e Limpeza de Base

Descrição: Simule a extração de dados de três tabelas (`transações`, `produtos` e `clientes`). O foco é implementar um pipeline de limpeza que identifique problemas comuns, como valores nulos, duplicatas, e inconsistências de formato. Grande parte da extração e validação dos dados envolve a verificação de campos de texto, como e-mails, números de documentos, códigos de produtos e outros formatos específicos.

Desafios:

- O candidato deverá criar funções reutilizáveis para realizar as seguintes operações de limpeza:
    - Tratamento de valores nulos: Substituição ou remoção de entradas.
    - Deduplicação: Identificação e remoção de dados duplicados.
    - Validação de campos de texto: Utilizar expressões regulares (Regex) para validar campos importantes, como e-mails, CEPs, números de telefone, códigos de produtos, etc.
    - Correção de inconsistências de formato: Ajustar dados que não seguem o formato esperado, como datas fora de padrão ou campos numéricos com caracteres não numéricos.

Dicas:

- **Uso de Regex:** Dê atenção especial à validação de dados textuais. Utilize expressões regulares para garantir que campos como e-mails, CEPs e produtos estejam formatados corretamente.
- Estruture o pipeline de modo que ele seja adaptável para novas verificações ou integrações com outros tipos de campos.

Observação: Não será necessária a conexão a um banco de dados real. O objetivo é demonstrar a capacidade de criação de funções genéricas que lidem com problemas comuns de qualidade de dados.

2. Conexão a Storages (Fictício)

- **Descrição:** Desenvolva uma simulação de conexão a diferentes tipos de storages, como SFTP e APIs. O foco é validar a conexão e garantir que os arquivos seriam enviados de forma segura e eficiente.
- **Desafios:** Criar um código que, hipoteticamente, conecte-se a um storage seguro, utilizando chaves públicas e privadas para SFTP, e tokens de API para acessos RESTful.

Exemplos de Parâmetros de Conexão:

- **SFTP:**
  - **Chave pública:** `izI32-vwq5-pPv1-Hr68`
  - **Chave privada:** `J8lM-ld9e-4fg8-5VEH`
- **API:**
  - **Usuário:** `aiqfome`
  - **Senha:** `x9eYJIl0tERhj9BerXdG80C7`

Dicas:

- Mostre como trataria erros de conexão ou falhas de envio.
- Implemente boas práticas de segurança, como não armazenar senhas diretamente no código.

3. Automação do Processo

- **Descrição:** O código deverá orquestrar as etapas anteriores (extração e conexão), garantindo que os dados sejam processados de maneira automatizada e segura. Este passo deve garantir clareza na execução do pipeline e lidar com possíveis falhas.

Desafios:

- Demonstrar como o processo pode ser automatizado e executado periodicamente.
- Garantir que falhas em qualquer uma das etapas sejam tratadas adequadamente.

Dicas:

- Utilizar funções para modularizar o processo.
- Implementar logs e tratamento de exceções para monitorar a execução.

4. Guardrails e Monitoramento

- **Descrição:** Desenvolva uma camada de monitoramento e alertas (Guardrails) para identificar problemas no pipeline. O código deverá disparar alertas automáticos por e-mail caso ocorra algum erro na extração, manipulação ou envio dos dados.
    - **Utilize os e-mails:** correa.igor@aiqfome.com ,diego.pastrello@aiqfome.com, leonardo.artur@aiqfome.com

Desafios:

- Criar gatilhos que possam capturar erros e alertar o time de dados.
- Simular envios de e-mails para notificar falhas ou inconsistências nos dados.

Dicas:

- Mostre como configuraria a camada de alertas e notificações.
- Defina os tipos de erros que acionariam os alertas (ex: erro de conexão, falha de envio de dados, inconsistências detectadas).

Critérios de Avaliação:

- **Clareza e Organização:** O código deve ser limpo, bem documentado e modular.
- **Segurança:** É essencial garantir que o código simule práticas de segurança, como o uso adequado de chaves e senhas.
- **Automatização:** Capacidade de orquestrar os processos de forma eficiente e segura.
- **Monitoramento:** Implementação de alertas para garantir que problemas sejam detectados rapidamente.