PROVA PRÁTICA - PESQUISADOR DATA ENGINEER (ESPECIALISTA INTEGRAÇÃO)

O que queremos receber?

Um um arquivo zip contendo os arquivos da prova, incluindo o arquivo .json do fluxo exportado pelo n8n, nos e-mails: dgasilva@sfiec.org.br, elgomes@sfiec.org.br.

Como a prova deve ser feita?

Utilize todos os seus conhecimentos em engenharia de software, engenharia de dados e inteligência artificial para propor a melhor solução possível para os problemas apresentados. Não se acanhe, use sua criatividade, pense fora da caixa e demonstre todo o seu poder de desenvolvedor. Nosso time irá avaliar a sua capacidade através da solução que será entregue, por tanto, capriche, pois seus concorrentes irão caprichar. Ser ousado não tira pontos, pelo contrário, ajuda nossa equipe a avaliar suas habilidades.

Auto avaliação

Avalie suas habilidades nos requisitos de acordo com os níveis especificados.

Qual o seu nível de domínio nas técnicas/ferramentas listadas abaixo, onde:

- 0, 1, 2 - não tem conhecimento e experiência;
- 3, 4 ,5 - conhece a técnica e tem pouca experiência;
- 6 - domina a técnica e já desenvolveu vários projetos utilizando-a.

Tópicos de Conhecimento:

- Manipulação e tratamento de dados com Python e Pyspark: 5
- Desenvolvimento de data workflows em Ambiente Azure com databricks: 5
- Desenvolvimento de data workflows com Airflow: 0
- Desenvolvimento de data workflows com n8n: 6
- Manipulação de bases de dados com SQL: 6
- Web crawling e web scraping para extração de dados: 6
- Construção de APIs: REST, SOAP e Microservices: 4
- Integração de sistemas com n8n: 6
- Engenharia de prompts: 6
- Manipulação de dados com IA (LLM's): 6

Contexto

O Observatório da Indústria precisa automatizar o processo de coleta, tratamento e disponibilização de dados sobre **Acessos – Banda Larga Fixa** para apoiar análises estratégicas. A solução a ser desenvolvido deverá:

1. Capturar automaticamente os dados do site dados.gov.br.
2. Organizar e tratar os dados em um Data Lake.
3. Disponibilizar integrações e automações usando n8n.
4. Permitir que os mais de 500 usuários do Observatório solicitem análises sobre os dados de banda larga fixa em linguagem natural, com suporte de uma solução de IA/LLM.

Questões

Questão 1

Implemente uma rotina em **Python e Spark** que realize o **download automático** do conjunto de dados **Acessos – Banda Larga Fixa** a partir do site dados.gov.br.

- A rotina deve salvar os dados no Data Lake para permitir processamento posterior.

Questão 2

Utilizando **Spark**, trate os dados coletados para responder às perguntas:

1. Qual o total de acessos por região do Brasil no último ano disponível?
2. Qual a evolução do número de acessos por tecnologia (ex: fibra, cabo, rádio) nos últimos 3 anos?

Questão 3

Projete e implemente um **workflow no n8n** que:

- Automatize a execução da rotina de ingestão e transformação criada nas questões anteriores.
- Disponibilize uma **API REST** que permita consultar os resultados no data lake.

Questão 4

Implemente no próprio n8n um fluxo que permita ao usuário fazer perguntas em **linguagem natural** sobre os dados tratados (ex: *“Qual região teve o maior crescimento em acessos de fibra nos últimos 2 anos?”*).

- A solução deve utilizar um **LLM** (ex: OpenAI API ou similar).
- O LLM deve responder de forma clara ao usuário.

Questão 5

Explique como você:

1. Otimizou o fluxo para atender a quantidade de usuários informado no contexto.
2. Utilizou os conceitos de governança e segurança para manter a solução íntegra.
3. Implementaria um monitoramento para garantir que a coleta, atualização dos dados e interação com os usuários, ocorram corretamente.

- Você pode utilizar um bloco de comentário dentro do próprio fluxo do n8n para responder esta questão.