Avaliação Técnica – Engenheiro de Dados

Nome do Candidato: Felipe da Silva Faria

Data: 18-06-2025 E-mail: felipe.sfaria@hotmail.com

# Parte 1 – Questões de Múltipla Escolha

1. Sobre a arquitetura em camadas (Bronze, Silver e Gold), qual das opções descreve corretamente suas funções?

A. Bronze: dados limpos; Silver: dados crus; Gold: dados agregados

B. Bronze: dados estruturados; Silver: dados não estruturados; Gold: dados arquivados

**C**. **Bronze: ingestão de dados crus; Silver: dados refinados e com qualidade; Gold: dados prontos para consumo analítico**

D. Bronze: dados replicados; Silver: dados eliminados; Gold: dados transformados em imagens

E. Bronze: arquivos temporários; Silver: dados de machine learning; Gold: dados brutos

Gabarito : **C**

1. Durante o desenvolvimento de pipelines de dados, qual das práticas abaixo está mais alinhada com princípios de engenharia de dados escalável e sustentável?

A. Realizar transformações diretamente na camada de apresentação para reduzir latência

B. Evitar versionamento de dados para reduzir espaço em disco

C. Centralizar toda a lógica de negócios em dashboards analíticos

**D**. **Separar responsabilidades em etapas reutilizáveis e com controle de versionamento**

E. Construir pipelines únicos com lógica acoplada e parametrização mínima

Gabarito : **D**

1. Em um processo de ingestão de dados de fontes externas, qual prática contribui para maior resiliência e confiabilidade do pipeline?

A. Executar a carga manualmente para garantir precisão

B. Evitar logs para não gerar arquivos desnecessários

**C. Implementar mecanismos de monitoramento, retry e logging para falhas**

D. Eliminar checkpoints para reduzir a complexidade do código

E. Utilizar múltiplas fontes simultaneamente, sem controle de concorrência

Gabarito: **C**

1. Qual das opções a seguir representa uma vantagem do uso de formatos de dados orientados a colunas (como Parquet ou ORC) em ambientes analíticos?

**A**. **Melhor desempenho em leitura seletiva de colunas e compressão eficiente**

B. Facilidade para edições linha a linha em arquivos

C. Compatibilidade com arquivos XML sem necessidade de conversão

D. Estruturação ideal para uso com bancos de dados relacionais

E. Suporte automático à normalização de dados

Gabarito: **A**

1. Em relação ao conceito de data lineage, qual das alternativas melhor descreve sua utilidade em um ambiente de dados?

A. Minimizar o uso de metadados em ambientes de produção

B. Automatizar a modelagem relacional dos dados

**C**. **Mapear a origem, transformação e destino dos dados para auditoria e governança**

D. Otimizar diretamente a performance de consultas em camadas analíticas

E. Armazenar logs de acesso a dashboards para métricas de uso.

Gabarito: **C**

# Parte 2 – Questão Discursiva

Explique, com suas próprias palavras, como você estruturaria um pipeline de dados em arquitetura do tipo "medalhão", considerando todas as etapas de:

* Extração de dados (batch e streaming, se aplicável)
* Transformação e limpeza dos dados
* Armazenamento em camadas (Bronze, Silver e Gold)
* Orquestração do processo
* Disponibilização dos dados para consumo (API, dashboards, relatórios, camadas analíticas)

Utilize exemplos práticos e mencione os principais componentes. A resposta deve ter no máximo uma página.

# Parte 3 – Desafio Prático – Construção de Bot de Dados

1. Desenvolva um bot (robô) em Python que atenda aos critérios abaixo:

* Utilize a linguagem python (versão 3 ou superior)
* Obtenha os dados do IPCA em: https://sidra.ibge.gov.br/Ajax/JSon/Tabela/1/1737?versao=-1
* Coloque os dados no formato tabular (estruturado) e grave um arquivo com este conteúdo no formato “parquet”
* Construa ao menos 3 funções (ou métodos) e as utilize no código
* Documente as etapas do processo dentro do próprio código
* Disponibilize o(s) código(s) e o arquivo final gerado pelo bot (parquet) em um projeto do GitHub (repositório público - <https://github.com/>)

2. No exemplo do bot que você construiu, o conjunto de dados necessários estava disponível no site diretamente por meio de um link já definido. No entanto como você resolveria o problema da captura dos dados caso fosse necessário antes navegar no site (executando passos e cliques por meio de menus, login, botões, links) para se chegar ao arquivo alvo (se não existisse um link direto para o conjunto de dados).

Como resposta para este item, crie um arquivo (txt), disponibilize-o no referido projeto do GitHub e inclua no seu conteúdo um texto explicativo (com suas próprias palavras) que descreva uma proposta de solução.