Avaliação Técnica – Engenheiro de Dados

Nome do Candidato: Tiago Novak

Data: 0/08/2025 E-mail: tiagonovaksbs@gmail.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

Resposta Questão 1: Letra 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

Resposta Questão 2: Letra 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

Resposta Questão 3: 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

Resposta Questão 4: Letra 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

Resposta Questão 5: A

# 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.

Primeiramente é necessário identificar a fonte dos dados que serão ingeridos. Logo após isso, saber o formato dos dados para a melhor abordagem técnica e quais etapas serão utilizadas para a manipulação e o processamento dos dados. Para o modelo em bath, avaliar o tamanho desse lote para melhor performance de processamento e armazenamento(arquivos grandes têm processamentos demorados). Em Streaming, também é necessário saber a origem dos dados, como APIs de vendas, sensores, transações financeiras que necessitam de velocidade de processamento para que a ingestão e disponibilidade desses dados de imediato, ou seja, dados em tempo real que Kafka ou Event Hub podem ser utilizados. Schemas dos dados devem ser avaliados a depender da área de negócio para que a performance dos dados seja a mais normalizada possível(se aplicável) para uma melhor busca e rapidez de dados.   
Após essa parte deve-se avaliar o formato dos dados que foram carregados na extração, como o sigilo necessário, pois precisa-se saber qual área de negócio eles serão disponibilizados. Pois há situações em que dados sensíveis devem ser mascarados com algum tipo de hash para a confidencialidade e segurança. Na limpeza, deve-se garantir que não haja duplicidade, dados nulos, formatação de datas, valores inteiros, decimais, entre outros. Por isso a importância de schemas ou regras de qualidade dos dados como boa prática.  
O armazenamento é importante para usar a arquitetura medalhão, aqui é feita a organização desses dados, passando pela camada bruta(Bronze) com versionamento dos dados para auditoria e recuperação. Depois pela (Silver), momento esse em que é feita a limpeza, formatação para para inúmeras demandas. Na camada Gold, apenas os dados necessários, bem organizados para velocidade de consumo, com dados já alinhados a cada situação de negócios específicos. Esse armazenamento e organização dos dados é uma boa prática na aplicação de formato Parquet, Delta Lake, pois são formatos muito utilizados em Big Data, devido a sua maior compressão e eficiente poder de leitura.  
Na orquestração o uso de ferramentas de código aberto, como Airflow, que permite a orquestração, agendamento e monitoramento dos processos de ETL ou ELT, assim como Data Factory. Cada etapa desse processo, pode ser feita para um chamada de conversão de arquivos, leitura, gravação controle de dependência e reexecução em eventos falhos. As etapas de jobs, onde é possível ver cada uma delas e monitorar seu desempenho se ela está ocorrendo como foi projetado, além de alertas para possíveis falhas e em seguida resolvê-las para a entrega dos dados aos setores designados.  
A disponibilização desses dados pode se dar por meio de data warehouse ou data lakes com as pastas organizadas com camadas de segurança de permissões de acesso. A depender da necessidade podem ser consumidos por Dashboards para analistas, APIs para consumo por sistemas ou aplicativos, camadas analíticas para Cientistas de dados. Documentar e catalogar os dados é uma etapa importante para o consumo correto de dados e sua governança alinhada.

# 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.  
  
Eu farei um webscraping para coletar esses dados e logo em seguida um processamento desses dados levando em conta, o processo de ETL ou ELT a depender dos dados.