Teste de Data Engineer

AME DIGITAL

PROPÓSITO

A AME Digital está sempre em busca das melhores pessoas para compor o grupo que está revolucionando a maneira como as pessoas se relacionam com o dinheiro. Para tal, nosso time de dados está trabalhando duro para fortalecer a cultura “data-driven” dentro da nossa companhia e estamos em busca de um Engenheiro de Dados para nos ajudar nessa missão.

O desafio a seguir tem como objetivo avaliar seus conhecimentos e experiências com dados e habilidade de resolver problemas. Ao solucioná-lo, você nos mostrará:

Sua capacidade de extrair dados de uma fonte, processá-los e transformá-los em informações.

- Seu entendimento sobre tecnologias de Big Data.
- Seu conhecimento em SQL.

SOBRE O DESAFIO

A seguir você encontrará links para dois conjuntos de dados que contém requisições HTTP para os servidores da NASA - Kennedy Space Center para os períodos de Julho e Agosto de 1995.

Kennedy Space Center - Julho : **Nasa Kennedy Server - July**

Kennedy Space Center - Agosto : **Nasa Kennedy Sever - August**

Os logs contém as seguintes informações:

- Host que está realizando a requisição
- Timestamp do momento em que a requisição aconteceu
- Requisição

- Código de retorno da chamada HTTP
- Total de bytes retornados

Seu desafio é nos ajudar a identificar, utilizando tecnologias de Big Data, algumas informações dentro desse conjunto de dado. Separamos o desafio em duas partes:

1. Perguntas teóricas sobre Big Data.
2. Execução das consultas nas bases de dados que passamos para sanar nossas dúvidas.

BIG DATA - TEÓRICO

---

1. Considerando que a Ame possui diferentes aplicações, resultando em diferentes fontes de dados (como bancos relacionais e noSQL), de que maneira você construiria uma arquitetura para realizar a ingestão desses dados em uma plataforma de Big Data?

   Descreva as tecnologias que você escolheria para realizar a ingestão, bem como o fluxo de dados entre elas (lembrando que o objetivo é disponibilizar as informações o mais próximo de "real-time" possível).

2. Ao utilizar ferramentas de processamento distribuído como Spark ou Hive, é muito comum enfrentar problemas relacionados à má distribuição de dados entre as máquinas do cluster, diminuindo drasticamente a performance das aplicações, principalmente em operações relacionadas a agregação ou join. Utilizando seus conhecimentos e experiências, descreva uma possível solução para o problema em questão.

3. O dia a dia de um engenheiro de dados, dentre outras tarefas, é disponibilizar as informações em alta performance (próximos a real-time) para Analistas e Cientistas de Dados de modo a possibilitar à análise e criação de modelos estatísticos. De que modo e quais tecnologias você usaria para disponibilizar os dados para estas pessoas.

4. Por fim, tendo em mente o crescimento exponencial dos dados e utilização massiva da plataforma de Big Data, quais métodos de organização e/ou governança você implementaria para manter o ambiente sustentável?

BIG DATA - PRÁTICA

Queremos que você nos ajude a responder as questões abaixo e, para isso, queremos que você trate e estruture os dados utilizando ***Spark*** e realize as consultas para responder às questões abaixo utilizando ***SparkSQL***.

1. Número de HOSTs únicos.
2. O total de erros 404 dentro do período.
3. Quais dias do período especificado tiveram o maior número de erros 404.
4. O total de bytes retornados no período, com uma visão acumulada. Por exemplo, se no dia 1 tivemos 50 bytes, dia 2 tivemos 100 bytes e dia 3 mais 150 bytes, sua resposta deverá ser: Dia 1 = 50 bytes, Dia 2 = 150 bytes, Dia 3 = 300 bytes.

Obs. Utilize GMT-3 nas suas análises e as datas retornadas devem seguir o formato 'DD-MM-YYYY'.

O QUE ESPERAMOS VER NO FINAL?

Nosso time está curioso para ver o seu projeto. Esperamos que seu entregável final contenha os seguintes ítens:

1. Um documento com as respostas às perguntas teóricas e práticas sobre Big Data.
2. Os arquivos com o seu código utilizado para ler, interpretar e estruturar as informações das bases recebidas (de preferência em um Jupyter Notebook ou similar).

Caso você tenha dificuldade de finalizar o seu projeto, nós o encorajamos fortemente a nos enviar a sua evolução (código, descrição de como resolveria o problema, etc.).

VEM PRA AME! =)