Desafio Técnico

Processo Seletivo – Santander – Big Data – São Carlos

Jéssica Naiara B. de Farias

Questão 01: Utilizando suas palavras, defina Big Data e seus "4V's".

Big Data é um fenômeno que surge com a popularização do acesso à tecnologia e informação. Com o aumento dos smartphones, smartTV, smartwatch, internet banking, redes sociais, entre outros, a quantidade de dados gerados cresceu significativamente e virou uma fonte de informações que indicam padrões e/ou tendências. Por esse motivo, a análise desses dados se tornou um desafio com grande potencial de retorno financeiro para inúmeros segmentos de mercado. O termo Big Data é direcionado a essas grandes volumes de dados obtidos, bem como, uma área de conhecimento que estuda como coletar, tratar e obter informações por meio desses dados.

Alguns especialistas definiram 4 conceitos principais que deixam mais clara a ideia de Big Data. São eles:

- Volume: refere-se ao grande volume de dados
- Variedade: os dados são de diferentes tipos, como imagens, áudios, vídeos, entre outros
- Velocidade: o tratamento desses dados deve ser feito em tempo hábil e muitas vezes em tempo real
- Veracidade: é necessário estratégias e processos para garantir a consistência dos dados

Questão 02: Cite e explique as principais diferenças entre Hadoop e Spark.

Hadoop e Spark são tecnologias diferentes que auxiliam no processamento de dados de forma distribuída. O Hadoop é um framework para o desenvolvimento de aplicações de forma distribuída que agrega diversas ferramentas e por isso, é considerado um ecossistema.

Dentre as tecnologias pertencentes ao ecossistema Hadoop, o HDFS, um sistema distribuído de armazenamento de arquivos muito grandes que utiliza diferentes clusters, se assemelha mais com a estrutura do Spark, pois as duas tecnologias são estruturas de Big Data

Dentre as tecnologias pertencentes ao ecossistema Hadoop, duas são importantes para contrastar com o Spark, são elas:

- HDFS: um sistema distribuído de armazenamento de arquivos muito grandes que utiliza diferentes clusters
- MapReduce: É a estrutura que gerencia o processamento de dados em larga escala que estão em diferentes clusters.

O Spark é um framework para processamento de Big Data de forma distribuída.

Algumas das principais diferenças das estruturas são, o Hadoop MapReduce tem bom desempenho para processamento linear de grandes conjuntos de dados, enquanto o Spark tem um bom

desempenho considerando o processamento iterativo, análise em tempo real, aprendizado de máquina, entre outros. O Spark pode chegar em uma performance de até 100 vezes mais rápido.

Uma vantagem do Spark é o funcionamento compatível com o ecossistema Hadoop, o que possibilita o uso das duas tecnologias em conjunto.

Questão 03: Suponha que você possui a sua disposição três bancos de dados para armazenar dados tradicionais: HBase, Hive e ScyllaDB. Qual seria sua escolha para esse tipo de dado? Justifique sua resposta.

Considerando que os dados são tradicionais, eu escolheria o SGBD Hive, pois lida com grande volume de dados de forma distribuída, faz parte do ecossistema Hadoop e é um banco de dados relacional, diferente dos outros dois, que possuem estrutura de armazenamento orientada a colunas.

Questão 04: A utilização de Docker no contexto de Big Data pode ser feita de diversas maneiras a fim de ajudar a lidar com o grande volume de dados. Cite e explique resumidamente pelo menos um caso de uso de Docker e Big Data.

O Docker possibilita que exista um isolamento das ferramentas de big data, de forma que desenvolvedores possam realizar processamentos com diferentes ferramentas sem que ocorra problemas de dependência entre elas.

Outro uso interessante de docker é a capacidade de agendamento de tarefas de manipulação de dados de forma automatizada. Essa é uma característica muito útil para Big Data, para manter o ritmo de execução de tarefas, sem que seja necessário uma configuração manual em cada nó dos clusters de big data.

O uso de docker também permite a criação de um cluster de vários nós em uma única máquina principal, replicando a configuração típica de Big Data. Sendo muito útil para estudo e realização de testes.

Questão 05: Quais pontos de abstração você julga importante para compor um projeto de Big Data?

A abstração é essencial para lidar com big data, por conta dos 4 V's, pois, a variedade dos tipos de dados que devem ser processados e retornados em tempo útil exige que as estruturas sejam flexíveis para trabalhar com essas diferenças. É primordial pensar em estruturas lógicas que lidem melhor com o surgimento de situações inesperadas, além da facilidade de alteração da estrutura decorrente de algum novo erro ou necessidade.

Desafio Técnico

Parte 2 - Desafio de codificação

Disponível no repositório

Parte 3 - Desafio Engenharia de Dados

Com o objetivo de criar um Data Lake on premise os seguintes passos foram realizados:

- Coleta de dados do twitter considerando a #Santander, presente no arquivo *Parte2/CapturaTwitter.ipynb:* Para realizar a coleta de dados é necessário alterar o arquivo *twitter-tokens.txt* adicionando os tokens pessoais de developer fornecidos pelo twitter
- Instalação da Oracle Virtual Machine
- Cloudera Quick Start VM
- Instalação do MongoDB
- Estruturação da arquitetura