

Prezado candidato.

Gostaríamos de fazer um teste que será usado para sabermos a sua proficiência nas habilidades para a vaga. O teste consiste em algumas perguntas e exercícios práticos sobre Spark e as respostas e códigos implementados devem ser armazenados no GitHub. O link do seu repositório deve ser compartilhado conosco ao final do teste.

Quando usar alguma referência ou biblioteca externa, informe no arquivo README do seu projeto. Se tiver alguma dúvida, use o bom senso e se precisar deixe isso registrado na documentação do projeto.

Qual o objetivo do comando **cache** em Spark?

Salvar os resultados parciais intermediários para que possam ser reutilizados nos estágios subsequentes.

O mesmo código implementado em Spark é normalmente mais rápido que a implementação equivalente em MapReduce. Por quê?

O Spark implementa as estruturas de dados na memória, dessa forma os algoritmos podem interagir com esses dados várias e várias vezes de forma eficiente.

Qual é a função do **SparkContext**?

É o ponto inicial para ter acesso a todo o ambiente do Spark, é o coração da aplicação.

Explique com suas palavras o que é **Resilient Distributed Datasets (RDD)**.

Tabela de dados que aceita qualquer tipo de dado que são armazenados e processados de forma distribuída.

GroupByKey é menos eficiente que **reduceByKey** em grandes dataset. Por quê?

O GroupByKey tenta mover as chaves iguais para mesma partição isso produz um shuffle porque ele faz isso para cada dado de cada partição, logo se as partições não estiverem na mesma máquina, isso causará muito tráfego na rede além disso pode causar problemas de memória.

Explique o que o código Scala abaixo faz.

```
val textFile = sc.textFile("hdfs://...")
val counts = textFile.flatMap(line => line.split(" "))
                      .map(word => (word, 1))
                      .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```

Conta a quantidade de palavras existentes no arquivo carregado.

HTTP requests to the NASA Kennedy Space Center WWW server

Fonte oficial do dataset: <http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html>

Dados:

- [Jul 01 to Jul 31, ASCII format, 20.7 MB gzip compressed](#), 205.2 MB.
- [Aug 04 to Aug 31, ASCII format, 21.8 MB gzip compressed](#), 167.8 MB.

Sobre o dataset: Esses dois conjuntos de dados possuem todas as requisições HTTP para o servidor da NASA Kennedy Space Center WWW na Flórida para um período específico.

Os logs estão em arquivos ASCII com uma linha por requisição com as seguintes colunas:

- **Host fazendo a requisição.** Um hostname quando possível, caso contrário o endereço de internet se o nome não puder ser identificado.
- **Timestamp** no formato "DIA/MÊS/ANO:HH:MM:SS TIMEZONE"
- **Requisição (entre aspas)**
- **Código do retorno HTTP**
- **Total de bytes retornados**

Questões

Responda as seguintes questões devem ser desenvolvidas em Spark utilizando a sua linguagem de preferência.

1. Número de hosts únicos.
2. O total de erros 404.
3. Os 5 URLs que mais causaram erro 404.
4. Quantidade de erros 404 por dia.
5. O total de bytes retornados.