Desafio Semantex Keiko Barbosa Ioshitake

Pergunta 1: Qual o objetivo do comando cache em Spark?

Armazenar datasets na memória cache, para que assim o acesso possa ocorrer de forma rápida, principalmente para dados que serão acessados repetidas vezes.

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

Enquanto MapReduce realiza as ações de em etapas de forma procedural, fazendo leitura, operação, escrita do agrupamento e assim por diante, o Spark pode realizar as operações utilizando processamento em memória, em vez de usar o disco rígido.
Ele lê o agrupamento, realiza todas as operações e escreve os resultados de uma vez.

3. Qual é a função do SparkContext?

O SparkContext configura serviços internos e estabelece uma conexão com um ambiente de execução do Spark. Pode-se criar RDDs, acumuladores e variáveis de difusão, acessar serviços do Spark e executar jobs.

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

RDD é a abstração principal do Spark.
É um conjunto de dados que não pode ser modificado. Este é particionado e distribuído para diferentes nós do cluster para realizar o processamento paralelo.

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

O groupByKey pode causar problemas de falta disco quando os dados são enviados pela rede e coletados nos workers de redução.
Enquanto que reduceByKey os dados são juntados em cada partição, e somente um dado por chave é enviado via rede, isso força que os dados transformados sejam enviados como sendo o mesmo tipo de dados.

Explique o que o código Scala abaixo faz.

Primeiro carrega do dataset como um arquivo texto.
Depois conta a quantidade de palavras faz um dataset com string e int,
E salva como arquivo em HDFS.

Questões do log da NASA:

1. Número de hosts únicos.
   157.043
2. O total de erros 404.
   20.872
3. Os 5 URLs que mais causaram erro 404.

4. Quantidade de erros 404 por dia.
5. O total de bytes retornados.