Qual o objetivo do comando cache em Spark?

É armazenar um dataframe que foi materializado em memória para um uso mais rápido ao realizar outra ação que esteja relacionado. Se o dataframe não estiver em cache o Spark fara todas as transformações novamente ao chamar o dataframe.

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

Isso ocorre porque o MapReduce realizava o processamento em cima dos dados armazenados no datalake, por isso era mais utilizado para o processamento de grandes volumes de informações que já estavam armazenados. Já o Spark como veio para permitir o processamento de dados em tempo real ele utiliza a memória RAM para realizar esse processamento, por ser mais rápida.

Qual é a função do SparkContext?

A função do SparkContext é armazenar as configurações do ambiente de Big Data em que o Spark atua para execução de algumas tarefas. Um exemplo disso é a inicialização do HiveContext que recebe o SparkContext e após ser inicializado permite a execução de comandos HiveQL pelo Spark.

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

RDDs são datasets baseados na execução de funções do tipo map e reduce para o processamento de dados. Sendo que a função map serve para “organizar” os dados de maneira que faça sentido para a ação que quer realizar e reduce para agrupar esses valores mapeados. Ao executar essas funções o RDD pode ser visualizado como uma tabela.

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

Isso ocorre porque o groupByKey no processo de shuffle envia chaves repetidas sem nenhum tratamento para o próximo nó, enquanto o reduceByKey antes de efetuar o envio agrupa essas chaves. Sendo assim o reduceByKey envia menos informações e mais relevantes.

Explique o que o código Scala abaixo faz.

Ele carrega um arquivo de texto que está no file system do hadoop.

Realiza a separação das palavras pelo espaço entre elas (“ “) , mapeia a palavra com o número 1 e, por fim, agrupa as palavras iguais. A partir de agora se uma palavra se repetir terá o número de repetições no lugar do um.

Salva um arquivo com as palavras e a quantidade de repetições no HDFS.