Qual o objetivo do comando cache em Spark?

A função do cache é armazenar na memória conjuntos de dados. Utilizar este comando é útil para fazer operações em dados que são acessados repetidamente, diminuindo o tempo da operação.

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

Além de possuir a função de armazenar em cache os dados que serão utilizados com frequência, ele tem a vantagem de não necessitar trazer o dataset completo sempre que necessita realizar uma operação, o que acontece com o MapReduce.

Qual é a função do SparkContext?

O SparkContext necessita ser criado, a partir de um SparkConf, para permitir o acesso ao cluster pelo Spark. Ele também é necessário para a criação de RDDs, acumuladores e variáveis dentro do cluster.

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

RDD é uma estrutura de dados utilizada no Spark. Ele é o responsável por armazenar resultados intermediários das operações e evita a necessidade de utilizar o dataset completo para os próximos comandos, melhorando muito o desempenho do sistema.

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

O groupByKey e o reduceByKey trazem o mesmo resultado, porém o primeiro faz a operação de comparação no dataset completo antes de fazer o agrupamento, enquanto o reduceByKey faz as comparações em cada partição, organizando as informações preliminarmente, e traz os dados já pré-agrupados para o dataset.

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://...")
```

Ele está acessando um arquivo de texto no endereço ("hdfs://...") e separando palavra por palavra (line.split(“ ”)). A seguir, está fazendo o agrupamento de palavras iguais utilizando o reduceByKey no formato:

(“palavra”, quantidade) + (“palavra”, quantidade) + (“palavra”, quantidade) + ...

e salvando o resultado em um novo arquivo de texto (saveAsTextFile).

Questões utilizando Spark

Infelizmente, não consegui fazer o build do Spark no meu computador. Ocorreram diversos erros ao utilizar o spark_shell e ao tentar a instalação pelo PySpark. Procurei na documentação no site da empresa e em fóruns relacionados, porém, não obtive sucesso na identificação do problema.