***Respostas***

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

**R:** *cache()* serve para você utilizar um nível de armazenamento, que por padrão é MEMORY\_ONLY. Contudo, existe a opção de declarar outros níveis, como MEMORY\_AND\_DISK, MEMORY\_ONLY\_SER, dentre outros. Esses níveis de memória o ajudam a fornecer compensações entre o uso da memória e a eficiência da CPU.

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

**R:** O Spark executa seu processamento em memória sem a utilização de escrita/leitura em disco. Spark também faz uso do RDDs, onde implementa estruturas de dados em cluster. Por estes motivos o Spark é mais eficiente, já que ele foi projetado para ter uma performance maior em trabalho iterativos, enquanto que o MapReduce foi projetado para trabalhar melhor em lotes.

* **Qual função do SparkContext?**

**R:** SparkContext é um objeto que deve ser criado para informar para o Spark como acessar um cluster. Ele pode se conectar a vários tipos de gerenciadores de cluster.

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

**R:** RDD é uma coleção de elementos particionada através dos nós do cluster que podem ser operados em paralelo.

É possível imaginar o RDD como uma tabela de banco que pode guarda diversos tipos de dados. O Spark armazena os dados do RDD em diferentes partições.

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

**R:** Pois quando se trata de grandes datasets, no reduceByKey os pares na mesma máquina com a mesma chave são combinados antes que os dados sejam embaralhados e logo após, sua função é chamada novamente para reduzir os valores de cada partição para produzir um resultado final. Já o GroupByKey todos os pares de valores-chaves são alternados, onde são muitos dados desnecessários para serem transferidos pela rede.

* **Explique o que o código abaixo faz.**

**R:** O código recebe um arquivo em “textFile”, realiza algumas transformações para construir um dataset, em “counts”, e por fim salva em um arquivo.