Respostas dos testes propostos por Semantix

1 - Qual o objetivo do comando cache em Spark?

É um mecanismo que visa acelerar aplicativos que acessam várias vezes o mesmo RDD. Um RDD que não é armazenado em cache é reavaliado toda vez que uma ação é invocada neste RDD.

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

Com Spark os dados são mantidos em memória RAM entre os 'steps' e não gravados no disco rígido como no caso do MapReduce e isso torna a execução várias vezes mais rápida.

3- Qual é a função do SparkContext?

O contexto do Spark configura os serviços internos e estabelece uma conexão com um ambiente de execução do Spark.

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

A sigla RDD siginifica algo como 'Conjuntos de Dados Distribuídos Resilientes' e é uma espécie de abstração para representar os dados. De maneira mais formal, é uma coleção de registros particionados, somente leitura, sob o qual se pode trabalhar com a utilização e APIs. É uma solução com bom desepenho para o processamento de grandes volumes de dados em estruturas de computação em cluster.

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

A maior parte da documentação de Spark recomenda a transformação reduceByKey. A razão é que ela implementa um combinador do lado do mapa que executa alguma agregação na memória do lado do mapa. Isso reduz a quantidade de dados embaralhados e evita possíveis exceções de falta de memória, algo que pode ocorrer com a transformação reduceByKey.