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

Armazenar em memória o comando de transformação dado até um comando de ação for executada.

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

Pois o Spark realiza as ações e comandos utilizando a memória principal dos nodes, evitando I/O desnecessários dos discos. Devido também por utilizar os RDDs para processamento em memória, fazendo com que seja muito mais rápido que o MapReduce.

**Qual é a função do SparkContext?**

É o ponto de entrada para as funcionalidades do Spark, inciando a conexão com o Spark Cluster. Através dele que é possível chamar outras funções do Spark.

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

É o “core” do Spark, representa uma coleção de elementos particionados que, por meio dele, podem processar/executar ações em paralelo e otimizados. Acaba sendo o diferencial perante outras ferramentas, devido ao seu poder de processamento, por exemplo, em comparação ao Map Reduce.

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

O “ReduceByKey” agrupa os dados em “common Keys” entre cada uma das partições de memória e depois distribui nos nodes, reduzindo a quantidade de informações trabalhadas.