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

O comando cache tem como objetivo persistir os dados com storage level memory\_only.

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

Porque o Spark faz basicamente todo o trabalho em memoria enquanto o MapReduce precisar ler e gravar em disco.

**Qual é a função do SparkContext?**

A função do SparkContext é ser o ponto de entrada da sessão, pode ser usado para criar RDDs,

acumuladores e variáveis de transmissão no cluster.

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

Uma coleção distribuida imutável de objetos tolerantes a falhas que podem ser operados em paralelo. São criados a partir de outro RDD ou de um conjunto externo de dados.

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

O GroupByKey não usa parâmetros e agrupa tudo. Sobrecarrega as partições com transferências de dados e pode resultar em problemas de falta de memória.

CREATE EXTERNAL TABLE logs\_nasa

(host string,data\_a string ,requisicao string ,retorno\_http int ,

total\_bytes int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

LOCATION '/user/cloudera/nasa/';