**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.

**O seguinte código transforma um conjunto de dados e salva em um arquivo.**

// Lê os dados armazenados no HDFS em um RDD de String.

**val textFile = sc.textFile("hdfs://...")**

//Pega o RDD de linhas e transforma em um RDD de palavras.

**val counts = textFile.flatMap(line => line.split(" "))**

//Transforma o RDD de palavras em RDD de valor-chave

**.map(word => (word, 1))**

//cada palavra chave é reduzida e todos os valores são somados

**.reduceByKey(\_ + \_)**

//Salva um arquivo de texto novamente no HDFS

**counts.saveAsTextFile("hdfs://...")**

Os scripts estão no arquivo **respostas.py**

**Questões**

1. Número de hosts únicos.

R: 137979

1. O total de erros 404.

R: 20725

1. Os 5 URLs que mais causaram erro 404.

+-----------------------------+-------+

| requisicao|count|

+-----------------------------+-------+

|GET/pub/winvn/rea...| 2004|

|GET/pub/winvn/rel... | 1732|

|GET/shuttle/missi... | 682 |

|GET/shuttle/missi... | 426 |

|GET/history/apoll... | 384 |

+-----------------------------+-------+

1. Quantidade de erros 404 por dia.

+-----------------+-----+

| data|count|

+-----------------+-----+

|20/Jul/1995 | 428|

|04/Jul/1995 | 357|

|22/Aug/1995| 283|

|06/Aug/1995| 372|

|02/Jul/1995 | 288|

|20/Aug/1995| 312|

|19/Jul/1995 | 639|

|04/Aug/1995| 344|

|17/Jul/1995 | 404|

|19/Aug/1995| 203|

|15/Jul/1995 | 253|

|17/Aug/1995| 265|

|13/Jul/1995 | 524|

|16/Aug/1995| 256|

|11/Jul/1995 | 470|

|30/Aug/1995| 565|

|14/Aug/1995| 284|

|27/Jul/1995 | 333|

|12/Aug/1995| 190|

|29/Aug/1995| 411|

+-------------------+-----+

1. O total de bytes retornados.

R: 65253787975