Qual o objetivo do comando cache em Spark?

É armazenar um dataframe que foi materializado em memória para um uso mais rápido ao realizar outra ação que esteja relacionado. Se o dataframe não estiver em cache o Spark fara todas as transformações novamente ao chamar o dataframe.

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

Isso ocorre porque o MapReduce realizava o processamento em cima dos dados armazenados no datalake, por isso era mais utilizado para o processamento de grandes volumes de informações que já estavam armazenados. Já o Spark como veio para permitir o processamento de dados em tempo real ele utiliza a memória RAM para realizar esse processamento, por ser mais rápida.

Qual é a função do SparkContext?

A função do SparkContext é armazenar as configurações do ambiente de Big Data em que o Spark atua para execução de algumas tarefas. Um exemplo disso é a inicialização do HiveContext que recebe o SparkContext e após ser inicializado permite a execução de comandos HiveQL pelo Spark.

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

RDDs são datasets baseados na execução de funções do tipo map e reduce para o processamento de dados. Sendo que a função map serve para “organizar” os dados de maneira que faça sentido para a ação que quer realizar e reduce para agrupar esses valores mapeados. Ao executar essas funções o RDD pode ser visualizado como uma tabela.

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

Isso ocorre porque o groupByKey no processo de shuffle envia chaves repetidas sem nenhum tratamento para o próximo nó, enquanto o reduceByKey antes de efetuar o envio agrupa essas chaves. Sendo assim o reduceByKey envia menos informações e mais relevantes.

Explique o que o código Scala abaixo faz.

Ele carrega um arquivo de texto que está no file system do hadoop.

Realiza a separação das palavras pelo espaço entre elas (“ “) , mapeia a palavra com o número 1 e, por fim, agrupa as palavras iguais. A partir de agora se uma palavra se repetir terá o número de repetições no lugar do um.

Salva um arquivo com as palavras e a quantidade de repetições no HDFS.

Questões praticas. (Estará ao lado de todas as questões o csv gerado que o corresponde a resposta)

Número de hosts únicos (unique\_hosts.csv):

137978

O total de erros 404 (error\_404\_count.csv):

20901

As 5 URLs que mais causaram erro 404 (error\_404\_count\_top\_urls.csv):

ts8-1.westwood.ts.ucla.edu/images/Nasa-logo.gif - 37

nexus.mlckew.edu.au/images/nasa-logo.gif - 34

203.13.168.17/images/nasa-logo.gif - 25

203.13.168.24/images/nasa-logo.gif - 25

onramp2-9.onr.com/images/nasa-logo.gif – 22

Quantidade de erros 404 por dia (error\_404\_count\_by\_day.csv):

1995-07-27 - 336

1995-07-13 - 532

1995-08-24 - 420

1995-07-07 - 570

1995-07-12 - 471

1995-07-06 - 640

1995-07-17 - 406

1995-08-11 - 263

1995-07-01 - 316

1995-08-03 - 304

1995-07-22 - 192

1995-07-23 - 233

1995-08-15 - 327

1995-07-02 - 291

1995-08-07 - 537

1995-07-24 - 328

1995-08-27 - 370

1995-07-16 - 257

1995-08-09 - 279

1995-08-01 - 243

1995-08-10 - 315

1995-08-29 - 420

1995-07-14 - 413

1995-08-22 - 288

1995-08-08 - 391

1995-07-11 - 471

1995-08-13 - 216

1995-07-19 - 639

1995-07-25 - 461

1995-08-31 - 526

1995-07-05 - 497

1995-08-18 -256

1995-07-04 - 359

1995-08-04 - 346

1995-08-19 - 209

1995-07-18 - 465

1995-08-05 - 236

1995-08-26 - 366

1995-08-12 - 196

1995-07-09 - 348

1995-08-16 - 259

1995-08-28 - 410

1995-07-21 - 334

1995-07-15 - 254

1995-08-25 - 415

1995-07-26 - 336

1995-07-08 - 302

1995-07-28 - 94

1995-08-20 - 312

1995-07-20 - 428

1995-07-03 - 474

1995-08-14 - 287

1995-08-23 - 345

1995-08-17 - 271

1995-08-21 - 305

1995-07-10 - 398

1995-08-30 - 571

1995-08-06 – 373

O total de bytes retornado (total\_bytes.csv):

65524314915