RAPHAEL CASTILHO GIL

Perguntas descritivas

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

1 – O objetivo do comando cache seria para armazenar os dados em memória RAM e utilizar em comandos posteriores, evitando assim tempo de demora durante o acesso do HD e aproveitar a velocidade de acesso com a RAM.

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

2 – A velocidade é diferente devido a forma de processamento dos dados, além de poder manter na memória ram os dados necessários, sem a necessidade de coletar em diferentes clusters, o mapreduce também precisa preparar a JVM para cada task, o que consome alguns segundos. No caso do spark isso não ocorre visto que o spark mantêm em cada nó uma JVM rodando.

Qual é a função do **SparkContext** ?

3 – SparkContext serve para criar conexões aos ambientes de execução do servidor SPARK, sendo assim um cliente que ficará disponível até a finalização do job.

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

4 – RDD significa uma massa de dados que são imutáveis, que podem ser distribuídos em pequenas partições para 1 ou mais clusters com o fim de realizar paralelismo caso haja alguma necessidade de transformação ou leitura desse dado;

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

5 – ReduceByKey é mais eficiente, ficando mais evidente em base de dados grandes, já que o groupbykey não combina os dados antes de realizar o shuffle, ou seja, o tráfego de dados é extremamente superior durante o shuffle dos dados o que compromete no desempenho da aplicação.

Explique o que o código Scala abaixo faz.

Primeira linha pega o endereço onde será coletado os dados, provindo de um sistema distribuido(haddop).

Na segunda até quarta linha ele separa as linhas e conta as palavras e as quantidades de palavras que foram coletadas no arquivo. Para isso o reducebykey serve para criar um índex sobre a palavra e conta essas palavras disponíveis.

Na última linha, o sistema salva o arquivo em um sistema distribuído (haddop)

HTTP requests to the NASA Kennedy Space Center WWW server

Número de hosts únicos.

Numero de hosts únicos: 137978

O total de erros 404.

Quantidade de erros com 404: 20890

Os 5 URLs que mais causaram erro 404.

5 URLS com maiores logs 404 sao: [(u'hoohoo.ncsa.uiuc.edu', 251), (u'piweba3y.prodigy.com', 157), (u'jbiagioni.npt.nuwc.navy.mil', 132), (u'piweba1y.prodigy.com', 114), (u'www-d4.proxy.aol.com', 91)]

O total de bytes retornados.

Soma dos bytes: 65524261001

Quantidade de erros 404 por dia.

Compilado da resposta:

Dia	Quantidade de logs 404
01/07/1995	316
02/07/1995	291
03/07/1995	474
04/07/1995	359
05/07/1995	497
06/07/1995	640
07/07/1995	570
08/07/1995	299
09/07/1995	348
10/07/1995	398
11/07/1995	471
12/07/1995	467
13/07/1995	530
14/07/1995	413
15/07/1995	254
16/07/1995	257
17/07/1995	406
18/07/1995	465
19/07/1995	639
20/07/1995	428
21/07/1995	334
22/07/1995	192
23/07/1995	233

24/07/1995	328
25/07/1995	461
26/07/1995	336
27/07/1995	336
28/07/1995	94
29/07/1995	0
30/07/1995	0
31/07/1995	0
01/08/1995	243
02/08/1995	0
03/08/1995	303
04/08/1995	346
05/08/1995	236
06/08/1995	373
07/08/1995	537
08/08/1995	391
09/08/1995	279
10/08/1995	314
11/08/1995	263
12/08/1995	196
13/08/1995	216
14/08/1995	287
15/08/1995	327
16/08/1995	259
17/08/1995	271
18/08/1995	256
19/08/1995	209
20/08/1995	312
21/08/1995	305
22/08/1995	288
23/08/1995	345
24/08/1995	420
25/08/1995	415
26/08/1995	366

27/08/1995	370
28/08/1995	410
29/08/1995	420
30/08/1995	571
31/08/1995	526

Resposta do sistema:


```
[(datetime.datetime(1995, 7, 1, 0, 0), 316),
 (datetime.datetime(1995, 7, 2, 0, 0), 291),
 (datetime.datetime(1995, 7, 3, 0, 0), 474),
 (datetime.datetime(1995, 7, 4, 0, 0), 359)
, (datetime.datetime(1995, 7, 5, 0, 0), 497),
 (datetime.datetime(1995, 7, 6, 0, 0), 640),
 (datetime.datetime(1995, 7, 7, 0, 0), 570),
(datetime.datetime(1995, 7, 8, 0, 0), 299),
 (datetime.datetime(1995, 7, 9, 0, 0), 348),
 (datetime.datetime(1995, 7, 10, 0, 0), 398),
 (datetime.datetime(1995, 7, 11, 0, 0), 471),
 (datetime.datetime(1995, 7, 12, 0, 0), 467),
 (datetime.datetime(1995, 7, 13, 0, 0), 530),
(datetime.datetime(1995, 7, 14, 0, 0), 413),
 (datetime.datetime(1995, 7, 15, 0, 0), 254),
 (datetime.datetime(1995, 7, 16, 0, 0), 257),
 (datetime.datetime(1995, 7, 17, 0, 0), 406),
 (datetime.datetime(1995, 7, 18, 0, 0), 465),
 (datetime.datetime(1995, 7, 19, 0, 0), 639),
 (datetime.datetime(1995, 7, 20, 0, 0), 428),
(datetime.datetime(1995, 7, 21, 0, 0), 334),
 (datetime.datetime(1995, 7, 22, 0, 0), 192),
 (datetime.datetime(1995, 7, 23, 0, 0), 233),
 (datetime.datetime(1995, 7, 24, 0, 0), 328),
 (datetime.datetime(1995, 7, 25, 0, 0), 461),
 (datetime.datetime(1995, 7, 26, 0, 0), 336),
```

(datetime.datetime(1995, 8, 1, 0, 0), 243),
(datetime.datetime(1995, 8, 3, 0, 0), 303),
(datetime.datetime(1995, 8, 4, 0, 0), 346),
(datetime.datetime(1995, 8, 5, 0, 0), 236),
(datetime.datetime(1995, 8, 6, 0, 0), 373),
(datetime.datetime(1995, 8, 7, 0, 0), 537),
(datetime.datetime(1995, 8, 8, 0, 0), 391),
(datetime.datetime(1995, 8, 9, 0, 0), 279),
(datetime.datetime(1995, 8, 10, 0, 0), 314),
(datetime.datetime(1995, 8, 11, 0, 0), 263),
(datetime.datetime(1995, 8, 12, 0, 0), 196),
(datetime.datetime(1995, 8, 13, 0, 0), 216),
(datetime.datetime(1995, 8, 14, 0, 0), 287),
(datetime.datetime(1995, 8, 15, 0, 0), 327),
(datetime.datetime(1995, 8, 16, 0, 0), 259),
(datetime.datetime(1995, 8, 17, 0, 0), 271),
(datetime.datetime(1995, 8, 18, 0, 0), 256),
(datetime.datetime(1995, 8, 19, 0, 0), 209),
(datetime.datetime(1995, 8, 20, 0, 0), 312),
(datetime.datetime(1995, 8, 21, 0, 0), 305),
(datetime.datetime(1995, 8, 22, 0, 0), 288),
(datetime.datetime(1995, 8, 23, 0, 0), 345),
(datetime.datetime(1995, 8, 24, 0, 0), 420),
(datetime.datetime(1995, 8, 25, 0, 0), 415),
(datetime.datetime(1995, 8, 26, 0, 0), 366),
(datetime.datetime(1995, 8, 27, 0, 0), 370),
(datetime.datetime(1995, 8, 28, 0, 0), 410),
(datetime.datetime(1995, 8, 29, 0, 0), 420),
(datetime.datetime(1995, 8, 30, 0, 0), 571),
(datetime.datetime(1995, 8, 31, 0, 0), 526)]