Universidade Federal do Amazonas
Projeto de implementação 2 - Banco de dados II
Hilton dos Santos Costa Neto - 21751599

Para a execução do projeto foi utilizado o ambiente Spark, assim como consultas geradas a partir da manipulação de RDD's regulares. No decorrer do projeto, a extração dos dados não será realizada a partir de um RDD previamente organizado para todas as consultas, as consultas possuem similaridade no processo de extração de dados. Ou seja, o arquivo de entrada não foi transformado em único RDD visando a praticidade da manipulação dos dados para a realização de todas as consultas estipuladas no projeto. O motivo para a adoção dessa postura se deu por conta de problemas envolvendo memória. É verdade existir um comando para alocação de memória, "***spark.execute.memory",*** a ser utilizada, contudo o problema persistiu. As consultas se encontram no arquivo "***PI2_Hilton_Costa.py"*** e se encontram em grandes blocos comentados no arquivo.

A primeira consulta consiste em listar 5 comentários mais úteis e com maior avaliação assim como 5 comentários mais úteis e com menor avaliação dado um produto. Para gerar essa consulta foi utilizado o produto de id=15 (assim como para as demais consultas), produto dado de exemplo no projeto. A princípio foram gerados 2 RDD's povoados com dados ASIN e Reviews a partir da filtragem de linhas do RDD da Amazon. A solução utiliza o artifício de que cada linha no RDD de ASIN pertence a mesma linha -1 no RDD de review. Apenas 2 atributos não estão mapeados na mesma posição quando gerado um RDD para os demais atributos: ASIN, clientes nos comentários. O motivo dos dados de ASIN não estarem mapeados precisamente com os demais consiste no fato de que o produto de Id 0 não possuir os demais atributos, apenas o ASIN, já os comentários dos clientes variam muito de acordo com o produto.

Outro RDD criado no processo foi o com comentários, que filtrou o "rating: " nas linhas do arquivo de entrada. Como as linhas dos RDD's gerados a partir de ASIN e review são mapeáveis, foi gerado a lista contendo valores ASIN e coletado o índice do ASIN pertencente ao produto de id 15, a partir da função **"*lista.index()"***. Tendo em conhecimento o índice do produto em ASIN, o mesmo índice subtraindo 1 é utilizado para encontrar a review correspondente na lista de reviews. A review localizada com o índice de ASIN pertence ao mesmo produto e tem importante função como cabeçalho, para localizar o começo do comentário na lista de comentários. Como dito antes, gerar o RDD de comentários se baseou na filtragem das linhas a partir da palavra "rating:" que está presente tanto nas linhas de reviews como nos próprios comentários. Como os comentários possuem um "cabeçalho" que são as reviews, utilizando o index é possível localizar a posição do comentário de determinado produto e dar um salto. As reviews indicam a quantidade de comentários, logo percorre-se a quantidade necessária de linhas e coleta os comentários do produto em questão. A partir dos comentários foi gerado uma lista de tuplas com a data, cliente, rating e helpfull e gerado um RDD, que posteriormente sofre uma ordenação na coluna de rating, retornando comentários com boas avaliações e úteis assim como outra ordenação na coluna 3, com comentários úteis e com menor avaliação.

+helpful/+rating			
Data	**Cliente**	**Rating**	**Helpful**
2002-5-13	A2IGOA66Y6O8TQ	5	2
2002-6-17	A2OIN4AUH84KNE	5	1
2004-2-24	A2C5K0QTLL9UAT	5	2
2004-10-13	A5XYF0Z3UH4HB	5	1
2003-6-7	A2FDJ79LDU4O18	4	1
+helpful/-rating			
2002-5-13	A2IGOA66Y6O8TQ	5	2
2004-2-24	A2C5K0QTLL9UAT	5	2
2002-6-17	A2OIN4AUH84KNE	5	1
2003-1-2	A2HN382JNT1CIU	1	1
2003-6-7	A2FDJ79LDU4O18	4	1

Para a segunda consulta, foi sugerido que dado produto, listar os produtos similares com maiores vendas do que o próprio produto. Na consulta foi gerado RDD's como salesrank, ASIN e similar. A mesma ideia anteriormente descrita acima foi utilizada. A partir do index na lista ASIN descobre-se a localização de determinado produto (índice-1) nos demais atributos exceto para os comentários. Este índice gerado coleta o valor de salesrank do produto em questão e a lista de ASIN's disposta no atributo "similar: ". Todos os produtos similares estão dispostos numa string grande, necessitando quebrar a string em palavras. Com toda essa informação um loop foi gerado com base na quantidade de produtos similares existentes e todo o processo se repete, descobre-se o índice do ASIN em questão, coleta o salesrank e compara com o produto da consulta e gera uma lista final com os resultados, processo bem simples.

ASIN (1559362022)	**Salesrank (518927)**
1559361247	275936
1559360828	283821
1559361018	396273

“Dado um produto, mostrar a evolução diária das médias de avaliação ao longo do intervalo de tempo coberto no arquivo de entrada”.
Com base nas características dessa consulta, pode-se perceber semelhança na execução da primeira consulta, onde devemos entrar com um código ASIN e retornar a evolução diária das avaliações, tirar média dos rating baseado nos anos. Basicamente o mesmo processo para encontrar os comentários da consulta “a” foi utilizado, diferenciando na forma como a lista de tuplas são geradas, optando por pegar como chave os anos e o próprio rating dos comentários como valor. A partir da geraração da lista de tuplas, os dados são transformados em RDD e passam por algumas manipulações como o reduceByKey e o próprio groupByKey, onde todos os rating ligados a uma chave são somados e todos os rating ligados a uma chave são agrupados num objetoRDD. Como os dados gerados possuem o mesmo tamanho, um loop foi introduzido como forma de gerar as médias, dividindo o somatório com o tamanho do objetoRDD.

ANO	**AVG**
2002	5,0
2003	3,0
2004	3,66

A consulta a seguir consiste em listar os 10 produtos líderes de venda de cada grupo de produto: Book, Music, DVD e VIDEO. A consulta é simples e consiste em gerar uma lista de tuplas baseado em RDD’s de groups, salesrank e ASIN. A tupla possui em sua constituição 3 colunas com elementos dos RDD’s mencionados, como o ASIN, o nome do grupo e o valor de salesrank como inteiro. A lista é transformada em um RDD e passa por uma filtragem devido possuir valor de salesrank igual a -1, tendo em vista que -1 foi considerado um valor de marcação para produtos que não tiveram vendas. Caso os valores de -1 não fossem retirados, apenas valores que não tiveram vendas seriam os mais vendidos, já que o produto mais próximo de zero possui um status de muitas vendas, logo não faz sentido manter o -1. Por fim, o RDD gerado passa por mais 4 filtragens a fim de pegar os grupos além da utilização de uma função chamada takeOrdered que gera uma action definida para retornar os 10 elementos de forma ordenada crescentemente.

ASIN	grupo	salesrank
1590520319	Book	0
0743406842	Book	0
0821774565	Book	0
B00006IJJB	Book	0
0714837598	Book	0
1584652071	Book	0
0312266596	Book	0
076030596X	Book	0
1558703616	Book	0
1583500480	Book	0

ASIN	grupo	salesrank
B00005A9ZA	Music	0
1930928416	Music	0
B000003JXB	Music	0
0672314894	Music	27
0874860660	Music	33
0811212564	Music	42
9589393268	Music	46
1574533924	Music	53
0471130486	Music	55
0375705104	Music	62

ASIN	grupo	salesrank
B00005M2C8	DVD	0
B00005M2C9	DVD	0
B00005M2CZ	DVD	0
0465026087	DVD	28
6302294274	DVD	47
B000009CRT	DVD	49
0812931394	DVD	55
081983078X	DVD	85
0520203240	DVD	85
0553582062	DVD	88

ASIN	grupo	salesrank
078900190X	VIDEO	0
052158812X	VIDEO	1
B00000JWW9	VIDEO	2
0201877562	VIDEO	6
1571515291	VIDEO	7
0874172861	VIDEO	8
0787952842	VIDEO	12
0823203719	VIDEO	14
0671536648	VIDEO	16
0851158455	VIDEO	17

Neste ponto, a mesma abordagem no processo da consulta anterior devido compartilhar característica semelhante de trabalhar com os grupos e, portanto, a aplicação do mesmo princípio é utilizada. A consulta atual pede os 10 produtos com maior média de avaliações úteis positivas por produto. A geração de RDD's baseado na filtragem de group, ASIN e reviews são necessárias para a construção do RDD disposto de linhas formado por tuplas com o grupo, ASIN e um float indicando a média de avaliações retirado de reviews. A partir desse ponto o mesmo processo feito na consulta anterior foi realizado. Ao todo 4 filtragens contendo o nome dos grupos nas linhas do RDD seguido de uma ordenação decrescente e resultando em 10 tuplas para cada grupo como resultado.

grupo	ASIN	AVG
Book	0827229534	5,0
Book	0486287785	5,0
Book	0871318237	5,0
Book	3895780812	5,0
Book	0895872218	5,0
Book	0439240751	5,0
Book	1573221740	5,0
Book	0590568833	5,0
Book	0944708498	5,0
Book	0613100093	5,0

grupo	ASIN	AVG
Music	B000007R0T	5,0
Music	B00004W1WK	5,0
Music	B000000HFH	5,0
Music	B000003ONE	5,0
Music	B000003ONP	5,0
Music	B000000HF0	5,0
Music	B00005LAPN	5,0
Music	0062514547	5,0
Music	0873525825	5,0
Music	1879773058	5,0

grupo	ASIN	AVG
DVD	B00000IC8Z	5,0
DVD	B00007L4N1	5,0
DVD	1892629011	5,0
DVD	0764315854	5,0
DVD	B00005YTR7	5,0
DVD	B000008NVU	5,0
DVD	0767901231	5,0
DVD	B00002MY6L	5,0
DVD	0486294412	5,0
DVD	0788190202	5,0

grupo	ASIN	AVG
VIDEO	B0000060T5	5,0
VIDEO	B00000IC8N	5,0
VIDEO	6304733542	5,0
VIDEO	6301045734	5,0
VIDEO	1885593694	5,0
VIDEO	1551802775	5,0
VIDEO	0486284999	5,0
VIDEO	0486420116	5,0
VIDEO	0966880706	5,0
VIDEO	8408044133	5,0