

## Universidade Federal do Amazonas

### Projeto de implementação 2: Parte C - Banco de dados II

Hilton dos Santos Costa Neto - 21751599

Para a parte C do projeto de implementação, foi passado como foco desenvolver consultas utilizando dataframes e SQL. Apesar da utilização de dataframes, a estruturação dos dados para a realização das consultas continua sendo a mesma, utilizando RDD's regulares. Cada consulta gera um parser e estrutura os dados a serem trabalhados conforme a consulta. Diferentemente da parte A, do projeto de implementação 2, produtos diferentes foram utilizados como entrada para a realização das consultas. Mesma estrutura do script da parte A: consultas estão em blocos no script que se encontram comentados.

Ao listar 5 comentários mais úteis com maior avaliação e os mais úteis de menor avaliação, uma certa confusão predomina. Por exemplo, na parte A, foi decidido tratar esse problema ordenando pelo campo *"rating"*: para comentários de melhor avaliação. Assim como ordenar pelo campo *"helpful"*: para comentários mais úteis, já que comentários de menor avaliação poderiam aparecer normalmente, tendo em vista que nenhuma restrição foi imposta a eles. O mesmo pode acontecer ordenando por *"rating"*: produtos menos úteis podem aparecer. O ponto é que dificilmente algo bem avaliado não seja útil, assim como algo bem útil seja péssimo avaliado. Portanto, o critério da consulta é atendido de alguma forma. Para a parte C foi utilizado um novo critério para a realização da consulta, gerando resultado diferente.

Os dados são gerados baseando-se nas filtragens dos atributos relativos a consulta, de forma independente e julgando a existência de um mapeamento entre o índice das diferentes filtragens. Por exemplo, filtrando por ASIN, todas as linhas com ASIN são coletadas, filtrando por salesrank todas as linhas com salesrank são coletadas. Pegando as duas primeiras linhas das estruturas que armazenam os dados filtrados, podemos perceber que na posição inicial 0, o produto destacado possui apenas o Id e ASIN e mais nenhum outro atributo, logo a estrutura ASIN não está mapeado totalmente com salesrank, devido o produto inicial ter um código ASIN enquanto não possui salesrank. O mapeamento então fica: ASIN na posição um com a posição zero em salesrank, gerando duas informações de um mesmo produto.

Utilizando esse mapeamento, foi possível localizar o índice do produto procurado a partir da estrutura ASIN. Utilizando esse índice-1, foi possível localizar a review gerada a partir da filtragem de *"review"*, com a review do produto em questão foi possível localizar os comentários gerados a partir da filtragem de *"rating"*, que inclui as reviews como cabeçalho dos comentários. O mesmo procedimento da parte A foi feito.

Depois de todo o processo de coleta vem a organização dos dados em tuplas, com as colunas *"Data"*, *"Cliente"*, *"Rating"* e *"Helpful"*. Logo após gerado um dataframe com colunas de mesmo nome. Com o dataframe pronto, as consultas utilizando dataframe consistem em: gerar tabela filtrando a coluna *"Rating"* apenas com valores iguais a 5, para comentários mais úteis e maior avaliados; gerar tabela com *"Rating"* menor que 5, para comentários de menor avaliação.

Em ambas as tabelas, ordenar por “Helpful” de forma decrescente. Esse foi o novo critério utilizado para a realização da consulta. A imagem a seguir representa o resultado da consulta realizada sobre o produto de **Id=6**, **ASIN=0486220125**:

```
# 5 comentários mais úteis e com maior avaliação
```

Data	Cliente	Rating	Helpful
2003-10-21	A2WSI8HOWHFDOT	5	25
2002-11-8	A2NJO6YE954DBH	5	21
2000-4-28	A2CHULHA03A9BY	5	16
1998-10-11	AUEZ7NVOEHYRY	5	12
1997-7-4	ATVPDKIKX0DER	5	11

```
# 5 comentários mais úteis e com menor avaliação
```

Data	Cliente	Rating	Helpful
2000-1-4	AJYG6ZJUQPZ9M	4	10
2003-10-26	A3BGC9MSXGM0WH	4	10
2002-12-4	A393PYR83LT7R8	1	8
2003-10-10	A32Z5HQGTG5V49	4	4
2004-5-1	A30JTDN020MAJB	4	4

*“Dado um produto, listar os produtos similares com maiores vendas do que ele.”*

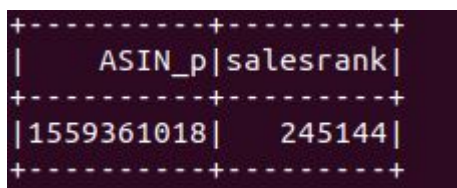
Aqui, foram gerados dois RDD's filtrando linhas com ASIN, salesrank e similar dos produtos em questão. O objetivo é gerar duas tabelas, uma com o ASIN e seu respectivo salesrank e outra com ASIN e seu respectivo ASIN do produto similar. Gerar os dados para as tabelas foi um ponto preocupante por conta de usar loops que por sua vez demoram muito, rodam cerca da quantidade de produtos existentes, assim como também loop com loop interno variando muito, como no caso da segunda tabela mencionada.

Com os dados, foram criados os dois dataframes, *“asin\_sim”* e *“asin\_sales”*. A consulta utilizando dataframe foi um tanto trabalhosa. Devido a pouca experiência com consultas em dataframes, a tentativa constante de forçar o formato de uma consulta SQL foi várias vezes experimentada e com resultado frustrante. Existem momentos onde mudar o nome de uma coluna é necessário, apesar de declarar variáveis ser uma saída para contornar o problema, o método não teve êxito, logo foi tentado usar o **“AS”** para renomear, e também sem êxito. Houve também a tentativa de usar o **“from”** para trabalhar com as duas tabelas.

Para a consulta com o dataframe, foi utilizado um join, mesclando as duas tabelas e pegando apenas as linhas onde os ASIN de ambas as tabelas fossem iguais. A tabela gerada é mais um pouco refinada, agora com linhas contendo ASIN igual ao do produto

pesquisado, gerando todos os ASIN dos produtos similares assim como a coluna salesrank, do produto pesquisado, renomeada para “**produto\_cod**”.

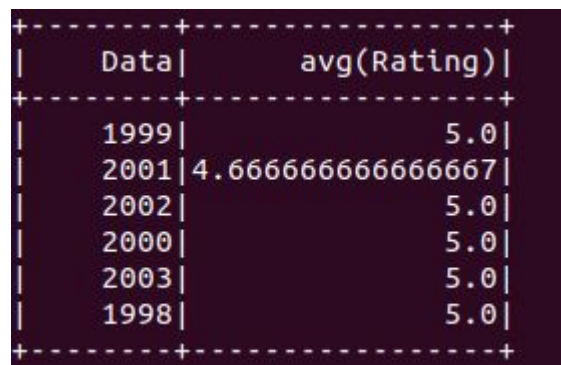
Essa tabela gerada passa por um novo join, filtrando os ASIN similares com os ASIN da primeira tabela que contém os salesrank dos produtos, gerando uma penúltima tabela e passando por uma última filtragem, comparando salesrank com “**produto\_cod**” e pegando os salesrank menor que os produto\_cod. O mesmo processo ocorre com a consulta SQL, a diferença está nas funcionalidades de cada ferramenta de pesquisa. A imagem a seguir representa o resultado da consulta para para o produto de **id=15** e **ASIN = 1559362022**, mesmo produto utilizado na consulta da parte A, porém com o resultado correto, tendo em vista que na linha 72 faltou colocar o “-1”, fazendo alusão ao mapeamento irregular do ASIN com os demais atributos, onde o primeiro produto não possui os demais atributos.



ASIN_p	salesrank
1559361018	245144

*“Dado um produto, mostrar a evolução diária das médias de avaliação ao longo do intervalo de tempo coberto no arquivo de entrada”*

Utilizando os dados gerados pelos RDD's regulares da parte A do projeto de implementação 2, os dados utilizados para compor o dataframe são compostos por tuplas com o ano do comentário e a avaliação correspondente. Os dados gerados pelos RDD's regulares separam antecipadamente os comentários do produto pesquisado. Com isso a consulta consistiu em agrupar a data e fazer a média baseado nela. A imagem a seguir é resultado da consulta utilizando o produto de **Id=18** e **ASIN=B000007R0T**.



Data	avg(Rating)
1999	5.0
2001	4.666666666666667
2002	5.0
2000	5.0
2003	5.0
1998	5.0

*“Listar os 10 produtos líderes de venda em cada grupo de produtos”*

O dataframe para essa consulta possui três colunas “ASIN”, “Grupo” e “salesrank”. Esses dados foram coletados pelos RDD's regulares a partir das filtrações do ASIN, group e salesrank podendo assim serem agrupados em tuplas e servindo de base para o dataframe. As consultas foram simples, ao todo quatro consultas para cada (dataframes e SQL) devido possuírem quatro grandes grupos de produtos.

A consulta consistiu em gerar tabelas filtrando os grupos de interesse e com um salesrank diferente de -1, já que esse valor indica que o produto não teve vendas. A seguir as tabelas são ordenadas de forma ascendente:

ASIN	Grupo	salesrank
B000003JXB	Music	0
1930928416	Music	0
B00005A9ZA	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
0743406842	Book	0
6305901155	Book	0
0821774565	Book	0
0714837598	Book	0
1583500480	Book	0
1584652071	Book	0
1558703616	Book	0
076030596X	Book	0
0312266596	Book	0
B00006IJJB	Book	0

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
0061013846	Video	17

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

*“Listar os 10 produtos com a maior média de avaliações úteis positivas por grupo de produto”*

Por ser bastante semelhante a consulta anterior, todos os RDD's serão mantidos com exceção de salesrank, dando a vez para a média de avaliações do produto. O RDD com a média de avaliações é obtido por meio da filtragem de reviews, gerando sa reviews de todos os produtos. As tuplas serão geradas para compor o dataframe, que por sua vez é composto pelas colunas “Grupo”, “ASIN” e “AVG\_rat”. As consultas são relativamente simples, muito igual aos da consulta realizada anteriormente, diferenciando na coluna a ser ordenada e na forma decrescente de disposição dos dados.



Grupo	ASIN	AVG_rat
Book	0613100093	5.0
Book	0471346608	5.0
Book	0393971694	5.0
Book	0944708498	5.0
Book	0966723821	5.0
Book	3895780812	5.0
Book	1571515429	5.0
Book	0486287785	5.0
Book	0810114275	5.0
Book	1573221740	5.0

Grupo	ASIN	AVG_rat
Music	B00005LAPN	5.0
Music	0262561077	5.0
Music	0062514547	5.0
Music	B000007R0T	5.0
Music	0873525825	5.0
Music	B00004W1WK	5.0
Music	1879773058	5.0
Music	0140433562	5.0
Music	0415143551	5.0
Music	B000003ONE	5.0

Grupo	ASIN	AVG_rat
DVD	B000008NVU	5.0
DVD	0813810426	5.0
DVD	0767901231	5.0
DVD	0764315854	5.0
DVD	B00002MY6L	5.0
DVD	1892629011	5.0
DVD	0486294412	5.0
DVD	1887591400	5.0
DVD	B000005IZS	5.0
DVD	0807730017	5.0

Grupo	ASIN	AVG_rat
Video	0486420116	5.0
Video	0918346177	5.0
Video	0966880706	5.0
Video	B0000060T5	5.0
Video	8408044133	5.0
Video	6301045734	5.0
Video	B00004T22P	5.0
Video	1551802775	5.0
Video	0813432030	5.0
Video	026802765X	5.0