

Data Analytics



BEATRIZ CARDOSO CUNHA

EVERSON DANIEL DA SILVA

GABRIEL SOLER BELMONTE

HARYEL COSTA ASSENÇÃO

WESLEY DE SOUZA

FASE 4 – MBA ENGENHARIA DE SOFTWARE

TURMA 4AOJO

MAIO DE 2021

INTRODUÇÃO

O desafio dessa fase é a transformação de dados semiestruturados em estruturados, considerando que dados de fluxo contínuo com rastros de navegação web são valiosos para análise da audiência dos sites.

A correlação entre as visualizações de produtos e as conversões, a sazonalidade dos acessos, horário e dias da semana, sistemas operacionais, entre outros KPIs, podem criar, orientar, melhorar, as diversas ações da organização.

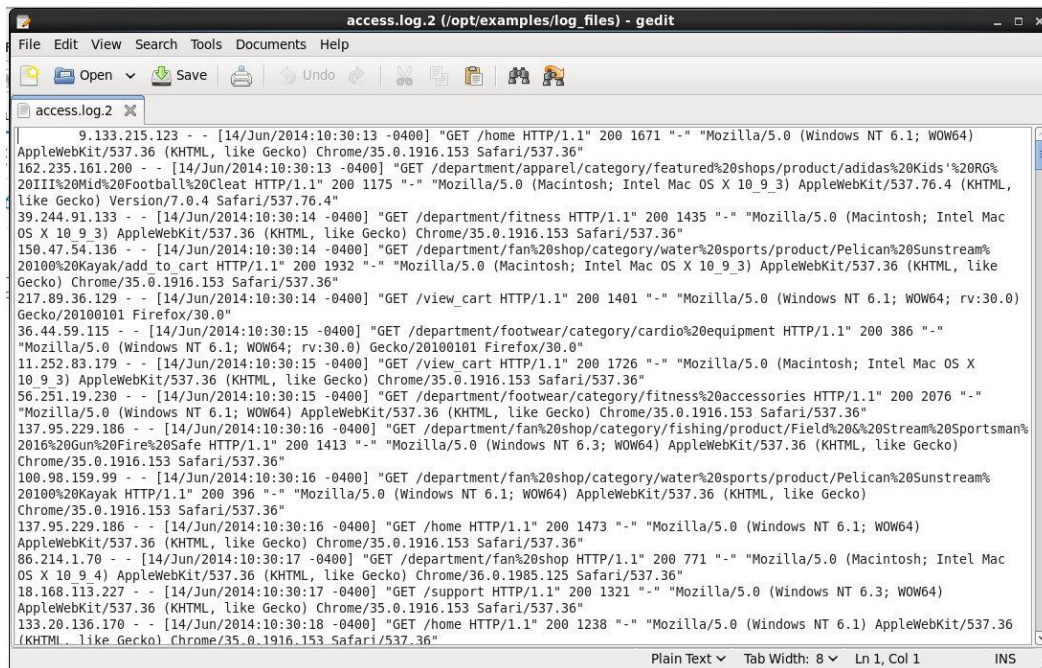
Para isso é necessário extrair informação de grandes bancos de dados, de grandes arquivos de log e tabulá-los de modo visual. Faremos isso usando a VM Cloudera, para uso do Hadoop, Hive e Impala.

Finalmente, a apresentação dos dados através de dashboards, os conhecidos painéis na área de visualização de dados, torna a interpretação dos mesmos - Data Analytics - muito ágil, concisa, trazendo inúmeros pontos para análise, inclusive em tempo real. O que certamente apoia e abre caminho para a tomada de decisões e implementação da cultura data driven, com os quatro tipos de análise de dados possíveis: descritiva, preditiva, prescritiva e diagnóstica, com todas as possibilidades que nos dá o Big Data. Vamos explorar o tema com esse desafio.

DESENVOLVIMENTO

❑ O Desafio

O primeiro passo do desafio é utilizar a VM Cloudera, localizar o arquivo de log de um e-commerce na pasta opt e analisá-lo.



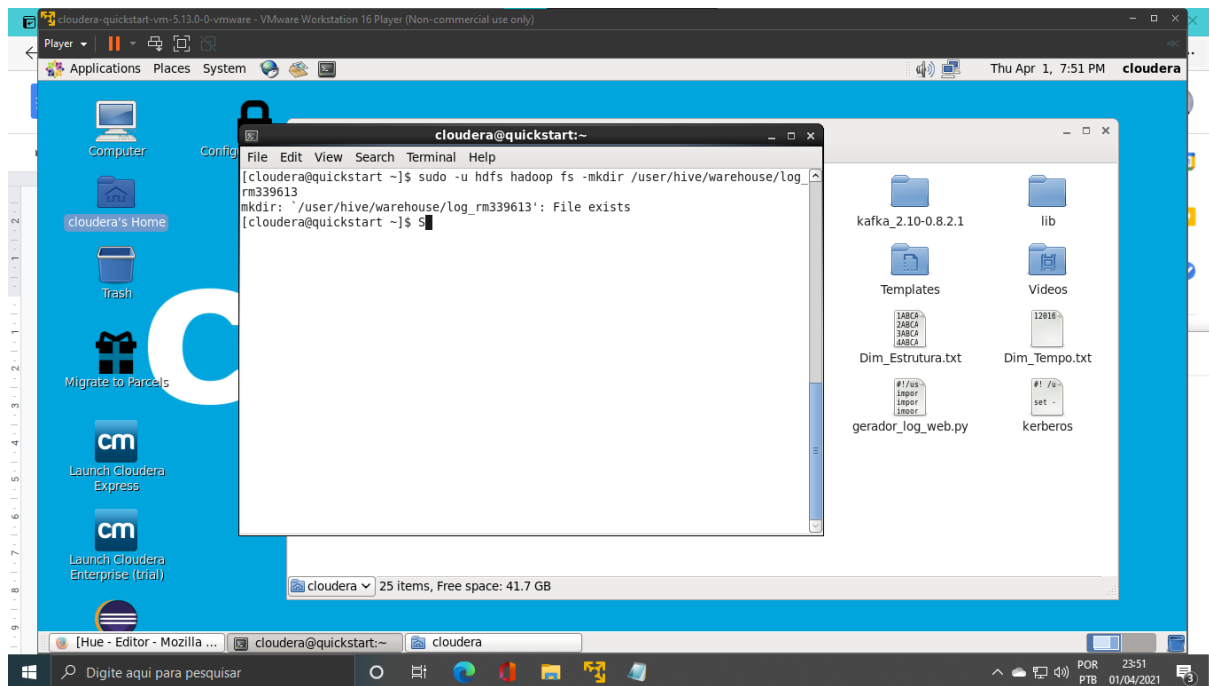
Devemos agora seguir um Road Map seguindo um padrão de acrescentar o número de RM de um dos integrantes do grupo nos arquivos, tabelas, diretórios e scripts. Evidenciar a realização deste Hands On com capturas de tela de cada etapa, montando um arquivo em Word com o passo a passo, o script utilizado no HIVE e no IMPALA.

Para facilitar o entendimento formatamos as instruções do desafio com o // precedendo-as, em cor verde e a seguir o comando em azul. Ademais das capturas de tela, fizemos comentários à medida que encontrávamos dificuldades e nossas soluções para elas (troubleshooting).

❏ O Passo a Passo

//Crie no HDFS, dentro de /user/hive/warehouse um diretório "log" para receber o arquivo;

sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/log_rm339613



Concedemos acesso igual ao material didático

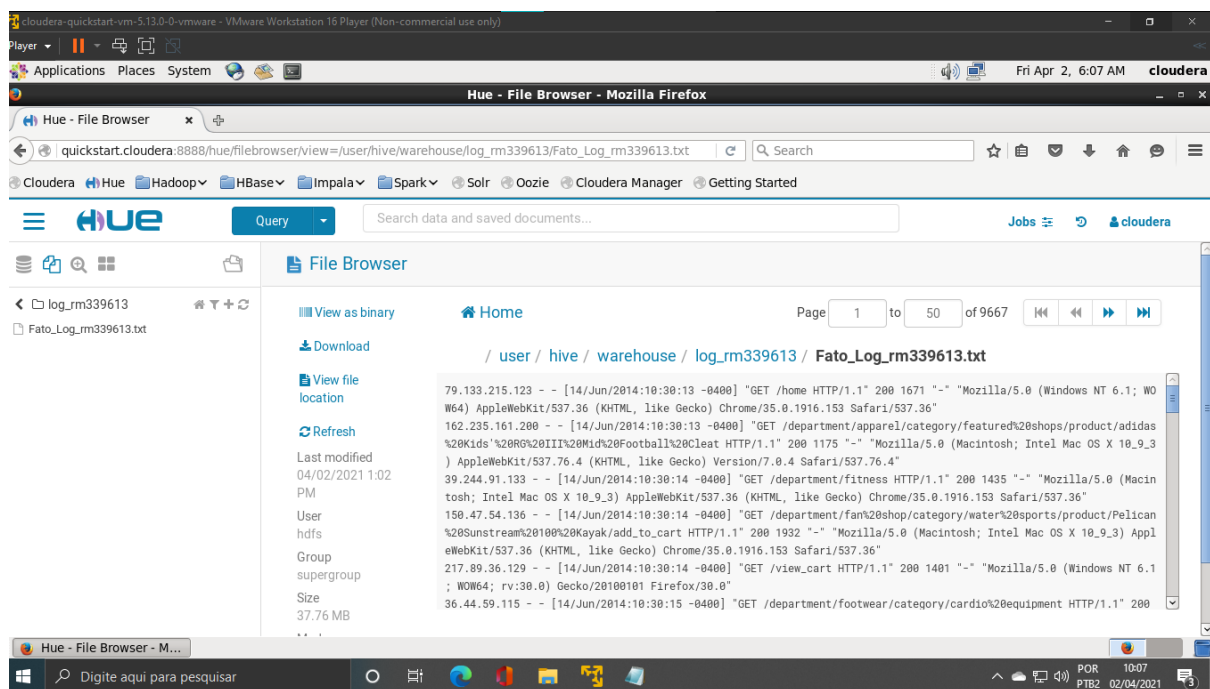
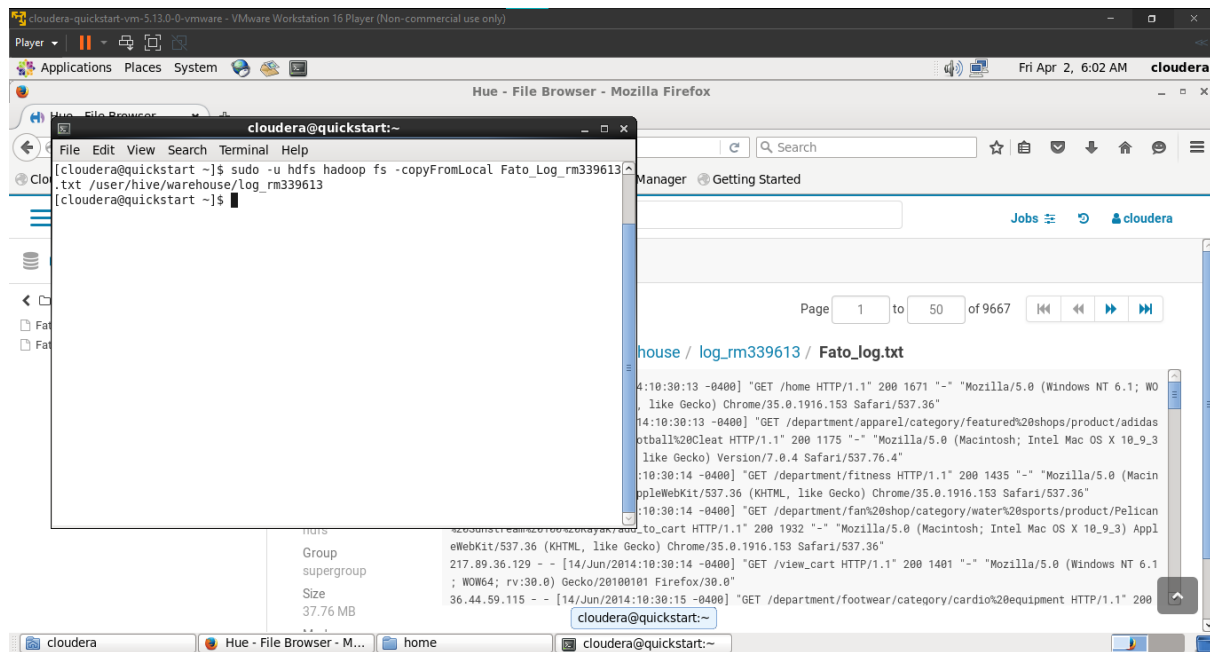
```
chmod 777 -R /home/cloudera/
```

Observação: trocamos o nome do arquivo de log para Fato_Log_rm339613

//Faça a ingestão do arquivo de log para o diretório criado. Você pode fazer a carga do dado via Bulk Upload pelo terminal

```
sudo -u hdfs hadoop fs -copyFromLocal Fato_Log_rm339613.txt  
/user/hive/warehouse/log_rm339613
```

```
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal Fato_Log_rm339613  
.txt /user/hive/warehouse/log_rm339613  
copyFromLocal: `Fato_Log_rm339613.txt': No such file or directory  
[cloudera@quickstart ~]$
```



//Crie uma tabela dentro no editor Hive a partir do log e aplique expressão regular para tabular os dados;

3 minutes ago



```
CREATE EXTERNAL TABLE log_intermediario_rm339613 (  
  ip STRING, data STRING, metodo STRING, url STRING,  
  http_versao STRING, codigo1 STRING, codigo2 STRING,  
  traco STRING, Sistema_operacional STRING)  
ROW FORMAT SERDE  
  'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' WITH  
SERDEPROPERTIES ( 'input.regex' = '([^\ ]*) - - \\[([^\ ]*\]  
\\])*\]\] \"([^\ ]*) ([^\ ]*) ([^\ ]*)\" (\\d*) (\\d*)  
\"([^\"]*)\" \"([^\"]*)\"',  
  
  'output.format.string' = \"%1$$$ %2$$$ %3$$$ %4$$$ %5$$$ %6$$$  
%7$$$ %8$$$ %9$$$\" )  
LOCATION '/user/hive/warehouse/log_rm339613'
```

  default

Tables

(1)    log_intermediario_rm339613

ip (string)
data (string)
metodo (string)
url (string)
http_versao (string)
codigo1 (string)
codigo2 (string)
traco (string)
sistema_operacional (string)

Tivemos que deletar a tabela pois o nome “data” é uma palavra reservada e gerava erro durante a utilização do Impala

“Drop table log_intermediario_rm339613”

The screenshot shows the Hive query editor interface. At the top, there's a header with the Hive logo, a refresh button, and fields for "Add a name..." and "Add a description...". To the right are icons for saving, opening, settings, and help. Below the header, a toolbar shows "0s", "default", "text", and other icons. The main area contains a SQL query:

```
1 DROP TABLE IF EXISTS log_intermediario_rm339613;
2
3 CREATE EXTERNAL TABLE log_intermediario_rm339613 (
4
5 ip STRING,
6
7 dt STRING,
8
9 metodo STRING,
10
11 url STRING
12 )
```

Below the query, a status bar indicates "Success." with a green checkmark. At the bottom, there's a "Query History" section showing a recent query executed "a few seconds ago":

```
Create EXTERNAL TABLE log_intermediario_rm339613 (
ip STRING, dt STRING, metodo STRING, url STRING,
```

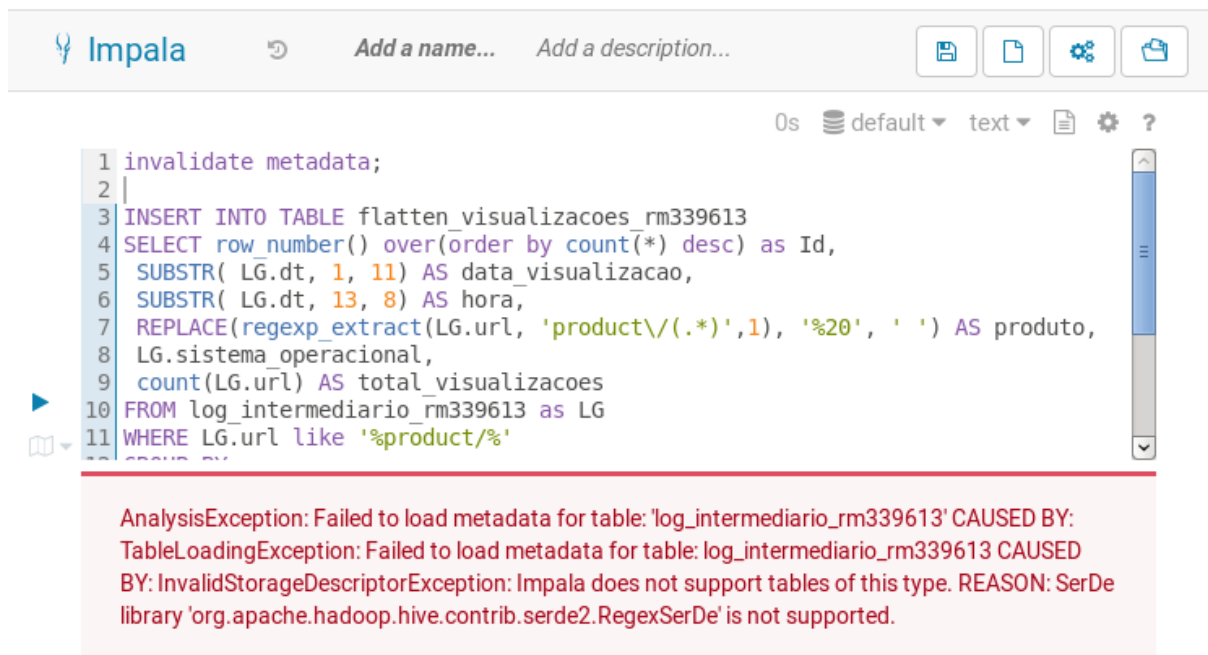
Criamos a tabela no Impala

The screenshot shows the Impala query editor interface. At the top, there's a header with the Impala logo, a refresh button, and fields for "Add a name..." and "Add a description...". To the right are icons for saving, opening, settings, and help. Below the header, a toolbar shows "0s", "default", "text", and other icons. The main area contains a SQL query:

```
1 invalidate metadata;
2
3 Create Table Flatten_Visualizacoes_rm339613
4 (
5   id bigint,
6   data_visualizacao STRING,
7   hora STRING,
8   produto STRING,
9   sistema_operacional STRING,
10  total_visualizacoes BIGINT
11 )
```

Below the query, a status bar indicates "Success." with a green checkmark. At the bottom, there's a "Query History" section showing a recent query executed "a few seconds ago":

```
Create Table Flatten_Visualizacoes_rm339613 ( id bigint,
data_visualizacao STRING, hora STRING, produto STRING,
```



The screenshot shows the Impala SQL interface. At the top, there's a header with the Impala logo, a refresh button, and fields for "Add a name..." and "Add a description...". Below the header, there's a toolbar with icons for saving, opening, settings, and help. The main area displays a SQL query with line numbers 1 through 11. The query starts with "invalidate metadata;" and then inserts data from "log_intermediario_rm339613" into "flatten_visualizacoes_rm339613". The query uses "row_number()" for ordering and "SUBSTR" for date and time extraction. Below the query, a red error message is displayed, stating: "AnalysisException: Failed to load metadata for table: 'log_intermediario_rm339613' CAUSED BY: TableLoadingException: Failed to load metadata for table: log_intermediario_rm339613 CAUSED BY: InvalidStorageDescriptorException: Impala does not support tables of this type. REASON: SerDe library 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' is not supported."

```
1 invalidate metadata;
2 |
3 INSERT INTO TABLE flatten_visualizacoes_rm339613
4 SELECT row_number() over(order by count(*) desc) as Id,
5 SUBSTR( LG.dt, 1, 11) AS data_visualizacao,
6 SUBSTR( LG.dt, 13, 8) AS hora,
7 REPLACE(regexp_extract(LG.url, 'product\/(.*)',1), '%20', ' ') AS produto,
8 LG.sistema_operacional,
9 count(LG.url) AS total_visualizacoes
10 FROM log_intermediario_rm339613 as LG
11 WHERE LG.url like '%product/%'
```

AnalysisException: Failed to load metadata for table: 'log_intermediario_rm339613' CAUSED BY: TableLoadingException: Failed to load metadata for table: log_intermediario_rm339613 CAUSED BY: InvalidStorageDescriptorException: Impala does not support tables of this type. REASON: SerDe library 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' is not supported.

Para corrigir o erro acima, fizemos uma cópia da tabela, depois disso conseguimos executar uma query intermediária para testar se o insert pelo Impala funcionava.



The screenshot shows the Hive SQL interface. At the top, there's a header with the Hive logo, a refresh button, and fields for "Add a name..." and "Add a description...". Below the header, there's a toolbar with icons for saving, opening, settings, and help. The main area displays a SQL query with line numbers 1 through 4. The query creates a new table "copy_of_table_rm339613" stored as Parquet, and then selects all data from "log_intermediario_rm339613" into it. Below the query, there's a "Query History" section with a search icon and a "Saved Queries" section with a search icon and a refresh icon.

```
1 CREATE TABLE copy_of_table_rm339613 STORED AS PARQUET AS
2 |
3 SELECT * FROM log_intermediario_rm339613
4 |
```

Query History Saved Queries

Somente para testes inserimos os dados na tabela flatten.

```
INSERT INTO TABLE flatten_visualizacoes_rm339613
SELECT row_number() over(order by count(*) desc) as Id,
SUBSTR( LG.dt, 1, 11) AS data_visualizacao,
SUBSTR( LG.dt, 13, 8) AS hora,
LG.url AS produto,
LG.sistema_operacional,
count(LG.url) AS total_visualizacoes
FROM copy_of_table_rm339613 as LG
WHERE
  LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart'
GROUP BY
  data_visualizacao,
  hora,
  produto,
  sistema_operacional
ORDER BY COUNT(*) DESC;
```


Agora precisamos utilizar uma expressão Regex para extrair os produtos e contá-los e extrair o nome do sistema operacional.

```

1 SELECT DISTINCT
2   REPLACE(
3     REPLACE(
4       regexp_extract(
5         regexp_extract(
6           regexp_extract(LG.sistema_operacional, '\(.*\)') , 0)
7         , '[^;]+(;|\)|)' , 0)
8       , '[^\)]+' , 0)
9     , ',' , ';' )
10    , '(' , ')' ) AS sistema_operacional
11 FROM
12   copy_of_table_rm339613 as LG
13 WHERE
14   LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart'
15

```

Query History Saved Queries Results (4)

	sistema_operacional
1	Windows NT 6.3
2	Windows NT 6.1
3	Macintosh
4	X11

O “regexp_extract” mais interno é para remover o navegador da string e extrair o que está entre parênteses.

```

"Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Ge

```

O segundo mais interno limpa o “WOW64; rv:30.0” da string de exemplo “(Windows NT 6.1; WOW64; rv:30.0)”

E o “regexp_extract” mais externo limpa essa string abaixo:

```

87 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153

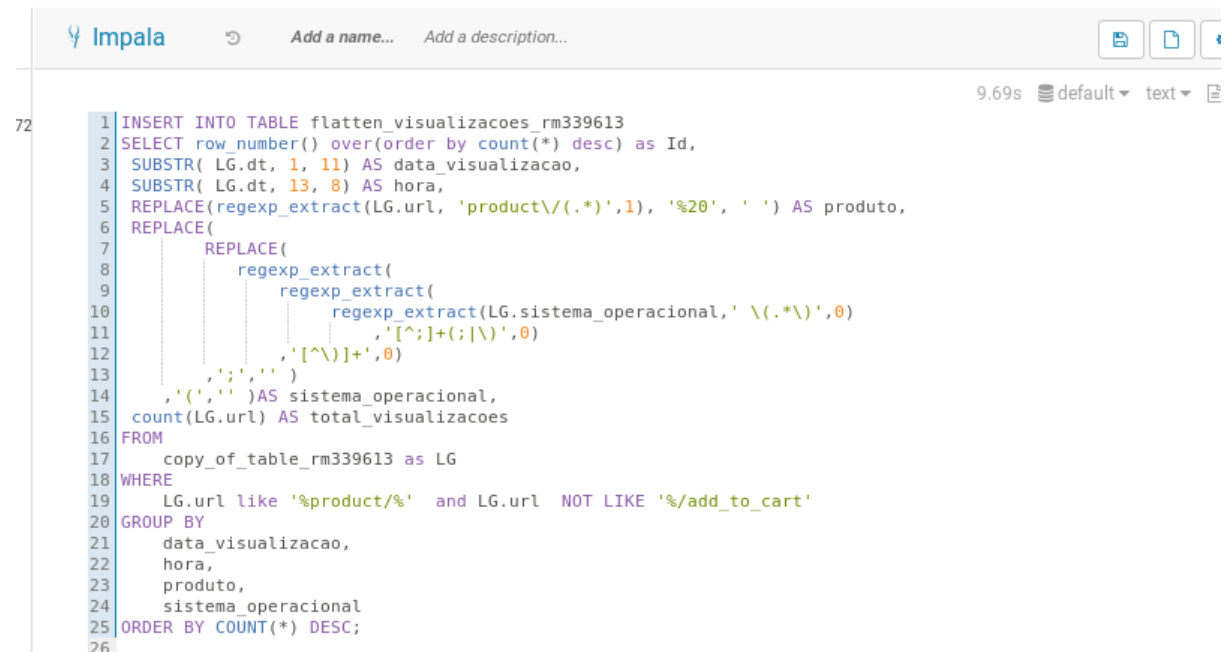
```

Limpamos a tabela de testes.

2 minutes ago	✓	TRUNCATE table flatten_visualizacoes_rm339613
7 minutes ago	✓	SELECT REPLACE(REPLACE(regexp_extract(regexp_extract(LG.sistema_operacional, '\(.*\)') , 0) , '[^;]+(; \))' , 0) , '[^\)]+' , 0) , ',' , ';') , '(' , ')' AS sistema_operacional FROM copy_of_table_rm339613 as LG WHERE LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart'

//Sumarize os dados, filtrando apenas os cliques cujo usuário foi até o nível de produto na URL e conte o total de visualizações. Faça a sumarização no Impala.

Executamos o insert com todos os tratamentos.



The screenshot shows the Impala SQL editor interface. At the top, there's a header bar with the Impala logo, a refresh icon, and fields for 'Add a name...' and 'Add a description...'. On the right, there are icons for saving, running, and other actions. Below the header, the main area displays a SQL query. The query is a multi-line INSERT statement. It starts with 'INSERT INTO TABLE flatten_visualizacoes_rm339613'. The SELECT clause includes 'row_number() over(order by count(*) desc) as Id', 'SUBSTR(LG.dt, 1, 11) AS data_visualizacao', 'SUBSTR(LG.dt, 13, 8) AS hora', and a REPLACE function for 'produto' that uses 'regexp_extract' to pull a product name from the URL. Another REPLACE function follows for 'sistema_operacional', which uses multiple 'regexp_extract' calls to parse a complex system identifier. The final part of the SELECT is 'count(LG.url) AS total_visualizacoes'. The FROM clause is 'copy_of_table_rm339613 as LG'. The WHERE clause filters for 'LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart''. The GROUP BY clause lists 'data_visualizacao', 'hora', 'produto', and 'sistema_operacional'. The query ends with 'ORDER BY COUNT(*) DESC;'. The line numbers 72 through 26 are visible on the left margin.

```
1 INSERT INTO TABLE flatten_visualizacoes_rm339613
2 SELECT row_number() over(order by count(*) desc) as Id,
3 SUBSTR( LG.dt, 1, 11) AS data_visualizacao,
4 SUBSTR( LG.dt, 13, 8) AS hora,
5 REPLACE(regexp_extract(LG.url, 'product\\/(.*)',1), '%20', ' ') AS produto,
6 REPLACE(
7     REPLACE(
8         regexp_extract(
9             regexp_extract(
10                 regexp_extract(LG.sistema_operacional, '\\(.*\\)',0)
11                 , '[^;]+(:|\\)',0)
12                 , '[^\\)]+' ,0)
13             , ';' , '' )
14             , '(' , '' ) AS sistema_operacional,
15 count(LG.url) AS total_visualizacoes
16 FROM
17     copy_of_table_rm339613 as LG
18 WHERE
19     LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart'
20 GROUP BY
21     data_visualizacao,
22     hora,
23     produto,
24     sistema_operacional
25 ORDER BY COUNT(*) DESC;
26
```

```
INSERT INTO TABLE flatten_visualizacoes_rm339613
SELECT row_number() over(order by count(*) desc) as Id,
SUBSTR( LG.dt, 1, 11) AS data_visualizacao,
SUBSTR( LG.dt, 13, 8) AS hora,
REPLACE(regexp_extract(LG.url, 'product\\/(.*)',1), '%20', ' ') AS produto,
REPLACE(
    REPLACE(
        regexp_extract(
            regexp_extract(
                regexp_extract(LG.sistema_operacional, '\\(.*\\)',0)
                , '[^;]+(:|\\)',0)
                , '[^\\)]+' ,0)
            , ';' , '' )
            , '(' , '' ) AS sistema_operacional,
count(LG.url) AS total_visualizacoes
FROM
    copy_of_table_rm339613 as LG
WHERE
    LG.url like '%product/%' and LG.url NOT LIKE '%/add_to_cart'
GROUP BY
    data_visualizacao,
    hora,
    produto,
    sistema_operacional
```

ORDER BY COUNT(*) DESC;

id	data_visualizacao	hora	produto	sistema_operacional	total_visualizacoes
1	14/Jun/2014	22:57:10	Perfect Fitness Perfect Rip Deck	Windows NT 6.1	6
2	14/Jun/2014	21:53:59	adidas Kids' RG III Mid Football Cleat	Macintosh	6
3	14/Jun/2014	19:42:39	Nike Men's Fingertrap Max Training Shoe	Windows NT 6.1	4
4	14/Jun/2014	21:02:58	adidas Youth Germany Black/Red Away Match Soc	Windows NT 6.1	4
5	14/Jun/2014	20:37:18	Nike Men's Comfort 2 Slide	Macintosh	4
6	14/Jun/2014	11:19:21	TYR Boys' Team Digi Jammer	Windows NT 6.3	4
7	14/Jun/2014	22:57:56	adidas Kids' RG III Mid Football Cleat	Windows NT 6.1	4
8	14/Jun/2014	23:24:16	Perfect Fitness Perfect Rip Deck	Macintosh	4
9	14/Jun/2014	20:50:44	adidas Kids' RG III Mid Football Cleat	Macintosh	4
10	14/Jun/2014	22:17:05	The North Face Women's Recon Backpack	Windows NT 6.1	4

Como o resultado das colunas retornou diferente da foto da atividade, foi feita uma validação da primeira linha pelo Hive e depois abrimos o arquivo de origem para ver se realmente estava certo.

	log_intermediario_rm339613.ip	log_intermediario_rm339613.dt	log_intermediario_rm339613.metodo	log_intermediario_rm339613.ur
1	72.162.136.240	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl
2	72.162.136.240	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl
3	66.211.61.116	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl
4	72.162.136.240	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl
5	72.162.136.240	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl
6	66.211.61.116	14/Jun/2014:22:57:10 -0400	GET	/department/apparel/category/cl

Preparando arquivos de saída.

Em seguida mudamos para o Hive e criamos uma tabela de saída.

```

1 CREATE EXTERNAL TABLE flatten_trade_hive_rm339613 (
2   Id BIGINT,
3   data_visualizacao STRING,
4   hora STRING,
5   produto STRING,
6   sistema_operacional STRING,
7   total_visualizacoes BIGINT)
8 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
9 LOCATION '/user/hive/warehouse/flatten_visualizacoes_rm339613';
10
11 ADD JAR /usr/lib/hive/lib/hive-contrib.jar;
12
13 INSERT OVERWRITE TABLE flatten_trade_hive_rm339613

```

Para que a tabulação com vírgula aconteça, precisamos chamar uma biblioteca Java dentro do Hive.

```

9 LOCATION '/user/hive/warehouse/flatten_visualizacoes';
10
11 ADD JAR /usr/lib/hive/lib/hive-contrib.jar;
12
13 INSERT OVERWRITE TABLE flatten_trade_hive_rm339613
14 SELECT Id,

```

Inserimos os dados na tabela de saída.

```

11 ADD JAR /usr/lib/hive/lib/hive-contrib.jar;
12
13 INSERT OVERWRITE TABLE flatten_trade_hive_rm339613
14 SELECT Id,
15 data_visualizacao,
16 hora,
17 produto,
18 sistema_operacional,
19 total_visualizacoes
20 FROM flatten_visualizacoes_rm339613;

```

Gerou o arquivo “000000_0”, em seguida baixamos.

000000_0
_impala_insert_staging

[Download](#)
[View file location](#)
[Refresh](#)
Last modified
05/18/2021 10:09 PM
User
cloudera
Group
supergroup
Size
4.91 MB
Mode
100777

```

1,14/Jun/2014,22:57:10,Perfect Fitness Perfect Rip Deck, Windows NT 6.1 ,6
2,14/Jun/2014,21:53:59,adidas Kids' RG III Mid Football Cleat, Macintosh ,6
3,14/Jun/2014,23:30:43,Under Armour Hustle Storm Medium Duffle Bag, Macintosh ,4
4,14/Jun/2014,20:32:03,Nike Men's CJ Elite 2 TD Football Cleat, Windows NT 6.1 ,4
5,14/Jun/2014,19:38:54,Nike Men's CJ Elite 2 TD Football Cleat, Windows NT 6.1 ,4
6,14/Jun/2014,22:08:40,Pelican Sunstream 100 Kayak, Windows NT 6.1 ,4
7,14/Jun/2014,23:18:07,Perfect Fitness Perfect Rip Deck, Windows NT 6.1 ,4
8,14/Jun/2014,19:58:50,adidas Men's Germany Black Crest Away Tee, Macintosh ,4
9,14/Jun/2014,22:18:41,Pelican Sunstream 100 Kayak, Windows NT 6.1 ,4
10,14/Jun/2014,20:45:42,adidas Kids' RG III Mid Football Cleat, Windows NT 6.1 ,4
11,14/Jun/2014,21:52:37,adidas Youth Germany Black/Red Away Match Soc, Windows NT 6.1 ,4
12,14/Jun/2014,19:32:08,Nike Men's Free 5.0+ Running Shoe, Windows NT 6.1 ,4
13,14/Jun/2014,21:05:04,Nike Men's Free 5.0+ Running Shoe, Windows NT 6.1 ,4
14,14/Jun/2014,23:14:36,O'Brien Men's Neoprene Life Vest, Windows NT 6.1 ,4
15,14/Jun/2014,20:55:56,adidas Brazuca 2014 Official Match Ball, Windows NT 6.1 ,4
16,14/Jun/2014,20:48:04,Nike Men's Fingertrap Max Training Shoe, Windows NT 6.1 ,4
17,14/Jun/2014,20:58:52,adidas Kids' RG III Mid Football Cleat, Windows NT 6.1 ,4
18,14/Jun/2014,23:17:55,adidas Kids' RG III Mid Football Cleat, Macintosh ,4
19,14/Jun/2014,23:24:16,Perfect Fitness Perfect Rip Deck, Macintosh ,4
20,14/Jun/2014,23:05:25,Nike Men's Dri-FIT Victory Golf Polo, Windows NT 6.1 ,4
21,14/Jun/2014,23:16:28,adidas Youth Germany Black/Red Away Match Soc, Windows NT 6.1 ,4
22,14/Jun/2014,22:55:48,Nike Women's Free 5.0 TR FIT PRT 4 Training S, Macintosh ,4
23,14/Jun/2014,18:38:10,Nike Men's Free 5.0+ Running Shoe, Windows NT 6.3 ,4
24,14/Jun/2014,23:25:39,adidas Kids' RG III Mid Football Cleat, Macintosh ,4
25,14/Jun/2014,21:17:06,Titleist Pro V1x High Numbers Personalized Go, Windows NT 6.1 ,4

```

Adicionamos o cabeçalho, já preparando a saída para o Power BI.

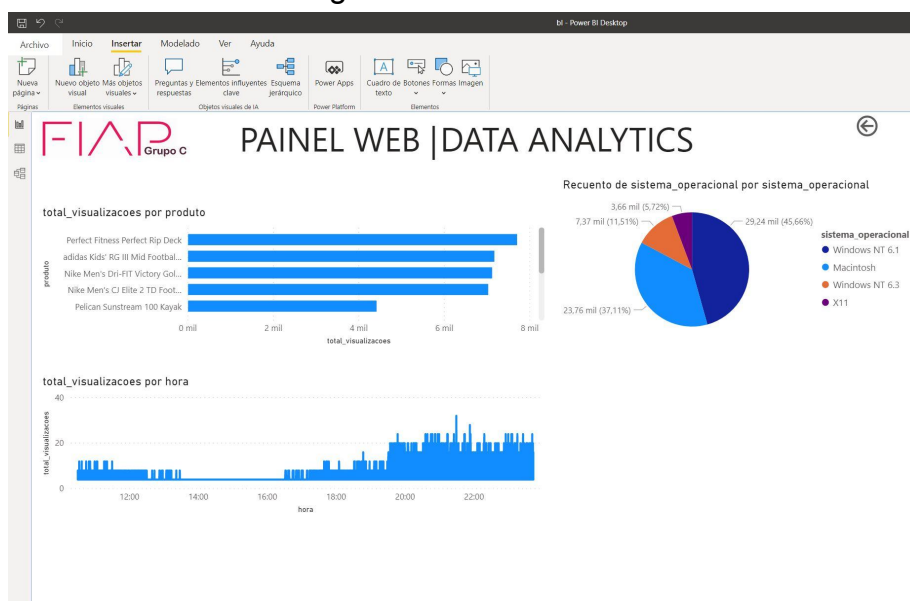
Id	data_visualizacao	hora	produto	sistema_operacional	total_visualizacoes
1	14/Jun/2014	22:57:10	Perfect Fitness Perfect Rip Deck	Windows NT 6.1	6
2	14/Jun/2014	21:53:59	adidas Kids' RG III Mid Football Cleat	Macintosh	6
3	14/Jun/2014	23:30:43	Under Armour Hustle Storm Medium Duffle Bag	Macintosh	4
4	14/Jun/2014	20:32:03	Nike Men's CJ Elite 2 TD Football Cleat	Windows NT 6.1	4
5	14/Jun/2014	19:38:54	Nike Men's CJ Elite 2 TD Football Cleat	Windows NT 6.1	4
6	14/Jun/2014	22:08:40	Pelican Sunstream 100 Kayak	Windows NT 6.1	4
7	14/Jun/2014	23:18:07	Perfect Fitness Perfect Rip Deck	Windows NT 6.1	4
8	14/Jun/2014	19:50:50	adidas Men's Germany Black Crest Away Tee	Macintosh	4
9	14/Jun/2014	22:18:41	Pelican Sunstream 100 Kayak	Windows NT 6.1	4
10	14/Jun/2014	20:45:42	adidas Kids' RG III Mid Football Cleat	Windows NT 6.1	4
11	14/Jun/2014	21:52:37	adidas Youth Germany Black/Red Away Match Soc	Windows NT 6.1	4
12	14/Jun/2014	19:32:00	Nike Men's Free 5.0+ Running Shoe	Windows NT 6.1	4
13	14/Jun/2014	21:05:04	Nike Men's Free 5.0+ Running Shoe	Windows NT 6.1	4
14	14/Jun/2014	23:14:36	O'Brien Men's Neoprene Life Vest	Windows NT 6.1	4
15	14/Jun/2014	20:55:56	adidas Brazuca 2014 Official Match Ball	Windows NT 6.1	4
16	14/Jun/2014	20:48:04	Nike Men's Fingertrap Max Training Shoe	Windows NT 6.1	4
17	14/Jun/2014	20:58:52	adidas Kids' RG III Mid Football Cleat	Windows NT 6.1	4
18	14/Jun/2014	23:17:55	adidas Kids' RG III Mid Football Cleat	Macintosh	4
19	14/Jun/2014	23:24:16	Perfect Fitness Perfect Rip Deck	Macintosh	4
20	14/Jun/2014	23:05:25	Nike Men's Dri-FIT Victory Golf Polo	Windows NT 6.1	4
21	14/Jun/2014	23:16:28	adidas Youth Germany Black/Red Away Match Soc	Windows NT 6.1	4
22	14/Jun/2014	22:55:48	Nike Women's Free 5.0 TR FIT PRT 4 Training S	Macintosh	4
23	14/Jun/2014	18:38:10	Nike Men's Free 5.0+ Running Shoe	Windows NT 6.3	4
24	14/Jun/2014	23:25:39	adidas Kids' RG III Mid Football Cleat	Macintosh	4
25	14/Jun/2014	21:17:06	Titleist Pro Vlx High Numbers Personalized Go	Windows NT 6.1	4
26	14/Jun/2014	23:01:45	Nike Men's CJ Elite 2 TD Football Cleat	Windows NT 6.1	4
27	14/Jun/2014	21:40:53	Nike Men's CJ Elite 2 TD Football Cleat	Windows NT 6.1	4
28	14/Jun/2014	23:27:25	Pelican Sunstream 100 Kayak	Windows NT 6.1	4
29	14/Jun/2014	22:17:05	The North Face Women's Recon Backpack	Windows NT 6.1	4
30	14/Jun/2014	11:19:21	TYR Boys' Team Digi Jammer	Windows NT 6.3	4
31	14/Jun/2014	18:11:40	Diamondback Women's Serene Classic Comfort Bi	Macintosh	4
32	14/Jun/2014	22:24:37	adidas Kids' RG III Mid Football Cleat	Macintosh	4
33	14/Jun/2014	20:05:32	Nike Men's Free TR 5.0 TB Training Shoe	Macintosh	4
34	14/Jun/2014	20:02:08	adidas Kids' RG III Mid Football Cleat	Windows NT 6.1	4

❑ VISUALIZAÇÃO DE DADOS

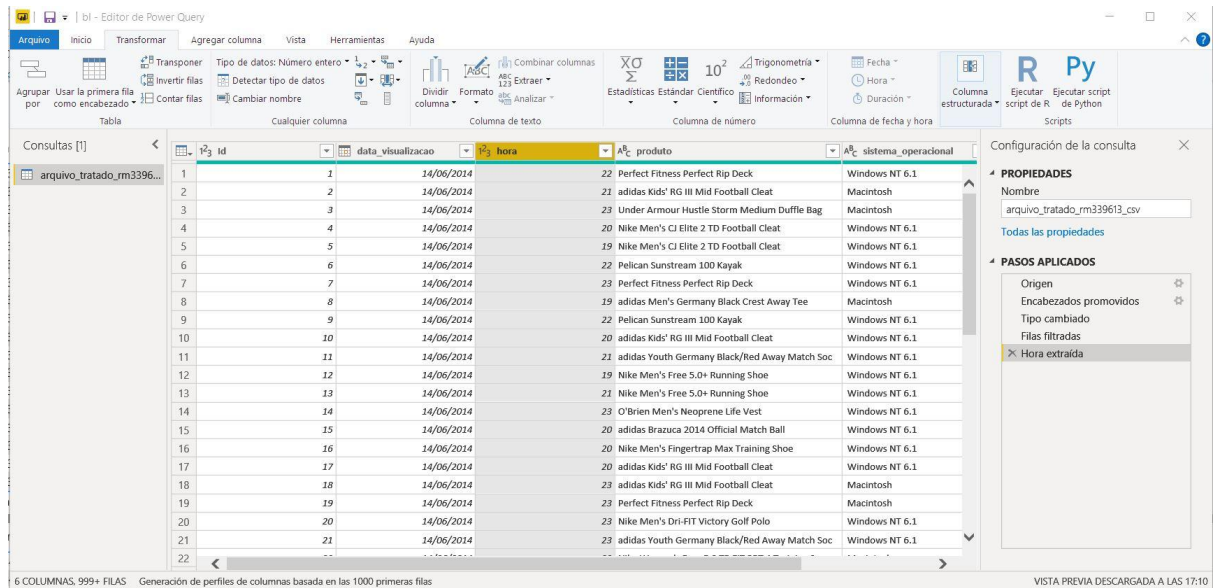
Utilizamos o Power BI para a saída visual dos dados, também conhecida como painel ou “dashboard”.

Procuramos chegar na primeira saída dada pelo desafio e evoluir com os conceitos da visualização de dados para um dashboard mais interativo.

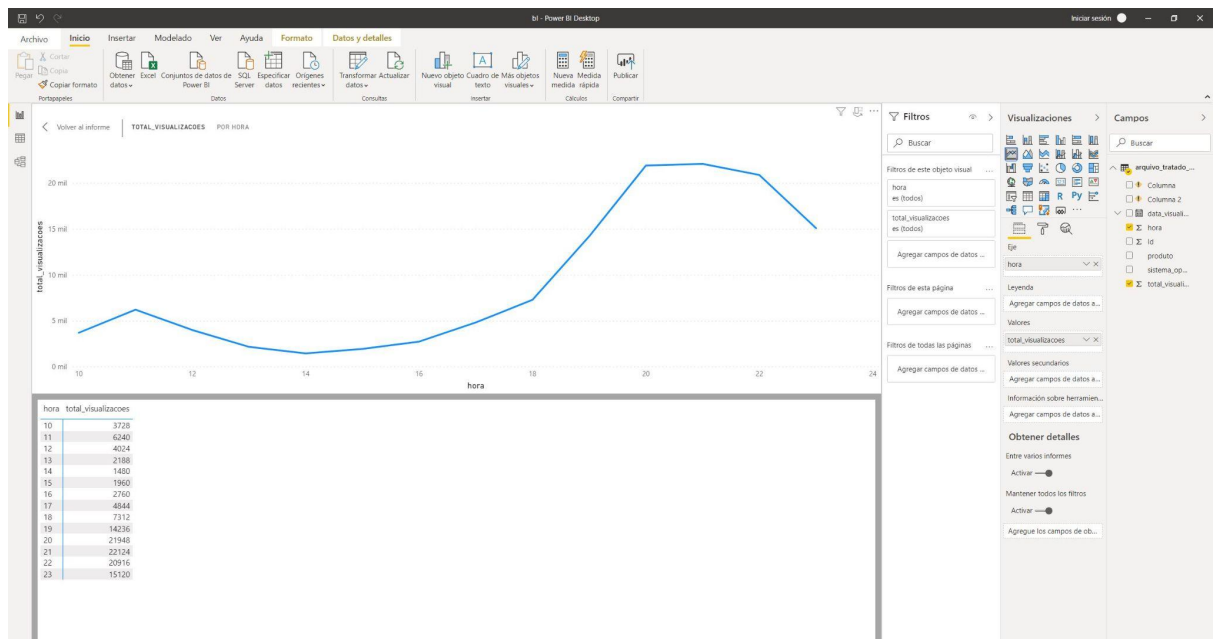
O problema encontrado foi com o gráfico de linhas hora x total_visualizações com a formatação “hora” que se apresentava em 00:00:00. A apresentação gráfica contabilizando cada segundo tornou a leitura caótica:



Precisamos tratar esta coluna no Power Query formatando para hora unicamente = 00, desconsiderando minuots e segundos. Do padrão 00:00:00 para o 00

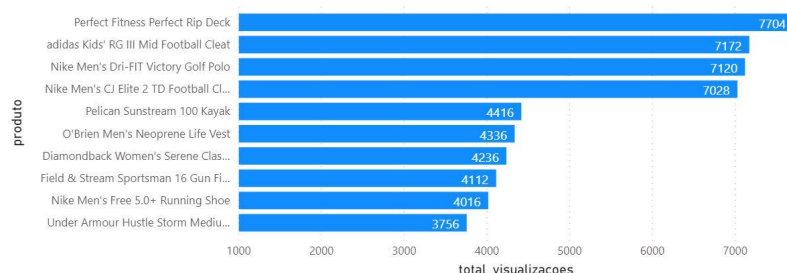


Esse ajuste permitiu a correta leitura do gráfico:

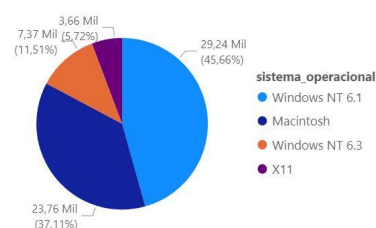


Assim chegamos no painel proposto:

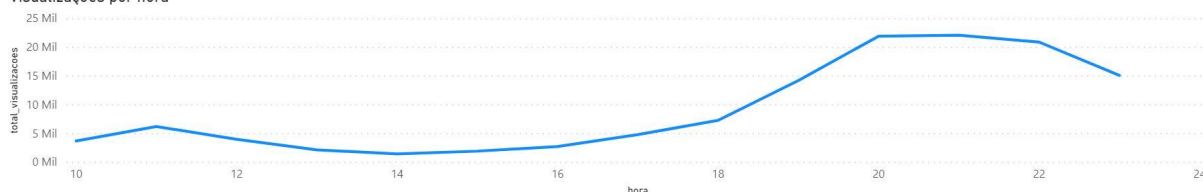
Top 10 Produto



Participação do sistema_operacional



Visualizações por hora



CONCLUSÃO

O aprendizado nesse hands on foi bem intenso, já que os desafios no uso das Vm's são vários, desde limitações com conexão, hardware e o que é preciso descifrar na própria ferramenta.

Seguir a sugestão de Road Map trazia a cada etapa um *troubleshooting* diferente, no qual a colaboração entre os integrantes deste grupo, com os demais alunos do curso foi muito importante, com o suporte da Cloudera e sua comunidade de ajuda.

O Regex - expressão regular - mostrou-se imprescindível para resolver a correta formatação, a sintaxe, extração e filtragem de informação para tabular os dados corretamente.

Sem dúvida a extração de dados de um arquivo log até sua visualização em ferramentas de Data Discovery é essencial no trabalho com Big Data.

REFERÊNCIAS BIBLIOGRÁFICAS

- Cloudera Community - Disponível em: <<https://community.cloudera.com/>> Acesso em 21 de maio de 2021
- Atividade Data Analytics - FIAP Acesso em abril/maio de 2021
- Material Fase 4 - FIAP Acesso em abril/maio de 2021