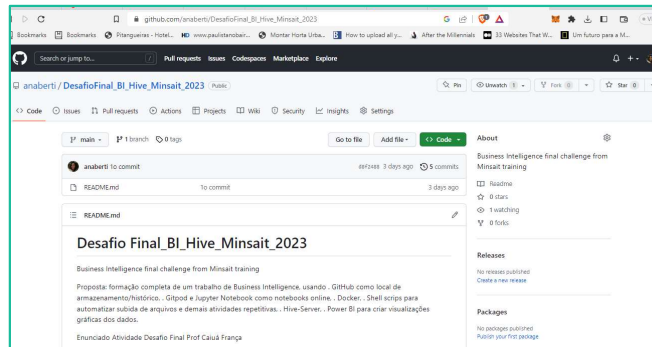


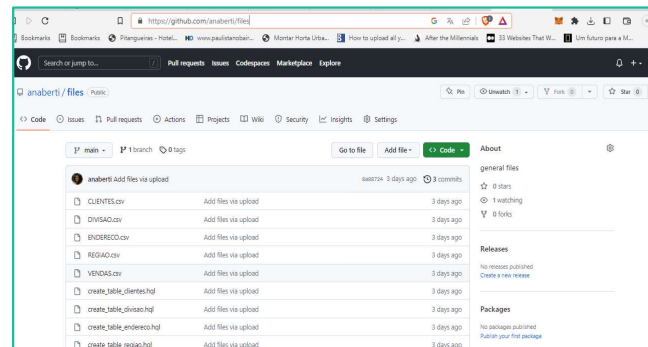
Etapas de Desenvolvimento do Desafio

1. Criação novo repositório no GitHub	2. Criação de repositório do GitHub para arquivos de fonte de dados	3. Criação de novo dashboard no gitpod	4. Clone da pasta bigdata_docker. Iniciar docker e hive	5. Criar diretórios e importar arquivos para dir de input	6. Criar pastas em hdfs e copiar para elas os arquivos de dados usando sh	7. Criar database e tabelas em beeline usando sh
8. Acessar Jupyter Notebook criar dataframes	9. Criar colunas de dia, mês, ano e quarter com base em invoice_date	10. Tratar campos brancos, nulos e Duplicados. Substituir pontos por vírgulas em numerais	11. Criar stage, colunas de consultas. Criar Fato e dimensões	12. Salvar dfs no hdfs	13. Download dos arquivos do gitpod	14. Upload das dfs no PowerBI
15. 2º tratamento/ transformação dos dados	16. Testes PBI e Jupyter	17. Montagem Dashboard	18. Salvar arquivos na pasta para entrega zip e Github			

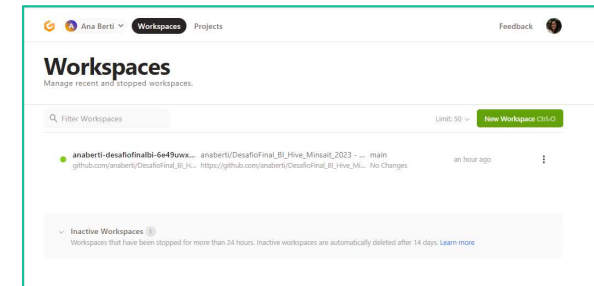
Preparo de ambientes – etapas 1 a 4



1. Criação novo repositório no GitHub
https://github.com/anaberti/DesafioFinal_BI_Hive_Minsait_2023



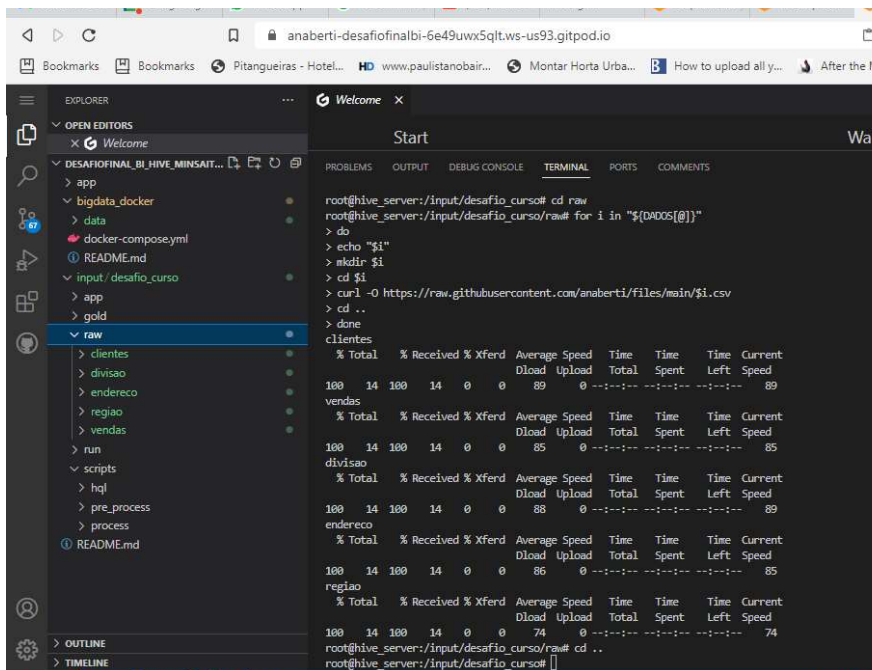
2. Criação de repositório do GitHub para arquivos de fonte de dados
<https://github.com/anaberti/files>



3. Criação de novo dashboard no gitpod

4. Clone da pasta bigdata_docker e iniciar docker e hive
git clone ambiente-curso -b https://github.com/caiuafraanca/bigdata_docker.git
cd bigdata_docker
docker-compose up -d
docker exec -it hive-server bash

Preparo de diretorios e arquivos – etapas 5 a 7



```
root@hive_server:/input/desafio_curso# cd raw
root@hive_server:/input/desafio_curso/raw# for i in "${DADOS[@]}"
do
> echo "$i"
> mkdir $i
> cd $i
> curl -O https://raw.githubusercontent.com/anaberti/files/main/$i.csv
> cd ..
done
clientes
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 14 100 14 0 0 89 0 --:--:-- --:--:-- --:--:-- 89
vendas
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 14 100 14 0 0 85 0 --:--:-- --:--:-- --:--:-- 85
divisao
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 14 100 14 0 0 88 0 --:--:-- --:--:-- --:--:-- 89
endereco
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 14 100 14 0 0 86 0 --:--:-- --:--:-- --:--:-- 85
regiao
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 14 100 14 0 0 74 0 --:--:-- --:--:-- --:--:-- 74
root@hive_server:/input/desafio_curso/raw# cd ..
root@hive_server:/input/desafio_curso#
```

```
# Cria estrutura de diretorios do desafio
mkdir desafio_curso
cd desafio_curso

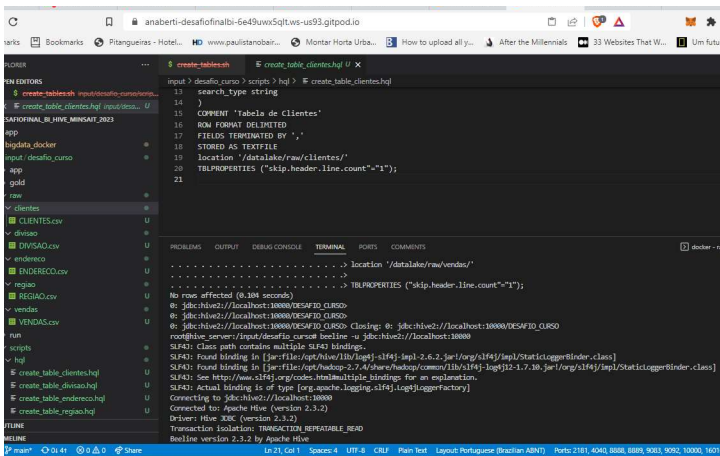
#!/bin/bash
PASTAS=("app" "gold" "raw" "run" "scripts")
for i in "${PASTAS[@]}"
do
echo "i"
mkdir $i
done

# Cria demais subdiretorios
#!/bin/bash
cd scripts
mkdir pre_process
mkdir process
mkdir hql
cd hql
TABELAS=("clientes" "vendas" "divisao" "endereco" "regiao")
for i in "${TABELAS[@]}"
do
curl -O https://raw.githubusercontent.com/anaberti/files/
main/$create_table_$i.hql
done
cd ..
cd ..
```

```
# importa os arquivos .csv
DADOS=("clientes" "vendas" "divisao" "endereco" "regiao")
#!/bin/bash
cd raw
for i in "${DADOS[@]}"
do
echo "$i"
mkdir $i
cd $i
curl -O
https://raw.githubusercontent.com/anaberti/files/main/$i.csv
cd ..
done
cd ..

# Transferência arquivos para hdfs
cd raw
DADOS=("clientes" "vendas" "divisao" "endereco" "regiao")
for i in "${DADOS[@]}"
do
echo "$i"
cd $i
hdfs dfs -mkdir -p /datalake/raw/$i
hdfs dfs -copyFromLocal $i.csv /datalake/raw/$i
cd ..
done
```

Preparo de diretórios e arquivos – etapas 5 a 7



```
anaberti-desafiofnatbi-6e49uws5qtlws-us93.gitpod.io
$ create_table.sh
input > desafio_curso > scripts > hql > $ create_table_clientes.hql
10 search_type string
11 }
12
13 COMMENT 'Tabela de Clientes'
14 ROW FORMAT DELIMITED
15 FIELD TERMINATED BY ','
16 STORED AS TEXTFILE
17 LOCATION '/datalake/raw/clientes/'
18 TBLPROPERTIES ("skip.header.line.count"="1");
19
20
21
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS COMMENTS
$ docker -ra
+-----+
| tab_name |
+-----+
| tbl_clientes |
| tbl_divisao |
| tbl_endereco |
| tbl_regiao |
| tbl_vendas |
+-----+
5 rows selected (0.024 seconds)
0: jdbc:hive2://localhost:10000>
```

```
#!/bin/bash
```

```
# Cria e usa database
```

```
beeline -u jdbc:hive2://localhost:10000 -e 'create database if not exists DESAFIO_CURSO;'
```

```
beeline -u jdbc:hive2://localhost:10000 -e 'use DESAFIO_CURSO;'
```

```
# Criacao tabelas
```

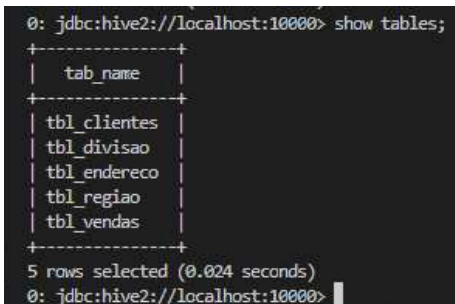
```
DADOS=("CLIENTES" "VENDAS" "DIVISAO" "ENDERECO" "REGIAO")
```

```
for i in "${DADOS[@]}"
```

```
do
```

```
beeline -u jdbc:hive2://localhost:10000 -f/DESAFIO_CURSO ../input/desafio_curso/scripts/hql/create_table_${i}.hql
```

```
done
```



```
0: jdbc:hive2://localhost:10000> show tables;
+-----+
| tab_name |
+-----+
| tbl_clientes |
| tbl_divisao |
| tbl_endereco |
| tbl_regiao |
| tbl_vendas |
+-----+
5 rows selected (0.024 seconds)
0: jdbc:hive2://localhost:10000>
```

Trabalhando os dados – stage e dimensões – 8 a 13

```
In [19]:  
dim_local = df_stage.select("city", "country", "state", "region_code", "region_name").distinct()
```

```
In [20]:
```

```
dim_local.show(5)
```

```
+-----+-----+-----+-----+-----+  
| city|country|state|region_code|region_name|  
+-----+-----+-----+-----+-----+  
| Morton| US| IL| 4| Central|  
| Maxwell| US| CA| 1| Western|  
| Meridian| US| MS| 2| Southern|  
| St Louis Park| US| MN| 4| Central|  
| Houston| US| TX| 2| Southern|  
+-----+-----+-----+-----+-----+  
only showing top 5 rows
```

```
fato_vendas.show(5)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
| order_number|invoice_number|customer_key|sales_rep|item_number|item|item_class|line_number|list_price|date_key|promised_delivery_date|actual_delivery_date|region_code|sales_quantity|sales_price|discount_amount|sales_amount|sales_amount_based_on_list_price|sales_cost_amount|  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
| 202427| 104794| 10000472| 119| 28761| Ebony Fuji Apples| P01| 4| 1000| 157,76|09/06/2018| 363,48| 09/06/2019| 631,04| 189,23| 4| 315975| 127,84|06/01/2019| 31/12/2019| 37574| Walrus Light Beer| P01| 12| 131,675| -46,02| 1580,1| 47803| Red Spade Foot-Lo...| 21/11/2018| 1| 317842| 221032| 10025298| 185| 37441| Atomic Mint Choco...| 15/01/2020| 0| 1000| 1254,1899|15/01/2019| 744,08| 510,1099| 744,08| 320895| 767,75|08/03/2019| 07/03/2020| 406,75325| 14439,87| 16270,13| 406,75325| 14439,87| 16270,13|  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
only showing top 5 rows
```

```
dim_cliente.show(5)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
| customer_key|customer|customer_type|phone|address_number|customer_address_1|customer_address_2|customer_address_3|customer_address_4|division|division_name|line_of_business|business_family|business_unit|regional_sales_mgr|search_type|zip_code|  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
| 10021223| PacificServ Shop| G2|816-455-8733| 10021223|Michigan Truck Plant|3830| 3| Michigan Av...| Não informado| Não informado| 2| Domestic| Não informado| | | | |
| 10025571| Unitec Maxistore| G2|816-455-8733| 10025571| PO Box 3283| Domestic| Não informado|  
| 10016784| IBEX Shop| G2|816-455-8733| 10016784| PO Box 129| Domestic| Não informado|  
| 10009647| Dayton Supermarket| G2|816-455-8733| 10009647|3248 Auburn Boule...| Domestic| Não informado|  
| 10002412| Accior Shop| G1|816-455-8733| null| null| null| null| null| null| null| null| null| null| null| null| null| null|  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
only showing top 5 rows
```

Conforme demonstrado em process.py

```
In [23]:
```

```
dim_tempo = df_stage.select("date_key", "invoice_date", "invoice_day", "invoice_month", "invoice_year", "quarter")
```

```
In [24]:
```

```
dim_tempo.show(5)
```

```
+-----+-----+-----+-----+-----+  
| date_key|invoice_date|invoice_day|invoice_month|invoice_year|quarter|  
+-----+-----+-----+-----+-----+  
| 23/04/2018| 25/04/2018| 25| 04| 2018| 2|  
| 11/05/2018| 13/05/2018| 13| 05| 2018| 2|  
| 01/02/2019| 03/02/2019| 03| 02| 2019| 1|  
| 31/12/2017| 02/01/2018| 02| 01| 2018| 1|  
| 08/09/2017| 10/09/2017| 10| 09| 2017| 3|  
+-----+-----+-----+-----+-----+  
only showing top 5 rows
```

Dashboard – etapas 14 a 18

Dados carregados no Power Bi, corrigindo campos que por algum motivo ainda apresentaram problemas

Table: ReplaceValue("Valor Substituído", "", "Não informado", Replacer.ReplaceValue, ("zip_code"))

customer_key	customer	customer_type	phone	address_number	customer_ad
10021223	PacificServ Shop	G2	816-455-8733	10021223	Michigan Truc
10025571	Unitex Maxistore	G2	816-455-8733	10025571	PO Box 3283
10016784	IBEX Shop	G2	816-455-8733	10016784	PO Box 129
10009647	Dayton Supermarket	G2	816-455-8733	10009647	3248 Auburn C
10002412	Acclor Shop	G1	816-455-8733	0	Não informado
10019617	Kool-Sea Shop	G2	816-455-8733	10019617	3445 FM 2500
10020668	Kyrates Shop	G1	816-455-8733	10020668	Ast Reg Del Ci
10020672	Oni Shop	G1	816-455-8733	10020672	Brooksketen
10022746	Rdado Shop	G1	816-455-8733	10022746	Landtrasse 18
10010897	Edify Shop	G2	816-455-8733	10010897	PO Box 588
10020448	Namas Shop	G3	816-455-8733	10020448	1155 Mackay I
10010862	Edin Supermarket	G2	816-455-8733	0	Não informado
10020320	ActiCard Shop	G1	816-455-8733	0	Não informado
10000601	Radio Market	G2	816-455-8733	10000601	1550 Aerel Av
10021996	Pinnacle Supermarket	G2	816-455-8733	10021996	8 N Industrial I
10012927	Galaxy Supermarket	G2	816-455-8733	0	Não informado
10022454	R&R Store	G2	816-455-8733	10022454	PO Box 958
10022725	Razorfish Shop	G2	816-455-8733	10022725	1723 Webb Dr
10020221	Revel Supermarket	G2	816-455-8733	10020221	Attention Acc
10023524	Satin Gossipce	G2	816-455-8733	10023524	PO Box 1006
10013572	GeerSource Shop	G2	816-455-8733	0	Não informado
10002114	Accol Supermarket	G2	816-455-8733	0	Não informado
10023552	Unison Medistore	G2	816-455-8733	10023552	Order Process

Alterando os tipos dos dados

Table: TransformColumnTypes("Cabeçalhos Promovidos", ("order_number", Int64.Type))

order_number	invoice_number	customer_key	sales_rep	item_number	item	
202427	104794	10000472	119	28763	Elbony Fuji Asa	
2	151575	220154	160	37574	Walrus Light B	
3	114882	118272	10009631	150	47801	Red Spade Foc
4	117842	221032	10025298	185	37441	Atomic Mine C
5	100895	225396	10011618	180	43869	Red Spade Low
6	100895	225396	10011618	180	28761	Elbony Fuji Asa
7	117202	102076	10019617	170	20910	Moms Sliced T
8	208784	111236	10025039	104	65063	Nationnel Fud
9	119524	103972	10023524	103	28500	Elbony Squash
10	109154	111630	10023524	103	47350	Red Spade Tur
11	1008846	112386	10023524	134	67350	Discover Mani
12	100676	111326	10023524	134	61560	Best Wheat Pl
13	116875	104833	10008993	144	64008	Washington G
14	205183	106186	10012235	175	28829	Nationnel Potz
15	116929	102634	10013538	170	64008	Washington G
16	207801	109766	10023251	112	60035	Golden Waffle
17	120620	102569	10012161	109	64008	Washington G
18	106396	108447	10000456	163	265032	BBB Best Grap
19	115480	107177	10000456	163	10405	Moms Chicker
20	204480	112645	10021160	119	65659	Gurilla String C
21	225903	102748	10021160	119	25300	Fast Dried App
22	220832	126144	10022755	118	26502	Bravo Large C
23	119367	102367	10022444	173	17801	Better Fancy C

Dashboard – etapas 14 a 18

Acrescentando colunas e medidas interessantes para a análise dos dados

	id	sales_amount	sales_amount_based_on_list_price	sales_cost_amount	sales_amount_after_sales_cost
1	267,56	363,48	631,04	289,23	174,25
2	-46,02	1.580,10	1.534,08	1.118,13	461,97
3	690,21	937,63	1.637,84	475,75	461,88
4	510,11	744,08	1.254,19	344,28	399,8
5	14.439,87	16.270,13	30.710,00	11.526,38	4743,75
6	7.269,53	8.190,95	15.460,48	4.679,01	3511,94
7	368,79	456,17	824,96	260,98	195,19
8	0,00	212,85	212,85	180,54	32,31
9	2.099,15	2.135,57	4.234,72	1.238,58	896,99
10	1.150,72	1.170,68	2.321,40	658,10	532,58
11	9.120,33	11.281,27	20.401,60	5.523,21	5758,06
12	260,14	294,34	554,48	116,22	178,12
13	1.932,84	2.162,16	4.095,00	1.460,37	701,79
14	475,22	645,58	1.120,80	230,57	415,01
15	8.155,57	8.207,93	16.365,50	4.523,51	3684,42
16	1.829,87	1.829,88	3.659,75	1.344,95	484,93
17	308,94	345,60	654,54	180,94	164,66
18	1.514,38	1.557,62	3.072,00	931,86	625,76
19	258,57	245,03	501,60	78,10	166,93
20	309,90	286,08	596,00	190,39	95,79
21	279,07	319,16	596,23	167,81	151,35
22	6.760,97	7.732,00	14.492,97	4.371,11	3460,89
23	2.156,87	2.136,82	4.293,69	1.349,08	787,74
24					

Na fato_vendas:

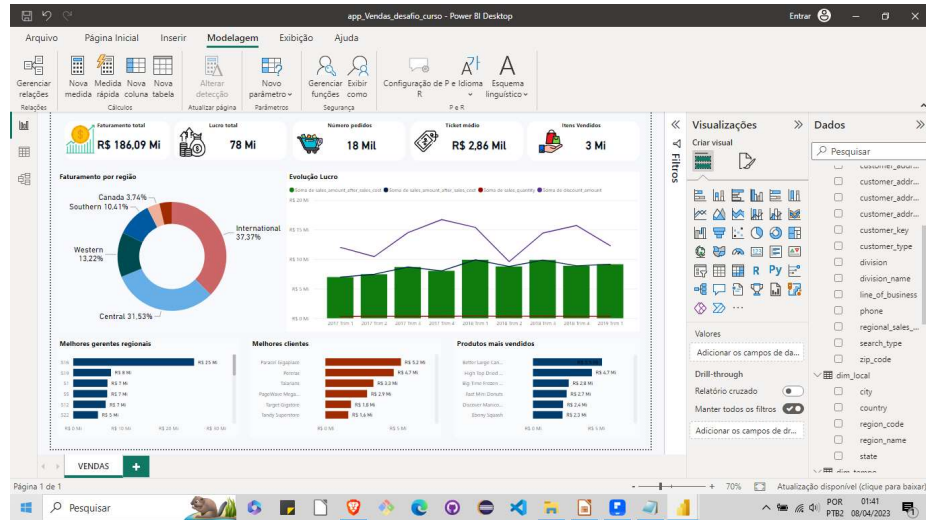
- sales_amount_after_sales_costs (valor de venda após custos = lucro)
- discount_percentage

Nova tabela medidas:

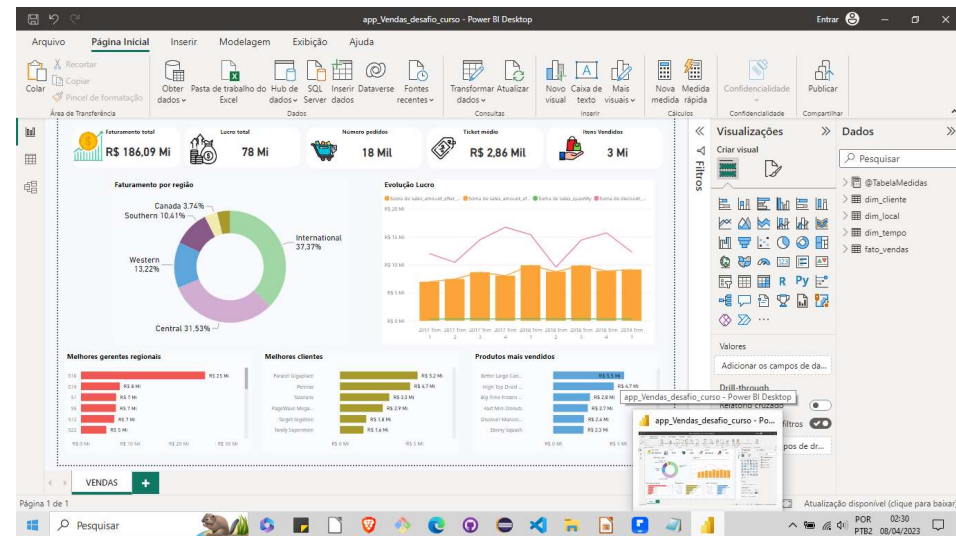
- total de vendas
- lucro total
- ticket médio
- ...



Dashboard – etapas 14 a 18



Montando layout do dashboard



Dashboard – etapas 14 a 18

Testando os resultados dos cálculos do Power BI x Jupyter

Power BI



Jupyter

```
fato_vendas = fato_vendas.withColumn("sales_amount", fato_vendas["sales_amount"].cast('double'))
```

```
fato_vendas.select(sum("sales_amount")).show(truncate=False)
```

```
+-----+
|sum(sales_amount)|
+-----+
|2.6407834E7      |
+-----+
```



```
: fato_vendas = fato_vendas.withColumn("sales_quantity", fato_vendas["sales_quantity"].cast('double'))
```

```
: fato_vendas.select(sum("sales_quantity")).show(truncate=False)
```

```
+-----+
|sum(sales_quantity)|
+-----+
|2942610.0          |
+-----+
```



```
In [21]: fato_vendas.select('order_number').distinct().count()
```

```
_Out[21]: 17732
```

Montando layout do dashboard

