



MVP Engenharia de Dados

Pós-Graduação em Ciências de Dados e Analytics

Sprint 03

Aluno: Gabriel Prata

Data: 14/08/2023

1. Objetivo

O mercado de banda larga fixa vem crescendo cada vez mais no Brasil, gerando uma grande concorrência entre empresas de telecomunicações.

Cada vez mais, os brasileiros desejam ter em casa uma conexão de alta velocidade e de grande estabilidade, e esse cenário é um efeito da modernização da infraestrutura de telecomunicações no país.

Trata-se de um movimento cujo início beneficiou principalmente grandes centros urbanos, mas que foi expandindo gradualmente para cidades pequenas e bairros mais afastados.

Não resta dúvida hoje em dia, que a banda larga mais eficaz é a Fibra óptica.

A ANATEL(Agência Nacional de Telecomunicações) divulgou em seu portal de dados, que em 2022 o Brasil registrou 44,9 milhões de acessos de banda larga fixa, e que 70% desses acessos, são de Fibra Óptica.

Com um mercado tão aquecido, tendo um crescimento de 6.7% em relação a 2021, a empresa Oi, nos pede uma análise do panorama do mercado de Fibra Óptica no Brasil.

2. Busca pelos Dados

Os dados foram coletados do sítio da Agência Nacional de Telecomunicações.

<https://informacoes.anatel.gov.br/paineis/acessos>

Arquivos:

Acessos_Banda_Larga_Fixa_2022.zip

3. Coleta

Os dados foram coletados para uma máquina local, e posteriormente descompactados.

Para armazenar os dados, escolhemos a plataforma de nuvem AWS(Amazon Web Services).

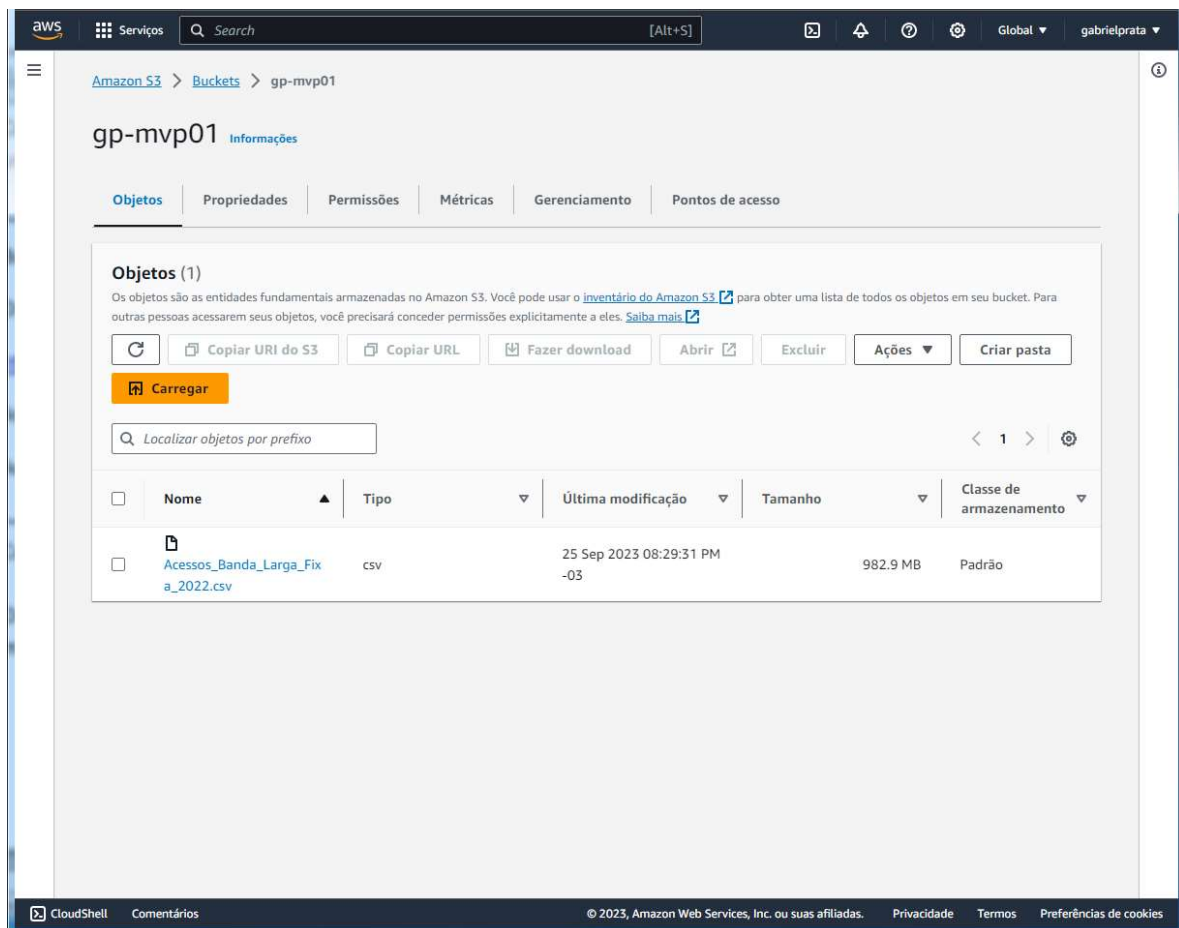
A AWS é a nuvem mais abrangente e amplamente adotada do mundo, que inclui ofertas de infraestrutura como serviço (IaaS) e plataforma como serviço (PaaS).

Oferecendo soluções escaláveis para computação, armazenamento, banco de dados, análises e muito mais.

Para armazenar o arquivo, escolhemos o AWS S3(Simple Storage Service), que é um serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e performance. O Amazon S3 pode armazenar e proteger qualquer volume de dados para uma variedade de casos de uso, como data lakes, sites, aplicações móveis, backup e restauração, arquivamento, aplicações corporativas, dispositivos IoT e análises de big data.

O AWS S3 armazena dados como objetos em buckets. Um objeto é um arquivo e quaisquer metadados. Sendo assim, um bucket é um contêiner de objetos.

Criamos um bucket chamado de gp-mvp01, e em seguida carregamos o arquivo .



4. Modelagem

Para esse trabalho, estamos utilizando um arquivo FLAT, onde todos os campos necessários já estão na tabela.

Construímos o catálogo de dados, utilizando o AWS Glue, ficando da seguinte maneira:

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

▼ Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

▼ Data Integration and ETL

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Interactive Sessions

Data classification tools

Sensitive data detection

Record Matching

Triggers

Workflows (orchestration)

Blueprints

Security configurations

► Legacy pages

What's New

Documentation

AWS Marketplace

Enable compact mode

Enable new navigation

CloudShell

Comentários

AWS

Serviços

Q Search

[Alt+S]

Ohio

gabrielprata

AWS Glue

×

AWS Glue > Tables > anatel-2022

Last updated (UTC)
September 27, 2023 at 23:42:27

↺

Version 1 (Current version)

▼

Actions

Table overview

Data quality New

Table details

Advanced properties

Name

anatel-2022

Description

-

Database

anatel_mvp

Classification

CSV

Location

s3://gp-mvp01/

Connection

-

Deprecated

-

Last updated

September 27, 2023 at 23:42:27

Input format

org.apache.hadoop.mapred.T
extInputFormat

Output format

org.apache.hadoop.hive.ql.io
.HiveIgnoreKeyTextOutputFo
rmat

Serde serialization lib

org.apache.hadoop.hive.serd
e2.lazylazySimpleSerDe

Schema

Partitions

Indexes

Schema (16)

View and manage the table schema.

Edit schema as JSON

Edit schema

Q Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	ano	bigint	-	-
2	mês	bigint	-	-
3	grupo econômico	string	-	-
4	empresa	string	-	-
5	cnpj	bigint	-	-
6	porte da prestadora	string	-	-
7	uf	string	-	-
8	município	string	-	-
9	código ibge município	bigint	-	-
10	faixa de velocidade	string	-	-
11	velocidade	string	-	-
12	tecnologia	string	-	-
13	meio de acesso	string	-	-
14	tipo de pessoa	string	-	-
15	tipo de produto	string	-	-
16	acessos	bigint	-	-

© 2023, Amazon Web Services, Inc. ou suas afiliadas.

Privacidade

Termos

Preferências de cookies

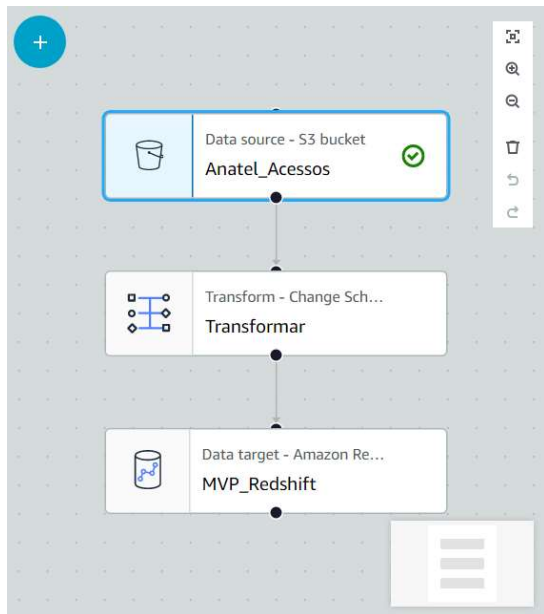
Nome da coluna	Tipo	Descrição
Ano	bigint	Variável com valores numéricos, representando o ano
Mês	bigint	Variável com valores numéricos, representando o mês
Grupo Econômico	string	Nome do grupo de empresas de Telecomunicações.
Empresa	string	Nome da empresa de telecomunicações.
CNPJ	bigint	Número do documento da empresa.
Porte da Prestadora	string	Tipo do porte da empresa de telecomunicações.
UF	string	Unidade federativa de instalação do acesso.
Município	string	Nome do município da instalação do acesso.
Código IBGE Município	bigint	Código do município pelo IBGE.
Faixa de Velocidade	string	Faixa de velocidade contratada.
Velocidade	string	Velocidade contratada.
Tecnologia	string	Tecnologia de conexão.
Meio de Acesso	string	Meio pelo qual o usuário faz o seu acesso. Ex.: Fibra, Cabo Coaxial
Tipo de Pessoa	string	Tipo de pessoa jurídica. Ex.: Pessoa física, Pessoa jurídica
Tipo de Produto	string	Tipo de produto contratado.
Acessos	bigint	Quantidade de acessos instalados.

5. Carga

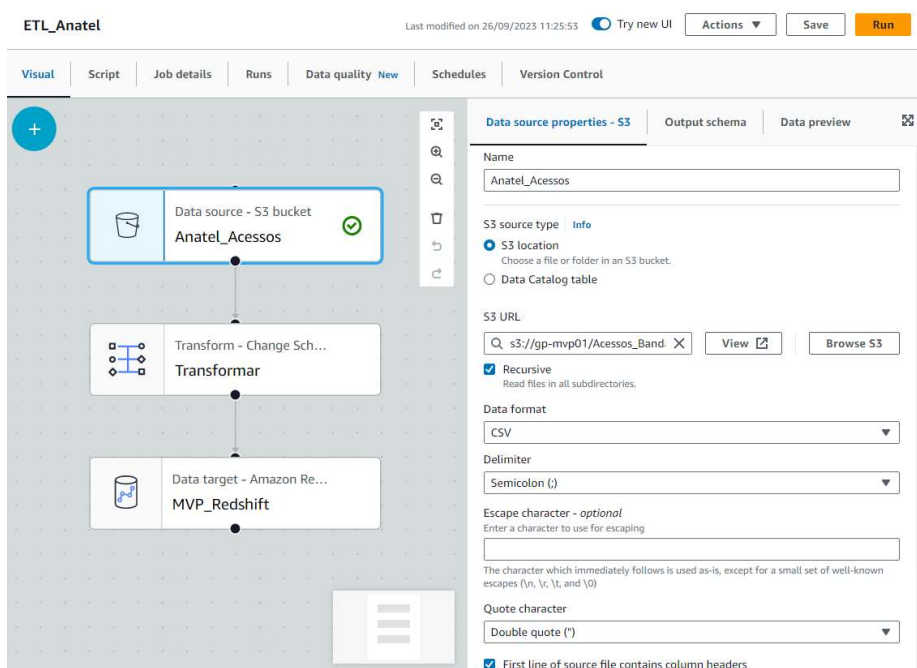
Nessa etapa desenvolvemos uma ETL(Extract, transform, load), que é um processo de combinação de dados de várias fontes em um grande repositório central. O processo de ETL usa um conjunto de regras de negócios para limpar e organizar dados brutos e prepará-los para armazenamento, análise de dados e machine learning (ML). Podendo assim, atender a necessidades específicas de business intelligence por meio da análise de dados (como prever o resultado das decisões de negócios, gerar relatórios e painéis, reduzir a ineficiência operacional e muito mais).

Para realizar esse processo, escolhemos a ferramenta AWS Glue, que é um serviço de integração de dados com tecnologia sem servidor que facilita aos usuários de análise a descoberta, preparação, transferência e integração de dados de várias fontes.

Utilizamos a sua ferramenta Visual ETL, da seguinte maneira:



Na primeira etapa do ETL, escolhemos a origem da informação, que no nosso projeto, é o arquivo bruto, sem tratamento, armazenado no bucket do S3.



Na segunda etapa, iremos transformar os dados, renomeando o nome dos campos retirando caracteres especiais e os espaços entre as palavras.

Determinamos também o tipo dos campos, e excluimos os campos que não serão necessários para as análises.

ETL_Anatel

Last modified on 26/09/2023 11:25:53 Try new UI Actions Save Run

Visual Script Job details Runs Data quality New Schedules Version Control

Visual

Transform

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
Ano	anos	int	<input type="checkbox"/>
Mês	mes	int	<input type="checkbox"/>
Grupo Econômico	grupo	string	<input type="checkbox"/>
Empresa	empresa	string	<input type="checkbox"/>
CNPJ			<input checked="" type="checkbox"/>
Porte da Prestadora	porte_prestadr	string	<input type="checkbox"/>
UF	UF	string	<input type="checkbox"/>
Município	municipio	string	<input type="checkbox"/>
Código IBGE Município	cod_ibge_mun	int	<input type="checkbox"/>
Faixa de Velocidade	faixa_velocidad	string	<input type="checkbox"/>
Velocidade	velocidade	string	<input type="checkbox"/>
Tecnologia	tecnologia	string	<input type="checkbox"/>
Meio de Acesso	meio_acesso	string	<input type="checkbox"/>
Tipo de Pessoa	tipo_pessoa	string	<input type="checkbox"/>
Tipo de Produto	tipo_produto	string	<input type="checkbox"/>
Acessos	acessos	int	<input type="checkbox"/>

Na terceira parte escolhemos o destino das informações transformadas, que nesse projeto optamos por utilizar o Amazon Redshift.

O Amazon Redshift é um serviço de data warehouse em escala de petabytes totalmente gerenciado na nuvem. O Amazon Redshift sem servidor permite acessar e analisar dados sem todas as configurações de um data warehouse provisionado.

ETL_Anatel

Last modified on 26/09/2023 11:25:53 Try new UI Actions Save Run

Visual Script Job details Runs Data quality New Schedules Version Control

Visual

Data target properties - Amazon Redshift

Data preview

Redshift access type

☒ Direct data connection - recommended

☐ Glue Data Catalog tables

Redshift connection

Choose the AWS Glue connection for Amazon Redshift, or create a new connection

connection_redshift

Connection

View properties

Database

dev

Schema

Choose your Amazon Redshift schema.

public

Table

Search and enter the name of the source Amazon Redshift table.

anatel

Handling of data and target table

☒ APPEND (insert) to target table

AWS Glue will append data to existing columns of the table and discard any extra columns.

☐ MERGE data into target table

AWS Glue will either update or append data to the table based on a set of conditions.

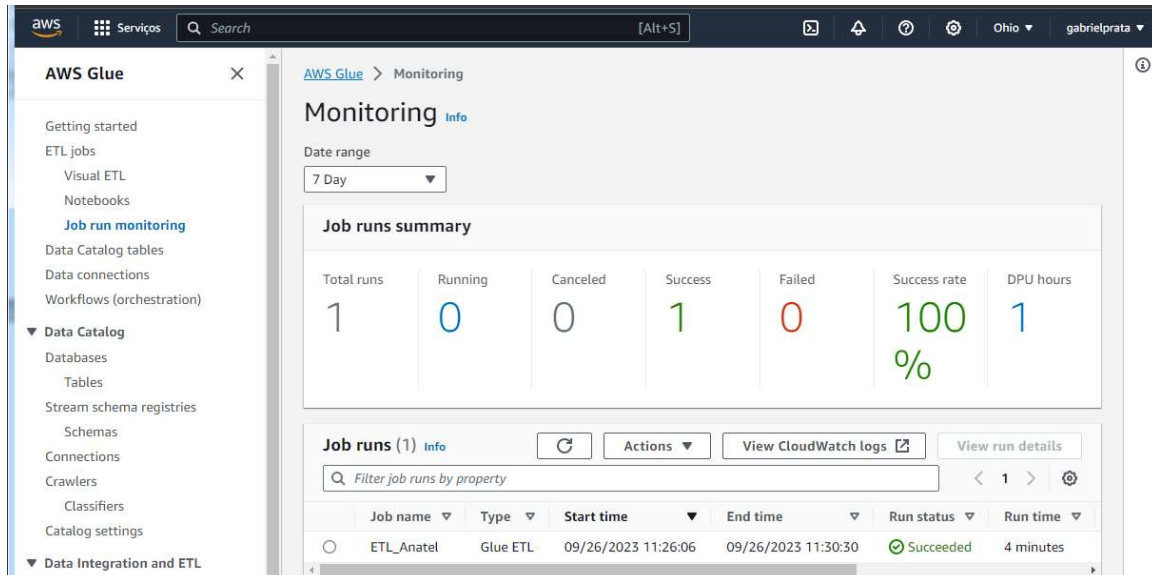
☐ TRUNCATE target table

Same as Append, except AWS Glue will first clear the contents of the table.

☐ DROP and recreate target table

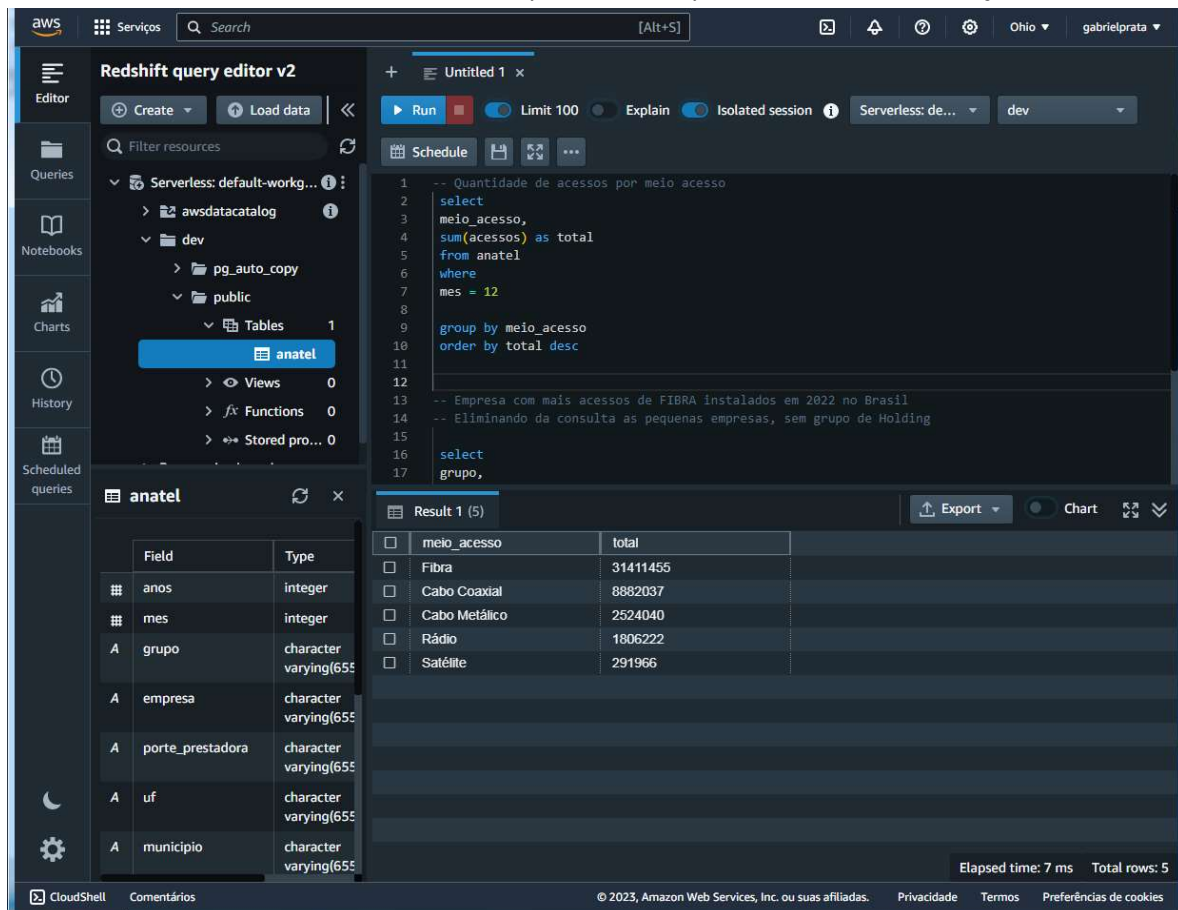
AWS Glue will delete and recreate the table with the schema from the source data.

Após as configurações do ETL, a execução do Job foi realizada com sucesso.



Sendo assim, as informações já estão disponíveis no Amazon Redshift.

Testamos a conexão e executamos uma simples consulta para verificar as informações.



6. Análise

Nessa etapa do projeto eu gostaria de usar o Python, que é o meu foco de aprendizado hoje, porém, não foi possível aprender uma ferramenta na nuvem em tão pouco tempo, que integrasse com o AWS.

Em uma sessão de dúvidas, perguntei se poderia ser feito as análises no Google Colab, e foi dito que sim.

Tendo em vista esse cenário, criei um ETL no AWS Glue, criando uma transformação de dados usando o SQL Query, criando um arquivo CSV apenas com os acessos de internet por FIBRA.

aws

Serviços

Search

[Alt+S]

Ohio

gabrielprata

anatel_py

Last modified on 28/09/2023 16:01:14

Try new UI

Actions

Save

Run

Unsaved job found
We found an unsaved job, do you wish to restore it?

Restore

Visual

Script

Job details

Runs

Data quality New

Schedules

Version Control

Data source - S3 bucket

Anatel

Transform - SQL Query

SQL Query

Data target - S3 bucket

anatel_py

Data source properties - S3

Output schema

Data preview

Name

Anatel

S3 source type

Info

S3 location

Choose a file or folder in an S3 bucket.

Data Catalog table

S3 URL

s3://gp-mvp01/Acessos_Band

X

View

Browse S3

Recursive

Read files in all subdirectories.

Data format

CSV

Delimiter

Semicolon (;)

Escape character - optional

Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

Quote character

CloudShell

Comentários

© 2023, Amazon Web Services, Inc. ou suas afiliadas.

Privacidade

Termos

Preferências de cookies

aws

Serviços

Search

[Alt+S]

Ohio

gabrielprata

anatel_py

Last modified on 28/09/2023 16:01:14

Try new UI

Actions

Save

Run

Unsaved job found
We found an unsaved job, do you wish to restore it?

Restore

Visual

Script

Job details

Runs

Data quality New

Schedules

Version Control

Data source - S3 bucket

Anatel

Transform - SQL Query

SQL Query

Data target - S3 bucket

anatel_py

Transform

Output schema

Data preview

Name

SQL Query

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Anatel
S3 - DataSource

Associate an alias with each input source

Info

Edit the aliases used for the inputs to this node.

Input sources

SQL aliases

Anatel

myDataSource

SQL query

Enter a SQL statement to add to your job.

1 select * from myDataSource
2 where "Meio de Acesso" = 'Fibra'

CloudShell

Comentários

© 2023, Amazon Web Services, Inc. ou suas afiliadas.

Privacidade

Termos

Preferências de cookies

6.1 Qualidade de dados

Explorando o as informações do arquivo, que foi carregado em um Pandas Dataframe.

```
#Informações do dataframe
acessos_fibra.info(memory_usage='deep')

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3905690 entries, 0 to 3905689
Data columns (total 17 columns):
#   Column                Dtype
---  -
0   Unnamed: 0             int64
1   Ano                   int64
2   Mês                   int64
3   Grupo Econômico       object
4   Empresa               object
5   CNPJ                  int64
6   Porte da Prestadora  object
7   UF                    object
8   Município             object
9   Código IBGE Município int64
10  Faixa de Velocidade    object
11  Velocidade             object
12  Tecnologia             object
13  Meio de Acesso         object
14  Tipo de Pessoa         object
15  Tipo de Produto        object
16  Acessos                int64
dtypes: int64(6), object(11)
memory usage: 3.0 GB

[13] #Quantidade de linhas e colunas
acessos_fibra.shape

(3905690, 17)
```

O Dataframe possui 17 colunas e 3,90 milhões de registros.

Não possui nenhum registro NULO.

```
#Quantidades de Nulos/NaN
acessos_fibra.isnull().sum()

Unnamed: 0      0
Ano              0
Mês              0
Grupo Econômico  0
Empresa          0
CNPJ             0
Porte da Prestadora 0
UF              0
Município        0
Código IBGE Município 0
Faixa de Velocidade 0
Velocidade       0
Tecnologia        0
Meio de Acesso    0
Tipo de Pessoa    0
Tipo de Produto   0
Acessos           0
dtype: int64
```

Conhecendo os primeiros registros do Dataframe.

```
#Exibindo as primeiras linhas do dataframe
acessos_fibra.head()
```

	Unnamed: 0	Ano	Mês	Grupo Econômico	Empresa	CNPJ	Porte da Prestadora	UF	Município	Código IBGE Município	Faixa de Velocidade	Velocidade	Tecnologia
0	0	2021	12	OUTROS	Click Networks Tis Ltda	43046927000175	Pequeno Porte	MS	Três Lagoas	5008305	> 34Mbps	80,000000	FTTH
1	1	2021	12	OUTROS	TURBO NET TELECOM LTDA	9366952000106	Pequeno Porte	SP	Embaúba	3514957	12Mbps a 34Mbps	25,000000	FTTH
2	2	2021	12	OUTROS	TURBO NET TELECOM LTDA	9366952000106	Pequeno Porte	SP	Palmares Paulista	3535101	> 34Mbps	200,000000	FTTH
3	3	2021	12	OUTROS	IMPACTNET INSTALADORA DE EQUIPAMENTOS DE COMUN...	22007662000126	Pequeno Porte	PR	Curitiba	4106902	> 34Mbps	76,700000	FTTH
4	4	2021	12	OUTROS	NETLINE TECNOLOGIA EM TELECOMUNICAÇÕES LTDA	6292667000191	Pequeno Porte	PB	São José de Piranhas	2514503	> 34Mbps	100,000000	FTTH

Agora iremos verificar a distribuição dos principais atributos. Para ver se existe a necessidade de tomar alguma ação de transformação na etapa de preparação dos dados.

Grupo Econômico

```
acessos_fibra.groupby('Grupo Econômico').size().sort_values(ascending=False)
```

```
Grupo Econômico
OUTROS                2448471
OI                    520232
TELECOM AMERICAS     258832
ALGAR (CTBC TELECOM) 189466
EB FIBRA              88753
TELEFÔNICA            83253
VERO                  69497
UNIFIQUE              68542
LIGGA TELECOM         67301
TRIPLE PLAY          31659
BRISANET              31302
TELECOM ITALIA        20876
BT                    16527
AZZA TELECOM          10259
TELEFONICA             445
SKY/AT&T              171
DATORA                104
dtype: int64
```

Porte da Prestadora

```
acessos_fibra.groupby('Porte da Prestadora').size().sort_values(ascending=False)
```

```
Porte da Prestadora
Pequeno Porte    3021881
Grande Porte     883809
dtype: int64
```

Faixa de velocidade e Velocidade

```
acessos_fibra.groupby('Faixa de Velocidade').size().sort_values(ascending=False)
```

```
Faixa de Velocidade
> 34Mbps                2567019
2Mbps a 12Mbps          521848
12Mbps a 34Mbps         507780
512kbps a 2Mbps         223239
0Kbps a 512Kbps         85804
dtype: int64
```

```
[20] acessos_fibra.groupby('Velocidade').size().sort_values(ascending=False)
```

```
Velocidade
100,000000    327218
50,000000     246693
200,000000    221046
300,000000    189675
10,000000     180891
...
578,120000      1
578,86          1
100,22          1
100,210000      1
155,65          1
Length: 8241, dtype: int64
```

Removendo as colunas que não iremos utilizar e renomeando, retirando caracteres especiais e os espaços.

Remover Colunas que não serão utilizadas

```
[21] # removendo as colunas
acessos_fibra.drop(['CNPJ',
                    #'Porte da Prestadora',
                    'Tipo de Pessoa'
                    #'Código IBGE Município'
                    ], axis=1, inplace=True)
```

```
[22] #Renomeando as colunas
acessos_fibra.rename(columns={'Município': 'municipio',
                              'Faixa de Velocidade': 'faixa_velocidade',
                              'Tipo de Produto': 'tipo_produto',
                              'Empresa': 'empresa',
                              'Código IBGE Município': 'codigo_ibge',
                              'Grupo Econômico': 'grupo_economico',
                              'Meio de Acesso': 'meio_acesso',
                              'Mês': 'mes',
                              'Ano': 'ano'
                              },
                      inplace=True)
```

Durante o ano de 2022, algumas empresas foram compradas por outras de outro grupo econômico, sendo assim iremos fazer esses ajustes.

Ajustes dos dados

```
[23] #A empresa "EB FIBRA" passa a se chamar "ALLOHA" e comprou a empresa "XP SERVICOS DE COMUNICACAO LTDA"

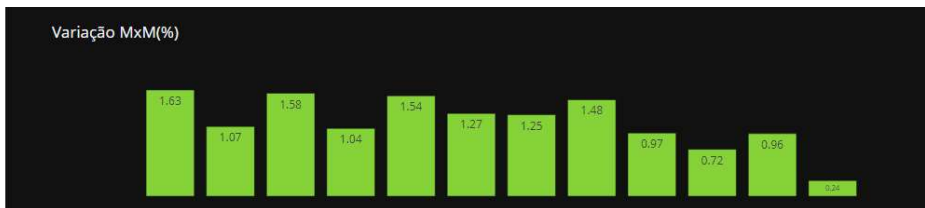
acessos_fibra.loc[acessos_fibra.empresa=="XP SERVICOS DE COMUNICACAO LTDA",'grupo_economico']="ALLOHA"
acessos_fibra.loc[acessos_fibra.empresa=="EB FIBRA",'grupo_economico']="ALLOHA"
acessos_fibra.loc[acessos_fibra.empresa=="EB FIBRA",'empresa']="ALLOHA"
acessos_fibra.loc[acessos_fibra.empresa=="XP SERVICOS DE COMUNICACAO LTDA",'empresa']="ALLOHA"

# simplificar o nome de algumas empresas para facilitar na exibição dos gráficos
acessos_fibra.loc[acessos_fibra.empresa=="Desktop - Sigmanet Comunicacao Multimidia S.a.",'grupo_economico']="Desktop"
acessos_fibra.loc[acessos_fibra.empresa=="ALGAR (CTBC TELECOM)",'grupo_economico']="ALGAR TELECOM"
acessos_fibra.loc[acessos_fibra.empresa=="Desktop - Sigmanet Comunicacao Multimidia S.a.",'empresa']="Desktop"
acessos_fibra.loc[acessos_fibra.empresa=="ALGAR (CTBC TELECOM)",'empresa']="ALGAR TELECOM"
```

6.2 Analise Final – Solução do problema

A Oi fechou o ano de 2022 com **4,23 milhões** de acessos em banda larga por fibra ótica em todo o país.

```
[43] fig1.show()
fig2.show()
```



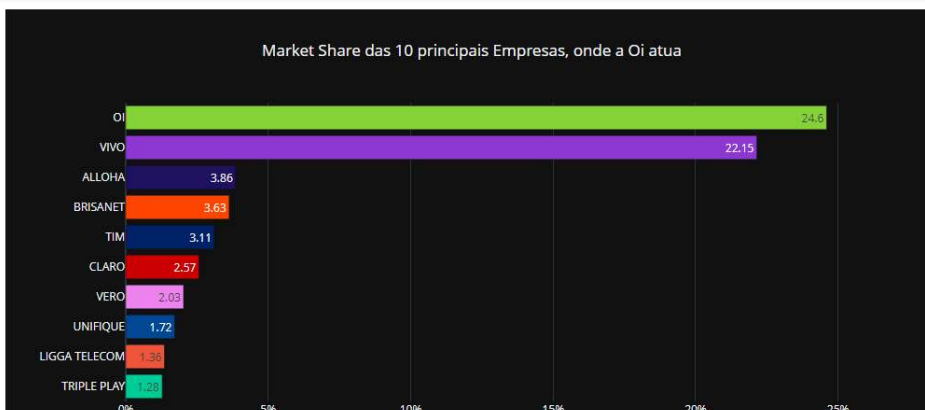
Durante o ano passado, foram conectados **540 mil novos acessos**, o que representa um **crescimento de 14,6%** em comparação com 2021.

```
[ ] cresc[(cresc['ano_mes'] == '202212')]
```

	ano_mes	Acessos	202112	novos_acessos	crescimento%
12	202212	4.23	3.69	0.54	14.6

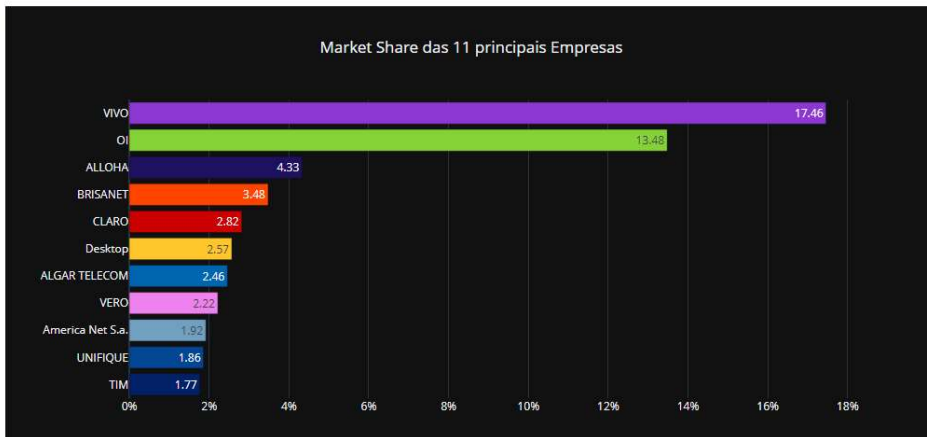
Considerando a tecnologia fibra, os acessos de Oi Fibra representam **25,3% do market share nas cidades onde atua com fibra**, o que evidencia o resultado positivo da estratégia da companhia de investir na expansão dos serviços digitais e de conexão por fibra ótica.

```
[44] fig3.show()
```



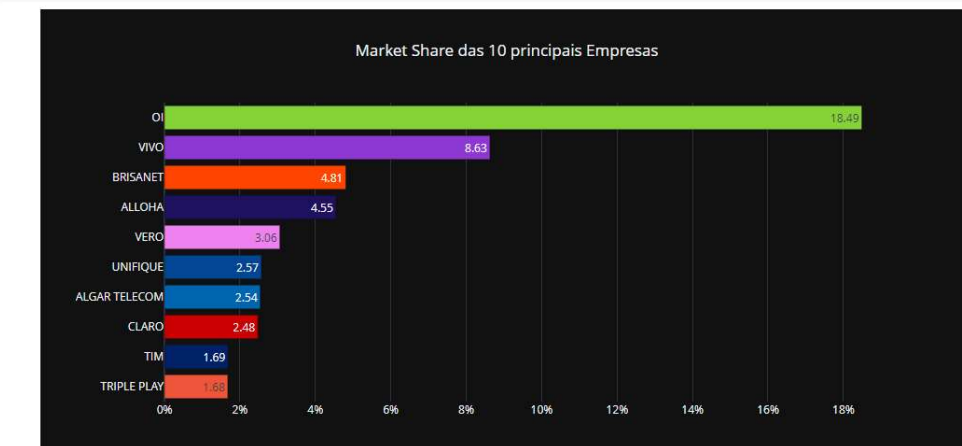
Considerando todas as cidades Brasileiras a Oi fica com **13,48%** do market share.

fig4.show()



Desconsiderando o estado de SP, a Oi fica com **18,49%** do market share. (A concessão de serviço em SP é mandatória para a Telefonica)

[50] fig5.show()



A Oi é líder em acessos em FIBRA em 84 cidades brasileiras.

```
tb_cidades_rk[(tb_cidades_rk['rank'] == 1) & (tb_cidades_rk['empresa'] == 'OI')].sort_values(by=['Acessos'], ascending=False)
```

	empresa	MUNICIPIO	Acessos	rank
5984	OI	RIO DE JANEIRO	285484	1.0
5797	OI	BRASILIA	177742	1.0
5839	OI	CURITIBA	168559	1.0
5792	OI	BELO HORIZONTE	159991	1.0
5992	OI	SALVADOR	129880	1.0
...
5929	OI	MORRINHOS	1980	1.0
5900	OI	JUINA	1819	1.0
5916	OI	MARACAJU	1530	1.0
5836	OI	CORUMBA	1145	1.0
5982	OI	RIBEIRAO DO PINHAL	542	1.0

84 rows x 4 columns

```
[ ] tb_gross_cresc.sort_values(by='market_share', ascending=False)
```

```
[ ] tb_gross_cresc.sp.sort values(by='market share', ascending=False)
```

```
[ ] # Somente OI
tb regiao cresc.sort values(by='percentual', ascending=False)
```

```
# Todas as operadoradoras
tb.regiao.crescx.sort_values(by='percentual', ascending=False)
```

[illegible]

Mercado por Velocidade

Oi Fibra também avançou em relação à velocidade da conexão por fibra, apresentando em 2022 o maior crescimento na velocidade média, entre as prestadoras de grande porte. A Oi Fibra cresceu 42,3% na velocidade média, bem acima do crescimento de 26,7% da segunda colocada e de 26,2% da terceira colocada.

```
[ ] pv_faixa.style.background_gradient(cmap='Greens').format("{:,}")
```

								Acessos
Velocidade_Agrupada	< 34Mbps	Entre 34 e 50Mbps	Entre 50 e 100Mbps	Entre 100 e 200Mbps	Entre 200 e 300Mbps	Entre 300 e 400Mbps	Entre 400 e 500Mbps	500 +
empresa								
CLARO	99,994	3,783	6,893	125,524	158,926	42,036	394,547	53,174
OI	279,734	76,229	239,445	1,467,951	734	1,499,668	611,266	59,000
SKY/AT&T	0	0	0	398	0	294	0	0
TIM	4,571	681	32,822	271,317	152,525	57	69,960	23,398
VIVO	45,399	160,162	380,671	2,228,824	2,447,409	0	24,273	197,182

Concluimos que a Oi teve um bom resultado no ano de 2022, comparando as mesmas cidades de atuação da concorrência.

Investir em manutenção da rede para que possa ser líder entre clientes de alta velocidade, dominando as faixas de 300 a 500 megas.

Indicamos que a empresa OI, deve investir mais nos estados do Ceará, Alagoas, Rio Grande do Norte e Paraíba, aonde a empresa BRISANET, está dominando o mercado.