

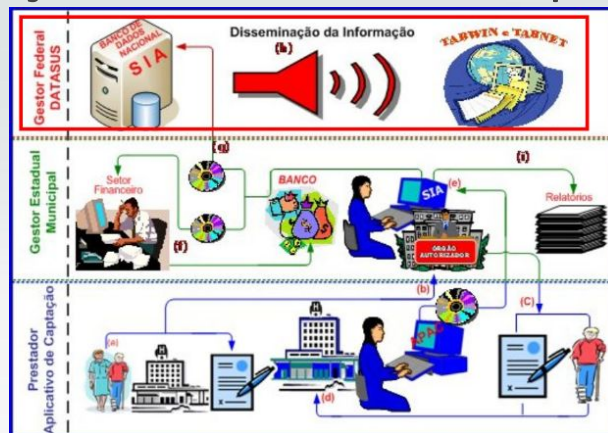
ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

Allan Silva



INTRODUÇÃO

- Disseminação da informação em arquivo no formato DBF/DBC.
- 154 milhões de entradas de produção ambulatorial no quarto trimestre de 2023.



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

INTRODUÇÃO

- TabWin - Programa disponibilizado pelo DATASUS para realizar a leitura dos arquivos.
- Não permite análise de dados em grande escala.
- Muitos arquivos estão particionados por Mês/Ano.

F:\FAPLIC\TABTABMORT\WANGR38.DBF									
Reg DO	CARTORIO	REGISTRO	DATA REG	TIPO REG	DATA OBTO	ESTO CIVIL	SEXO	DATA NASC	
1	0554485	0203	037533	980218	2	980217	4	2	19350126
2	05120590	0202	053112	980823	2	980822	4	2	19321205
3	05710560	1302	058776	980810	2	980808	2	1	19410510
4	05723646	0801	161439	981010	2	981010	1	2	19770203
5	05739727	0205	048033	981120	2	981119	1	1	19550415
6	05740036	0202	053849	981121	2	981120	2	2	19491117
7	05747411	0301	043798	981211	2	981209	2	1	19541222
8	05753974	0801	162667	981222	2	981222	1	1	19520816
9	05582508			2	980128	2	1	19401016	
10	05582596			2	980203	3	2	19830128	
11	05619406			2	980511	2	1	19380524	
12	05707852			2	980720	9	2	19350320	
13	05683030			2	980714	2	1	19250407	
14	06745698	0001	022643	980302	2	980301	2	2	19640604
15	05693293	0002	034766	980801	2	980801	2	1	19370406
16	05720006	0003	036517	981103	2	981102	1	1	19900426
17	05575795	0001	048251	980822	2	980821	2	1	19460604



PROBLEMA

- Arquivos DBC/DBF não são suportados nas ferramentas de análise de dados do mercado.
- Curva de aprendizado para utilizar alternativas ao TabWin existentes, como as bibliotecas PySUS e read.dbc.



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

PROBLEMA

producao_ambulatorial

QUERY

SHARE

COPY

SNAPSHOT

SCHEMA

DETAILS

PREVIEW

LINEAGE

DATA PROFILE

DATA QUALITY

Table info

Table ID

puc-tcc-412315.informacoes_ambulatoriais.producao_ambulatorial

Created

Feb 11, 2024, 6:13:46 PM UTC-3

Last modified

Feb 13, 2024, 1:30:08 AM UTC-3

Table expiration

NEVER

Data location

US

Default collation

Default rounding mode

ROUNDING_MODE_UNSPECIFIED

Case insensitive

false

Description

Labels

Primary key(s)

Tags

Storage info

Number of rows

154,158,549

Total logical bytes

65.17 GB

Active logical bytes

65.17 GB

Long term logical bytes

0 B

- Outubro/2023
Novembro/2023
Dezembro de 2023;
- 154 milhões de registros, distribuídos em;
- 93 arquivos DBF/DBC.



OBJETIVO

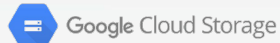
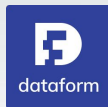
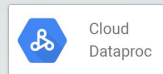
- Prover uma biblioteca de código aberto, criada em uma linguagem de programação interoperável, que converte o formato DBF/DBC em formato Parquet, suportado pela maioria das ferramentas de análise de dados do mercado.
 - O gestor federal do DATASUS poderia escolher integrar essa facilidade no fluxo de disseminação, liberando ao público tanto os arquivos DBF/DBC, quanto os arquivos parquet correspondentes.
 - O cidadão que quer analisar os dados de saúde, não precisaria ter a curva de aprendizado relacionada aos formatos DBF/DBC.
- Demonstrar através de um pipeline de engenharia de dados, ponta a ponta, a viabilidade do uso da biblioteca

OBJETIVO

- Prover uma biblioteca de código aberto, criada em uma linguagem de programação interoperável, que converte o formato DBF/DBC em formato Parquet, suportado pela maioria das ferramentas de análise de dados do mercado.
 - O gestor federal do DATASUS poderia escolher integrar essa facilidade no fluxo de disseminação, liberando ao público tanto os arquivos DBF/DBC, quanto os arquivos parquet correspondentes.
 - O cidadão que quer analisar os dados de saúde, não precisaria ter a curva de aprendizado relacionada aos formatos DBF/DBC.
- Demonstrar através de um pipeline de engenharia de dados, ponta a ponta, a viabilidade do uso da biblioteca

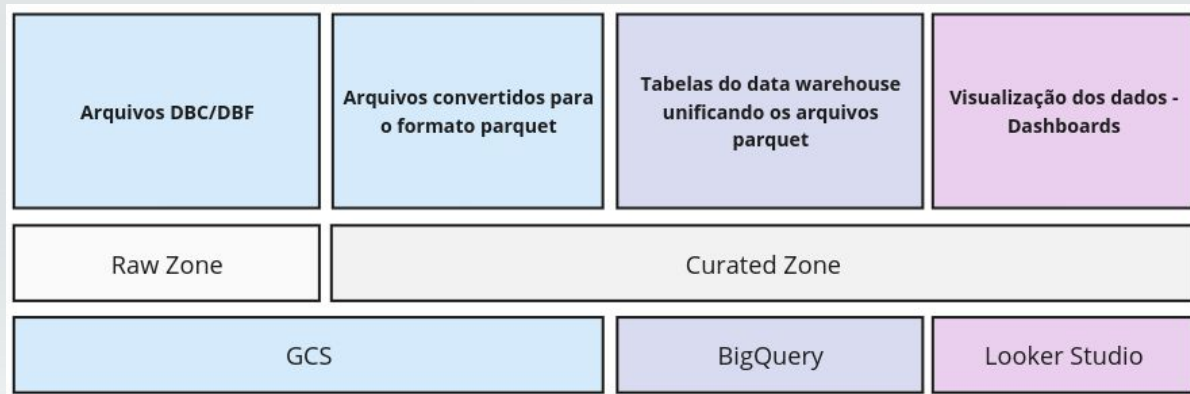
ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TECNOLOGIAS



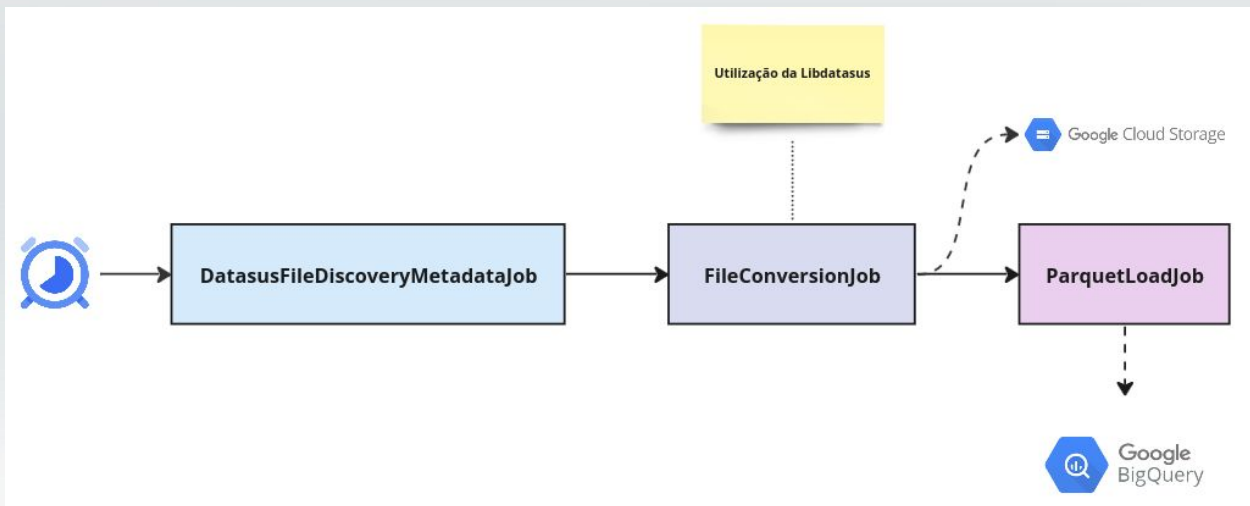
ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

ARQUITETURA



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

INGESTÃO DE DADOS



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

INGESTÃO DE DADOS - ORQUESTRAÇÃO

Workflow instance details

79746d92-5a4e-4ec9-83d1-81434397e55a

Template: [informacoes-ambulatoriais](#)

Job details

Filter Filter jobs

Step ID	Status	Job ID	Dependencies
file-discovery	Completed	file-discovery-aq4co62pl73ek	No dependencies
file-conversion	Completed	file-conversion-aq4co62pl73ek	file-discovery
parquet-load	Completed	parquet-load-aq4co62pl73ek	file-conversion

Job ID	file-conversion-aq4co62pl73ek
Job UUID	19381563-7b16-4b44-a8b3-eb09e7b8f715
Type	Dataprocc Job
Status	Succeeded
Output	<div>LINE WRAP: OFF</div> <div>24/02/19 08:25:41 INFO FileConversionJob: Read session 'readSessionName' for project 'readSessionName' - 'projects/pjct-id-12345' location 's3://session/04100f8b7b9d07b1d0c0e0a0b0' -</div> <div>24/02/19 08:25:41 INFO ReadSessionCreator: Requested 20000 new partitions, but only received 1 from the Rigbary Storage API for session project/pjct-id-12345</div> <div>24/02/19 08:25:41 INFO RigbaryDataSourceReaderContent: Get read session for DataSourceInfo[datasetId, projectId, tableId], [datasetId=ingestion_info</div> <div>-----</div> <div>[source_file_uri] [parquet_file_uri] [source converted_date] [success/error_message]</div> <div>-----</div> <div>gs://informacoes-ambulatoriais-raw/PAPR2311.dbc gs://informacoes-ambulatoriais-curated/PAPR2311.dbc.parquet STA 2024-02-19 True NULL </div> <div>gs://informacoes-ambulatoriais-raw/PAPR2310.dbc gs://informacoes-ambulatoriais-curated/PAPR2310.dbc.parquet STA 2024-02-19 True NULL </div> <div>gs://informacoes-ambulatoriais-raw/PAPR2310.dbc gs://informacoes-ambulatoriais-curated/PAPR2310.dbc.parquet STA 2024-02-19 True NULL </div>

```
jobs {
  step_id = "file-discovery"
  spark_job {
    args = [var.raw_bucket, var.source_system]
    main_class = "br.dev.contrib.gov.sus.opendata.jobs.DatasusFileDiscoveryMetadataJob"
    jar_file_uri = ["${var.job_bucket}/datasussparkjobs_2.12-0.1.0-SNAPSHOT.jar"]
  }
}

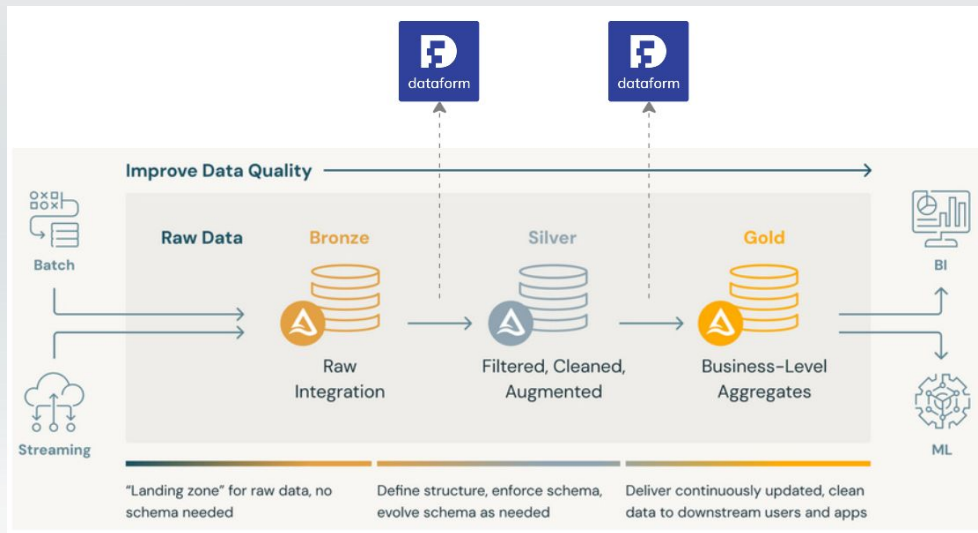
jobs {
  step_id = "file-conversion"
  prerequisite_step_ids = ["file-discovery"]
  spark_job {
    args = [var.source_system, var.curated_bucket, var.worker_instances]
    main_class = "br.dev.contrib.gov.sus.opendata.jobs.FileConversionJob"
    jar_file_uri = ["${var.job_bucket}/datasussparkjobs_2.12-0.1.0-SNAPSHOT.jar"]
    properties = {
      "spark.jars.packages" : "br.dev.contrib.gov.sus.opendata:libdatasus-parquet-dbf:1.0.7"
    }
  }
}

jobs {
  step_id = "parquet-load"
  prerequisite_step_ids = ["file-conversion"]
  spark_job {
    args = [var.source_system]
    main_class = "br.dev.contrib.gov.sus.opendata.jobs.ParquetLoadJob"
    jar_file_uri = ["${var.job_bucket}/datasussparkjobs_2.12-0.1.0-SNAPSHOT.jar"]
  }
}
```



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TRATAMENTO DE DADOS



<https://www.databricks.com/glossary/medallion-architecture>



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TRATAMENTO DE DADOS

← datasus CODE COMPILED GRAPH EXECUTIONS START EXECUTION ▾

Files + ▾ <|

Workspace is up to date

Q Type to search ?

- definitions
 - bronze
 - gold
 - estabelecimentos
 - producao_ambulatorial
 - gold_producao_ambul...
 - silver
 - includes
 - .gitignore
 - dataform.json
 - package-lock.json ?
 - package.json

definitions/gold/producao_ambulatorial/gold_producao_ambulatorial.sqlx

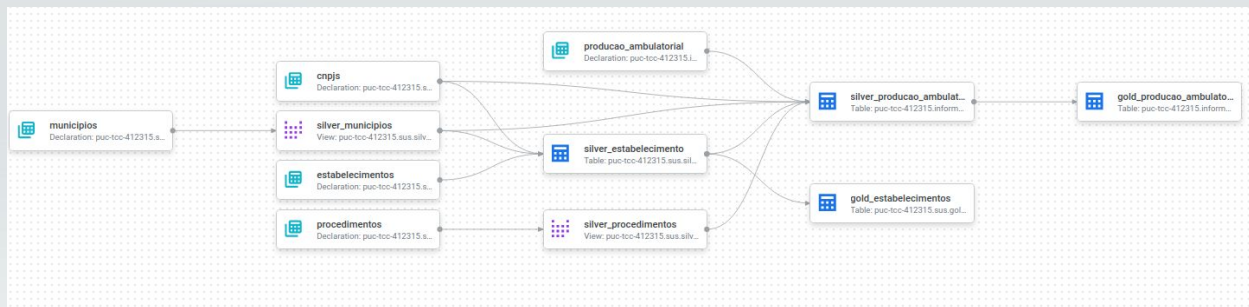
Press Alt+F1 for Accessibility Options.

```
1 config {
2   type: "table",
3   database: "puc-tcc-412315",
4   schema: "informacoes_ambulatoriais",
5   dependencies: ["silver_producao_ambulatorial"]
6 }
7
8 SELECT
9   'Brasil' AS Brasil,
10  pa.uf_municipio_estabelecimento,
11  pa.municipio_estabelecimento,
12  pa.cnpj_producao,
13  pa.razao_social_producao,
14  pa.complexidade_procedimento,
15  pa.descricao_procedimento,
16  pa.tipo_financiamento,
17  pa.situacao_producao,
18  pa.origem_informacao,
19  SUM(pa.valor_declarado_procedimento) AS valor_declarado_procedimento,
20  SUM(pa.valor_aprovado_procedimento) AS valor_aprovado_procedimento,
21  SUM(pa.divergencia_valor_declarado) AS divergencia_valor_declarado_tabela,
22  SUM(CASE
23    WHEN pa.paciente_reside_mesmo_estado_estabelecimento THEN 1
24    ELSE
25    0
26  END
```



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TRATAMENTO DE DADOS



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TRATAMENTO DE DADOS - ORQUESTRAÇÃO

← puc-tcc-datasus

DEVELOPMENT WORKSPACESWORKFLOW EXECUTION LOGSRELEASES & SCHEDULINGSETTINGS

Release configurations let you configure how Dataform should compile the code of your repository. If your repository is connected to a remote git repository, you can create release configurations from different branches. Dataform will pull code from your remote git repository before compiling it. [Learn more](#)

Release configurations [CREATE](#) [START EXECUTION](#)

Filter Enter name

Name ↑	Remote branch	Cron schedule	Last updated	Status
production	datasus	Every day at 7:00 AM UTC	Mar 2, 2024, 4:00:02 AM	✓ Active



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

TRATAMENTO DE DADOS - ORQUESTRAÇÃO

← Mar 2, 2024, 4:00:01 AM

CANCEL WORKFLOW

⌂ REFRESH

Details

Start time

Mar 2, 2024, 4:00:01 AM

Status

✔ Success

Duration

1 minute 35 seconds

Source type

Workflow configuration

Source

prod-deployment

Actions

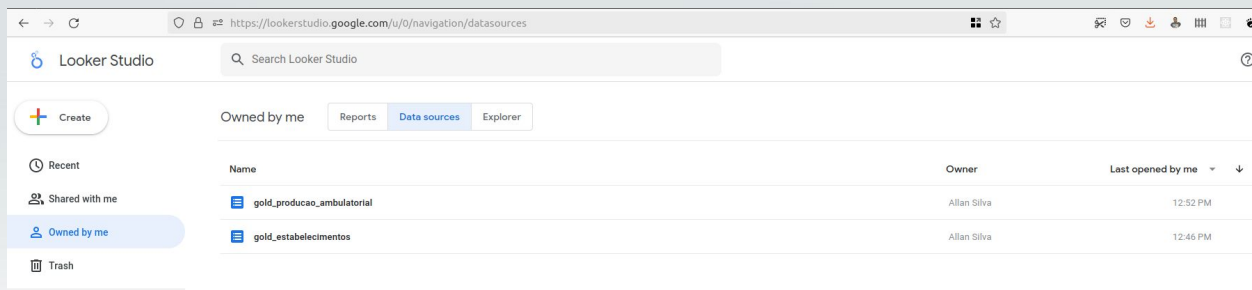
Filter

Enter property name or value

Status	Start time ↑	Duration	Action	Destination	Details
✔	Mar 2, 2024, 4:00:02 AM	2 seconds	silver_municipios	puc-tcc-412315.sus.silver_municipios	VIEW DETAILS
✔	Mar 2, 2024, 4:00:02 AM	2 seconds	silver_procedimentos	puc-tcc-412315.sus.silver_procedimentos	VIEW DETAILS
✔	Mar 2, 2024, 4:00:06 AM	32 seconds	silver_estabelecimento	puc-tcc-412315.sus.silver_estabelecimento	VIEW DETAILS
✔	Mar 2, 2024, 4:00:40 AM	42 seconds	silver_producao_ambulatorial	puc-tcc-412315.informacoes_ambulatoriais.silver_producao_amb...	VIEW DETAILS
✔	Mar 2, 2024, 4:00:40 AM	6 seconds	gold_estabelecimentos	puc-tcc-412315.sus.gold_estabelecimentos	VIEW DETAILS
✔	Mar 2, 2024, 4:01:24 AM	12 seconds	gold_producao_ambulatorial	puc-tcc-412315.informacoes_ambulatoriais.gold_producao_ambu...	VIEW DETAILS

ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

VISUALIZAÇÃO DE DADOS

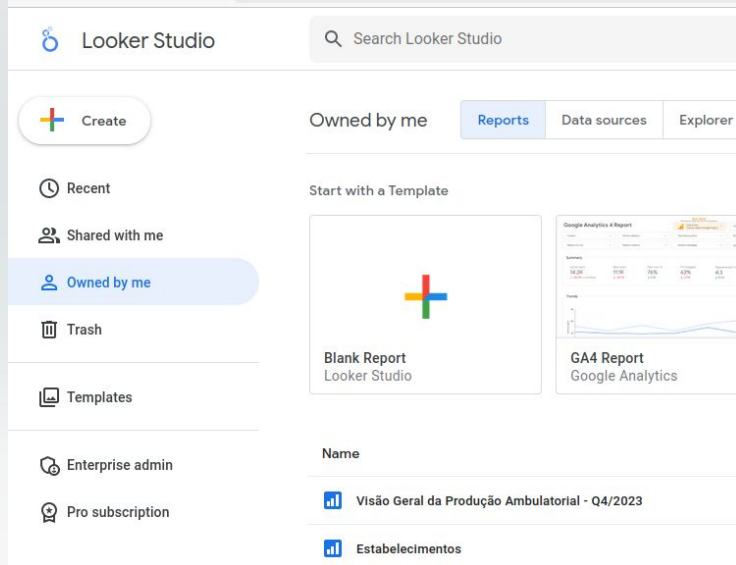
A screenshot of the Looker Studio web interface. The browser address bar shows 'https://lookerstudio.google.com/u/0/navigation/datasources'. The interface includes a search bar, a 'Create' button, and a sidebar with navigation options: 'Recent', 'Shared with me', 'Owned by me' (selected), and 'Trash'. The main content area shows a table of data sources owned by the user.

Name	Owner	Last opened by me
gold_producao_ambulatorial	Allan Silva	12:52 PM
gold_estabelecimentos	Allan Silva	12:46 PM



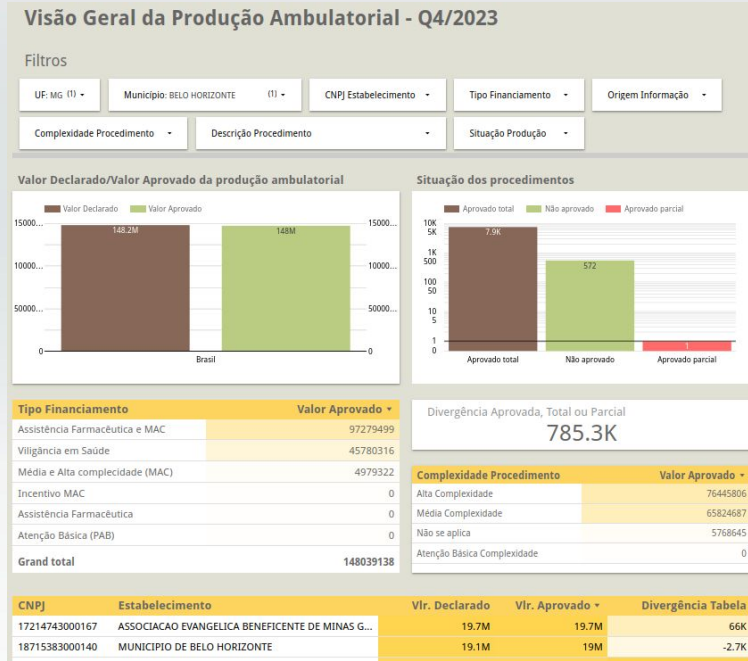
ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

VISUALIZAÇÃO DE DADOS



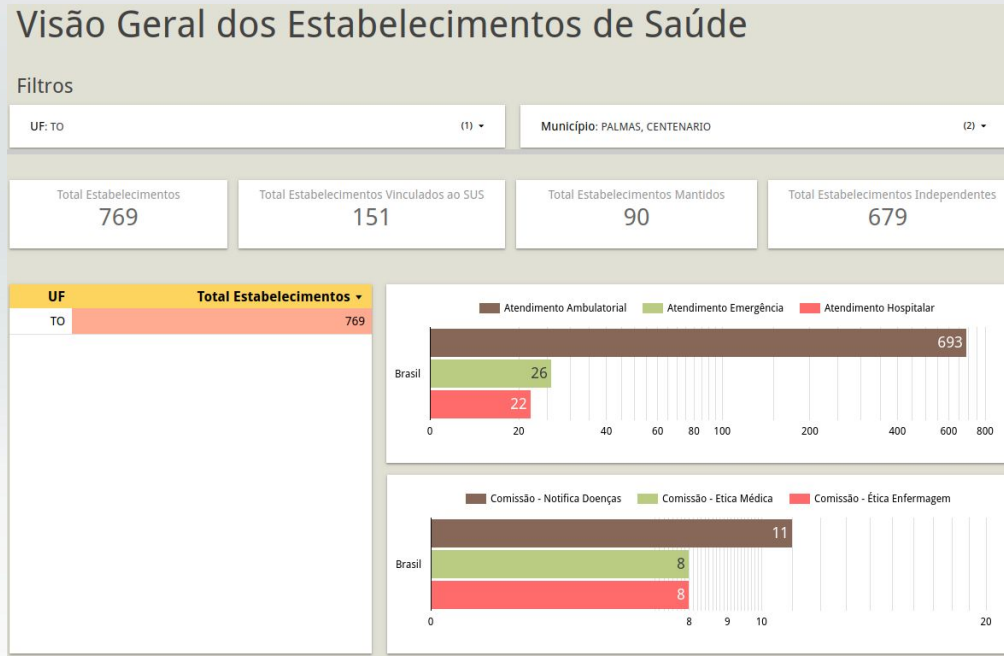
ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

VISUALIZAÇÃO DE DADOS



ENTREGANDO DADOS PÚBLICOS DOS SISTEMAS DO DATASUS EM FORMATO INTEROPERÁVEL COM AS FERRAMENTAS DE ANÁLISE DE DADOS DE MERCADO

VISUALIZAÇÃO DE DADOS



LINKS

Repositório TCC:

<https://github.com/allan-silva/DE-puc-tcc>

<https://github.com/allan-silva/DE-puc-tcc-Dataform/tree/datasus> (gerenciado pelo dataform).

Libdatasus:

<https://github.com/allan-silva/libdatasus>

<https://mvnrepository.com/artifact/br.dev.contrib.gov.sus.opendata/libdatasus-parquet-dbf>

Contato:

allan@allansilva.com.br

www.linkedin.com/in/allan-t-silva

