Rio de Janeiro, 26 de Janeiro de 2023

CONTROLE DE VERSÃO			
Autor	**Versão**	**Data**	**Descrição**
Thiago Wilian de Barros Ferreira	1.0	26/01/2023	Criação do documento

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Esta documentação tem por finalidade detalhar as necessidades apresentadas no Laboratório 4 de Azure, bem como expor o desenvolvimento das soluções solicitadas.

2 Solicitação

O Instituto Pocco de Artes Visuais (IPAV) está empenhado em apoiar e impulsionar programas culturais no Brasil. Para isso foi solicitado à Blueshift a realização do mapeamento das instituições, eventos e projetos distribuídos no país. A presidente do instituto, Stefani Germanotta, deseja ter, em um dashboard, uma visão geral das entidades e projetos culturais no Brasil, divididos por estado e região, bem como um cronograma com os principais eventos.

3 Premissas da solução

Origem e especificação dos dados

Os dados sobre as instituições culturas e os eventos são extraídos da API do MuseusBR no formato JSON através das seguintes URL's:

- **Instituições culturas (Museus)**

  - `http://museus.cultura.gov.br/api/space/find?@select=id,name,location,shortDescription,acessibilidade,acessibilidade_fisica,mus_acessibilidade_visual,capacidade,endereco,En_Municipio,En_Estado,telefonePublico,emailPublico,site,mus_ingresso_cobrado,horario`

- **Eventos**

  - `http://museus.cultura.gov.br/api/event/find?@select=*`

  - `http://museus.cultura.gov.br/api/event/find?@select=occurrences.*`

Ambiente de desenvolvimento

O cliente disponibilizará, ao time BlueShift, os acessos necessários ao ambiente de desenvolvimento da Microsoft Azure, um serviço de computação em nuvem, e às suas ferramentas, apontadas na arquitetura proposta na imagem 1, e citadas abaixo.

- **Visual Studio Code** - O VS Code, editor de código, é utilizado para a extração dos dados da API do MuseusBR, através da linguagem de programação Python. Em seguida, os dados são limpos, transformados e convertidos para arquivos do tipo Parquet. Por fim, esses arquivos são enviados para o Azure Blob Storage.

- **Azure Blob Storage** - Recebe os arquivos Parquet, enviados pelo VS Code, e os armazena em um contêiner.

- **Azure Databricks** - É utilizado o Azure Databricks para a extração dos dados armazenados em um contêiner do Azure Blob Storage. Em seguida, ocorre a finalização do processo de limpeza e transformação dos dados, para possibilitar a gravação desses dados nas tabelas temporárias no Azure SQL.

- **Azure SQL** - Criação e armazenamento das tabelas. É também, utilizada uma Stored Procedure para inserir os dados nas tabelas Fato e nas tabelas Dimensão.
- **Azure Data Factory** - Ferramenta utilizada para orquestrar todas as atividades citadas anteriormente (exceto o Visual Studio).
- **Power BI** - Consome os dados das tabelas Fato e Dimensão, a fim de construir um dashboard para obter de uma visão geral das entidades e projetos culturais no Brasil.

4 Modelo da arquitetura sugerida

A Figura abaixo apresenta a arquitetura da solução proposta, levando em consideração o levantamento de requisitos e entendimento do negócio.

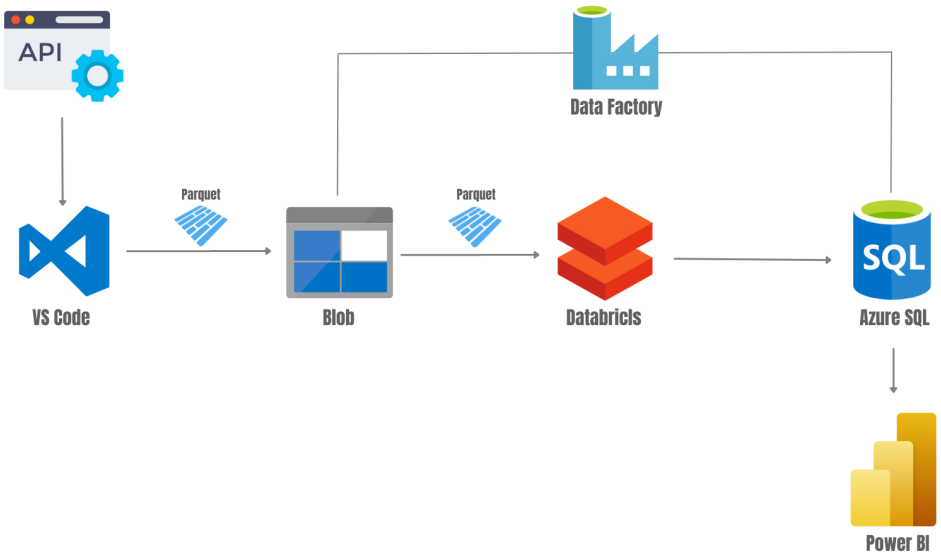

Figura 1: Arquitetura do projeto.

As ferramentas utilizadas necessárias para a construção e desenvolvimento do projeto estão enumeradas abaixo:

- **Visual Studio Code** - Utilizado como editor de código para realizar a extração, limpeza, transformação, conversão e transferência dos dados obtidos na API do MuseusBR.
- **Azure Blob Storage** - É utilizado para armazenamento dos arquivos Parquet em um contêiner, que será utilizado no projeto.

- **Databricks** - O Databricks é utilizado para criação de um notebook para limpeza e transformação dos dados dos arquivos Parquet, localizado no Blob Storage.
- **Azure SQL** - Armazena dentro de tabelas os dados gravados pelo notebook Databricks.
- **Azure Data Factory** - Utilizado para orquestrar todas as atividades anteriores (exceto o VS Code), através de um Pipeline.
- **Power BI** - Obtendo os dados das tabelas do Azure SQL, o Power BI é a ferramenta utilizada com a finalidade de gerar uma visão ampla das instituições culturais e dos eventos no Brasil.

5 Dicionário de dados

Segue abaixo a estrutura das tabelas:

5.1 Museus

- **Tabela [STAGE_thiago_wilan4].[STAGE_museus]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Nome	VARCHAR	MAX	NOT NULL	Nome do museu
Desc_curta	VARCHAR	MAX	NOT NULL	Curta descrição sobre o museu
Regiao	VARCHAR	20	NOT NULL	Localização regional do museu
Estado	VARCHAR	2	NOT NULL	Localização estadual do museu
Municipio	VARCHAR	100	NOT NULL	Localização municipal do museu
Endereco	VARCHAR	MAX	NOT NULL	Endereço do museu
Horario	VARCHAR	MAX	NOT NULL	Horário de funcionamento
Ingresso_cobrado	VARCHAR	40	NOT NULL	Obrigatoriedade de ingresso
Sites	VARCHAR	MAX	NOT NULL	Site do Museu
Telefone_pub	VARCHAR	MAX	NOT NULL	Telefone público do museu
Email_pub	VARCHAR	MAX	NOT NULL	Email público do museu
Acessibilidade	VARCHAR	20	NOT NULL	Acessibilidade do museu
Acess_fisica	VARCHAR	MAX	NOT NULL	Acessibilidade física do museu
Acess_visual	VARCHAR	MAX	NOT NULL	Acessibilidade visual do museu
Latitude	FLOAT	-	NOT NULL	Latitude do museu
Longitude	FLOAT	-	NOT NULL	Longitude do museu

Tabela 1: Tabela [STAGE_thiago_wilan4].[STAGE_museus]

- **Tabela [DW_thiago_wilian4].[Fato_museus]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Nome	VARCHAR	MAX	NOT NULL	Nome do museu
Desc_curta	VARCHAR	MAX	NOT NULL	Curta descrição sobre o museu
Municipio	VARCHAR	100	NOT NULL	Localização municipal do museu
Endereco	VARCHAR	MAX	NOT NULL	Endereço do museu
Horario	VARCHAR	MAX	NOT NULL	Horário de funcionamento
Sites	VARCHAR	MAX	NOT NULL	Site do Museu
Telefone_pub	VARCHAR	MAX	NOT NULL	Telefone público do museu
Email_pub	VARCHAR	MAX	NOT NULL	Email público do museu
Acess_fisica	VARCHAR	MAX	NOT NULL	Acessibilidade física do museu
Acess_visual	VARCHAR	MAX	NOT NULL	Acessibilidade visual do museu
Latitude	FLOAT	-	NOT NULL	Latitude do museu
Longitude	FLOAT	-	NOT NULL	Longitude do museu

Tabela 2: Tabela [DW_thiago_wilian4].[Fato_museus]

- **Tabela [DW_thiago_wilian4].[DIM_Regiao]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Regiao	VARCHAR	20	NOT NULL	Localização regional do museu

Tabela 3: Tabela [DW_thiago_wilian4].[DIM_Regiao]

- **Tabela [DW_thiago_wilian4].[DIM_Estado]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Estado	VARCHAR	2	NOT NULL	Localização estadual do museu

Tabela 4: Tabela [DW_thiago_wilian4].[DIM_Estado]

- **Tabela [DW_thiago_wilian4].[DIM_Ingresso_cobrado]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Ingresso_cobrado	VARCHAR	40	NOT NULL	Obrigatoriedade de ingresso

Tabela 5: Tabela [DW_thiago_wilian4].[DIM_Ingresso_cobrado]

- **Tabela [DW_thiago_wilian4].[DIM_Acessibilidade]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do museu
Acessibilidade	VARCHAR	20	NOT NULL	Acessibilidade do museu

Tabela 6: Tabela [DW_thiago_wilian4].[DIM_Acessibilidade]

5.2 Eventos

- **Tabela [STAGE_thiago_wilian4].[STAGE_eventos]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do evento
Id_museu	INT	-	NOT NULL	ID do museu
Nome_evento	VARCHAR	MAX	NOT NULL	Nome do Evento
Curta_desc	VARCHAR	MAX	NOT NULL	Curta descrição
Longa_desc	VARCHAR	MAX	NOT NULL	Longa descrição
Subtitulo	VARCHAR	MAX	NOT NULL	Subtítulo do evento
Preco	VARCHAR	MAX	NOT NULL	Preço do ingresso
Inicio	DATE	-	NOT NULL	Data inicial do evento
Fim	DATE	-	NOT NULL	Data final do evento
Hora_inicio	VARCHAR	10	NOT NULL	Hora inicial do evento
Hora_fim	VARCHAR	10	NOT NULL	Hora final do evento
Descript	VARCHAR	MAX	NOT NULL	Descrição do evento
Class_etaria	VARCHAR	30	NOT NULL	Classificação etária
Frequencia	VARCHAR	20	NOT NULL	Frequência do evento
Telefone_pub	VARCHAR	MAX	NOT NULL	Telefone do evento
Sites	VARCHAR	MAX	NOT NULL	Site do evento
Trad_libras	VARCHAR	MAX	NOT NULL	Tradução em Libras
Desc_sonora	VARCHAR	MAX	NOT NULL	Descrição sonora
Obs	VARCHAR	MAX	NOT NULL	Observações do evento

Tabela 7: Tabela [STAGE_thiago_wilian4].[STAGE_eventos]

- **Tabela [DW_thiago_wilian4].[Fato_eventos]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do evento
Id_museu	INT	-	NOT NULL	ID do museu
Nome_evento	VARCHAR	MAX	NOT NULL	Nome do Evento
Curta_desc	VARCHAR	MAX	NOT NULL	Curta descrição
Longa_desc	VARCHAR	MAX	NOT NULL	Longa descrição
Subtitulo	VARCHAR	MAX	NOT NULL	Subtítulo do evento
Preco	VARCHAR	MAX	NOT NULL	Preço do ingresso
Inicio	DATE	-	NOT NULL	Data inicial do evento
Fim	DATE	-	NOT NULL	Data final do evento
Hora_inicio	VARCHAR	10	NOT NULL	Hora inicial do evento
Hora_fim	VARCHAR	10	NOT NULL	Hora final do evento
Descript	VARCHAR	MAX	NOT NULL	Descrição do evento
Telefone_pub	VARCHAR	MAX	NOT NULL	Telefone do evento
Sites	VARCHAR	MAX	NOT NULL	Site do evento
Trad_libras	VARCHAR	MAX	NOT NULL	Tradução em Libras
Desc_sonora	VARCHAR	MAX	NOT NULL	Descrição sonora
Obs	VARCHAR	MAX	NOT NULL	Observações

Tabela 8: Tabela [DW_thiago_wilian4].[DIM_Class_etaria]

- **Tabela [DW_thiago_wilian4].[DIM_Class_etaria]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do evento
Class_etaria	VARCHAR	30	NOT NULL	Classificação etária

Tabela 9: Tabela [DW_thiago_wilian4].[DIM_Class_etaria]

- **Tabela [DW_thiago_wilian4].[DIM_Frequencia]**

Campo	Tipo	Tamanho	Restrição	Descrição
Id	INT	-	PRIMARY KEY	ID do evento
Frequencia	VARCHAR	20	NOT NULL	Frequência do evento

Tabela 10: Tabela [DW_thiago_wilian4].[DIM_Frequencia]

6 Processo de desenvolvimento até etapa final

Segue abaixo os processos que são realizados:

6.1 VS Code

São criados dois arquivos do tipo Python dentro do editor de código VS Code. Através desses arquivos, os dados sobre os museus e os eventos são extraídos da API do MuseusBR. Feito isso, esses dados são organizados em dois DataFrames do Pandas para que possam ser explorados, analisados, limpos, transformados e por último, convertidos em dois arquivos Parquet para que, em seguida, sejam enviados para um contêiner no Azure Blob Storage.

6.2 Notebook Databricks

Criação de um notebook no Databricks, extraindo os dados localizados em um contêiner do Azure Blob Storage. Posteriormente, já com os dados carregados e transformados em dois DataFrames do Spark, foram realizados alguns processos de limpeza nos dados, como:

- Conversão do tipo de dado das colunas de data, que estavam no formato de texto(*string*) para o formato de data(*date*) e para a estrutura "*dia/mês/ano*".
- Conversão do tipo de dado das colunas numéricas, anteriormente no formato numérico *float* para o formato numérico *int*.
- Conversão do tipo de dado das colunas numéricas, anteriormente no formato de texto(string) para o formato numérico(*float*)

6.3 Azure SQL

Criação de dois Schemas (**[STAGE_thiago_wilian4]** e **[DW_thiago_wilian4]**) para a criação e armazenamento das tabelas abaixo:

6.3.1 Schema [STAGE_thiago_wilian4]

- **Tabelas [STAGE_museus] e [STAGE_eventos]** - A fim de armazenamento prévio e para servir de fonte para as tabelas Fato e Dimensão, as tabelas **[STAGE_museus]** e **[STAGE_eventos]**, criadas dentro do Schema [STAGE_thiago_wilian4], funcionam como tabelas temporárias e recebem os dados *tratados pelo Azure Databrick*.

6.3.2 Schema [DW_thiago_wilian4]

Segue abaixo as tabelas Fato e as Dimensão.

- **Tabela DIM_Regiao** - Armazenando os dados do ID do museu e da região geográfica de onde se localizam os museus, a tabela **DIM_Regiao**, criada dentro do Schema [DW_thiago_wilian4], exerce uma função de filtragem para as consultas no Power BI.
- **Tabela DIM_Estado** - Criada dentro do Schema [DW_thiago_wilian4], a tabela **DIM_Estado** armazena dados do ID do museu e do estado de onde esse museu se localiza, e funciona como um filtro para as consultas no Power BI.

- **Tabela DIM_Ingresso_cobrado** - Apresentada pelo Schema [DW_thiago_wilian4] a tabela **DIM_Ingresso_cobrado** contém informações sobre o ID do museu e as regras relacionadas à cobrança de ingresso. Essa tabela também serve como um mecanismo de filtragem para as consultas realizadas no Power BI.

- **Tabela DIM_Acessibilidade** - Dentro do Schema [DW_thiago_wilian4] foi criado a tabela **DIM_Acessibilidade** para armazenar detalhes a respeito do ID dos museus e se essa instituição possui acessibilidade para pessoas portadoras de deficiência, além de funcionar como um filtro no Power BI.

- **Tabela Fato_museus** - A tabela **Fato_museus**, criada dentro do Schema [DW_thiago_wilian4] contém o ID do museu e informações descritivas sobre cada museu, como: nome, curta descrição, horário de funcionamento, site, telefone público, email público, acessibilidade física e visual, município, endereço, latitude, longitude, e possui relacionamento com as tabelas Dimensões: DIM_Regiao, DIM_Estado, DIM_Ingresso_cobrado e DIM_Acessibilidade.

- **Tabela DIM_Class_etaria** - A tabela **DIM_Class_etaria** armazena o ID de cada evento e possui informações sobre as faixas etárias permitidas para participar do evento. Essa tabela visa aprimorar as consultas no Power BI, exercendo um papel de filtro, e localiza-se dentro do Schema [DW_thiago_wilian4].

- **Tabela DIM_Frequencia** - A tabela **DIM_Frequencia** reune informações sobre o ID do evento e a frequência da realização do evento (uma vez, diário, semanal). Apresentada pelo Schema [DW_thiago_wilian4], essa tabela também funciona como um recurso para filtrar as pesquisas realizadas no Power BI.

- **Tabela Fato_eventos** - Presente no Schema [DW_thiago_wilian4], a tabela **Fato_eventos** contém o ID dos eventos, o ID dos museus de onde ocorrem e informações descritivas sobre cada evento, como: nome, curta descrição, longa descrição, subtítulo do evento, preço do ingresso, horário do evento, o período em que ele ficará disponível, o resumo do evento, o site, telefone público, se possui tradução em libras e descrição sonora, e observações gerais. A tabela **Fato_eventos** possui relacionamento com as tabelas Dimensões: DIM_Class_etaria e DIM_Frequencia.

6.3.3 Stored Procedure

O procedimento armazenado (Stored Procedure) nomeado de **stage_to_dw_thiago_wilian4**, criado no Azure SQL, dentro do Schema **[STAGE_thiago_wilian4]**, inicialmente, apaga todos os dados das tabelas presentes no Schema [DW_thiago_wilian4], para que possam receber os dados provenientes das tabelas STAGE_museus e STAGE_eventos. Em seguida, os dados são inseridos nas tabelas do Schema [DW_thiago_wilian4]. Por fim, todos os dados das tabelas STAGE_museus e STAGE_eventos são apagados, a fim de evitar dados duplicados, quando forem populados novamente.

6.4 Pipeline - Data Factory

Através da ferramenta Azure Data Factory, são criadas duas atividades no Pipeline: **Databricks** e **Procedure**. Toda vez que o Pipeline for depurado, a atividade "Databricks"(conforme explicado na subseção 6.2) executa todos os comandos do Notebook Databricks. Tendo êxito na atividade anterior, o procedimento armazenado stage_to_dw_thiago_wilian4 é executado na atividade "Procedure"(descrito no 6.3.3).

6.5 Power BI

6.5.1 Solicitações do cliente

No Power BI, é estabelecida uma conexão com o Azure SQL para obter os dados presentes nas tabelas do Schema [DW_thiago_wilian4] a fim de obter uma visão geral das entidades e projetos culturais no Brasil, separados por região e estado, e também um cronograma com os principais eventos.

6.5.2 Relacionamento

A visão multidimensional permite o uso mais intuitivo para o processamento analítico. É de grande importância uma boa modelagem multidimensional para permitir bom desempenho, intuitividade e escalabilidade em um DataWarehouse, e é por esse motivo que foi escolhido o esquema estrela como modelo. Segue abaixo o modelo do projeto:

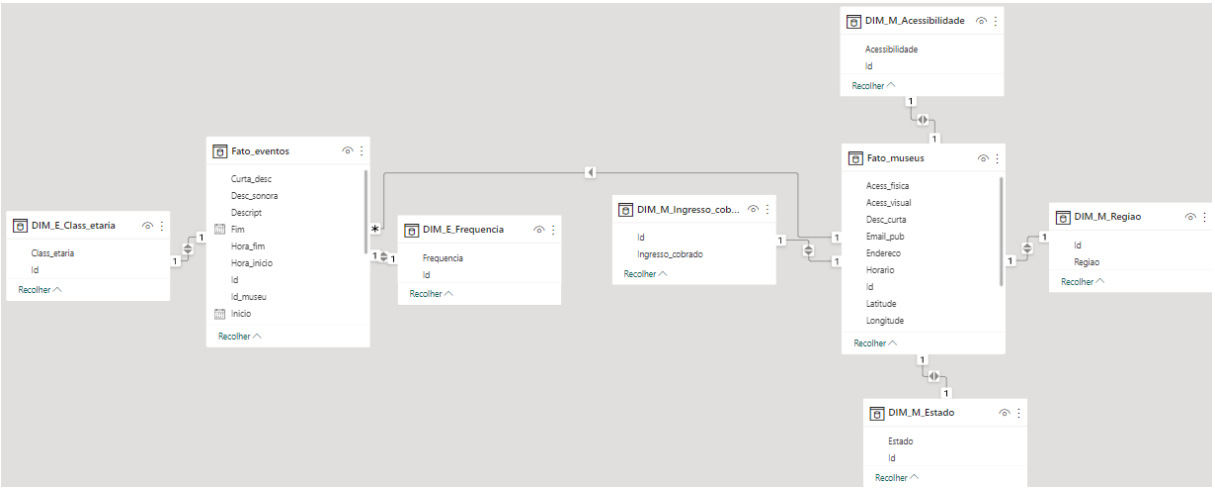

Figura 2: Modelo Esquema estrela

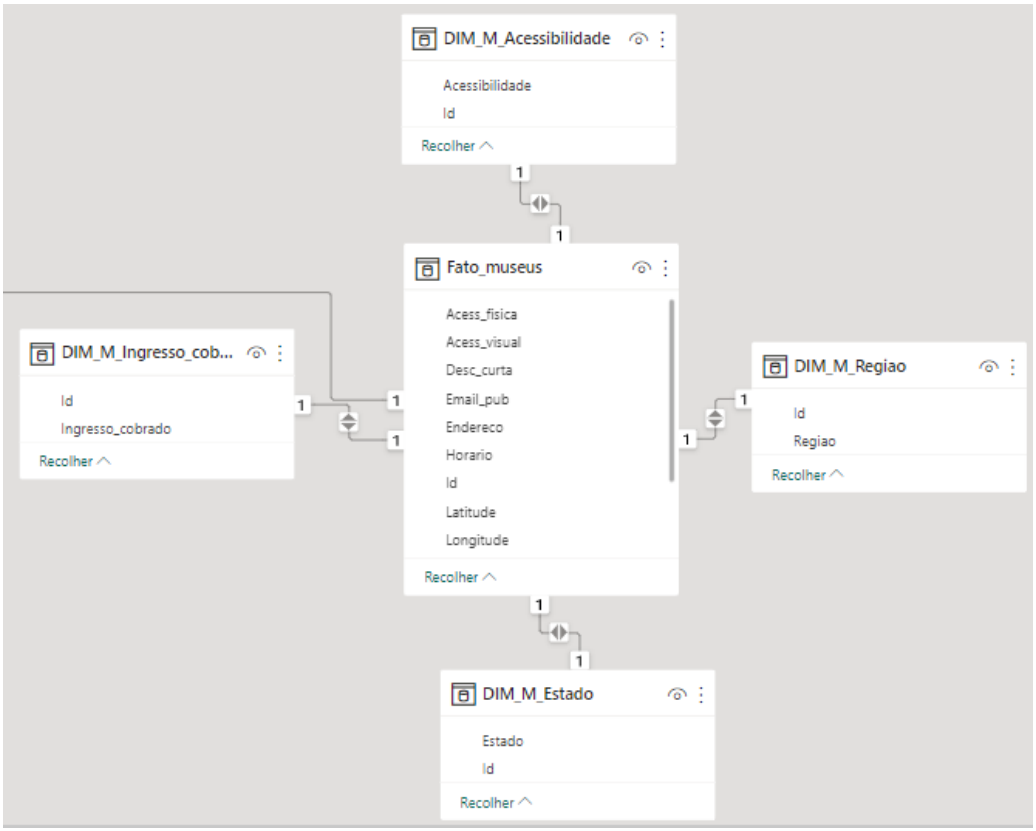

Figura 3: Relação museus

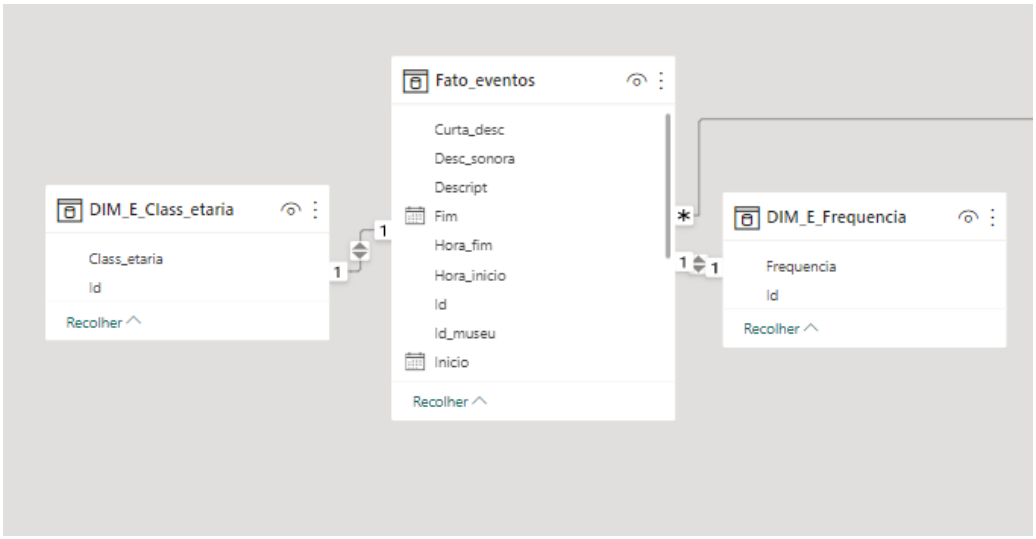

Figura 4: Relação eventos

6.5.3 Dashboard

Para acesso e manuseio do dashboard, basta acessar o link: *https://bit.ly/3JqHD4w*
Dito isso, segue abaixo as páginas do Dashboard e as questões solicitadas pelo cliente:

- Início do dashboard, onde possui os botões "Museus"e "Eventos"(há também botões na parte superior à esquerda). Ao clicar no botão "Museus", será redirecionado para a página que contém informações sobre os museus. E, ao clicar no botão "Eventos", será redirecionado para a página que contém informações sobre os eventos.

Figura 5: Dashboard - Início

- Página referente às informações descritivas das instituições culturais (Museus). Possui um visual com o nome dos museu que, ao colocar o cursor em cima, surge um pop-up com a descrição do museu, o horário de funcionamento e a informação quanto a obrigatoriedade de ingresso. Ao passar o cursor no visual que contém o endereço, surge um pop-up com as informações da região, estado e município do museu. No visual que contém a informação sobre a acessibilidade, ao colocar o cursor em cima, aparece um pop-up com informações sobre a acessibilidade física e visual oferecidas por aquele museu. Há visuais contendo o telefone, email e site do museu. Há também um visual que contém o nome dos museus em uma lista e uma barra de pesquisa, caso queira buscar um museu em específico. Por fim, existem quatro visuais exercendo a função de filtro para as pesquisas e são elas: Região, Estado, Ingresso cobrado e Acessibilidade. Abaixo dos filtros, há um mapa com a localização de todos os museus cadastrados.

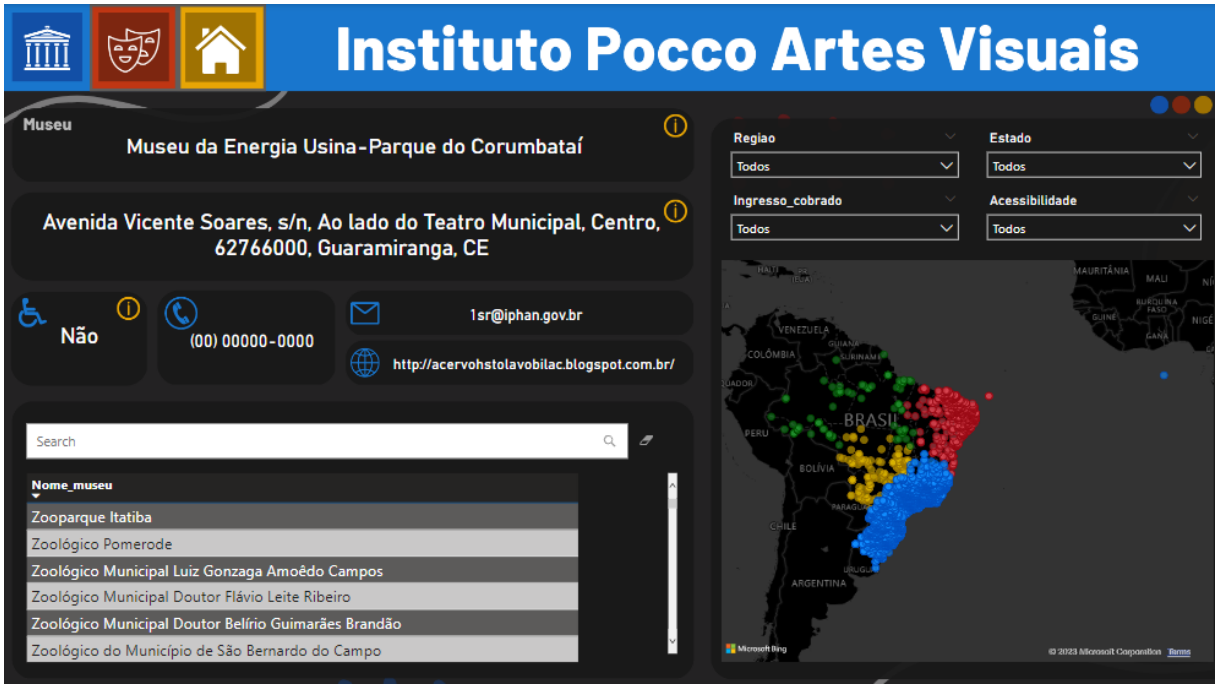

Figura 6: Dashboard - Museus

- Página referente às informações descritivas dos eventos. Esta página possui um visual com o nome do evento e, ao passar o cursor sobre o nome do evento, surge um pop-up com o nome do museu em que ocorre esse evento, um resumo do evento, o horário inicial e final, o preço do ingresso, o telefone e o site, além de observações sobre o evento. Abaixo há um visual contendo uma curta descrição sobre o evento. Há também um visual que contém o nome dos eventos em uma lista e uma barra de pesquisa, caso queira buscar um evento em específico. Possui um visual que contém as datas dos eventos, servindo como uma ordem cronológica e também como um mecanismo de filtragem para as consultas. Abaixo, existem quatro visuais exercendo a função de filtro para as pesquisas e são elas: Frequência, Descrição sonora, Classificação etária e Tradução em libras. E abaixo dos filtros, há também quatro visuais que contém informações sobre a frequência em que o evento ocorre, a faixa etária permitida para acesso ao evento, e se possui tradução em libras e descrição sonora para pessoas portadoras de deficiência.

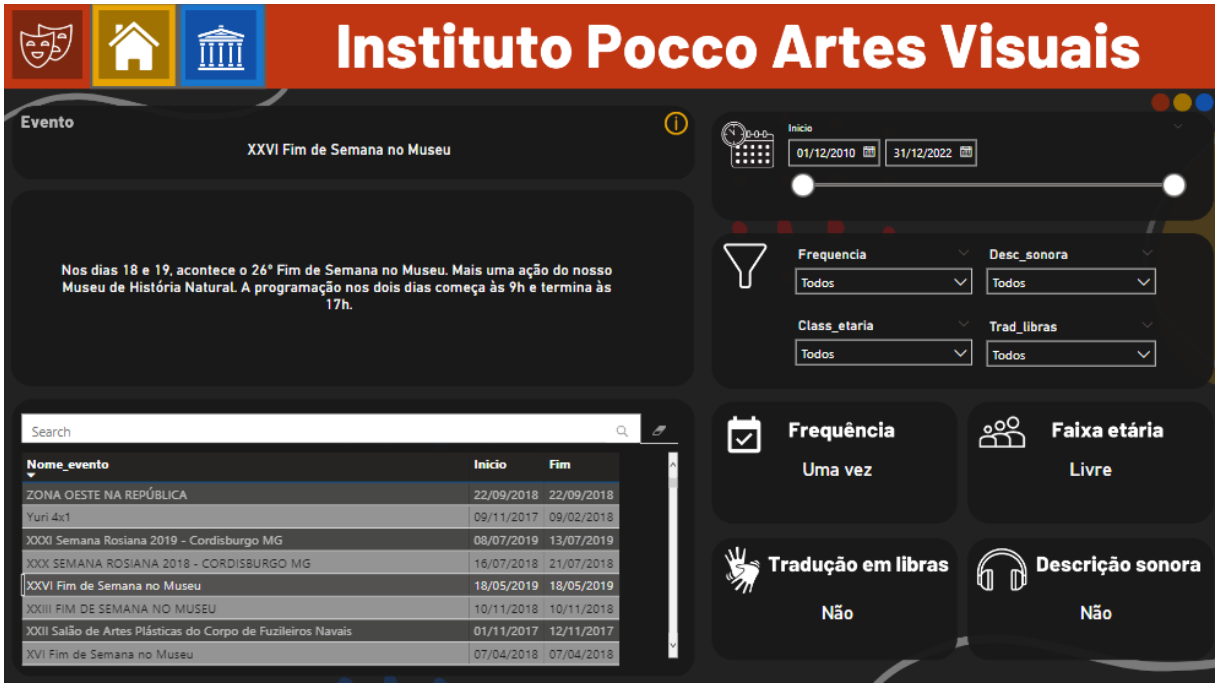

Figura 7: Dashboard - Eventos

