BlueShift

A BIG DATA COMPANY

São Paulo, 9 de agosto de 2022

CONTROLE DE VERSÃO			
Autor	**Versão**	**Data**	**Descrição**
Renan Da Silva Ramos	1.0	09/08/2022	Criação do documento

Sumário

Lista de Figuras

1 Introdução

Este documento visa detalhar tecnicamente as etapas utilizadas no cumprimento do projeto do cliente Museus Brasil, uma grande companhia ligada a todos os nossos museus em território nacional.

2 Solicitação

O cliente (Museus Brasil) deseja realizar um levantamento sobre algumas das principais informações referentes a cada museu localizado em território brasileiro. Para obter essas informações, o cliente firmou um contrato com a (Blueshift Brasil), onde ele disponibilizaria apenas uma (API) contendo todas as informações que os (Museus) parceiros forneceram. E então, a (Blushift Brasil) ficaria encarregada de organizar todas essas informações obtidas na (API) utilizando as melhores ferramentas e praticas do mercado, para então ao final do processo, dispor para o (Cliente) um (DashBoard e um Base de Dados final) , contendo informações específicas sobre o (Museu Especifico) e tambem informações sobre (Eventos) em que o museu está alocado.

3 Premissas da solução

Origem dos dados (API-JSON) :

- API Museus URL Base: http://museus.cultura.gov.br/api/
- API Museus endpoint com Museus: http://museus.cultura.gov.br/api/space
- API Museus endpoint com Eventos: http://museus.cultura.gov.br/api/event

- Exemplo de chamada dos 100 primeiros resultados de todas as colunas para os Mu- seus: http://museus.cultura.gov

- Parâmetros importantes:

Parâmetros Importantes	
@count	Retorna a quantidade de itens no JSON (linhas da query)
@select=	Parâmetro de seleção de colunas, usar separação por vírgula entre nome das colunas, "*"seleciona todos as colunas da tabela,usar "."para acessar colunas dentro de colunas-listas
@limit=	Limita a quantidade de itens que serão recebidos, útil quando usado em conjunto ao @page= ou @offset=
@page=	Usado em associação ao @limit, retorna a pagina requisitada tendo como base o numero de itens especificado no @limit=
@offset=	Similar ao page, mas o parâmetro deve ser dado em numero de itens à serem ignorados

Ambiente de desenvolvimento :

- **Microsoft Azure** *(Data Factory , Sql)*
- **Microsoft** *(Power - BI)*
- **Databricks** *(Pyspark)*
- **Jupyter Notebook** *(Pandas , Python)*

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.

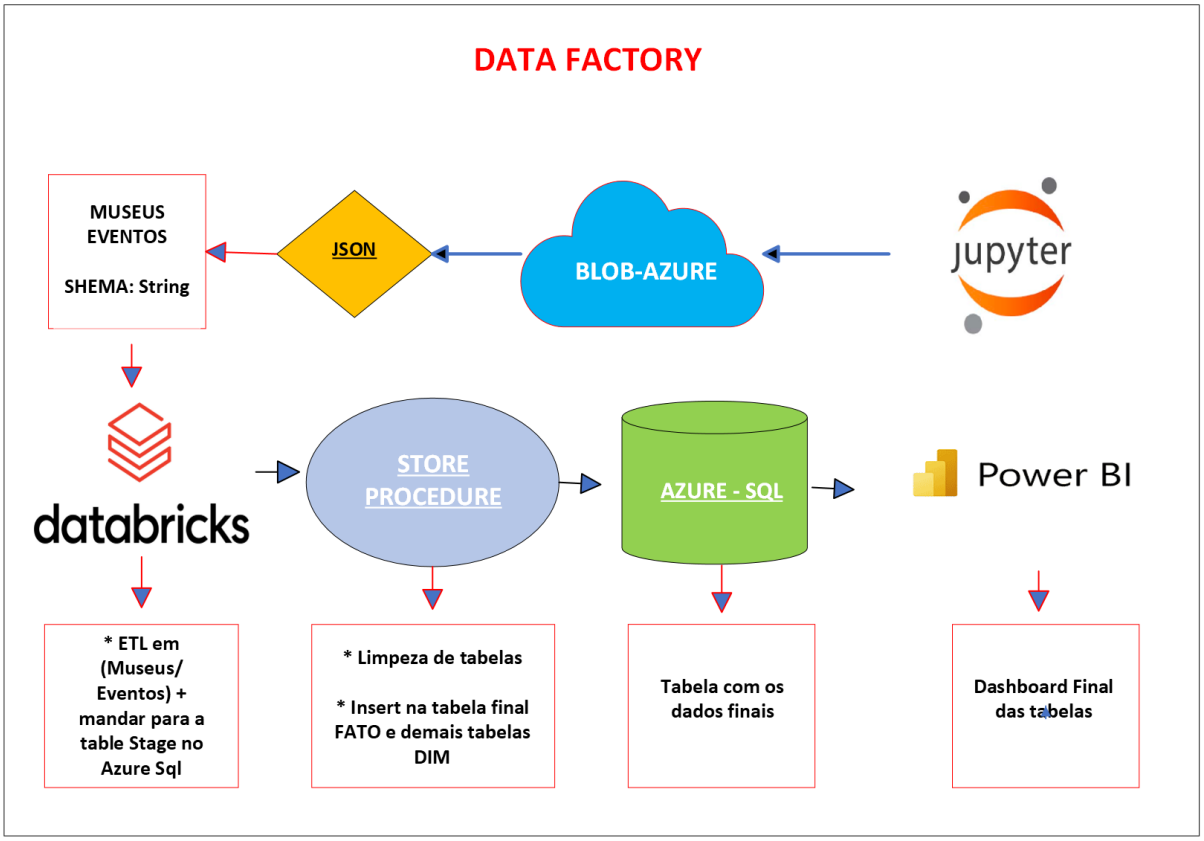


Figura 1: Arquitetura do projeto.

5 Dicionário de dados

Tabela Stage Museu

Campo	**Tipo**
Id	Int
Acessibilidade	varchar
Bairro	varchar
Cep	varchar
Descrição	varchar
Email	varchar
Endereço	varchar
Esfera	varchar
Estado	varchar
Horário	varchar
Latitude	float
Longitude	float
Museu	varchar
Região	varchar
Site	varchar
Temática	varchar

Tabela 1: Tabela Stage

Tabela DW Museu

Campo	**Tipo**
Id	int
Acessibilidade	varchar
Bairro	varchar
Cep	varchar
Descrição	varchar
Email	varchar
Endereço	varchar
Esfera	varchar
Estado	varchar
Horário	varchar
Latitude	float
Longitude	float
Museu	varchar
Região	varchar
Site	varchar
Temática	varchar

Tabela Stage Eventos :

Campo	**Tipo**
Id	int
Faixa$_{E}taria$	varchar
Evento	varchar
Descricao	varchar
Site	varchar
Telefone	varchar

Tabela DW Eventos :

Campo	**Tipo**
Id	int
Faixa$_{E}taria$	varchar
Evento	varchar
Descricao	varchar
Site	varchar
Telefone	varchar

Tabela DW Ocorrencias :

Campo	**Tipo**
Id	int
Frêquencia	varchar
Descricao	varchar
Duração	varchar
Preço	varchar
Hr-inicial	varchar
Hr-final	varchar
Data	Date
Id-evento	Int

6 Processo de desenvolvimento até etapa final

As atividades realizadas para desenvolvimento do projeto estão enumeradas a seguir.

1. Caderno do Jupyter Notebook (Obtendo API) :

   Figura-1

```
In [4]: museu = requests.get('http://museus.cultura.gov.br/api/space/findByEvents?@from=2016-05-01&@to=2016-05-
        m_json = json.loads(museu.content)
        dfm = pd.DataFrame(m_json)
```

   Figura-2

```
In [15]: dfm.info() # INFORMAÇÕES SOBRE A TABELA (MUSEUS)

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 5 entries, 0 to 4
         Columns: 134 entries, id to _relatedOpportunities
         dtypes: bool(1), int64(4), object(129)
         memory usage: 5.3+ KB
```

   Figura-3

```
#Enviando o arquivo "MUSEU-RENAN" para o blob azure

blob = BlobClient.from_connection_string(conn_str=string, container_name="renan-silva/lab4", blob_name=

with open("museu_re.json","rb") as data:

    blob.upload_blob(data, overwrite=True)
```

2. Caderno Databricks :

   Figura-1

```
1  # OBTENDO OS ARQUIVOS DO MUSEU EM JSON ==> 3 TABELAS Q ESTAO EM RENAN-SILVA/LAB4
2
3  museu = "dbfs:/mnt/renan-silva/lab4/museu_re.json"
4
5  eventos = "dbfs:/mnt/renan-silva/lab4/event_re.json"
6
7  ocorro = "dbfs:/mnt/renan-silva/lab4/ocoro_re.json"
```

   Figura-2

```
1  # CRIANDO DATAFRAME DOS ARQUIVOS EM JSON E LENDO NO SPARK
2
3  df_museu = spark.read.json(museu)
4
5  df_eventos = spark.read.json(eventos)
6
7  df_ocorro = spark.read.json(ocorro)
8
```

   Figura-3

	En_Bairro	En_CEP	En_Estado	En_Nome_Logradouro	En_Num	Formulário d
1	Centro	62766000	CE	Avenida Vicente Soares	s/n	não
2	Jardim Presidente	79015-350	MS	Rua Josefina Mingarelli	s/n	não
3	Ilha do Fundão	21949-909	RJ	UFRJ - Avenida Athos da Silveira Ramos	149	não
4	Centro	89820-000	SC	Rua Doutor José de Miranda Ramos	421	não
5	Vista Alegre	80820-200	PR	Rua Francisco Schaffer	s/n	não
6	Centro	95020-180	RS	Rua Visconde de Pelotas	586	não

   Figura-4

```
1  # ======================= > O QUE PRECISA ARRUMAR NA TABELA (MUSEU) ==>
2
3  #1) TROCAR OS NOMES DAS COLUNAS /
4  #2) TRANSFORMAR EM JSON (location: struct) E PEGAR AS CHAVES E TRANSFORMAR EM COLUNAS (LATITUDE) ; (LONGITUDE)
5  #3) AONDE ESTA NULL E VAZIO => COLOCAR (NÃO POSSUI)
6  #4) EXCLUIR AS COLUNAS ==> QUE NÃO SERÃO USADAS
7
```

Figura-5

```
# FAZENDO O 1°(ETL) DA TABELA (MUSEU) ==> RENOMEANDO COLUNAS / TRANSFORMANDO EM JSON(COLUNA) / EXCLUINDO COLUNAS DESNECESSARIAS / EXCLUINDO VALORES NULOS E VAZIOS
# CRIANDO TABELAS (LATITUDE) E (LONGITUDE)
dff_museu = df_museu.withColumnRenamed("id","Id")\
.withColumnRenamed("name","Museu")\
.withColumnRenamed("mus_tipo_tematica","Temática")\
.withColumnRenamed("site","Site")\
.withColumnRenamed("esfera","Esfera")\
.withColumnRenamed("geoEstado","Estado")\
.withColumnRenamed("emailPublico","Email")\
.withColumnRenamed("acessibilidade","Acessibilidade")\
.withColumnRenamed("En_Num","Numero")\
.withColumnRenamed("En_Nomo_Logradouro","Logradouro")\
.withColumnRenamed("En_CEP","Cep")\
.withColumnRenamed("En_Bairro","Bairro")\
.withColumnRenamed("endereco","Endereco")\
.withColumnRenamed("location","Localização")\
.withColumnRenamed("horario","Horario")\
.withColumnRenamed("shortDescription","Descricao")\
.withColumn("Localização",to_json(col("Localização")))\
.drop(*("En_Estado","Logradouro","Formulário de Visitação Anual - 2014","Formulário de Visitação Anual - 2015","Formulário de Visitação Anual - 2016"))\
.drop(*("mus_ingresso_valo","Formulário de Visitação Anual - 2017","Formulário de Visitação Anual - 2018","Museu Cadastrado"))\
.drop(*("Registro de Museus","createTimestamp","esfera_tipo","mus_ingresso_valo"))\
.withColumn('Site', when (col('Site').isNull(), lit('Não Possui') ).otherwise(col('site')))\
.withColumn('Descricao', when (col('Descricao').isNull(), lit('Não Possui') ).otherwise(col('Descricao')))\
.withColumn('Temática', when (col('Temática').isNull(), lit('Não Possui') ).otherwise(col('Temática')))\
.withColumn('Horario', when (col('Horario').isNull(), lit('Não Possui') ).otherwise(col('Horario')))\
.withColumn('Esfera', when (col('Esfera').isNull(), lit('Não Possui') ).otherwise(col('Esfera')))\
.withColumn('Museu', when (col('Museu').isNull(), lit('Não Possui') ).otherwise(col('Museu')))\
.withColumn('Bairro', when (col('Bairro').isNull(), lit('Não Possui') ).otherwise(col('Bairro')))\
.withColumn('Acessibilidade', when (col('Acessibilidade').isNull(), lit('Não Possui') ).otherwise(col('Acessibilidade')))\
.withColumn('Cep', when (col('Cep').isNull(), lit('Não Possui') ).otherwise(col('Cep')))\
```

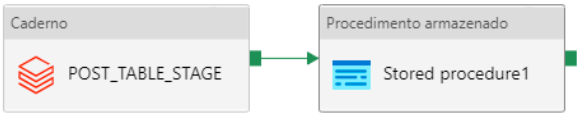
3. Procedimento Armazenado (SQL) :

   Figura-1

```
### CRIANDO PROCEDURE
CREATE PROCEDURE DW_renan_silva.sp_lab4

AS BEGIN

TRUNCATE TABLE DW_renan_silva.museu
TRUNCATE TABLE DW_renan_silva.eventos
TRUNCATE TABLE DW_renan_silva.ocorrencias

INSERT INTO DW_renan_silva.museu
SELECT Id ,Museu,Descricao,Bairro,Acessibilidade,Cep,Numero,Email,Endereco,Esfera,Estado,
Horario,Temática,Site,Longitude,Latitude,Região
FROM STAGE_renan_silva.museu

TRUNCATE TABLE STAGE_renan_silva.museu

INSERT INTO DW_renan_silva.eventos
SELECT Id,Faixa_Etaria,Evento,Descricao,Site,Telefone
FROM STAGE_renan_silva.eventos

TRUNCATE TABLE STAGE_renan_silva.eventos

INSERT INTO DW_renan_silva.ocorrencias
SELECT Id, Frêquencia,Descricao,Duração,Preço,Hr_inicial,Hr_final,
CAST(Data AS DATE) ,Id_evento

FROM STAGE_renan_silva.ocorrencias

TRUNCATE TABLE STAGE_renan_silva.ocorrencias

END
```

4. Data Factory (Pipile) :

   Figura-1



5. Dados na tabela final (DW) :

   Figura-1

```
select * from [DW_renan_silva].[museu]
```

Resultados | Mensagens

Pesquisar para filtrar itens...

Id	Museu	Descricao	Bairro	Acessibilidade	Cep	Numero
68	Memorial Juarez Barroso	O Memorial Juarez Barroso abri...	Centro	Não	62766000	Não Possui
6077	Parque Estadual Matas do Segr...	Este é outro remanescente de c...	Jardim Presidente	Não	79015-350	Não Possui
6078	Museu da Escola Politécnica da...	Museu da Escola Politécnica O ...	Ilha do Fundão	Não	21949-909	149

Power Bi :

Figura-1

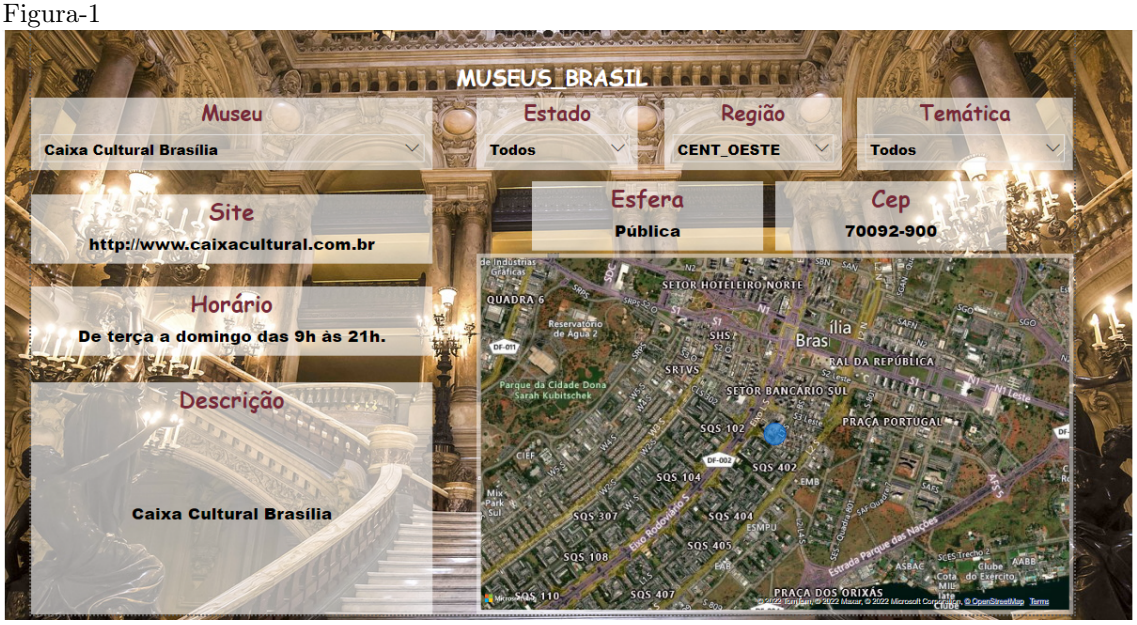


Figura-2