BlueShift

A BIG DATA COMPANY

Caieiras, 25 de agosto de 2022

CONTROLE DE VERSÃO			
Autor	**Versão**	**Data**	**Descrição**
Felipe Gustavo de Abreu	1.0	22/08/2022	Criação do documento

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Este documento visa detalhar as necessidades do cliente Instituto Pocco de Artes Visuais(IPAV), do ponto de vista técnico referente ao mapeamento de insituições, eventos e projetos no país, tendo como objetivo final o armazenamento dos dados e a visualização deles através de um Dashboard informativo contendo todas as instituições culturais do Brasil, separadas por estado e região, além dos principais eventos em uma linha do tempo.

2 Solicitação

O Instituto Pocco de Artes Visuais(IPAV) deseja uma solução para extração e processamento de dados para geração de um dashboard com informativo contendo todas as instituições culturais do Brasil, separadas por estado e região, além dos principais eventos em uma linha do tempo, e para isso eles necessitam de uma nova estruturação em sua pipeline de dados. A entrega do projeto deverá ser dividida em três partes. A primeira consistirá na extração dos dados de uma API via Databricks para o Azure SQL database. A segunda parte consistirá na construção de um Data Warehouse, via Stored Procedure dentro do Azure Data Factory. A terceira e última parte consistirá na cópia dos dados do Azure SQL database para um Dashboard construído no Power BI.

3 Premissas da solução

A seção abaixo apresenta as premissas da solução

Origem e especificação dos dados

- Os dados estão sendo disponibilizados por uma API pelo site MuseusBr do Governo Federal.

Ambiente de desenvolvimento

- O cliente deverá disponibilizar acessos aos ambientes de desenvolvimento em todas as ferramentas específicas na arquitetura proposta neste documento para a Blueshift 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.

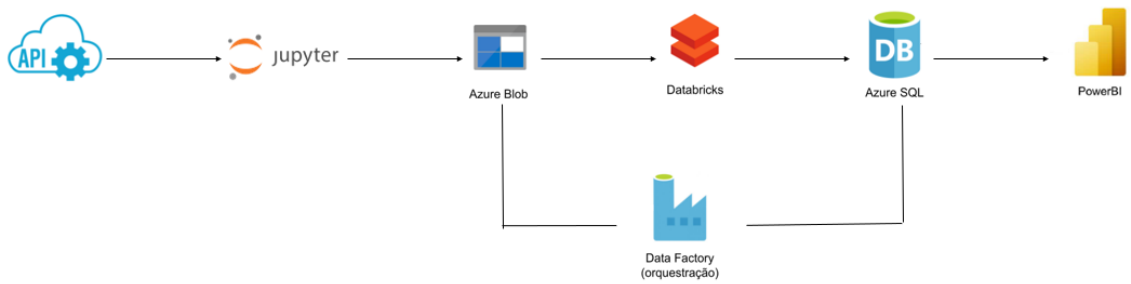

Figura 1: Arquitetura do projeto.

5 Extração dos dados da API

Serão extraidos os dados das APIs informados abaixo via Jupyter Notebook, na qual os arquivos serão armazenados os arquivos JSON no Blob Storage. Após será desenvolvido uma Pipeline no Azure Data Factory, e nele será inserido primeiramente um Notebook do Databricks, na qual será coletado os dados das ordens de vendas do arquivos JSON armazenados no Blob Storage, e a realização do ETL contendo a conversão e tratamento dos tipos dos dados, e após o consumo dos dados deste arquivo, essas informações serão inseridas no banco de dados criado no Azure SQL.

- API Espaços

  `https://museus.cultura.gov.br/api/space/find?@select=id,location,name,public,shortDescription,createTimesstamp,status,timezone_type,timezone,owner,endereco,En_Complemento,geoSetor_censitario,geoMicrorregiao,geoMunicipio,geoEstado,geoMesorregiao,En_CEP,En_Nome_Logradouro,En_Num,En_Complemento,En_Bairro,En_Municipio,En_Estado,esfera,esfera_tipo,type,acessibilidade,esfera,esfera_tipo,site,facebook,twitter,instagram,telefonePublico,capacidade,terms`

- API Eventos

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

- API Ocorrências

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

6 Modelagem dos Dados

Para este projeto estamos considerando duas tabelas de stage para armazenar os dados extraídos das APIs, e duas tabelas finais que irão receber os dados convertidos através da Storage Procedure. A estrutura das tabelas estão listadas abaixo.

- Tabela Museus

CAMPO	TIPO
Id_Museu	INT (PK) NOT NULL
Nome_Museu	VARCHAR(500) NULL
Bairro	VARCHAR(300) NULL
CEP	VARCHAR(50) NULL
Estado	VARCHAR(50) NULL
Regiao	VARCHAR(50) NULL
Municipio	VARCHAR(100) NULL
Logradouro	VARCHAR(50) NULL0
Latitude	FLOAT NULL
Longitude	FLOAT NULL
Acessibilidade	VARCHAR(50) NULL
Esfera	VARCHAR(300) NULL
Esfera_Tipo	VARCHAR(300) NULL
Descricao	VARCHAR(500) NULL
Instagram	VARCHAR(300) NULL
Site	VARCHAR(300) NULL
Telefone	VARCHAR(50) NULL

Tabela 1: Tabela Museus

- Tabela Eventos

CAMPO	TIPO
Id_Evento	INT (PK)
Evento	VARCHAR(500) NULL
Data_Criacao	DATE NULL
Classificacao	VARCHAR(200) NULL
Descricao	VARCHAR(MAX) NULL
Traducao_Libras	VARCHAR(50) NULL
Site	VARCHAR(300) NULL
Id_Museu	INT
Informacao	VARCHAR(500) NULL
Data_Inicio	DATE NULL
Data_Termino	DATE NULL
Horario	VARCHAR(50) NULL
Frequencia	VARCHAR(50) NULL
Preco	VARCHAR(300) NULL
Duracao	FLOAT NULL

Tabela 2: Tabela Eventos

7 Storage Procedure

Os dados extraídos dos arquivos armazenados no Blob Storage foram inseridos no banco de dados nas tabelas Stage já com a conversão dos dados feita no Databricks. Para realizar inserção dos dados nas tabelas do DW, será criado uma "STORED PROCEDURE" no Azure SQL Database, e introduzindo-a no Data Factory após o primeiro Copy Data. A Storage Procedure irá popular as tabelas DW conforme mostrado acima.

8 Dashboard no Power BI

A última etapa do projeto, será criado um Dashboard em Power BI, e será adicionado após o Stored Procedure na pipeline, sendo este Dashboard construído com as bases de dados do DW, através de uma conexão com o Azure SQL Server. O respectivo Dashboard deverá conter informações, sendo alguns deles:

- Mapa com as principais instituições por Região.
- Filtro e visualização de informações sobre cada instituição, contendo endereço, descrição, tipo da instituição, região em que a instituição está localizada.
- Linha do tempo com todos os eventos culturais que aconteceram e acontecerão nos próximos dias.

Com isto, o dashboard a ser desenvolvido irá apresentar todas as visões que atendam as necessidades apresentadas pela empresa.

