Caieiras, 7 de dezembro de 2022

CONTROLE DE VERSÃO			
Autor	**Versão**	**Data**	**Descrição**
Luiz Eduardo Sapori Gonçalves	1.0	22/11/2022	Criação do documento
Luiz Eduardo Sapori Gonçalves	1.1	29/11/2022	Atualização do documento
Luiz Eduardo Sapori Gonçalves	1.2	07/12/2022	Atualização do documento
Luiz Eduardo Sapori Gonçalves	1.3	07/12/2022	Atualização do documento

Sumário

Lista de Figuras

1 Introdução

Este documento visa detalhar do ponto de vista técnico as necessidades do projeto Instituto Pocco de Artes Visuais e as possíveis soluções, premissas e atividades a serem realizadas para a sua execução.

2 Solicitação

O Instituto Pocco de Artes Visuais (IPAV) é uma fundação que possui o objetivo de financiar iniciativas culturais no Brasil. Em parceria com a Blueshift eles desejam realizar o mapeamento de instituições, eventos e projetos no país, na finalidade de melhor alocarem seus recursos.

Espera-se ao final do projeto, a visualização em dashboard de todas as principais instituições culturais do Brasil, com suas geolocalizações, separadas por estados, regiões e seus respectivos eventos distribuídos em uma ordem cronólogica.

3 Premissas da solução

A partir da experiência dos técnicos Blueshift a solução proposta será apresentada nas seções a seguir.

3.1 Origem dos dados

Os dados são disponibilizados pelo governo brasileiro a partir da conexão com uma API no formato Json. Conforme suas funcionalidades, serão utilizadas para extração de dados brutos referente a museus, eventos e ocorrência dos eventos.

3.2 Ambiente de desenvolvimento

O cliente disponibilizará para o time da Blueshift Academy, acesso aos ambientes de desenvolvimento Azure Data Factory, Azure Blob Storage, Azure Databricks, Azure SQL e Microsoft Power BI, estando todos especificados a seguir:

1. Azure Data Factory: responsável por gerenciar o fluxo de trabalho, Api -> Blob Storage -> Databricks -> Banco de dados nas nuvens, através de uma pipeline (orquestração dos dados).

2. Azure Blob Storage: armazenará uma réplica da Api pública, dados não tratados, a ser utilizada no desenvolvimento do trabalho.

3. Azure Databricks: realizará a extração dos dados disponibilizados via Api, armazenado localmente e levá-los para um blob do Azure Storage Account seguindo uma estrutura de pastas pré-definida. A plataforma permitirá o tratamentos dos dados transformando-os em relacionais, logo, aptos para o carregamento no SQL.

4. Azure SQL: banco de dados nas nuvens, conforme a funcionalidade dividirá as informações em tabelas, alimentando o Microsoft Power BI.

5. Microsoft Power BI: modelagem dos dados relacionais. Disponibilizará através do Dashboard estatísticas pertinentes ao projeto.

4 Modelo da arquitetura sugerida

O Data Factory será responsável pela orquestração da pipeline, estruturando o processo de ETL (Extração - Tratamento - Loading / Carregamento) permitindo o tratamento dos dados da API, via Databricks e por fim, armazenamento em nuvem (SQL).

A figura 1 é uma representação geral da arquitetura proposta. Para o seu desenvolvimento considerou-se o levantamento de requisitos e entendimento do negócio.

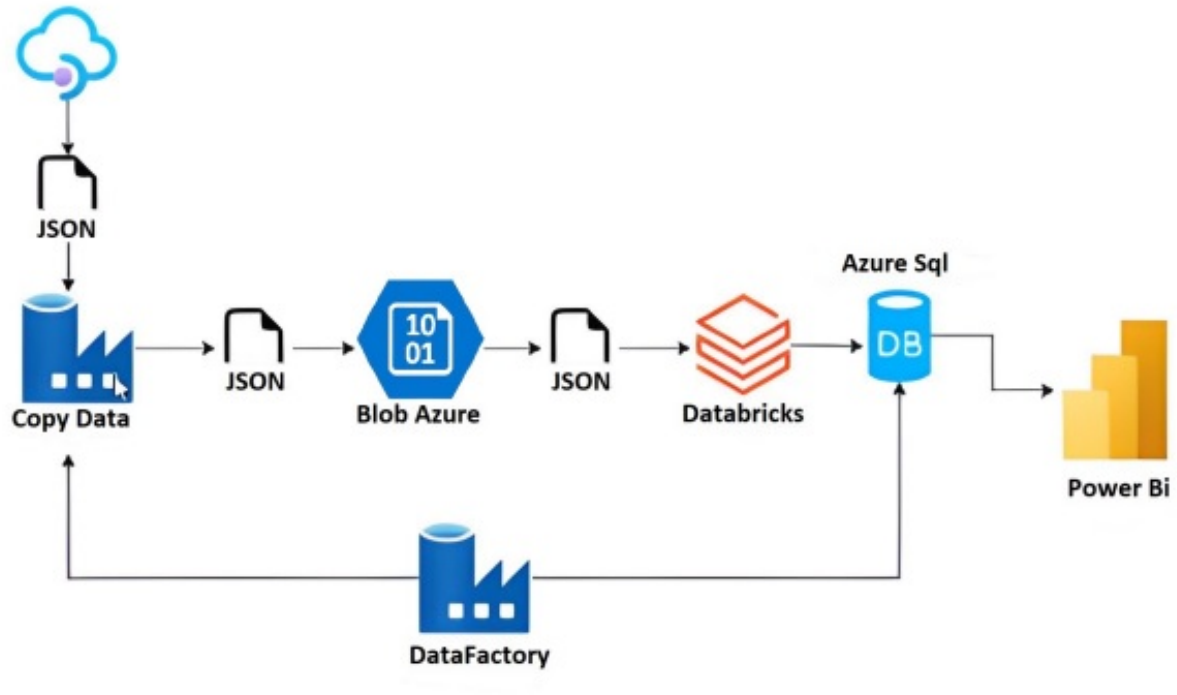

Figura 1: Arquitetura do projeto.

5 Dicionários de Dados

Para este projeto será utilizado um único Schema e conforme as aplicações, dois layouts principais de arquivos, (Museu e Eventos) que ao decorrer do processo alimentarão as tabelas, Sapori-Lab4.Sapori.Museus e Sapori-LAB4.Sapori.Eventos, no Azure SQL e nos dashboards do Power BI.

Nas tabelas 1 e 2, encontra-se o detalhamento dos dados necessários para aplicação da solução.

Campo	Descrição	Tipo	Restrição de domínio	Chave	Tamanho
En - Estado	Nome dos Estados	nvarchar	null	-	max
En - Municipio	Nome das cidades	nvarchar	null	-	max
En - Nome - Logradouro	Nome da rua	nvarchar	null	-	max
Esfera	referente ao ambiente se é público ou privado	nvarchar	null	-	max
Id	Identificador de registro	bigint	null	-	-
Name	Nome do Museu	nvarchar	null	-	max
ShortDescription	Descrição média do museu	nvarchar	null	-	max
TelefonePublico	Telefones dos Museus	nvarchar	null	-	max
Região	Estados divididos conforme regiões geográficas	nvarchar	null	-	max
Latitude	Medidas referente a distancia do globo terrestre	nvarchart	null	-	-
Longitude	Medidas referente a distancia do globo terrestre	nvarchar	-	-	

Tabela 1: Tabela de Sapori-Lab4.Sapori.Museus

Campo	Descrição	Tipo	Restrição de domínio	Chave	Tamanho
Occurrences - Id	Registro numérico de ocorrências	float	null	-	-
Duration	Duração em minutos dos eventos	float	null	-	-
Frequency	Frequência dos eventos	nvarchar	null	-	max
Price	Preço da entrada dos eventos	nvarcharr	null	-	max
StartsAt	Horário de inicio dos evento	nvarchar	null	-	max
StartsOn	Data de inicio dos eventos	nvarchar	null	-	max
ClassificacaoEtaria	Faixa Etária indicada para acesso aos eventos	nvarchar	null	-	max
Id	Registro numérico do evento	bigint	null	-	-
LongDescription	Descrição detalhada dos eventos	nvarcharr	null	-	max
Name	Nome dos eventos	nvarchar	null	-	max
ShortDescription	Breve descrição dos eventos	nvarcharr	null	-	max
Subtitle	Tema dos eventos	nvarchar	null	-	-
TelefonePublico	Telefone de contato dos eventos	nvarchar	null	-	max
TraducaoLibras	Eventos possuem interprete em Libras	nvarchar	null	-	max

Tabela 2: Tabela de Sapori-Lab4.Sapori.Eventos

6 Desenvolvimento dos dados

Responsável pela orquestração dos dados através do Data Factory, a pipeline será composta por três Copy Data e dois notebook do Databricks, **figura 2**, as qual desempenhará as seguintes atividades:

1. O acesso as informações brutas referentes a museus, eventos e ocorrêcias dos eventos, acontecerá através do Copy Data que acessará uma API pública do governo brasileiro através da url **museus.cultura.gov.br** e criará uma réplica dos dados a serem armazenada em um blob contêiner;

2. Os notebooks, conectados ao Databricks, realizarão a extração do blob, tratamento e carregamento no Azure SQL dividindo os dados higienizados (pós-tratamento) em duas tabelas, museus e eventos;

3. As tabelas seguem para ingestão do Power BI sendo modeladas no intuito de obtermos estatísticas relevantes ao projeto no formato de dashboards.

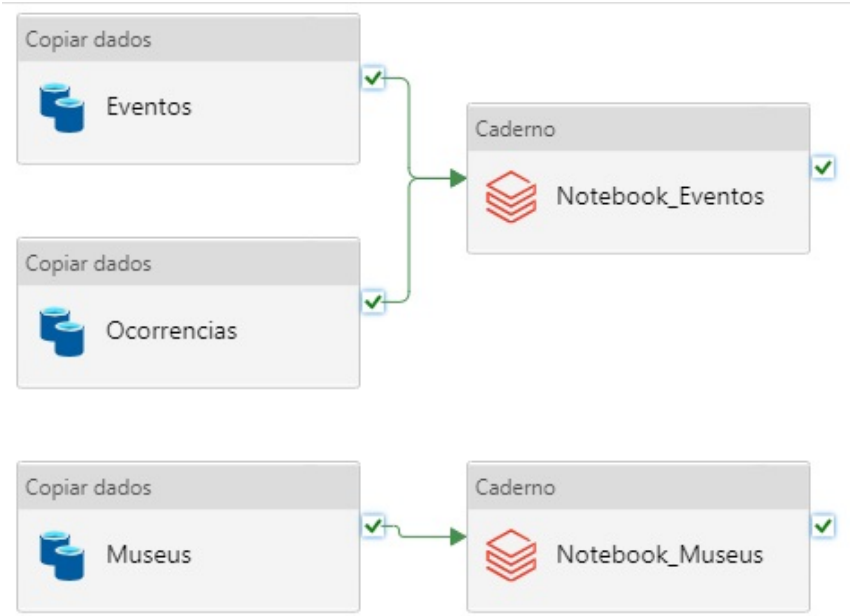

Figura 2: Pipeline do projeto contruída via Data Factory.

7 Power BI - Dashboard

Para este projeto será desenvolvido um dashboards com três páginas, uma capa de abertura e dois painéis consolidados, no Power BI que consumirá os dados disponibilizados nas tabelas museus e eventos do Azure SQL.

Capa - Instituto Pocco de Artes Visuais (IPAV)

A capa terá a função de introduzir de modo ilustrativo os paineís consolidados a serem apresentados. Optou-se por uma representação que configure os museus e as suas iniciativas culturais no Brasil, conforme **figura 3**.

Figura 3: Capa - Instituto Pocco de Artes Visuais (IPAV)

Painel Consolidado - Museus

A visualização do painel consolidado Museus, é alimentado pelos dados da tabela Sapori-Lab4.Sapori.Museus do Azure SQL, e retrará as principais intituições brasileiras, a distribuição geográfica e suas características. A **figura 4** representa o painel e as estatísticas extraídas.

Figura 4: Dashboard Museus - Power BI.

Painel Consolidado - Eventos

A visualização do painel consolidado Eventos, é alimentado pelos dados da tabela Sapori-Lab4.Sapori.Museus do Azure SQL, e retrará os principais eventos culturais que ocorrem nos museus brasileiros, em diferentes datas e suas características. A **figura 5** representa o painel e as estatísticas extraídas.

Figura 5: Dashboard Eventos - Power BI.

O Dashboard completo está disponível no link **https://app.powerbi.com/ groups/me/reports/ abf4bba2-e121-48c5-a4fb-c6eb1b10b083/ReportSection6f672cfbca0d993eb022**.

