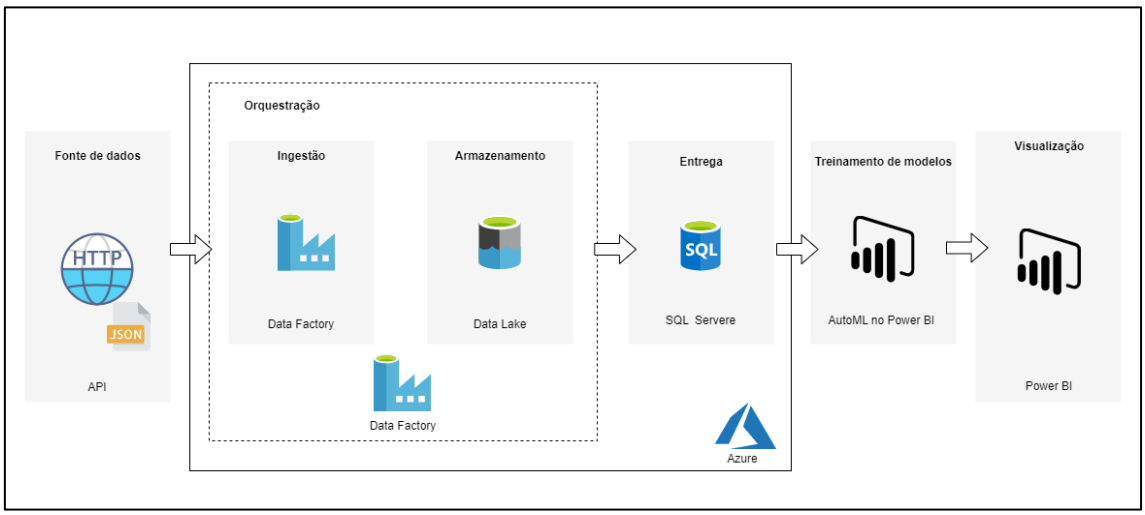
3 PRÉ-PROCESSAMENTO DE DADOS

De acordo com Corrêa (2020), é na fase de pré-processamento de dados que são realizadas as tarefas de seleção (coletar e reunir todos os dados que sejam relevantes para a resolução do problema de ciência de dados definido), limpeza (eliminar sujeira e informações irrelevantes) e transformação (converter os dados de origem para um outro formato, mais adequado) dos dados que serão utilizados.

3.1 Arquitetura

Para adequação do projeto, será utilizada a seguinte arquitetura (Figura 2):

FIGURA 2 - Arquitetura



Fonte: Elaborado pelos autores

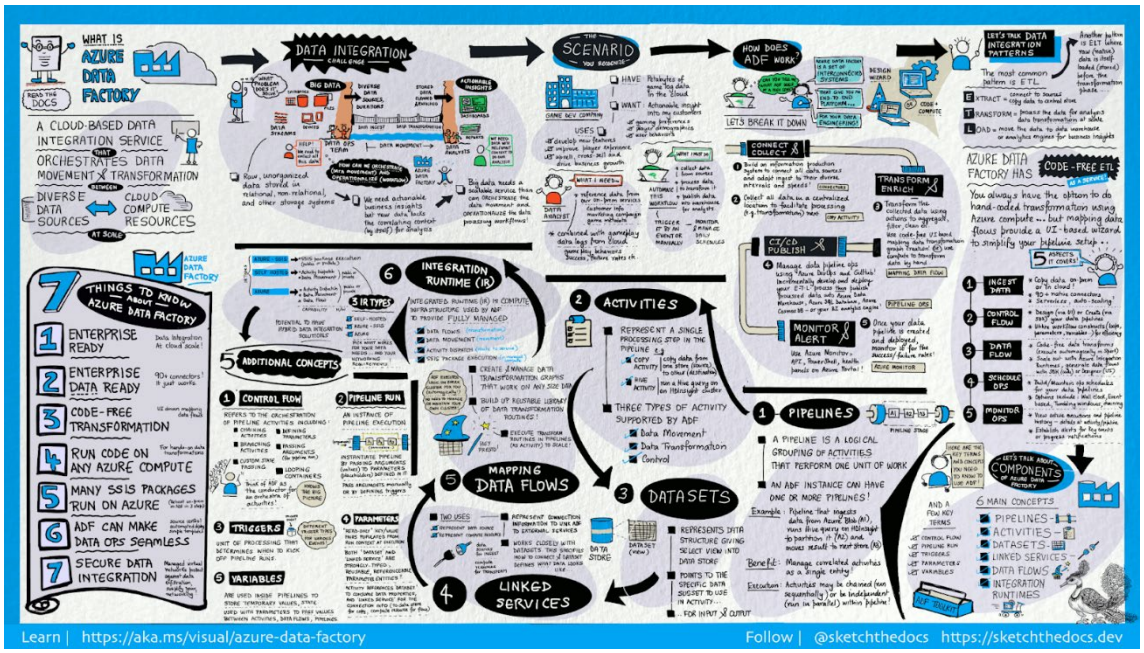
A arquitetura proposta tem em sua fonte os dados do openDataSUS sobre a vacinação contra a COVID-19 (disponíveis via API). A ingestão dos dados foi realizada no Data Factory diretamente a partir da API “Registros de vacinação COVID-19”. Foram utilizados a priori 10 mil registros, já que não foi possível retornar um número de registros maior que 10 mil devido a falhas ainda não solucionadas referentes ao parâmetro “?scroll=1m” e a chave "_scroll_id" utilizada para a requisição de um novo conjunto de elementos. A ferramenta

utilizada para o armazenamento é o Data Lake. A carga dos dados se dará no banco de dados Azure SQL Server. As operações de aprendizado de máquina serão realizadas no AutoML (machine learning automatizado) no Power BI. A visualização dos dados será disponibilizada a partir do Microsoft Power BI.

3.2 Escolhas de ferramentas

Optamos pelo uso do Azure Data Factory, um serviço de nuvem gerenciado que atende as necessidades de projetos de Big Data com o objetivo de “orquestrar e operacionalizar processos para refinar esses enormes repositórios de dados brutos em insights de negócio acionáveis”. (MICROSOFT, 2023a). Uma visão geral da arquitetura completa do Data Factory pode ser observada na Figura 3 a seguir.

Figura 3 - Azure Data Factory: guia visual



Fonte: MICROSOFT, 2023

Para o desenvolvimento e treinamento de modelos usaremos o AutoML diretamente no Power BI que é uma das ferramentas de Business Intelligence mais utilizadas no mercado atualmente. O AutoML dá suporte à criação de Modelos de Previsão Binária, Classificação e Regressão em fluxos de dados. (Figura 4).

Figura 4 - Ciclo de vida do projeto de aprendizado de máquina

Fonte: Elaborado pelos autores a partir do Microsoft Power BI