6 OTIMIZAÇÃO

Em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709 de 2018 - que regula as atividades de tratamento de dados pessoais - trabalhamos com dados abertos para consumo disponíveis no “openDataSUS”. Os dados utilizados foram registros de vacinação contra Covid-19, sem identificação do cidadão (anonimizados), contidos na Rede Nacional de Dados em Saúde (RNDS).

6.1 Otimizações adotadas ao longo do projeto

Ao observar falhas não solucionadas na API para consumo, realizamos a ingestão dos dados diretamente a partir dos “Registros de Vacinação COVID19 - Dados Completos” (disponíveis em CSV). Foram utilizados a priori a parte 1 dos dados completos disponíveis na “Campanha Nacional de Vacinação contra Covid-19”. Devido ao grande volume de dados (20 arquivos CSV no total, cada arquivo com aproximadamente 26 milhões de registros) e as restrições e limitações de crédito da conta *Azure for Students* utilizada, optamos - para adequação do projeto - por utilizar uma amostra de 10 mil registros, já que não foi possível retornar um número de registros maior que 10 mil devido a falhas referentes ao parâmetro “?scroll=1m” e a chave "_scroll_id" utilizada pela API para a requisição de um novo conjunto de elementos.

6.2 Propostas futuras para o projeto desenvolvido

Dentre as propostas futuras para o projeto desenvolvido entendemos que as seguintes soluções apresentadas abaixo possam mitigar alguns problemas observados e a sua adoção contribua para a otimização do projeto.

- **Problemas relacionados a API:** enfrentamos alguns problemas relacionados ao consumo de dados via API da Campanha Nacional de Vacinação da Covid-19 disponibilizada pelo Ministério da Saúde, que não retornava corretamente todas as informações de acordo com os parâmetros informados na documentação (Manual para Consumo da API). **Solução:** automatizar a coleta de dados. Criar uma rotina em *Python* para extração das bases de dados que estão disponíveis em CSVs (dados completos divididos em 20 arquivos), e essa rotina deverá acessar cada um dos CSV de forma automática e armazenar os dados no nosso *Data Lake*, como uma espécie de Robô.

- **Problemas relacionados a volumetria:** volume de dados acima do esperado. Usamos um banco de dados *SQL Server* padrão como camada de entrega e o *Azure Data Factory* como ferramenta de ingestão e processamento dos dados. Solução: **Solução:** Usar um *framework* para Big Data, como *Apache Spark*, e a feature PySpark (uma interface *Python* para *Spark*) onde seja possível escalar o processamento de toda a volumetria dos dados de modo performático, pois os dados utilizados equivalem a 20 arquivos CSVs de 11 GB de tamanho, com 26

milhões de registros, o que totaliza 110 GB de volume armazenado e aproximadamente 520 milhões de registros.

- **Problemas relacionados ao Banco de dados do SQL Server:** *Serverless* com base em DTU (Unidade de Transação do Banco de Dados) não seria suficiente para servir como camada de entrega de dados para o *Microsoft Power BI*. **Solução:** usar um banco de dados analítico de processamento distribuído, como *Azure Sinapse Analytics, Google Big Query, Amazon Redshift ou Snow Flake* para processar e entregar os dados com performance para serem consumidos pelo *Microsoft Power BI*.

- **Problemas relacionados a granularidade:** a alta volumetria quando considerada toda a granularidade no nível das doses e do paciente, poderia fazer com que o *Power BI* perdesse performance. **Solução:** usar uma solução de OLAP que proporcione condições de análise de dados on-line (por exemplo o *Azure Analysis Services*) e deixar os dados pré-processados em um cubo, onde o *Power BI* possa acessar e já com todas a médias criadas e relacionamentos, usar a técnica de MOLAP (*Multidimensional Online Analytical Processing*), a partir de processos de: *Drill Down, Drill UP, Drill Across, Drill Throught e Slice and Dice* para navegar de modo performático. Uma segunda solução seria utilizar o conceito de Fato agregada para diminuir a volumetria na modelagem dos dados sem a necessidade de ter uma ferramenta adicional de OLAP.