ITI/CNPq-MAI/DAI

Criação de um data lake com base em indicadores de dados abertos para uso em projetos de inovação social

Autores: Eduardo Eberhardt Pereira, Bianca Libardi, Ana Cristina Fachinelli Bertolini e Daniel Luis Notari

City Living Lab

INTRODUÇÃO / OBJETIVO

Um *data lake* é um repositório de dados centralizado que permite o armazenamento de uma multitude de dados, abrangendo dados estruturados, semi-estruturados e até não estruturados (SINGH; AHMAD, 2019).

Dados abertos são dados que podem ser usados, estudados e modificados sem restrições, e que podem ser copiados e redistribuídos com ou sem modificações igualmente sem restrição, ou com restrições que garantem que os recebedores seguintes também possam fazer todas essas ações (MURRAY-RUST, 2008).

Extract-Transform-Load (ETL) é o processo de busca, coleta, processamento e carregamento dos dados de forma separada. O processo de ETL garante qualidade e consistência dos dados, permite unir dados de diferentes fontes e entrega os dados em um formato pronto para apresentação em uma aplicação ou dashboard, permitindo que os usuários finais consigam usá-los para tomada de decisão (KIMBALL; CASERTA, 2004).

O objetivo da bolsa é auxiliar no desenvolvimento de um dashboard que permita a visualização do desenvolvimento das cidades brasileiras em indicadores de cidades inteligentes, seguindo os modelos dos Objetivos de Desenvolvimento Sustentável (ODS) (ONU, 2015) e da ABNT NBR ISO 37122 - Cidades e Comunidades Sustentáveis - Indicadores Para Cidades Inteligentes (ABNT, 2020). Para isso, a etapa inicial foca no uso de dados abertos, coletados e processados usando os métodos ETL para a criação de um *data lake*, o qual será em fase posterior usado para o desenvolvimento do dashboard.

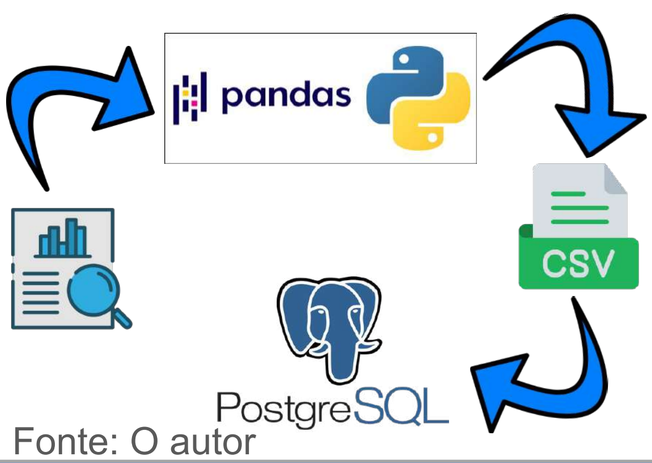
MATERIAL E MÉTODOS

Após análise dos documentos dos ODS e da ISO 37122, uma busca por dados abertos disponíveis na internet pelo governo federal brasileiro que fossem compatíveis com os indicadores definidos foi feita. Os dados eram baixados e organizados de forma a manter a fonte dos dados, o que eles significam, e a qual indicador pertencem.

Após a coleta dos dados abertos referentes a indicadores que poderiam ser atribuídos ao objetivo da pesquisa, o processamento desses dados foi feito utilizando bibliotecas de ciência de dados da linguagem de programação Python, como Pandas.

Em seguida, esses dados tratados foram importados para o Postgres, um banco de dados relacional. A imagem 1 demonstra visualmente o processo realizado.

Imagem 1 - Ferramentas do processo ETL



Fonte: O autor

RESULTADOS

Os comandos da biblioteca Pandas usados durante o processo variam bastante, pois cada conjunto de dados precisa ser tratado de maneira específica. Um exemplo de comando frequente, no entanto, é o “read_csv(‘path.csv)”, que lê um arquivo CSV e transfere os dados nele contidos para uma tabela dentro do programa, chamada de *dataframe*.

Diversos outros comandos e recursos são utilizados em diversos outros contextos, sendo que os comandos utilizados devem ser adequados à cada fonte de dados, que diferenciam-se entre si em diversas formas.

Já os comandos em SQL (*Structured Query Language*) são sempre os mesmos, visto que o trabalho feito em pandas sempre gera um arquivo CSV único que contém todos os dados. Os comandos se resumem a “CREATE TABLE x(item varchar)”, que cria uma nova tabela, inserindo os detalhes de cada coluna, e “COPY x FROM ‘path’”, que passa os dados do arquivo criado com pandas para o banco de dados.

O resultado parcial é um *data lake* com dados abertos, coletados visando coletar informações para gerar indicadores úteis para pesquisas posteriores. Os resultados podem ser visualizados na tabela 1:

Tabela 1 - Estatísticas sobre os dados

Número de indicadores	66
Número de Estudos (Fontes)	24
Número de Arquivos de Entrada	546
Número de Colunas de Entrada	14139
Número de tabelas geradas	37
Número de colunas geradas	1110

Fonte: O autor

CONSIDERAÇÕES FINAIS

Após a conclusão da criação do *data lake*, seguirá uma fase de análise dos indicadores armazenados e sua aderência ao projeto. Nesta fase, os indicadores serão avaliados e sua importância para o projeto será determinada.

REFERÊNCIAS BIBLIOGRÁFICAS

SINGH, A.; AHMAD, S. Architecture of Data Lake. I**nternational Journal of Scientific Research in Computer Science Engineering and Information Technology**, mai. 2019. V. 5, p. 411-414.

MURRAY-RUST, P. Open Data in Science. **Nat Prec, [S.l.]**, v. 2008, n. 1526, 2008. Disponível em: https://doi.org/10.1038/npre.2008.1526.1. Acesso em: 7 mai. 2024.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data**. Indianapolis: Wiley Publishing, Inc., 2004.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **ABNT NBR ISO 37122:** Cidades e Comunidades Sustentáveis - Indicadores Para Cidades Inteligentes. ABNT, 2020.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). **Agenda 2030 para o Desenvolvimento Sustentável: Transformando Nosso Mundo**. Nova York: ONU, 2015