XXXII Encontro de Jovens Pesquisadores e XIV Mostra Acadêmica de Inovação e Tecnologia UCS

# CRIAÇÃO DE UM DATA LAKE COM BASE EM INDICADORES DE DADOS ABERTOS PARA USO EM PROJETOS DE INOVAÇÃO SOCIAL

Eduardo Eberhardt Pereira (ITI/CNPq-MAI/DAI), Bianca Libardi. Daniel Luis Notari., Ana Cristina Fachinelli Bertolini (Orientador(a))

A análise de dados é uma das áreas de estudo mais utilizadas atualmente devido ao grande volume e à diversidade de dados disponíveis. Um *data lake* é uma das formas de organizar esse volume de dados, sendo definido como um repositório centralizado que armazena dados estruturados, semiestruturados e não estruturados. Uma das formas de dados variados é o uso de dados abertos disponibilizados por órgãos públicos, que podem ser usados, estudados, modificados e redistribuídos sem restrições. O presente trabalho explora a criação de um *data lake* a partir do uso de dados abertos, visando futuramente o desenvolvimento de um dashboard com informações sobre as cidades brasileiras sob a temática de cidades inteligentes. Inicialmente, foram analisados documentos da ISO 37122 para identificar os indicadores relevantes. Com os indicadores definidos, foi realizada uma busca por dados abertos em diversos órgãos públicos. Essas fontes de dados foram organizadas em tabelas de um Sistema Gerenciador de Banco de Dados Relacional (SGBDR). Cada tabela contém a origem da fonte de dados, seu significado e a quais indicadores esses dados correspondem. Em seguida, os dados passaram pelo processo de extração, transformação e carga (ETL), que garante a qualidade e consistência dos dados. Mais precisamente, os dados foram baixados da internet e organizados dentro de uma estrutura de pastas em um disco rígido (HD). Posteriormente, a transformação e o tratamento dos dados foram realizados utilizando a linguagem de programação Python e a biblioteca Pandas. Cada conjunto de dados gerou um programa diferente, uma vez que os dados vinham de 24 fontes diferentes e em formatos variados. O comando “read_csv('path.csv')” foi utilizado para processar os arquivos do tipo CSV (dados separados por ponto e vírgula) e exibi-los em um *dataframe*. Após o tratamento adequado utilizando o Pandas, os dados eram convertidos para um único arquivo CSV, que por sua vez era carregado no SGBDR PostgreSQL. Esse tipo de software utiliza uma linguagem padrão chamada SQL (Structured Query Language). Foi utilizado o comando “CREATE TABLE x(item varchar)” para criar as tabelas na estrutura do data lake e o comando “COPY x FROM 'path'” para realizar o carregamento dos dados do arquivo CSV para as tabelas. O resultado parcial foi a criação de um data lake com indicadores formados por dados abertos, os quais serão úteis para pesquisas posteriores referentes à inovação social de cidades inteligentes em cada município.

Palavras-chave: Data lake, Indicadores, Inovação social

Apoio: UCS, CNPq