*Avaliação de diferentes estratégias na engenharia de dados para a construção de um data warehouse com dados públicos voltado para cientistas sociais.*

Com o aumento acelerado de produção de dados e o amplo uso dentro dos mais diversificados cenários e objetivos, a engenharia de dados se destaca no sentido de manter e criar uma base sólida de dados, que então poderão ser consumidos, seja por cientistas de dados ou engenheiros de aprendizado de máquina ou analistas de dados. O engenheiro de dados tem entre suas atribuições a construção e manutenção de arquiteturas direcionadas a dados, tais como banco de dados e sistemas de processamento em grande escala. Dentro da gama de possibilidades, a ETL 一 do inglês, *Extract, Transform, Load* 一 ou extração, transformação e carregamento é uma das etapas que merece destaque.

Nessa etapa, a **extração** corresponde ao agrupamento das mais diversas fontes de dados. Nela os dados são recuperados em sua forma crua, seja em bancos relacionais, planilhas, arquivos, ou outras formas de armazenamento de dados. Por sua vez, a **transformação** visa tornar os dados aptos para a utilização. Assim, são utilizadas técnicas de limpeza, classificação 一 categóricos e contínuos 一, e normalização sobre os dados. Por fim, o **carregamento** dos dados dá um local de acesso ao consumidor dos dados (por exemplo, o cientista de dados). Exemplos de locais de acesso são API – Application Programming Interface – e *data warehouse.*.

Um *data warehouse* é um banco de dados estruturado voltado para consultas avançadas, com grande quantidade de dados históricos e direcionado para suportar análises. Nele são centralizados uma grande quantidade de dados advindos de diferentes fontes que passaram pelo processo de ETL.

Com relação ao ETL, há uma variação, o ELT (*Extract, Load, Transform*), onde inverte-se a ordem de execução da **carga dos dados (***Load***)** com a transformação *(Transform)*. Nesse modelo, os dados primeiro são salvos em algum lugar para o uso e depois ocorre a transformação e tratamento dos dados. Em nosso estudo, nos ocuparemos de criar um pipeline de dados voltado para o primeiro tipo, ETL, não sendo então tratadas diretamente maneiras de trabalhar em um processo de ELT de dados.

Sendo uma etapa tão importante e que tem uma massa de dados que pode chegar à casa de terabytes, foram surgindo diferentes ferramentas que auxiliassem no processo de ETL. Assim, temos desde ferramentas *open-source* a ferramentas proprietárias, mas todas voltadas para maximizar e melhorar as informações que são angariadas. Visando o uso de ferramentas open-source para a execução do ETL, há uma ausência de estudos que comparem as diferentes soluções existentes, seu prós e contras, bem como quais cenários cada ferramenta é indicada.

**Objetivo**

Este trabalho visa investigar a utilização de ferramentas *open-source* para a realização das etapas do ETL e armazenamento em data warehouse. Como prova de conceito, avaliaremos tais soluções em bases de dados utilizadas por estudantes e pesquisadores da área de ciências sociais.

**Metodologia**

Primeiramente, optamos por utilizar o Pentaho – suite completa de Business Intelligence e recursos de geração de relatórios, integração e armazenamento de dados, ETL, análise de informações, painéis para controle gerencial e mineração de dados – , que já é uma ferramenta bem conhecida no mercado e que vem sendo desenvolvida desde de 2004, contendo já muitos trabalhos que fundamentam sua utilização Caminhando em direção a linguagem em si, utilizaremos o Python para criar e automatizar o processo. Vale ressaltar que Python é hoje a segunda linguagem de programação mais utilizada no mundo e amplamente empregada na ciência de dados, graças a uma comunidade extremamente ativa que trabalha no desenvolvimento de suas bibliotecas.

Com relação aos cenários em que os processos de ETL serão empregados, observamos que as ciências sociais têm problemas para encontrar dados abertos que facilitem a análise de situações sociais. Assim, iremos trabalhar em três eixos básicos, mas que possibilitam englobar diversos temas dentro das ciências humanas. Assim, buscamos desconstruir uma barreira entre humanas e exatas que dificulta, em muitos aspectos, a interdisciplinaridade entre as áreas de conhecimento.

Trabalharemos a princípio com os eixos de política e economia. Vale observar que os eixos estão conectados e são interdependentes em algumas análises. Refinando os processos e provando sua viabilidade, poderemos então abrir o escopo e, assim, possibilitar a expansão para mais áreas como saúde, educação, população e outras possíveis.