Extração, Transformação e Carregamento de Dados: 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 pelo cientista de dados, engenheiro de aprendizado de máquina ou analista de dados. O engenheiro de dados tem entre seus atributos a construção e manutenção de arquiteturas direcionadas a dados, tais como banco de dados e sistemas de processamento de grande escala. Dentro da gama de possibilidades, a ETL 一 Extract, Transform, Load 一 ou extração, transformação e carregamento é uma das etapas que merece destaque.

Nessa etapa, a extração é o agrupamento das mais diversas fontes de dados. Nela os dados são recuperados em sua forma crua, seja em bancos relacionais ou não, planilhas, arquivos, e as mais diferentes formas de informações. Por sua vez, a transformação se ocupa em 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. Uma API, data warehouse, um repositório onde seja possível o acesso por qualquer usuário interessado. Com relação ao ETL há uma variação, o ELT, onde inverte-se a carga dos dados com a transformação. 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. Como premissa, utilizaremos apenas ferramentas opensource, permitindo então a replicabilidade sem custos da aquisição ou assinatura de softwares.

Assim sendo, propomos a utilização de três formas de trabalhar com os dados criando um pipeline de dados desde a extração até a carga deles em algum repositório de uso comum para futuros interessados.

Primeiramente optamos por utilizar o Pentaho, 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, como o comparativo feito por Lira Filho em “Análise comparativa das Ferramentas de ETL – Kettle e Talend” (2013), nos provou ser uma ferramenta que servirá de base no processo de ETL para os cenários de data warehouse objetivados.

Como representante brasileiro, iremos trabalhar com o JETL,:

ferramenta que tem como objetivo facilitar a automatização de processos de carga e integração de informações em banco de dados. Contando com diversos recursos como LOADER, ETL (transferência de informações entre banco de dados) e SQL que facilitam muito a execução de atividades em bancos de dados também dispões de recursos de transmissão utilizando os protocolos FTP, SFTP e Compartilhamento de arquivos do Windows além de recursos de monitoramento como e-mail e envio de SMS. (Documentação Oficial, 2016)

Caminhando em direção a linguagem em si, utilizaremos o Python para criar o processo de uma maneira direta. Vale ressaltar que Python é hoje a segunda linguagem de programação mais utilizada no mundo e amplamente empregada na ciência de dados, juntamente com 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. Ajudando assim, a desconstruir uma barreira entre humanas e exatas que dificulta, em muitos aspectos, a interdisciplinaridade entre as áreas de conhecimento.

Dessa forma, trabalharemos com os eixos de política, economia e educação à princípio. Refinando os processos e provando sua viabilidade, poderemos então abrir o escopo e, assim, possibilitar a expansão para mais áreas como saúde, sociedade, população e muitas outras possíveis.