ETL

O processo de ETL é a etapa em que os dados são extraídos da fonte inicial, seja elas um banco ou diversos bancos, com alta ou baixa granularidade, recebem então tratamento, se necessário, e é realizado o carregamento no repositório.

A extração é a primeira fase do ETL. Nela é realizada a extração de diversas fontes de dados. As diferentes fontes têm características próprias que são gerenciadas e só então realiza-se a extração. Deve-se observar a estrutura das fontes para que possam ser ajustadas ao repositório que as receberá. Os dados são recuperados em sua forma crua, seja em bancos relacionais ou não, planilhas, arquivos, e assim por diante. É um processo que, por sua vez, pode ser separado em dois outros, a importação dos dados em si e a atualização dos dados periodicamente.

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. Assim sendo, tudo o que não condiz com o que é esperado passam por “um processo de exclusão ou tratamento de acordo com as regras de negócio da aplicação de destino, solucionando-as para garantir confiabilidade ao processo de ETL” (Lira Filho, 2013, p. 24). Dados também podem ser convertidos, padronizados e tipificados para se adequarem às regras de negócio observadas pelo engenheiro de dados.

Por fim, a última fase do ETL, é o carregamento dos dados. Nela os dados são gravados em um repositório de destino que permite o acesso ao consumidor. No caso de um data warehouse, os dados são carregados em tabelas dimensão e tabelas fato, que fazem parte do cubo de dados OLAP.

Um cubo OLAP é uma estrutura de dados que supera as limitações dos bancos de dados relacionais, proporcionando rápida análise de dados. Os cubos podem exibir e somar grandes quantidades de dados enquanto fornecem aos usuários acesso pesquisável a quaisquer pontos de dados. Dessa forma, os dados podem ser acumulados, segmentados e incluídos, conforme necessário, para lidar com a maior variedade de perguntas relevantes à área de interesse de um usuário.(MICROSOFT, 2018)

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.

ELT extrai dados de um sistema-fonte, os carrega em um sistema de destino e, então, usa o poder de processamento do sistema-fonte para conduzir as transformações. Isso acelera o processamento de dados porque acontece onde os dados estão.(SAS)

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, embora em muitas oportunidades poderão ser aplicadas as mesmas tratativas.

O data warehouse será então o destino final dos dados processados pelo ETL, torna-se então importante entender o que é esse repositório a que nos referimos.

Bibliografia

EL-SAPPAGH, Shaker H. Ali; HENDAWI, Abdeltawab M. Ahmed; BASTAWISSY, Ali Hamed El. A proposed model for data warehouse ETL processes. **Journal Of King Saud University - Computer And Information Sciences**, [S.L.], v. 23, n. 2, p. 91-104, jul. 2011. Elsevier BV. Disponível em: https://www.sciencedirect.com/science/article/pii/S131915781100019X. Acesso em: 27 de maio de 2021.

LIRA FILHO, Hermanny Alexandre dos Santos. **Análise comparativa das Ferramentas de ETL – Kettle e Talend.** 2013. 79 f. TCC (Graduação) - Curso de Bacharelado em Sistemas de Informação, Universidade Federal da Paraíba, Rio Tinto, 2013. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/17132?locale=pt\_BR. Acesso em: 29 maio de 2021.

MICROSOFT. **Visão geral dos cubos OLAP do Service Manager para análise avançada**. 2018. Disponível em: https://docs.microsoft.com/pt-br/system-center/scsm/olap-cubes-overview?view=sc-sm-2019. Acesso em: 30 maio de 2021.

SAS. **ETL: o que é e qual sua importância?**. O que é e qual sua importância?. [20--]. Disponível em: https://www.sas.com/pt\_br/insights/data-management/o-que-e-etl.html. Acesso em: 30 maio 2021.