Fundamentos de Big Data

UNIDADE 08

ETL: Extract, Transform, Load

ETL

O que é a ETL?

O que acha de começarmos a nossa semana conversando um pouco sobre a ETL? O que significa esta sigla? O que ela faz? Qual é a ligação dela com a nossa disciplina?

Vídeo em breve

Pessoas que trabalham com engenharia, ciência e análise de dados trabalham frequentemente com um processo de ETL. Inclusive, este é o meu caso (quer dizer, o professor que estava digitando estas palavras e gravando estes vídeos). Para ficar mais claro, vou te dar um exemplo: vamos supor que você começou a trabalhar em uma grande *e-commerce* e precisa analisar os dados das compras dos clientes para entender melhor o comportamento deles e melhorar suas estratégias de venda. O problema disso é que empresas que trabalham com *big data* não possuem só um único *schema* do SQL contendo todas as suas tabelas. Os dados ficam espalhados entre vários microsserviços, sistemas e serviços internos e externos.

Empresas grandes possuem dados espalhados entre tabelas SQL, planilhas do Excel (sim, exatamente. Sempre terá uma planilha do Excel, não importa o ano), APIs externas, dados em arquivos JSON e CSV, e assim por diante. Logo, o desafio é *juntar e padronizar* informações. É aí que entra a ETL.

A ETL (***Extract, Transform, Load***) é como se fosse um processo mágico que te ajuda a fazer isso. O passo de "**extract**" (extração) é quando você pega os dados brutos das compras, como o valor gasto, os produtos comprados e a data da compra, e os extrai de onde estão armazenados, como um banco de dados da loja. É neste passo em que obtemos os dados daquelas diferentes fontes de dados que acabei de exemplificar.

Assim que temos todos estes dados vem o passo de "***transform***" (transformação). É aqui que analisamos, padronizamos e transformamos os dados. Como um exemplo, poderíamos agrupar os dados para obter as vendas que ocorreram durante um dia completo em vez de registrar cada venda que ocorreu individualmente. Também poderíamos remover as compras que foram canceladas, ou as compras cujo pagamento não tivesse sido aprovado.

Finalmente, o passo de "***load***" (carga) é quando você coloca esses dados já transformados em um lugar onde possa analisá-los facilmente, como um banco de dados especializado em análise, ou em um arquivo padronizado. Assim, você consegue ver padrões de compra, identificar produtos mais vendidos e até prever comportamentos futuros dos clientes, já que os dados estão padronizados.

Criando um exemplo de ETL com PySpark

Dito isso, o que acha de testarmos uma ETL simples usando tudo o que vimos até o momento? Criaremos neste vídeo um notebook em que

Vídeo em breve

faríamos um processo de ETL usando PySpark em um arquivo CSV. Vamos lá?

NA LITERATURA

A relevância da ETL

Você já deve ter entendido como funciona uma ETL ao executarmos os exemplos da videoaula. Dito isso, antes de finalizarmos o conteúdo da nossa disciplina recomendo mais uma leitura sobre o ETL. É o texto "Processo ETL, que se encontra entre as páginas 159 e 170 do livro *Data warehouse*, de Vida et al. (2021). Com essa leitura, a ideia é a de estendermos os conceitos de *extract, transform* e *load* um pouco mais. Vamos lá?

Data warehouse

Conclusão

Nesta unidade vimos o processo de ETL, relevante para qualquer operação de engenharia de dados. Ele é composto por três passos: a extração das fontes de dados; a sua transformação, agregamento e manipulação destas diferentes fontes de dados; e a carga dos dados transformados. Vimos, ainda, um exemplo de ETL utilizando Hadoop e Spark utilizando a linguagem Python.

Referências

BENGFORT, B.; KIM, J. **Data analytics with Hadoop**: an introduction for data scientists. Sebastopol: O'Reilly Media, 2016.

CARMO, C. A. M. **Dominando o PySpark**: guia prático e objetivo. [S.l.: s.n.], 2023.

VIDA, E. S. et al. **Data Warehouse**. Porto Alegre: Grupo A, 2021.