UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Extract Transform Load (ETL)

Wesley Pedrosa dos Santos 180029240

Brasília, DF

2023

A) Definição da tecnologia

O conceito do ETL foi introduzido na década de 1970 como um processo de integração e carregamento de dados em mainframes ou supercomputadores para computação e análise. Do final dos anos 1980 até meados dos anos 2000, ele foi o principal processo usado na criação de data warehouse, que oferece suporte a aplicações BI (IBM CLOUD EDUCATION, 2020).

O desafio desse ambiente ETL é fazer a integração, reorganização e consolidação de grandes volumes de dados em muitos sistemas, comumente utilizado em sistemas de business intelligence (BI) e em projetos de integração de dados, onde há a necessidade de consolidar e unificar dados de várias fontes para análise e tomada de decisões.

O processo ETL geralmente envolve as seguintes etapas:

Na fase de extração, o ETL identifica os dados e os copia de suas origens, de forma que possa transportar os dados para o armazenamento de dados de destino. Os dados podem ser provenientes de diversas fontes, como bancos de dados, planilhas, arquivos CSV, sistemas de terceiros, APIs, entre outros. Esses dados podem estar estruturados, ou seja, dados que possuem uma padrão pré-definido, uma estrutura bem definida e rígida que é pensada antes da existência do dado na base de dados. Como dados não estruturados, que são dados que não possuem um padrão pré-estabelecido, são dados flexíveis e dinâmicos, podendo ser compostos por diversos elementos diferentes dentro de um todo. Exemplos de fontes para esses dados podem ser documentos, emails, aplicações de negócios, data warehouses, bancos de dados, data lakes, equipamentos, sensores e muito mais.

Por diversas vezes o ETL lida com grandes volumes de dados, então essa extração de dados acaba sendo muito onerosa para o sistema, por isso é recomendado iniciar o ETL despejando os dados em um local chamado de área de preparação, ou intermediária, onde todas as atividades de ETL podem ser executadas (LEMAHIEU; BROUCKE; BAAESENS, 2018).

Na figura 1 pode se observar um exemplo onde os dados são retirados de fontes como, SGBD, dados externos, dados de um consórcio de indústria (CODASYL) e e-mail.

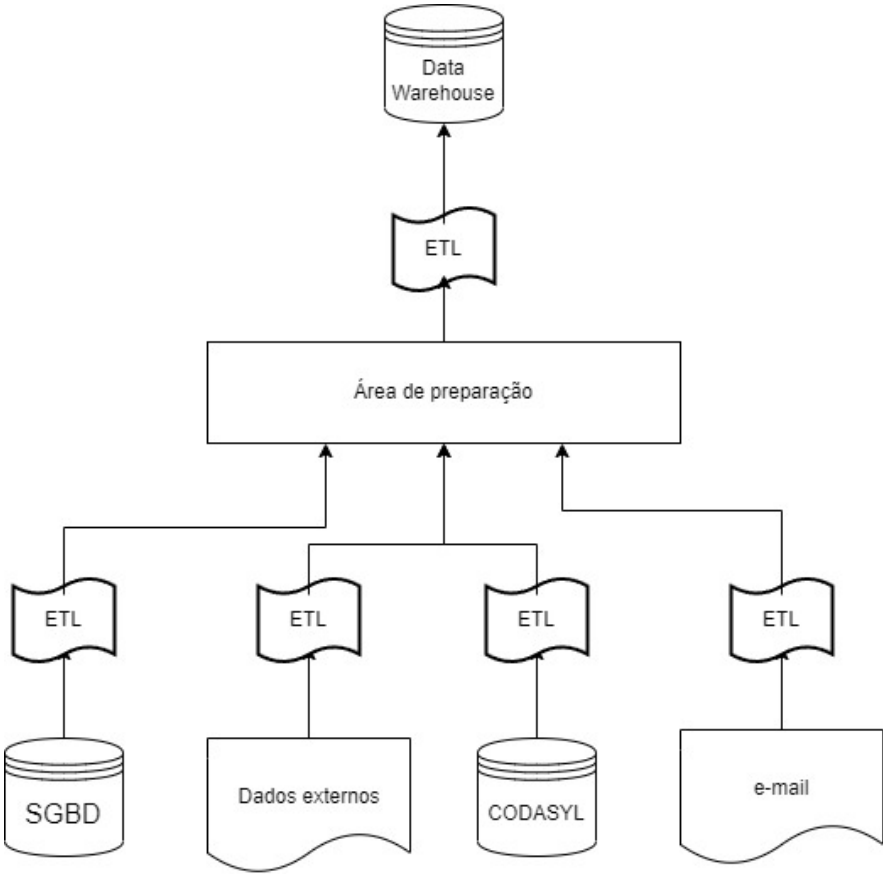


Figura 1.

Na área de preparação os dados brutos são transformados para serem úteis à análise e se ajustarem ao esquema do banco de dados do destino desejado. É durante essa fase que regras e regulamentos podem ser aplicados para garantir a qualidade e acessibilidade dos dados.

A transformação é realizada a transformação, que consiste em aplicar uma série de regras, limpezas e manipulações aos dados extraídos. As transformações podem incluir a filtragem e validação dos dados, a padronização de formatos, a correção de erros, a agregação de informações, a criação de novas colunas calculadas, a remoção de duplicatas, entre outras operações.

É importante que essas transformações ocorram na área de preparação, assim evitando um grande impacto de processamento no sistema de destino, além de reduzir também o risco possuir dados corrompidos no sistema de destino.

Essa etapa é considerada a mais importante do ETL, pois ela é

responsável para que os dados cheguem ao destino de maneira integra, compatíveis com as estratégias de BI que serão utilizadas sobre eles, e prontos para serem usados.

O último passo do ETL é o Carregamentos dos dados para o destino desejado, que pode ser um banco de dados, um data warehouse ou qualquer outro sistema de armazenamento de dados. Nessa fase, os dados são estruturados de acordo com o esquema do destino e são inseridos nas tabelas ou coleções apropriadas.

Outra maneira possível é o carregamento incremental, nessa opção os dados só serão carregados no destino se eles possuírem informações novas e exclusivas, com isso a manutenção do destino dos dados fique de fácil manutenção.

Uma estratégia de carregamento de dados bastante utilizada é um carregamento completo de todos os dados inicialmente, e carregamentos incrementais periodicamente, e em seguida com menos frequência manutenções na base de dados para apagar e substituir dados. O ETL também pode ser implantado com pipeline, onde logo após a extração dos dados, é feita a transformação, em paralelo a transformação pode ser feita a extração de novos dados, e com o carregamento dos dados pode ser feito em paralelo a transformação dos novos dados e a extração de mais dados, assim por diante, como é ilustrado na Figura 2.

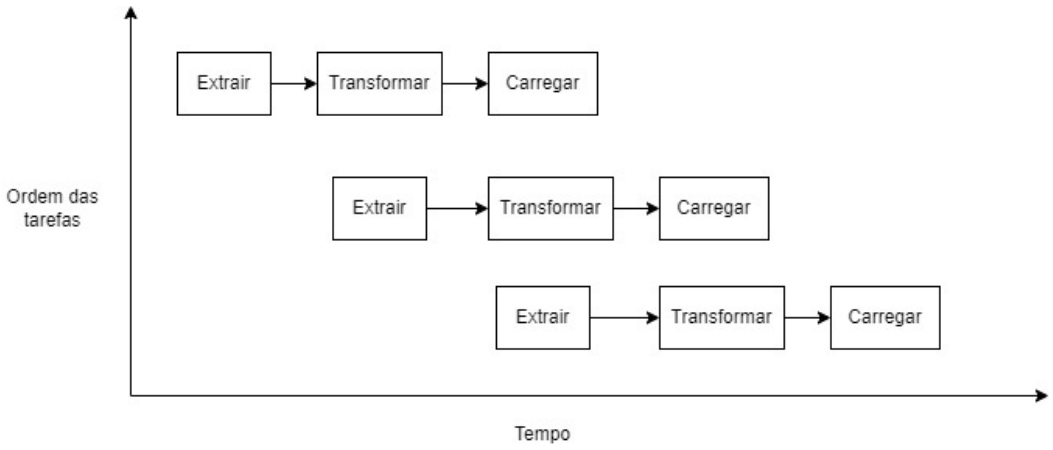


Figura 2.

B) Objetivos

Como mostrado na definição da tecnologia, um do principal uso do ETL é entregar dados a data warehouses corporativos que suportam aplicações de BI. Data warehouses esse que é projetado para representar uma fonte confiável de verdade sobre tudo o que está acontecendo em uma empresa em todas as atividades.

Um dos objetivos é a clareza nas informações. Durante a etapa de transformação do ETL, os dados passam por limpeza e união antes mesmo de serem transferidos para o banco de dados, para só assim ser possível fazer diferentes análises. Isso tudo vai permitir que os interessados manipulem informações com clareza e eliminem a ambiguidade desses dados que vieram das fontes, Isso impacta diretamente na qualidade dos dados, e consequentemente, no retorno das consultas desejadas que futuramente irão apoiar a tomada de decisão.

Outro ponto a ser alcançado pelo ETL é a integridade da informação. Quando temos uma pipeline ETL bem planejada e executada, por exemplo, vamos ter o máximo de fontes de negócios relevantes reunidas em um só local (o destino onde os dados serão carregados), contendo todas as informações completas. Isso vai impactar diretamente na qualidade das atividades de BI da empresa.

Com base no que foi exposto, podemos notar que a importância do ETL em uma empresa é diretamente proporcional ao quanto ela depende da data warehouse. Com a ajuda de funções ETL, é possível coletar e fazer leitura de grandes volumes de dados transformados, que antes eram brutos, inconsistentes, despadronizados e de baixa qualidade, oriundos de várias fontes de dados e diferentes plataformas. Com o ETL, há o processamento de dados para torná-los significativos com operações, como classificação, junção, reformatação, filtragem, mesclagem, agregação, entre outras. Como vimos a empresa passa a ter uma série de benefícios ao utilizar o ETL na data warehouse, principalmente no que diz respeito à qualidade e ao desempenho no retorno das consultas desejadas que irão apoiar a tomada de decisão de negócio da empresa.

C) Vantagens da tecnologia

O uso do processo Extract, Transform, Load (ETL) traz diversas vantagens para organizações que lidam com grandes volumes de dados e precisam consolidar informações de várias fontes. Exemplos de setores que pode desfrutar das vantagens do ETL:

- Consolidação de dados: O ETL permite a consolidação de dados dispersos em várias fontes. Isso possibilita a criação de um repositório centralizado de dados, como um data warehouse, onde as informações de diferentes sistemas e fontes podem ser integradas em um único local. Isso facilita o acesso e a análise dos dados de forma mais eficiente.
- Limpeza e padronização dos dados: O processo de transformação no ETL permite a limpeza e padronização dos dados extraídos. Isso inclui a identificação e correção de erros, a remoção de valores duplicados ou inconsistentes, a normalização de formatos e a aplicação de regras de negócio. Dessa forma, os dados ficam mais confiáveis e consistentes, melhorando a qualidade das análises e relatórios gerados.
- Integração de dados de diferentes fontes: Com o ETL, é possível integrar dados de diferentes fontes e sistemas, independentemente do formato ou estrutura em que se encontram. Isso permite a união de informações de bancos de dados, planilhas, arquivos de texto, sistemas legados, entre outros. A capacidade de integrar dados heterogêneos possibilita análises mais abrangentes e insights mais completos.
- Transformações e cálculos personalizados: O processo de transformação no ETL oferece flexibilidade para aplicar transformações e cálculos personalizados aos dados. É possível criar novas colunas calculadas, combinar dados de várias fontes, aplicar filtros e condições específicas, entre outras operações. Essa capacidade permite adaptar os dados às necessidades do negócio e gerar informações mais relevantes para análise.
- Performance e escalabilidade: Ao utilizar o ETL, é possível otimizar a performance do processo de carga de dados. As etapas de extração, transformação e carga podem ser projetadas de forma a melhorar o desempenho, reduzindo o tempo necessário para processar e carregar

grandes volumes de dados. Além disso, o ETL pode ser dimensionado para lidar com o aumento do volume de dados à medida que a empresa cresce.

- Segurança e governança dos dados: O ETL permite aplicar políticas de segurança e governança dos dados durante o processo de transformação e carregamento. Isso inclui a definição de permissões de acesso, a anonimização de dados sensíveis, a conformidade com regulamentações de proteção de dados, entre outros aspectos. A aplicação de medidas de segurança e governança ajuda a garantir a privacidade e a integridade dos dados.

D) Desvantagens do uso da tecnologia

Embora o processo Extract, Transform, Load (ETL) ofereça várias vantagens, também existem algumas desvantagens que devem ser consideradas.

- Latência nos dados: O processo ETL envolve várias etapas, como extração, transformação e carga. Dependendo do volume de dados e da complexidade das transformações aplicadas, pode haver um tempo significativo entre a extração dos dados das fontes originais e sua disponibilidade no destino final. Isso pode resultar em uma latência nos dados, ou seja, um atraso na disponibilidade de informações atualizadas.
- Complexidade e custo de desenvolvimento: A implementação de um processo ETL pode ser complexa e exigir conhecimentos técnicos avançados. É necessário criar pipelines de dados, definir transformações, lidar com possíveis erros e garantir a escalabilidade e a performance do processo. Isso pode exigir recursos especializados e aumentar os custos de desenvolvimento e manutenção.
- Dependência de estrutura pré-definida: O ETL geralmente requer uma estrutura pré-definida, como um data warehouse, onde os dados serão carregados. Isso implica em definir antecipadamente o esquema de dados e as transformações a serem aplicadas. Essa rigidez pode dificultar a adaptação a mudanças futuras nos requisitos de dados e

análises, exigindo ajustes significativos no processo ETL.

- Dificuldade em lidar com dados em tempo real: O ETL tradicionalmente opera em lotes, o que significa que os dados são processados em intervalos de tempo pré-determinados. Isso pode ser problemático quando há a necessidade de lidar com dados em tempo real, que precisam ser processados e disponibilizados imediatamente. Nesses casos, é preciso recorrer a abordagens complementares, como a integração de dados em tempo real (Real-time Data Integration) ou streaming de dados.
- Riscos de perda de dados ou inconsistências: Durante o processo ETL, há o risco de perda de dados ou introdução de inconsistências. Erros durante a extração, transformação ou carga podem resultar na exclusão ou corrupção de informações. Além disso, se as regras de transformação não forem definidas corretamente, podem ocorrer resultados inesperados ou dados inconsistentes, comprometendo a qualidade e a confiabilidade dos dados.
- Manutenção e evolução contínuas: À medida que os requisitos e as fontes de dados evoluem, é necessário atualizar e manter os processos ETL de forma contínua. Isso pode exigir esforços de manutenção e evolução para garantir que o processo continue funcionando corretamente e atendendo às necessidades em constante mudança da organização.

E) Exemplos de uso

FlightSafety International é uma empresa internacional que treina pilotos de avião profissionais e oferece voos de simulação para pilotos comerciais, pilotos militares e pilotos do governo.

Com a necessidade de treinar pilotos com estilos completamente diferentes, como pilotos mais agressivos ou mais seguros das decisões, a FlightSafety fez uma parceria com a IBM para desenvolver um sistema capaz de extrair informações prevenientes de simuladores em tempo real, que chagam a milhares de variáveis. Com a análise desses dados extraídos é possível observar os pontos em que o piloto foi excelente, e em quais pontos o piloto precisa de uma atenção adicional. Essa análise pode ser feita não somente sobre um perfil de pilotagem específica como também para toda a população de pilotos.

Para a realização desse cenário foi necessário um alto investimento em processamento para que as análises sejam produzidas quase que em tempo real, a análise já fica disponível para o piloto assim que ele termina seu voo de simulação.

Base de dados

Diagrama Entidade Relacionamento

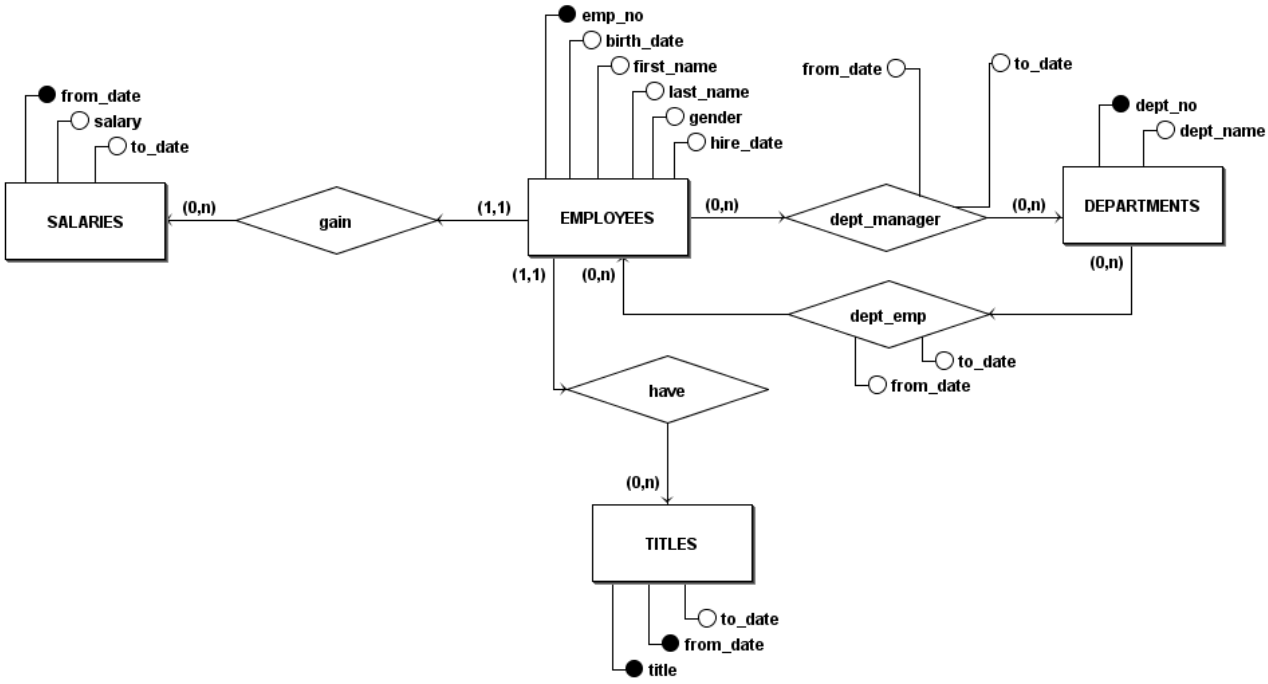
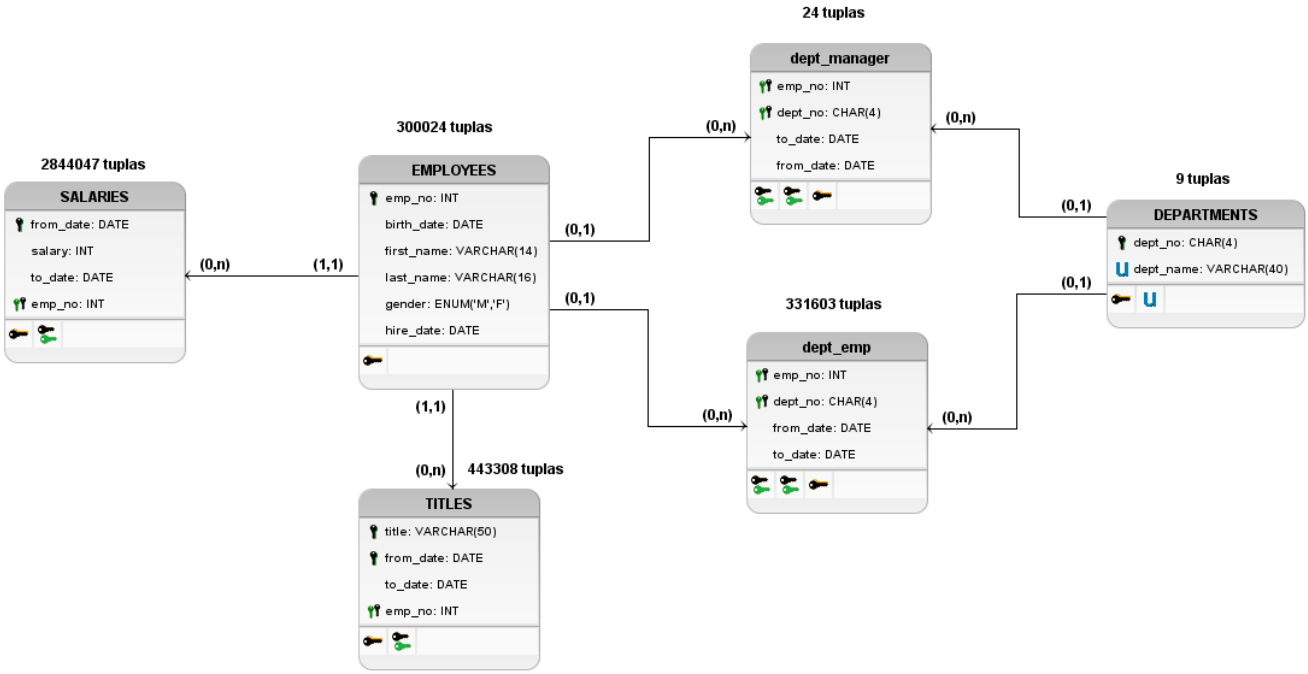


Diagrama Lógico de Dados



Para esse trabalho foi utilizada a base de dados encontrada nesse link:

https://launchpad.net/test-db/employees-db-1/1.0.6

Referências

- Edinilson da Silva Vida; Nicolli S. Rios Alves; Rafael G. Coimbra Ferreira; et al. Data warehouse. Porto Alegre: sagah, 2021.
- O que é ETL?. **Oracle**. Disponível em: < https://www.oracle.com/br/integration/what-is-etl/ >. Acesso em: 29, 08 e 2022.
- Pilots prepare for takeoff like never before. **IBM.** Disponível em < https://www.ibm.com/case-studies/flightsafety/ >. Acesso em: 29, 08 e 2022.