
Newton
Quem se prepara, não para.

Business Intelligence

4º período

Professora: Michelle Hanne

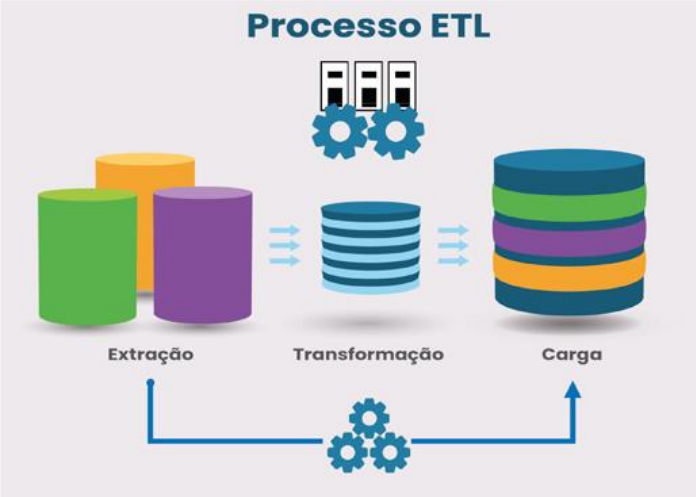
Newton
Quem se prepara, não para.

ETL
Extração, Transformação e Carga de dados

ETL

No ETL (Extração, Transformação e Load), você deve coletar dados de várias fontes, transformar os dados de acordo com as regras de negócio e carregar em um banco de dados de destino.

O trabalho de transformação no ETL ocorre em um mecanismo especializado e, geralmente, envolve o uso de tabelas de preparo para armazenar os dados temporariamente enquanto eles são transformados e, por fim, carregados em seu destino.



ETL

Newton
Quem se prepara, não para.

A **transformação** de dados geralmente envolve diversas operações, como *filtragem, classificação, agregação, junção de dados, limpeza de dados, eliminação de duplicação e validação de dados*

ETL

Operações da transformação de dados.
Fonte: adaptado de Kimball Ralph e Joe Caserta (2004).

Subsistemas ETL – Categoria Extração

Newton
Quem se prepara, não para.

- **Atualização por notificação:** se o sistema de origem dos dados é capaz de fornecer uma notificação de que um registro foi alterado e descrever a mudança. Essa é a maneira mais fácil de obter os dados. Assim, é possível coletar apenas os dados atualizados e inserir ou atualizar somente o necessário.
- **Extração incremental:** alguns sistemas podem não ser capazes de fornecer notificação de que uma atualização ocorreu, mas eles são capazes de identificar quais registros foram modificados e fornecer a extração de tais registros.
- **Extração integral:** alguns sistemas não são capazes de identificar quais dados foram alterados, então a extração completa é a única forma de obter os dados do sistema

Subsistemas ETL – Categoria Extração

Newton
Quem se prepara, não para.

Cada **sistema pode também utilizar um formato ou organização de dados que incluem: banco de dados relacionais e flat files** (também conhecidos como arquivos planos), mas podem incluir estruturas de bases de dados não relacionais, entre outras fontes de dados.

Outra maneira de realizar o processo de **ETL é utilizando dados online**, ou seja, conforme os dados vão chegando, através de um fluxo de dados, são tratados, transformados e carregados no sistema-alvo.

Subsistemas ETL – Categoria Transformação

Newton
Quem se prepara, não para.

1. **Ordenação** de um arquivo por uma determinada chave.
2. **Selecionar** somente algumas informações. Por exemplo, se o arquivo fonte tiver três tipos: nome, idade e nacionalidade, você deve escolher somente dois campos que poderiam ser nome e idade.
3. **Padronização**: como os arquivos podem vir de diversas fontes, você pode ter atributos que representam a mesma informação, mas com dados diferentes. Exemplo: em um arquivo você pode ter masculino igual a **1** e feminino igual **2** e em outro masculino igual a **M** e feminino igual a **F**.
4. **Derivação de um novo cálculo**. Por exemplo, **valor_venda = quantidade * preco_unitário.**

Subsistemas ETL – Categoria Transformação

Newton
Quem se prepara, não para.

5. Junção de múltiplas fontes de dados **por chaves de junção**. Você pode também efetuar duplicação de registros e agregar os registros para formar um dado estatístico. *Por exemplo, sumarizar múltiplas linhas de um arquivo de transações para ter o total de vendas por loja, total de vendas por região, entre outros*.
6. Transpor o arquivo original, isto é, linhas viram colunas e colunas viram linhas.
7. Dividir um arquivo em arquivos separados, um para cada coluna.
8. Validar e buscar dados em tabelas ou arquivos de referência.
9. Aplicar qualquer tipo de validação simples ou complexa no dado. *Por exemplo, validar se um determinado campo apresenta uma data válida. Caso haja alguma exceção poderá haver rejeição parcial ou completa.*

Subsistemas ETL – Categoria Carga (Loading)

Newton
Quem se prepara, não para.

Essas cargas, em sua maioria, são **realizadas periodicamente (diária, semanal ou mensalmente).** O DW é usualmente utilizado para guardar informações históricas. Conforme os dados são carregados no sistema, o banco de dados é responsável por **tratar as condições de integridade dos dados que estão sendo inseridos, como por exemplo, realizar uma validação de data, disparar um script de atualização de outra tabela ou tratar a unicidade de um dado.**

A entrega **ou a carga de dados** consiste em **estruturar fisicamente** os dados dentro do **modelo dimensional.** A periodicidade em que as informações devem ser gravadas ou até mesmo substituindo as informações dentro do data Warehouse pode variar de acordo com a necessidade do negócio.

Subsistemas ETL – Categoria Monitoramento

Newton
Quem se prepara, não para.

Criar políticas para suportar todos os componentes envolvidos no processo ETL: jobs, transformações, bases de dados intermediárias (stagings), sistema de arquivos, entre outros. E garantir que os códigos-fontes desenvolvidos sejam versionados.

Subsistemas ETL – Categoria Monitoramento

Existem dois tipos de incidentes que podem afetar um processo de ETL:

- **Uma falha no processo de ETL:** deve ser evitada a todo custo porque é o seu nome ou o nome da sua empresa que aparecerá no momento em que a falha ocorrer.
- **Uma falha em qualquer componente externo envolvido com o processo ETL:** interrupção elétrica, mau funcionamento nos discos de armazenamento de dados e/ou estruturas físicas comprometidas.

SCD – SLOWLY CHANGE DIMENSION

Newton
Quem se prepara, não para.

Você deve entender que cada dimensão é logicamente independente de todas as outras dimensões.

Em particular, supõe-se que as dimensões sejam independentes do tempo. Infelizmente, **esse não é o caso no mundo real**. Embora os atributos da tabela de dimensão sejam relativamente estáticos, eles não sofrem alterações.

SCD – SLOWLY CHANGE DIMENSION

Newton
Quem se prepara, não para.

- **Os atributos de dimensão mudam, embora de maneira bastante lenta, ao longo do tempo (SCDs)**
- Você pode preservar a estrutura dimensional independente com apenas ajustes relativamente pequenos para lidar com as mudanças.

- **Para cada atributo nas tabelas de dimensões, precisa ser especificada uma estratégia para lidar com as mudanças**. Em outras palavras, quando um valor de atributo muda no mundo dos sistemas transacionais, você pode decidir se precisa empregar uma combinação dessas técnicas em uma tabela de dimensão única.

Referências

- GONÇALVES, Glauber Rogério Barbieri. **Sistemas de informação**. Porto Alegre: SAGAH, 2017. ISBN digital 9788595022270.
- KIMBALL, Ralph; CASERTA, Joe. The Data Warehouse - ETL Toolkit **- Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. Indianápolis: Wiley, 2004. ISBN 0-764-57923-1.
- STAIR, Ralph M.; REYNOLDS, George W**. Princípios de Sistemas de Informação**. 11. ed. Cengage Learning,2016. ISBN digital 9788522124107.
- TURBAN, Efraim et al. **Business Intelligence**: um enfoque gerencial para a inteligência do negócio. Porto Alegre: Bookman, 2009. ISBN digital 9788577804252.