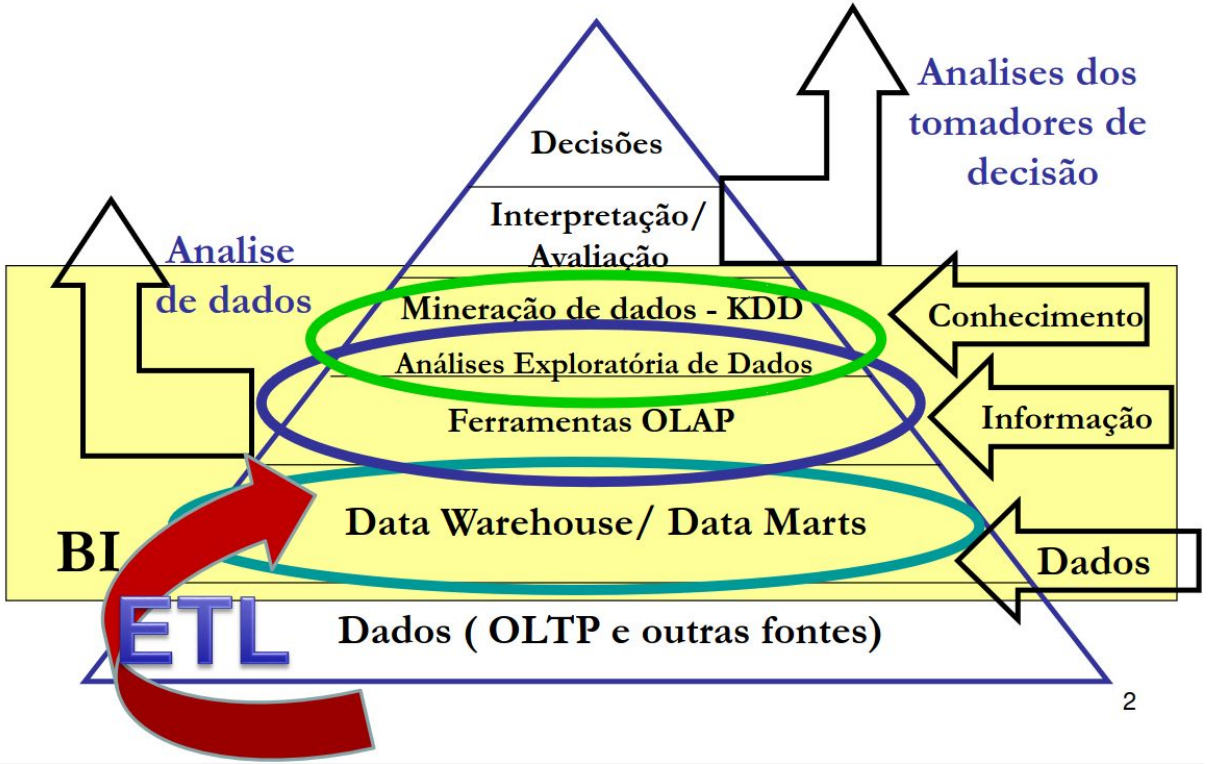
ETL

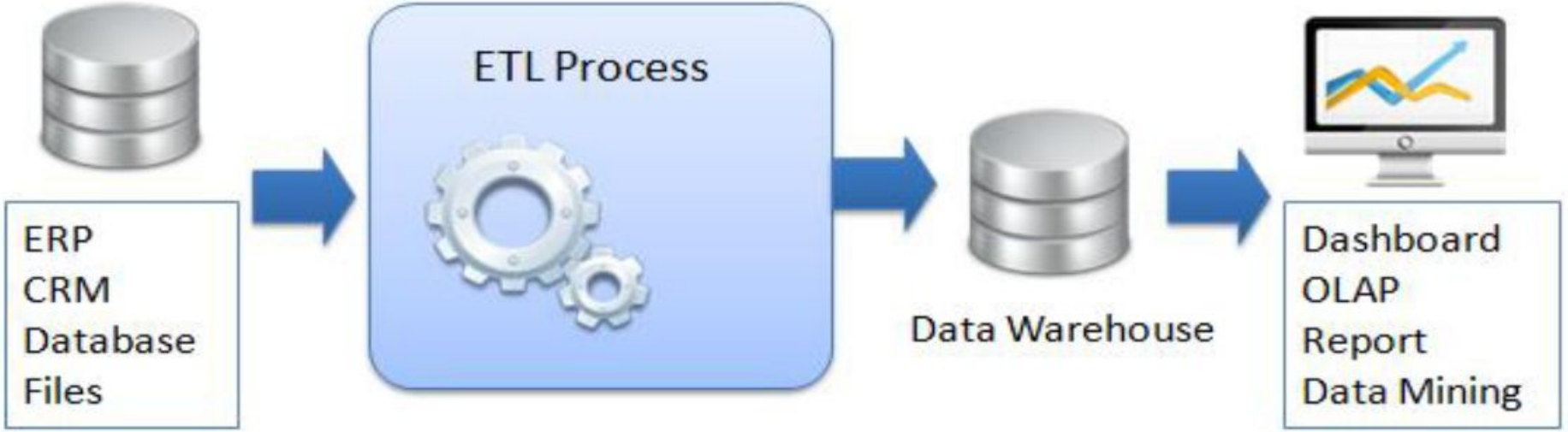
Extração, Transformação, Carregamento

Prof. Saulo Popov Zambiasi

Fundamentação da disciplina



Processo ETL - Extração, Transformação e Carregamento



ETL - Extração, Transformação e Carregamento

- O ETL é um sistema ou conjunto de scripts SQLs para extrair os dados das bases de operação e carregá-las no modelo dimensional construído depois da transformação
- **Objetivo da etapa ETL**
  - Fazer a integração de informações de fontes múltiplas e complexas, portanto, torna-se uma etapa bastante crítica, já que uma informação carregada erroneamente pode trazer consequências imprevisíveis nas fases posteriores.
- Esta etapa divide-se basicamente em três passos:
  - extração
  - transformação
  - carga.

ETL - Extração, Transformação e Carregamento

- Processo de combinação de dados de várias fontes em um grande repositório central - data warehouse.
- Faz uso de regras de negócios para limpar e organizar dados brutos e prepará-los para armazenamento
- Utiliza análise de dados e machine learning
- Pode atender a necessidades específicas de business intelligence por meio da análise de dados
    - prever o resultado das decisões de negócios
    - gerar relatórios e painéis
    - reduzir a ineficiência operacional, etc.

Importância do ETL

- Os dados das organizações podem estar disponíveis em várias fontes, como:
  - dados de clientes, funcionários e fornecedores
  - dados de estoques e fornecedores
  - Dados de dispositivos de IoT, etc.
- Com a aplicação de ETL, os dados brutos podem ser preparados em formato e estrutura para serem utilizados em análises, resultando em informações mais significativas. Exemplo:
  - varejistas online podem analisar dados de pontos de venda para prever a demanda e gerenciar o estoque. As equipes de marketing podem integrar os dados do CRM aos comentários do cliente nas mídias sociais para estudar o comportamento do consumidor.
- O ETL aprimora o business intelligence e a análise, tornando o processo mais confiável, preciso, detalhado e eficiente.

ETL

- Fornece um **contexto histórico** profundo aos dados da organização.
- Uma empresa pode combinar dados herdados com dados de novas plataformas e aplicações.
- É possível visualizar conjuntos de dados mais antigos juntamente com informações mais recentes, o que oferece uma visão de longo prazo dos dados.
- Fornece uma **visualização consolidada dos dados** para análises e relatórios detalhados.

ETL

- O gerenciamento de diversos conjuntos de dados exige tempo e coordenação, podendo resultar em ineficiências e atrasos.
- Combina bancos de dados e várias formas de dados em uma única visualização.
- O processo de integração de dados melhora a qualidade dos dados e economiza o tempo necessário para mover, categorizar ou padronizar dados. Isso facilita a análise, a visualização e a compreensão de grandes conjuntos de dados.

ETL

- Oferece uma **análise de dados mais precisa** para atender aos padrões regulatórios e de conformidade.
- Permite integrar ferramentas de ETL com ferramentas de qualidade de dados para alinhar, auditar e limpar dados, garantindo que os dados sejam confiáveis.
- **Automatiza tarefas** de processamento de dados que se repetem para uma análise eficiente.
- Ferramentas de ETL automatizam o processo de migração de dados e pode-se configurá-las para integrar alterações de dados periodicamente ou até mesmo durante uma execução.

Evolução do ETL

- Originou-se com o surgimento de bancos de dados relacionais que armazenavam dados na forma de tabelas para análises.
- As primeiras ferramentas de ETL tentaram converter dados de formatos de dados transacionais em formatos de dados relacionais para análise.

Evolução - ETL Tradicional

- Dados brutos armazenados em bancos de dados transacionais
  - ofereciam suporte para muitas solicitações de leitura e gravação, mas não ofereciam boas análises.
- Você pode imaginar isso como uma linha em uma planilha.
- **Por exemplo**, em um sistema de comércio eletrônico, o banco de dados transacional armazena o item comprado, os detalhes do cliente e os detalhes do pedido em uma transação.
- Ao longo do ano, uma longa lista de transações é apresentada com entradas repetidas para um mesmo cliente que comprou vários itens durante o ano.

Evolução - ETL Tradicional

- Dada a duplicação de dados, tornou-se trabalhoso analisar os itens mais populares ou as tendências de compra naquele ano.
- **Solução:** ferramentas de ETL acabam convertendo automaticamente esses dados transacionais em dados relacionais com tabelas interconectadas.
- Assim, os analistas puderam realizar consultas para identificar relações entre as tabelas, bem como padrões e tendências.

Evolução - ETL Moderno

- Com a evolução dos tipos de dados e fontes, a tecnologia ETL evoluiu
- A nuvem possibilitou a criação de grandes bancos de dados (coletores de dados)
- Permitiu receber dados de várias fontes e com alocação de recursos de hardware dinamicamente
- Ferramentas de ETL também se tornaram mais sofisticadas, podendo funcionar junto com os coletores de dados modernos.
- Elas podem converter dados de formatos de dados herdados para formatos de dados modernos (Data Warehouses e Data Lakes)

Bases de Dados Modernas

Data warehouses

- repositório central que pode armazenar vários bancos de dados
- em cada banco de dados pode-se organizar seus dados em tabelas e colunas
- o software de data warehouse funciona em diversos tipos de hardware de armazenamento (HDD, SSD, Nuvem)

Bases de Dados Modernas

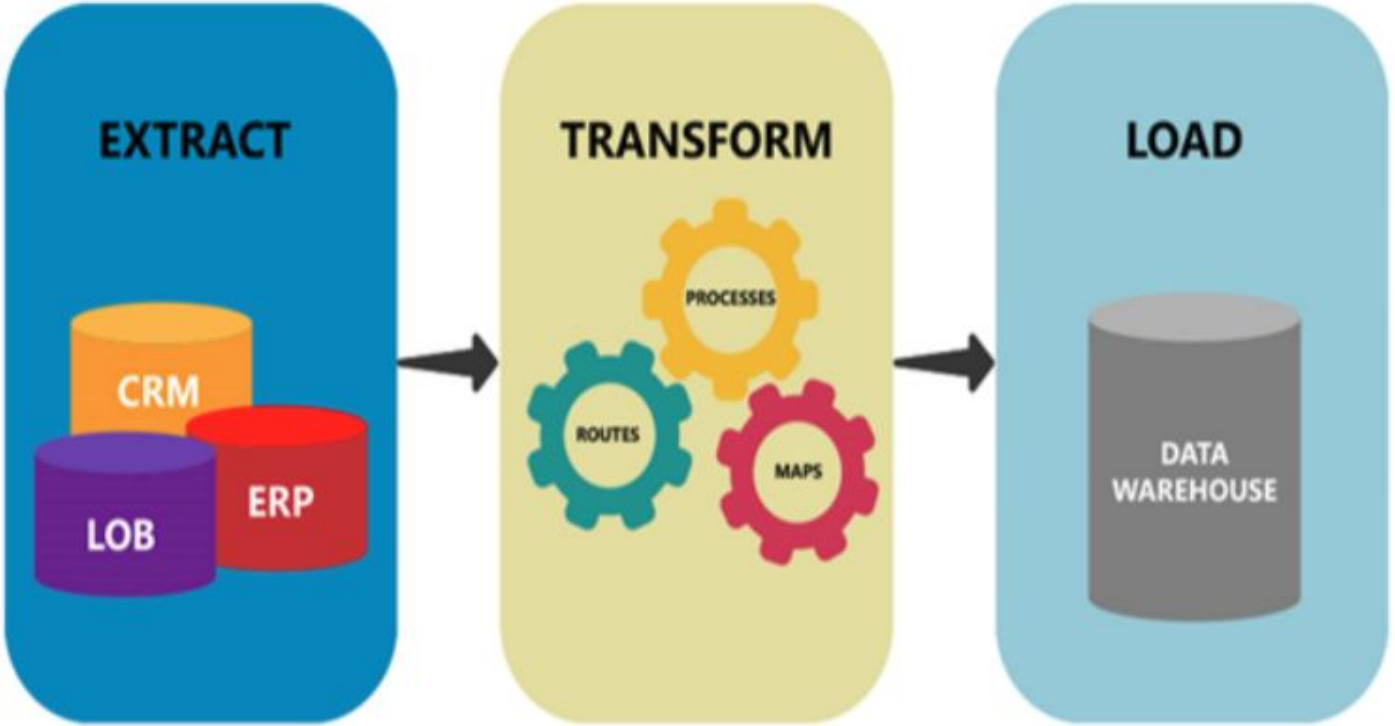
Data lakes

- permite armazenar dados estruturados e não estruturados em um repositório centralizado e em qualquer escala
- permite armazenar os dados como eles estão, sem a necessidade de estruturá-los com base em perguntas que possa ter no futuro
- permitem que possa executar diferentes tipos de análises em seus dados para orientar melhores decisões, como:
  - consultas SQL
  - análises de big data
  - pesquisas de texto completo
  - análises em tempo real e machine learning (ML)

Processo ETL

- Se dá pela movimentação de dados do sistema de origem para o sistema de destino em intervalos periódicos.
- Etapas:
  1. Extração dos dados relevantes do banco de dados de origem
  2. Transformação dos dados para que sejam mais adequados a análises
  3. Carregamento dos dados no banco de dados de destino

ETL - Extração, Transformação e Carregamento



ETL - Extração, Transformação e Carregamento

- Processo de extrair ou copiar dados brutos de diversas fontes e armazená-los em uma área de preparação.
- Uma **área de preparação** (zona de pouso) é uma área de armazenamento intermediária para armazenamento temporário dos dados extraídos.
- Áreas de preparação de dados geralmente são temporárias e seu conteúdo é apagado após a conclusão da extração de dados.

ETL - Extração, Transformação e Carregamento

- Também pode reter um arquivo de dados para fins de solução de problemas.
- A frequência com que o sistema envia dados da fonte de dados para o armazenamento de dados de destino depende do mecanismo de captura de dados de alterações subjacentes
- Normalmente, a extração de dados acontece por
  - Notificação de atualização
  - Extração Gradual
  - ou Extração completa

ETL - Extração, Transformação e Carregamento

Notificação de atualização

- O sistema de origem notifica quando um registro de dados é alterado.
- Assim, pode-se executar o processo de extração para essa alteração.
- A maioria dos bancos de dados e aplicações Web fornece mecanismos de atualização para oferecer suporte a esse método de integração de dados.

ETL - Extração, Transformação e Carregamento

Extração gradual

- Algumas fontes de dados não podem fornecer notificações de atualização, mas podem identificar e extrair dados que foram modificados em um determinado período.
- Nesse caso, o sistema verifica as alterações em intervalos periódicos, como uma vez por semana, uma vez por mês ou ao final de uma campanha. Depois é só extrair os dados que sofreram alterações.

ETL - Extração, Transformação e Carregamento

Extração completa

- Alguns sistemas não conseguem identificar alterações de dados ou fornecer notificações
- Assim, a única opção é realizar novamente o carregamento de todos os dados
- Esse método de extração exige que se mantenha uma cópia da última extração para verificar quais registros são novos.
- Envolve grandes volumes de transferência de dados, recomenda-se usá-la apenas para tabelas pequenas.

ETL - Extração, Transformação e Carregamento

- Transformam e consolidam os dados brutos na área de preparação para prepará-los para o data warehouse de destino
- A limpeza dos dados (uma forma de transformação) se dá porque os dados normalmente advêm de uma fonte muitas vezes desconhecida, concebida há muito tempo e contendo muito lixo e inconsistência.
- Operações de **remoção de ruídos**, de **atributos incompletos**, tratamento de **erros de digitação** ou **erros nos sistemas de captura de dados**, são tratados nesta etapa.
- Pode envolver os seguintes tipos de alterações de dados:
  - Transformação de dados básica
  - Transformação de dados avançada

ETL - Extração, Transformação e Carregamento

Transformação de dados básica

- Melhoram a qualidade dos dados ao remover erros, esvaziar campos de dados ou simplificar os dados. Exemplos:
- **Limpeza de dados**
  - Remove erros e mapeia os dados de origem para o formato de dados de destino.
  - Ex: mapear campos de dados vazios para o número zero, mapear o valor de dados “Parent” (Pai) para “P” ou mapear “Child” (Filho) para “C”.
- **Eliminação de duplicação de dados**
  - Serve para identificar e remover registros duplicados.
- **Revisão de formato de dados**
  - Converte dados (conjuntos de caracteres, unidades de medida e valores de data e horário) para um formato consistente.
  - Ex: uma empresa de alimentos pode ter diferentes bancos de dados de fórmulas com a medição de ingredientes em quilos e libras. O processo de ETL converterá tudo para libras.

ETL - Extração, Transformação e Carregamento

Transformação de dados avançada

- Utilizam regras de negócios para otimizar os dados a fim de facilitar a análise. Exemplo:
- **Derivação**
  - Aplica regras de negócios aos seus dados para calcular novos valores com base em valores existentes.
  - Ex.: converter receita em lucro subtraindo despesas ou calculando o custo total de uma compra e multiplicando o preço de cada item pelo número de itens pedidos.
- **Junção**
  - Vincula dados semelhantes de diferentes fontes de dados.
  - Ex.: encontrar o custo total de aquisição de um item adicionando o valor da aquisição de diferentes fornecedores e armazenando apenas o total final no sistema de destino.
- **Separação**
  - Pode-se dividir uma coluna ou um atributo de dados em diversas colunas no sistema de destino
  - Ex.: se a fonte de dados salvar o nome da cliente como “Jane John Doe”, você poderá realizar a separação em nome, nome do meio e sobrenome.

ETL - Extração, Transformação e Carregamento

Transformação de dados avançada (continuação)

- **Resumo**
  - Melhora a qualidade dos dados ao reduzir um grande número de valores de dados em um conjunto de dados menor.
  - Ex.: reduzir o atributo data com granularidade em dias para mês (redução de 365 para 12)
- **Encriptação**
  - Proteção dos dados confidenciais para cumprir as leis de dados ou a privacidade de dados adicionando encriptação antes que os dados sejam transmitidos para o banco de dados de destino.

ETL - Extração, Transformação e Carregamento

- Movem os dados transformados da área de preparação para o data warehouse de destino
- Em geral, esse processo é automatizado, bem definido, contínuo e orientado por lotes
- Os métodos podem ser por carregamento completo ou incremental
- Uma vez que a origem dos dados pode ser de sistemas diferentes, padronizam-se os diferentes formatos de modo que fiquem num formato uniforme, normalmente sugerido pelo próprio usuário.
- Com isso, a análise dos dados recuperados pela ferramenta OLAP (Online Analytical Processing) fica mais fácil, pois o usuário não estará vendo informações iguais em formatos diferentes

ETL - Extração, Transformação e Carregamento

Carregamento completo

- Todos os dados da origem são transformados e movidos para o data warehouse.
- Geralmente, ocorre na primeira vez que você realiza o carregamento de dados de um sistema de origem no data warehouse.

ETL - Extração, Transformação e Carregamento

Carregamento Incremental

- Realiza o carregamento do delta (ou diferença) entre os sistemas de destino e de origem em intervalos regulares
- A última data de extração é armazenada para que apenas os registros adicionados após essa data sejam carregados
- O carregamento incremental pode ser feito por transmissão ou por lotes

ETL - Extração, Transformação e Carregamento

Carregamento Incremental (continuação)

- **Por transmissão**
  - Para pequenos volumes de dados
  - transmite alterações de forma contínua através de pipelines de dados para o data warehouse de destino.
  - Quando a velocidade dos dados aumenta para milhões de eventos por segundo, pode-se usar o processamento de fluxo de eventos para monitorar e processar os fluxos de dados a fim de tomar decisões mais oportunas.
- **Em lotes**
  - Para grandes volumes de dados
  - Coleta alterações de dados de carregamento em lotes periodicamente
  - Durante esse período de tempo definido, nenhuma ação pode ocorrer no sistema de origem ou de destino à medida que os dados são sincronizados.

ETL - Dificuldades

- Várias origens para um mesmo dado: conflitos estruturais, de conteúdo e de formato nos dados
- Dados “faltantes”, dados com “erros”
- Não conformidade dos dados com as regras do negócio
- Dados significativos em campos de entrada livre
- Necessidade de normalização/desnormalização de dados
- Necessidade de “juntar/separar” atributos
- Diversos formatos de dados (xls, pdf, xml, etc.)
- Incompatibilidade entre ambientes operacionais diferentes