Data Warehouse

Após a extração dos dados, o processo usual envolve a transferência desses dados para **um sistema de armazenamento e gerenciamento de dados.** Neste sistema, ocorre a **integração de dados**, um processo essencial para combinar informações de diferentes fontes em um local único e centralizado. **Essa integração permite combinar dados de diferentes fontes em um local único e centralizado**. O objetivo é fornecer um conjunto de dados completo, preciso e atualizado para análise, inteligência de negócios e outras aplicações.

Um dos sistemas de armazenamento e gerenciamento de dados mais amplamente utilizados é o **Data Warehouse (DW).**

Um data warehouse é um hub central de dados usado para relatórios e análises. Os dados em um data warehouse geralmente são altamente formatados e estruturados para casos de uso analíticos. É uma das arquiteturas de dados mais antigas e bem estabelecidas. O data warehouse reúne e organiza dados de várias fontes (sistemas transacionais, APIs, etc.) em um formato otimizado para consultas analíticas.

A arquitetura organizacional de data warehouse tem uma característica principal:

- **Separar o processamento analítico online (OLAP) dos bancos de dados de produção (processamento de transações online - OLTP)**

Essa separação é fundamental à medida que as empresas crescem. Mover os dados para um sistema físico separado direciona a carga para longe dos sistemas de produção e melhora o desempenho das análises**.**

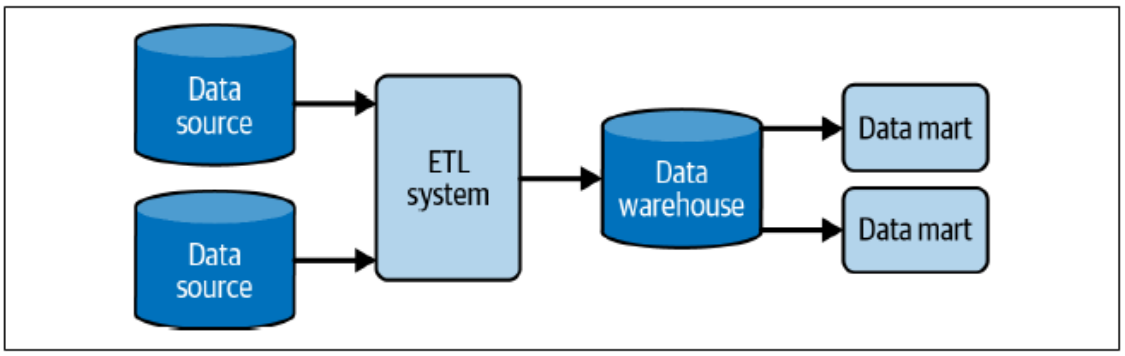
Desta forma, a separação entre o OLAP (Online Analytical Processing) e o OLTP (Online Transaction Processing) na arquitetura de um data warehouse é essencial porque eles servem a propósitos diferentes dentro de uma organização.

- **OLTP (Processamento de Transações Online):** é responsável por lidar com as transações diárias de uma empresa. Ele envolve o armazenamento e gerenciamento dos dados gerados por atividades de produção, como registrar uma venda, atualizar um inventário, ou gerenciar cadastros. Esse tipo de sistema é otimizado para transações rápidas e frequentes, com inserções, atualizações e consultas simples, como buscas de registros específicos. É o tipo de processamento que ocorre no banco de dados. Sistemas OLTP são frequentemente chamados de bancos de dados transacionais
- **OLAP (Processamento Analítico Online):** é responsável por executar consultas mais complexas para análise de dados e relatórios estratégicos. Diferente do OLTP, o OLAP é otimizado para consultas longas e complexas, que envolvem o processamento de grandes volumes de dados, agregações, e cálculos analíticos, como sumarizar vendas por região ao longo de vários anos.

Quando separamos os processamentos OLAP e OLTP, evitamos que o sistema de produção, que lida com as operações diárias (OLTP), fique sobrecarregado com consultas analíticas complexas (OLAP), o que poderia impactar negativamente a velocidade e a estabilidade das transações diárias.

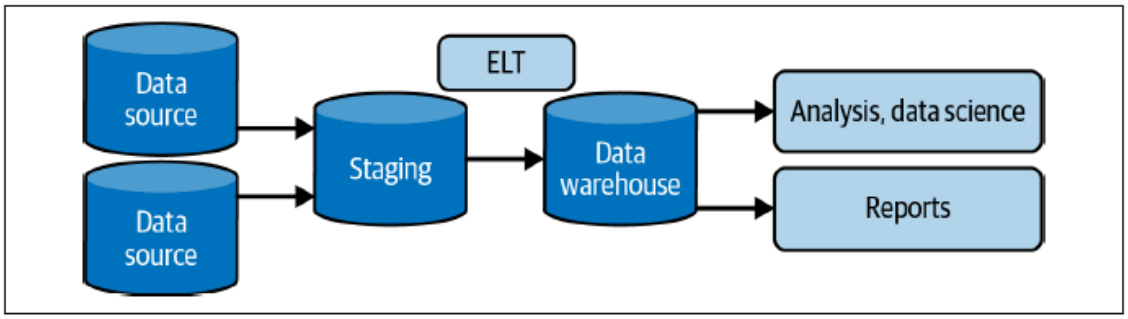
Ao mover os dados para um sistema físico separado (data warehouse), você cria um ambiente dedicado à análise de dados. Isso melhora o desempenho das consultas analíticas porque o sistema de data warehouse pode ser otimizado especificamente para esse tipo de processamento sem afetar a performance do sistema de transações.

Arquitetura de dados baseada no processo ETL:



- Os dados são extraídos de fontes, processados e transformados por um sistema ETL, armazenados em um Data Warehouse centralizado e, depois, organizados em Data Marts para atender a análises específicas de diferentes áreas de negócio.

Arquitetura de dados baseada no processo ELT:



- Com a arquitetura de data warehouse ELT, os dados são movidos mais ou menos diretamente dos sistemas de produção para uma área de preparação no data Warehouse. Em vez de usar um sistema externo, as transformações são feitas diretamente no data warehouse. A intenção é aproveitar o poder computacional massivo dos **data warehouses na nuvem** e das ferramentas de processamento de dados.

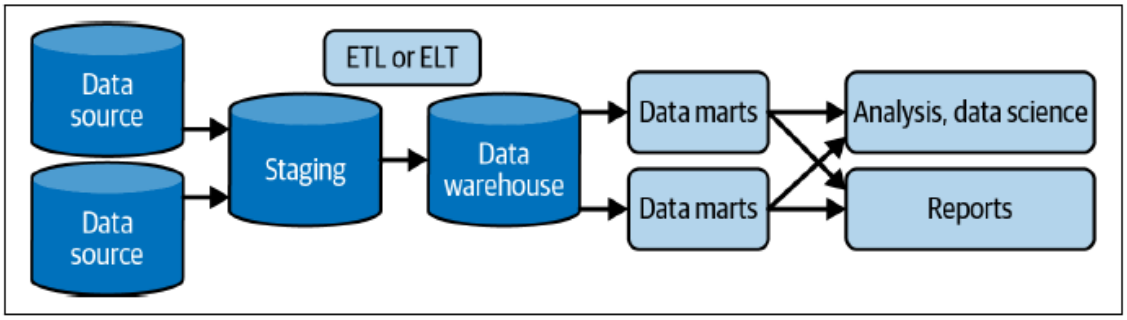
Os data warehouses na nuvem representam uma evolução significativa da arquitetura de data warehouse local e, por isso, trouxeram mudanças consideráveis para a arquitetura organizacional.

Data Mart

Um data mart é um subconjunto mais refinado de um data warehouse, projetado para atender às análises e relatórios, focado em uma única suborganização, departamento ou linha de negócios; cada departamento tem seu próprio data mart, específico para suas necessidades. Isso contrasta com o data warehouse completo, que atende a toda a organização ou negócio.

Um data mart tem uma etapa adicional de transformação além da que ocorre no data warehouse. Isso acontece porque, no data mart, os dados são organizados e processados de forma mais específica para atender às necessidades de um departamento ou área de negócios. Os data marts existem por duas razões. Primeiro, um data mart torna os dados mais facilmente acessíveis para analistas e desenvolvedores de relatórios. Segundo, os data marts fornecem uma etapa adicional de transformação além daquela realizada pelos pipelines iniciais de ETL ou ELT. Isso pode melhorar significativamente o desempenho se relatórios ou consultas analíticas exigirem junções e agregações complexas de dados, especialmente quando os dados brutos são volumosos. Os processos de transformação

podem preencher o *data mart* com dados já unidos e agregados, melhorando o desempenho para consultas em tempo real.



O processo de ETL/ELT com DW

1 - Área de Staging

Em um Data Warehouse (DW), a área de **staging** (ou camada de staging) é uma área temporária onde os dados são armazenados e preparados antes de serem transformados e carregados nas camadas finais do DW, como a área de dados integrados ou a camada de apresentação; A área de staging é uma etapa anterior ao **back room** em um Data Warehouse (DW). **É a primeira camada onde os dados são carregados diretamente das fontes de dados.**

Principais características da área de staging:

1. Armazenamento Temporário: Dados são carregados na área de staging apenas temporariamente, geralmente vindos de várias fontes.
2. Transformações Iniciais: Aqui podem ocorrer algumas transformações básicas, como limpeza de dados, remoção de duplicatas e padronização de formatos.
3. Segurança e Qualidade: Permite uma camada extra de segurança para validar e tratar os dados antes que eles cheguem à área de análise.
4. Isolamento: A área de staging geralmente não está disponível para usuários finais; é destinada apenas para o processo de ETL/ELT.

Por que a gente precisa usar a staging area?

- Alivia os sistemas de origem ao realizar limpezas básicas durante a conexão ativa com as fontes de dados, e, após a liberação do data source, permite transformações complexas, maximizando os recursos computacionais disponíveis.
- Os sistemas de origem têm janelas específicas para extração, e suas frequências de atualização nem sempre coincidem com o carregamento no Data Warehouse.

Não é necessário criar relacionamentos (joins) na etapa de stage. Por exemplo, ao integrar dois sistemas de origem com cadastros de clientes, as tabelas de cada sistema são carregadas separadamente na área de stage. A unificação e o ajuste são realizados posteriormente no Data Warehouse. Fazer a junção direto na carga pode ser problemático. Por exemplo, se o sistema A libera dados às 21h e o sistema B somente à 0h, a carga dependeria de ambos. Na abordagem recomendada, você traz todas as informações para a stage e realiza a junção após a disponibilidade de todos os dados.

2. Back Room

Após a área de staging, os dados passam para o **back room**, também conhecido como **Data Integration Layer**. Nessa camada, os dados são integrados, transformados em um formato comum e estruturados para que possam ser utilizados de forma mais consistente em relatórios e análises.

Nele ocorre a preparação e o processamento dos dados antes de serem disponibilizados para o uso analítico. Aqui, os dados são coletados, integrados e transformados. **Ele é focado na preparação e processamento dos dados. É onde ocorre a transformação dos dados brutos em dados prontos para análise.**

3. Front Room

O "front room" é a camada onde os dados já processados e preparados são acessados por usuários para fins de análise e tomada de decisão. Neste ambiente, os dados estão organizados de forma a facilitar o acesso e a análise, e são apresentados em estruturas de fácil leitura, como Data Marts e tabelas dimensionais, que permitem consultas rápidas e específicas. **Ele é focado no consumo dos dados para análise, onde os dados estão prontos e organizados para serem acessados pelos usuários finais de forma eficiente e intuitiva.**

A separação entre ambientes e o isolamento dos dados em diferentes camadas (como *staging*, *back room* e *front room*) é essencial para garantir a integridade, segurança e confiabilidade dos dados.

Data Warehouses na nuvem como Snow Flake e Google Big Query podem assumir outro tipo de arquitetura em camadas.

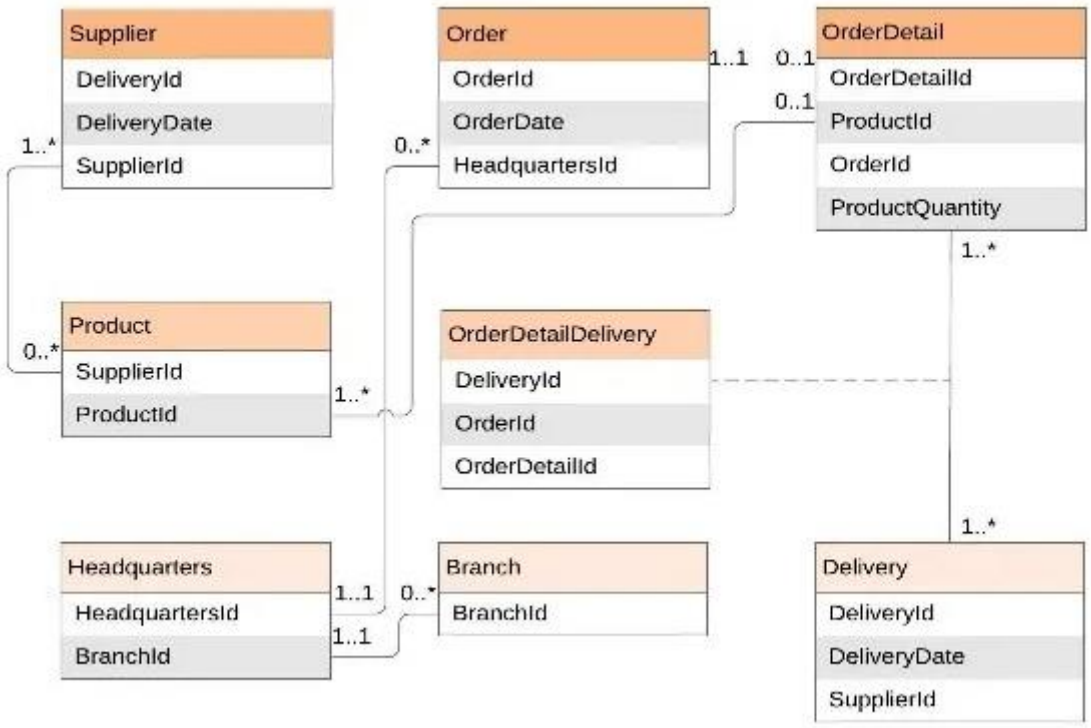
Modelo Relacional/ Modelo Dimensional

As diferenças entre o **modelo relacional** e o **modelo dimensional** estão diretamente relacionadas aos tipos de sistemas que eles suportam, como **Data Warehouse (DW)**, **OLAP (Processamento Analítico Online)** e **OLTP (Processamento de Transações Online)**. Cada modelo é otimizado para atender necessidades específicas de armazenamento, consulta e processamento de dados

Modelo Relacional

- Estrutura normalizada, geralmente até a 3ª Forma Normal (3FN), onde os dados são divididos em várias tabelas para reduzir redundância e garantir consistência.
- Mais usado em sistemas de OLTP (como sistemas bancários e ERP), que precisam de alto desempenho em transações diárias e garantir a integridade dos dados: Focado em operações de escrita rápidas.
- **A estrutura normalizada exige várias junções entre tabelas para consultas analíticas, tornando o modelo ineficiente para análise de grandes volumes de dados, comum em ambientes de Data Warehouse.**
- No modelo relacional**, os relacionamentos são definidos com o objetivo de garantir a integridade e consistência dos dados em operações transacionais.**
- Chaves Primárias e Estrangeiras: As tabelas se conectam através de chaves primárias e chaves estrangeiras. Cada tabela armazena um único tipo de entidade (ex.: clientes, produtos, pedidos), e os relacionamentos permitem unir essas entidades de forma controlada.

- As tabelas no modelo OLTP frequentemente se relacionam para permitir a consulta de informações integradas. Por exemplo, uma tabela Pedidos pode se relacionar com uma tabela Clientes para buscar informações do cliente associado ao pedido.



- O modelo Entidade-Relacionamento (ER) é uma ferramenta de modelagem conceitual usada para representar graficamente as entidades, atributos e relacionamentos que existem no domínio de dados de um sistema.
- O resultado da modelagem ER é geralmente um Diagrama Entidade-Relacionamento (DER), que mostra entidades (ex.: Cliente, Pedido, Produto) e relacionamentos (ex.: Cliente faz Pedido).
- **O modelo relacional implementa o modelo ER: Uma vez que o modelo ER está definido, ele é traduzido para um modelo relacional, que é implementado fisicamente no sistema de banco de dados.**

Modelo Dimensional

- Estrutura desnormalizada, com tabelas de fatos e tabelas de dimensões. A tabela de fatos armazena as medidas numéricas, enquanto as dimensões guardam os contextos (como data, produto, cliente).
- É amplamente usado em sistemas de OLAP e Data Warehouses, onde as consultas são mais analíticas e complexas, e o desempenho para leitura é essencial.
- Otimiza consultas analíticas, facilitando a navegação e a interpretação dos dados.
- O modelo dimensional, sendo desnormalizado, não é eficiente para operações transacionais e pode introduzir redundância de dados e inconsistência se usado em um sistema OLTP.
- **No modelo dimensional, os relacionamentos também existem, mas têm uma estrutura simplificada e desnormalizada com o objetivo de facilitar consultas rápidas e eficientes.**

Tabelas Fato

É onde estão armazenados os dados numéricos e quantitativos que representam eventos ou transações ocorridas em uma organização. Ela é o núcleo do modelo dimensional

e contém informações que podem ser medidas e analisadas, como valores de vendas, quantidade de itens vendidos, custo, lucro, etc. Elas contêm chaves estrangeiras que apontam para as tabelas de dimensão.

Tabelas de Dimensão

Contém dados descritivos que dão contexto às medidas na tabela fato. Em vez de números ou métricas, ela armazena informações qualitativas, como nomes, categorias, descrições e atributos detalhados sobre entidades do negócio.

Esquemas

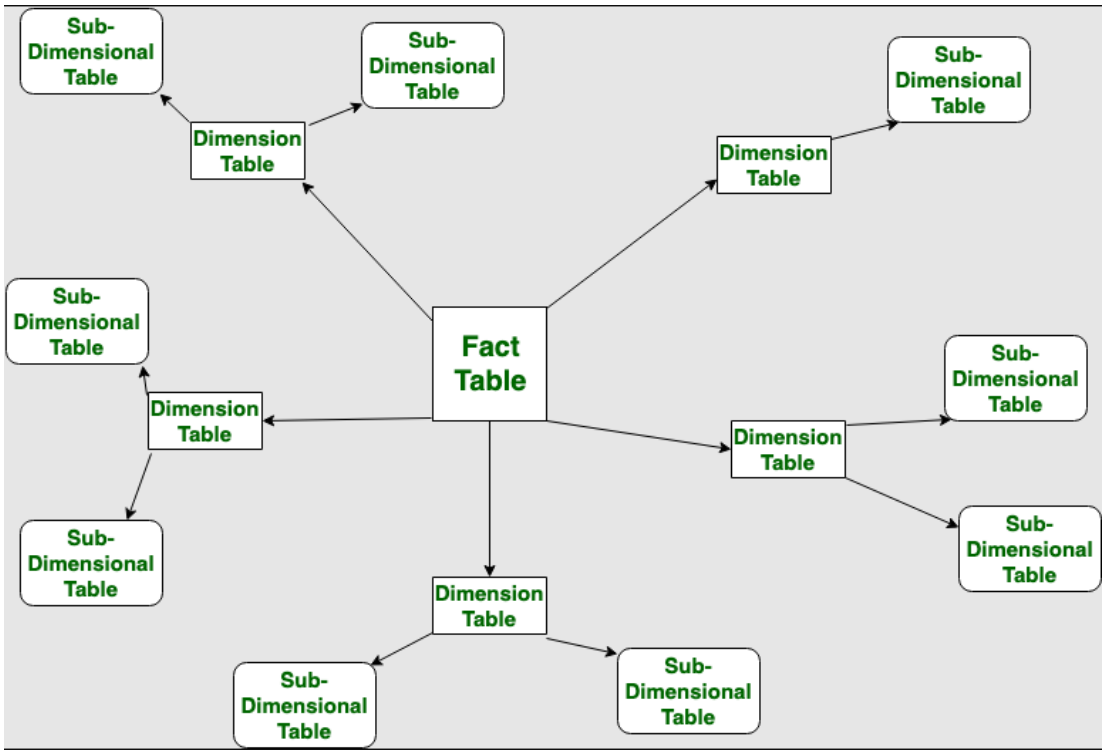
Os modelos dimensionais usam esquemas em estrela ou esquemas em floco de neve para representar esses relacionamentos de forma que as consultas analíticas sejam rápidas e fáceis de interpretar.

Modelo Estrela (Star Schema)

O modelo estrela é o mais simples e amplamente utilizado. Ele organiza os dados em uma tabela central (tabela de fatos), que é cercada por tabelas menores (tabelas de dimensões).

Modelo Floco de Neve (Snowflake Schema)

É uma extensão do modelo estrela, mas com as dimensões normalizadas em várias tabelas. Isso reduz redundâncias ao dividir as dimensões em várias tabelas menores relacionadas.


Sub-Dimensional Table
Sub-Dimensional Table
Dimension Table
Sub-Dimensional Table
Dimension Table
Sub-Dimensional Table
Fact Table
Sub-Dimensional Table
Dimension Table
Sub-Dimensional Table
Dimension Table
Sub-Dimensional Table
Sub-Dimensional Table
Dimension Table
Sub-Dimensional Table
Sub-Dimensional Table

.