Data Warehouse (DW):

Um Data Warehouse é como um grande depósito de informações em uma empresa. Ele coleta dados de muitas fontes e os organiza de forma que fiquem arrumados e prontos para serem usados. A ideia principal é ter todos os dados em um lugar só, para que as pessoas na empresa possam analisá-los e tomar decisões importantes. Com esses dados organizados, as empresas podem fazer relatórios, analisar questões complicadas e olhar para o que aconteceu no passado para entender como estão se saindo e quais são as tendências em seu negócio e poder tomar decisões mais assertivas.

Data Lake:

Um Data Lake é como um grande armazém que guarda muitos tipos diferentes de informações, como fotos, vídeos e documentos, do jeito que eles são, sem organização, incluindo aquelas que não são organizadas ou estruturadas. A principal ideia de um Data Lake é oferecer um lugar flexível para armazenar uma enorme quantidade de informações de muitas fontes diferentes sem precisar organizá-las antes. Isso é especialmente útil para informações "bagunçadas" como registros de servidores, dados de sensores, coisas de redes sociais e informações de máquinas. As empresas podem usar essas informações para fazer análises complicadas, como aprendizado de máquina e estudos de dados grandes, de forma mais rápida e fácil.

Diferenças entre Data Warehouse e Data Lake:

- **Estrutura dos Dados:** No Data Warehouse, os dados são submetidos a um processo de ETL (Extract, Transform, Load) antes de serem armazenados, o que envolve a transformação e estruturação. No Data Lake, os dados são armazenados em sua forma bruta, sem processamento prévio.
- **Escopo de Uso:** Data Warehouses são projetados principalmente para suportar relatórios e análises de negócios. Data Lakes são mais flexíveis e podem atender a uma ampla variedade de casos de uso, incluindo análises de big data, aprendizado de máquina e exploração de dados não estruturados.
- **Tempo de Processamento:** Data Warehouses são otimizados para consultas de alta velocidade devido à transformação prévia dos dados. Em contrapartida, o Data Lake pode demandar mais processamento durante as consultas, uma vez que a transformação ocorre posteriormente.
- **Custo e Complexidade:** A construção e manutenção de um Data Warehouse costumam ser mais dispendiosas e complexas devido ao processo de ETL. Os Data Lakes, por outro lado, podem ser mais econômicos na fase de ingestão, mas podem exigir mais esforço nas etapas de análise.

ETL vs. ELT:

- **ETL (Extract, Transform, Load):** No ETL, os dados são extraídos das fontes, transformados e depois carregados no Data Warehouse. A transformação ocorre antes do armazenamento, garantindo que os dados estejam prontos para análises.

- **ELT (Extract, Load, Transform):** No ELT, os dados são extraídos das fontes e carregados em seu estado bruto no Data Lake ou Data Warehouse. A transformação acontece posteriormente, conforme necessário. Isso oferece mais flexibilidade e a capacidade de transformar os dados de acordo com os requisitos específicos de análise.

Uso de Bancos de Dados Não-Relacionais:

Bancos de dados não relacionais, como bancos de dados NoSQL, são adequados para armazenar dados não estruturados ou semiestruturados. Eles podem ser usados em ambos os ambientes (Data Warehouse e Data Lake) para lidar com tipos de dados que não se encaixam bem em um modelo relacional tradicional. Isso inclui dados de documentos, dados de sensores, registros de eventos, registros de logs e outros tipos de informações não estruturadas.

Referências:

- TCC de Alexandre Tomás Hübner.
- TCC de Isabele Aurora Cândido Vitorino.
- "Data Warehouse passo a passo: o guia prático de como construir um Data Warehouse do zero" por Rafael Piton.