Data Warehouse (DW):

Conceito: Imagine um Data Warehouse como um grande arquivo ou depósito onde uma empresa guarda todos os tipos de informações importantes. Ele atua como um coletor de dados, reunindo informações de muitas fontes diferentes, como vendas, estoques, registros de clientes e outros. Em vez de deixar esses dados bagunçados, o Data Warehouse os organiza de uma forma especial para que fiquem mais fáceis de encontrar e usar.

Finalidade: O objetivo principal de um Data Warehouse é fornecer um local centralizado e organizado para todos os dados da empresa. Isso é útil para que as pessoas na empresa possam analisar esses dados e tomar decisões importantes com base neles. Quando os dados estão bem organizados, a empresa pode criar relatórios detalhados, responder a perguntas difíceis e até mesmo olhar para o que aconteceu no passado para entender como estão se saindo e quais direções o negócio está tomando. Isso ajuda a empresa a tomar decisões mais inteligentes e informadas, que podem levar a um melhor desempenho e ao entendimento das tendências no mercado.

Data Lake:

Conceito: Pense em um Data Lake como um grande armazém ou depósito onde as empresas armazenam todos os tipos de informações, como fotos, vídeos e documentos, sem se preocupar em organizá-los de antemão. Ele guarda essas informações do jeito que são, incluindo aquelas que não têm uma estrutura organizada, como mensagens de redes sociais, registros de servidores e dados de sensores.

Finalidade: A principal ideia de um Data Lake é oferecer um lugar flexível para armazenar uma enorme quantidade de informações de diversas fontes, sem a necessidade de organizá-las antes de guardá-las. Isso é especialmente útil para informações que podem ser confusas, desorganizadas ou sem um formato pré-definido, como registros de servidores, dados de sensores, publicações em redes sociais e informações de máquinas. As empresas podem usar essas informações para realizar análises avançadas, como aprendizado de máquina e estudos de dados grandes, de maneira mais rápida e fácil. O Data Lake serve como um grande reservatório de informações que podem ser exploradas e utilizadas para obter insights valiosos e tomar decisões informadas.

Diferenças entre Data Warehouse e Data Lake:

- **Estrutura dos Dados:** Em um Data Warehouse, os dados passam por um processo de ETL (Extract, Transform, Load) antes de serem armazenados. Isso significa que os dados são extraídos de várias fontes, depois são transformados e finalmente carregados no Data Warehouse. Durante a transformação, os dados são organizados em uma estrutura específica, o que envolve limpeza, padronização e, muitas vezes, agregação. Isso torna os dados prontos para consultas e relatórios analíticos. Em um Data Lake, os dados são armazenados em

sua forma bruta e original, sem processamento prévio. Isso significa que os dados são armazenados exatamente como são gerados ou coletados, sem qualquer alteração. Isso oferece flexibilidade para armazenar dados em qualquer formato, incluindo dados não estruturados, semiestruturados e estruturados. A transformação e estruturação ocorrem posteriormente, conforme necessário, durante a análise.

- **Escopo de Uso:** Data Warehouses são projetados principalmente para dar suporte a relatórios e análises de negócios. Eles são otimizados para consultas analíticas complexas e relatórios padronizados. Os Data Warehouses são ideais para responder a perguntas específicas sobre o desempenho de negócios, como vendas, estoques e tendências de mercado. Data Lakes são mais flexíveis e podem atender a uma ampla variedade de casos de uso. Eles são especialmente adequados para análises de big data, aprendizado de máquina e exploração de dados não estruturados. Isso permite que as empresas realizem análises mais avançadas, como processamento de linguagem natural, análise de sentimentos e análise de registros de servidores

- **Tempo de Processamento:** Data Warehouses são otimizados para consultas de alta velocidade devido à transformação prévia dos dados. Isso significa que os dados estão prontos para análises instantâneas, tornando os relatórios de negócios mais ágeis e responsivos. Em contrapartida, o Data Lake pode demandar mais processamento durante as consultas, uma vez que a transformação ocorre posteriormente. Isso significa que as análises podem levar mais tempo para serem concluídas, pois a transformação ocorre sob demanda. No entanto, a flexibilidade do Data Lake permite que as empresas analisem dados em sua forma bruta, o que pode ser valioso para cenários de big data.

- **Custo e Complexidade:** A construção e manutenção de um Data Warehouse tendem a ser mais dispendiosas e complexas devido ao processo de ETL, que envolve a extração, transformação e carregamento dos dados. Além disso, a estruturação prévia dos dados requer investimentos em tempo e recursos. Os Data Lakes podem ser mais econômicos na fase de ingestão de dados, uma vez que os dados são armazenados em sua forma bruta, sem a necessidade de uma transformação prévia intensiva. No entanto, os custos podem aumentar à medida que as empresas realizam análises mais avançadas, exigindo esforços para organizar e estruturar os dados conforme necessário.

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.