INTRODUÇÃO AO DATA LAKE

Alysson José Franca Ferreira, Anderson Richard da Silva, Jean Arthur Santos Almeida e Sylvio Cézar Rezende Pereira

UNIVAS
2024

1. INTRODUÇÃO

Nos últimos anos, o conceito de Big Data tem ganhado destaque como uma abordagem poderosa para o armazenamento, processamento e análise de grandes volumes de dados. Nesse contexto, a estrutura de Data Lake surgiu do desafio de se agrupar dados de diversas fontes e formatos em um único local, com o objetivo de utilizá-los futuramente para as mais diversas finalidades. Essa abordagem contrasta com o modelo tradicional de data warehouse, oferecendo maior flexibilidade, escalabilidade e capacidade de lidar com dados estruturados, semiestruturados e não estruturados.

Este artigo tem como objetivo explorar os fundamentos do data lake, desde sua definição e conceitos básicos até sua diferenciação em relação aos data warehouses, destacando quem deve utilizar essa tecnologia e quais os benefícios oferecidos por essa estrutura de dados. Ao compreender esses aspectos, as organizações podem tomar decisões mais informadas sobre como gerenciar e utilizar seus dados de forma eficaz para impulsionar a inovação e a tomada de decisões estratégicas.

2. FUNDAMENTAÇÃO

2.1 Definição e conceitos fundamentais

O termo data lake que em português significa lago de dados foi criado pelo CTO da Pentaho James Dixon. Descreve-se este tipo de repositório como um lago porque ele armazena um conjunto de dados sem qualquer tipo de alteração neste caso em seu estado natural, como um corpo d'água que não foi filtrado ou contido. Os dados fluem de diversas fontes onde são consolidadas para o Data Lake, sendo armazenados no formato original.

Sendo assim um data lake é um repositório centralizado que ingere e armazena grandes volumes de dados em sua forma original. Estes dados podem ser processados e usados para diversas necessidades. Devido à sua arquitetura aberta e escalonável, um data lake pode acomodar todos os tipos de dados de qualquer fonte, desde dados estruturados (tabelas de banco de dados, planilhas do Excel) até semiestruturados (arquivos XML, páginas da Web) e não estruturados (imagens, arquivos de áudio, tweets), tudo sem sacrificar a fidelidade.

Os arquivos com os dados normalmente são armazenados em zonas preparadas (brutos, limpos e coletados) para que diferentes tipos de usuários possam usar os dados em suas várias formas para atender às suas necessidades. Os data lakes fornecem consistência de dados em uma variedade de aplicativos, habilitando a análise de Big Data, o aprendizado de máquina, a análise preditiva e outras formas de ação inteligente.

Os dados são referidos como brutos quando eles ainda não foram processados para uma finalidade específica. Os dados em um Data Lake são definidos só após serem consultados. Os cientistas de dados possuem acesso às informações brutas por meio de modelagem preditiva ou ferramentas analíticas mais avançadas.

Data lakes modernos funcionam em três características principais:

- Uma zona de pouso (landing zone) para os dados brutos
- Uma zona preparatória (staging zone) na qual os dados são transformados tendo em mente um objetivo analítico
- Uma zona de exploração de dados (data exploration zone) onde os dados são utilizados por aplicativos e funções analíticas e alimentam os modelos de Machine Learning.

Sendo assim, a partir do data lake todas as informações são fornecidas a várias fontes, como funções analíticas e outros aplicativos de negócios, ou a ferramentas de Machine Learning para análise adicional.

2.2 Quem deve utilizar o Data Lake

Um Data Lake é ideal para organizações que precisam armazenar, processar e proteger grandes quantidades de dados. Antes de adotar essa abordagem, é

importante considerar os tipos de dados que sua organização lida, bem como os objetivos desejados. A complexidade do processo de aquisição e a organização dos dados também devem ser levadas em conta. Em resumo, o Data Lake oferece agilidade, escalabilidade e eficiência na gestão de dados, tornando possível extrair insights valiosos de forma rápida e econômica em comparação com o tradicional Data Warehouse.

Os data lakes ajudam em:

1. Simplificar o Gerenciamento de Dados
2. Acelerar Análises
3. Reduzir o custo total da propriedade

Os responsáveis que geralmente utilizam um data lake:

- Cientistas de Dados: Explora e analisa. grandes volumes de dados brutos, um exemplo seria usar as arquiteturas de um DL para acelerar a pesquisa e desenvolvimento de novos produtos de uma farmacêutica, analisando os vastos dados feitos por testes
- Analista de negócios: Ajuda na identificação de ineficiências operacionais e oportunidades de melhoria, permitindo uma análise aprofundada dos processos internos da empresa
- Profissionais de marketing: Analisam os padrões de compra dos usuários, otimizando as recomendações de produtos e personalizando as experiências de compra.
- Organizações com grandes volumes de dados: Organizações que precisam armazenar dados não estruturados ou semiestruturados para análise futura.

2.3 Diferença entre Data Lake e Data Warehouse

Enquanto um Data Warehouse armazena dados estruturados, um Data Lake é um repositório centralizado que permite armazenar qualquer dado em qualquer escala. Ou seja, por mais que sejam similares, possuem alguns pontos chave de diferença entre si, dentre os quais se destacam:

1. **Armazenamento de dados:**

a. DW: Contém dados estruturados que foram limpos e processados, prontos para análise estratégica com base em necessidades comerciais pré-definidas.

b. DL: Contém todos os dados de uma organização em uma forma bruta e não estruturada, podendo armazenar os dados indefinidamente - para uso imediato ou futuro.

2. **Usuários**:

a. DW: Os dados são normalmente acessados por gerentes e usuários comerciais que buscam obter insights de KPIs comerciais, pois os dados já foram estruturados para fornecer respostas a perguntas pré-determinadas para análise.

b. DL: Os dados são normalmente usados por cientistas de dados e engenheiros que preferem estudar dados em sua forma bruta para obter novos insights comerciais únicos.

3. **Análise:**

a. DW: Visualização de dados, BI, análise de dados.

b. DL: Análise preditiva, aprendizado de máquina, visualização de dados, BI, análise de big data.

4. **Esquema:**

a. DW: Definido antes do armazenamento dos dados. Isso alonga o tempo necessário para processar os dados, mas uma vez concluído, os dados estão prontos para uso consistente e confiável em toda a organização.

b. DL: Definido após o armazenamento dos dados, tornando o processo de captura e armazenamento dos dados mais rápido.

5. **Processamento:**

a. DW: ETL (Extrair, Transformar, Carregar). Os dados são extraídos de suas fontes(s), limpos e estruturados para que estejam prontos para análise de negócios.

b. DL: ELT (Extrair, Carregar, Transformar). Os dados são extraídos de sua fonte para armazenamento no data lake e estruturados apenas quando necessário.

6. **Custo:**

a. DW: Custam mais e exigem mais tempo para serem gerenciados, resultando em custos operacionais adicionais.

b. DL: Têm custos de armazenamento mais acessíveis e menos demorados para gerenciar, o que reduz os custos operacionais.

7. **Flexibilidade:**

a. DW: Menos flexível, pois é otimizado para armazenar dados estruturados e relacionais.

b. DL: Mais flexível, pois pode armazenar qualquer tipo de dado, incluindo dados estruturados, semiestruturados e não estruturados.

8. **Escalabilidade:**

a. DW: Também pode ser escalado, mas geralmente requer mais esforço e planejamento.

b. DL: Altamente escalável, pois pode lidar com grandes volumes de dados de diferentes fontes.

Em suma, DL e DW são duas alternativas diferentes para armazenamento e processamento de dados, cada uma com suas peculiaridades. Portanto, a escolha entre qual usar depende dos requisitos de negócios, da natureza dos dados e dos casos de uso específicos.

2.4 Os Benefícios de uma Estrutura de Data Lake

Conforme exposto anteriormente, em termos de estrutura, usabilidade e diferenças em relação aos Data Warehouses pode se destacar vários benefícios tais quais:

1. **Integração de Dados**: Data Lakes permitem que as organizações integrem diferentes tipos de dados e fontes de dados em um único repositório. Isso facilita a análise de dados correlacionados, mesmo que venham de fontes diferentes. Isso pode levar a uma visão mais completa do negócio e a uma melhor tomada de decisões.
2. **Flexibilidade de Dados**: Ao contrário dos sistemas tradicionais de gerenciamento de dados, os Data Lakes não exigem que os dados sejam processados antes de serem armazenados. Isso significa que você pode armazenar todos os seus dados, estruturados e não estruturados, sem a necessidade de saber como eles serão usados no futuro. Isso permite que as

organizações armazenem uma variedade mais ampla de dados, desde logs de servidores brutos, feeds de mídia social até dados de transações financeiras.

3. **Escalabilidade**: Data Lakes são projetados para fornecer armazenamento de baixo custo e alta capacidade. Isso permite que as organizações escalem seus sistemas de dados conforme necessário e de maneira econômica. Isso é especialmente útil dado que a quantidade de dados gerados está crescendo exponencialmente.
4. **Análise Avançada**: Com todos os seus dados armazenados em um Data Lake, as organizações têm a capacidade de realizar análises mais profundas. Isso inclui machine learning, previsão de dados, análise de dados em tempo real e muito mais. Isso pode levar a insights mais profundos e a uma melhor tomada de decisões.
5. **Governança e Segurança**: Embora a governança e a segurança sejam desafios para qualquer sistema de dados, os Data Lakes, mais especificamente os em nuvem, permitem uma governança de dados mais granular. Isso ocorre porque os dados podem ser armazenados em seu formato original, com segurança e controle de acesso em nível de dados. Isso permite que as organizações atendam aos requisitos regulatórios e de conformidade, ao mesmo tempo em que protegem seus dados.

3. CONCLUSÃO

Dadas as características de um Data Lake, que armazena de forma unificada diversos formatos de dados em seu estado bruto, sejam eles estruturados ou não, e os disponibilizam para uma ampla gama de usuários que vai desde técnicos, como desenvolvedores e cientistas de dados, à usuários de negócios como analistas e profissionais de marketing.

Essa amplitude só é possível devido às diferenças em relação aos Data Warehouses das quais se destacam a flexibilidade em relação aos tipos de dados e o processamento desses que ocorrem no momento do uso, o que traz diversos benefícios em termos de análises avançadas, integração, escalabilidade e de governança e segurança.

REFERÊNCIAS

Data Warehouse vs. Data lake data mart – Comparação entre soluções de armazenamento em nuvem – AWS. Disponível em: <https://aws.amazon.com/pt/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/>. Acesso em: 12 abr. 2024.

Data Lake vs Data Warehouse: 6 Key Differences. Disponível em: <https://www.qlik.com/us/data-lake/data-lake-vs-data-warehouse#:~:text=A%20data%20lake%20is%20a>. Acesso em: 12 abr. 2024.

Data Lake (DL) and Data Warehouse (DW) - SquareOne Technologies. Disponível em: <https://www.squareonemea.com/blogs/data-lake-dl-and-data-warehouse-dw/>. Acesso em: 12 abr. 2024.

Soluções de data lake storage | IBM. Disponível em: <https://www.ibm.com/br-pt/data-lake>. Acesso em: 12 abr. 2024.

O que é data lake? | SAP Insights. Disponível em: <https://www.sap.com/brazil/products/technology-platform/hana/what-is-a-data-lake.html>. Acesso em: 12 abr. 2024.

TOTVS, E. Data lake: o que é, vantagens, exemplos de aplicação e desafios. Disponível em: <https://www.totvs.com/blog/inteligencia-de-dados/data-lake/>. Acesso em: 12 abr. 2024.

O que é data lake? Disponível em: <https://www.redhat.com/pt-br/topics/data-storage/what-is-a-data-lake>. Acesso em: 12 abr. 2024.

O que é um data lake? — Introdução aos data lakes e análises — AWS. Disponível em: <https://aws.amazon.com/pt/what-is/data-lake/>.

O que é data lake? Disponível em: <https://cloud.google.com/learn/what-is-a-data-lake?hl=pt-br#section-3>. Acesso em: 14 abr. 2024.