Data Lakehouse

Data Lakes, como um conceito moderno de armazenamento de dados em formato nativo, foi apresentado como um substituto para *Data Warehouses*. No entanto, com o passar do tempo, ficou claro que *Data Lakes* pode ser utilizado em complementação ao *Data Warehouses* , não como substituto (OREŠČANIN e HLUPIĆ, 2021).

Data Warehouses (DW) surgiram com o intuito de auxiliar usuários de negócio a obterem *insights* analíticos através da centralização de dados provenientes de diversas fontes, principalmente de bancos de dados operacionais. Os dados em um DW são utilizados como suporte à tomada de decisão por meio de aplicações de *Business Intelligence* (BI). Os dados em um DW são armazenados em uma estrutura fixa, em *schemas* pré-definidos (*schema-on-write*), de modo a garantir segurança e performance. Este tipo de solução é chamado de plataforma de análise de dados de **primeira geração** (ARMBRUST, GHODSI, XIN e ZAHARIA, 2021).

Com o surgimento do *Big Data* (grande volume de dados, em grande velocidade e variedade de fontes e múltiplos formatos), os *Data Warehouses* passaram a não conseguir suprir as necessidades de algumas organizações, surgindo então a **segunda geração** de plataforma de análise de dados, os *Data Lakes* (ARMBRUST, GHODSI, XIN e ZAHARIA, 2021). *Data Lakes* utilizam formas de armazenamento de baixo custo (HDFS do Apache Hadoop, por exemplo) e, geralmente, com formatos de armazenamento abertos, como Apache Parquet e ORC. *Data Lakes* possuem arquitetura *schema-on-read*, ou seja, os dados não são armazenados em estruturas rígidas pré-definidas, e sim em seu formato original. Fornecendo flexibilidade e agilidade no armazenamento e acesso aos dados. Porém, por outro lado, a flexibilidade trazida por *Data Lakes* traz problemas de qualidade e governança de dados, como já apontado por Khine e Wang (2018) e Sawadogo e Darmont (2021).

De 2015 em diante surgiu uma nova forma de organização das plataformas de dados, com a utilização complementar de *Data Lakes* e *Data Warehouses*. *Data Lakes* em nuvem, como S3, ADLS e GCS têm substituído o Apache HDFS, com o armazenamento posterior de parcela dos dados em *Data Warehouses*, como Redshift e Snowflake (ARMBRUST, GHODSI, XIN e ZAHARIA, 2021). **Esta arquitetura de duas camadas, *Data Lake* + *Data Warehouse* é, segundo Armbrust, Ghodsi, Xin e Zaharia (2021), dominante no mercado (aparentemente utilizado por todas as empresas do ranking Fortune 500).**

Esta arquitetura de duas camadas possui alta complexidade de manutenção. Primeiramente os dados são carregados em um *Data Lake*, e depois em um *Data Warehouse*, exigindo portanto duas etapas de ETL até a sua disponibilização para análise por usuários de BI. A nova arquitetura, portanto, aumentou a complexidade, os atrasos e os pontos de falha se comparada à estratégia de inclusão de dados diretamente no *Data Warehouse*, por exemplo (ARMBRUST, GHODSI, XIN e ZAHARIA, 2021).

Diante dos problemas ainda não solucionados por *Data Warehouses* e/ou *Data Lakes*, surge um novo conceito de arquitetura denominado *Data Lakehouse*, com o objetivo de superar tais problemas (OREŠČANIN e HLUPIĆ, 2021).

Um *Data Lakehouse* busca combinar as vantagens das arquiteturas presentes em *Data Lakes* e *Data Warehouses*, contendo tanto estruturas de dados pré-definidas (*schema-on-write*), quanto não pré-definidas (*schema-on-read*) (OREŠČANIN e HLUPIĆ, 2021).

Conceitualmente, um *Data Lakehouse* promete conciliar as demandas de armazenamento de dados heterogêneos de modo eficiente e econômico, permitindo o acesso aos mesmos por meio de diversos tipos de aplicações, desde aquelas baseadas em aprendizado de máquina àquelas baseadas em SQL. De modo geral, *Data Lakehouses* buscam solucionar os problemas de confiabilidade e obsolescência de dados, suporte limitado a análises avançadas e diminuição do custo total de propriedade das arquiteturas de dados (BEGOLI, GOETHERT E KNIGHT, 2021).

Definição

Armbrust, Ghodsi, Xin e Zaharia (2021) definem um *Data Lakehouse* como um sistema de gerenciamento de dados baseado em armazenamento de baixo custo e de acesso direto, que também fornece recursos de desempenho e gerenciamento analíticos tradicionais de SGBDs, como transações ACID, controle de versão de dados, auditoria, indexação, armazenamento em cache e otimização de consultas. Deste modo, *Lakehouses* combinam os principais benefícios de *Data Lakes* e *Data Warehouses*: armazenamento de baixo custo em formatos abertos e acessíveis, e poderosos recursos de gerenciamento e otimização. A questão-chave é se é possível combinar esses benefícios de maneira eficaz: em particular, o suporte de *Lakehouses* para acesso direto aos dados significa que eles abrem mão de alguns aspectos de independência de dados, que tem sido um dos pilares do design de SGBDs relacionais.

Begoli, Goethert e Knight (2021) afirmam que o conceito de *Data Lakehouse* ainda é vago, não materializado. De acordo com os autores, não há ainda uma solução completa definitiva, “fora da prateleira”, para o conceito de *Data Lakehouse*. Ou seja, é ainda um conceito em evolução, que parte das seguintes premissas:

- *Data Lakes* oferecem suporte a uma gestão de dados eficiente;
- Formatos abertos (Apache Parquet, OCR, etc.) provêem suporte a ferramentas e bibliotecas de inteligência artificial / aprendizado de máquina;
- O estado da arte em performance SQL pode ser reproduzido sobre formatos abertos.

Direcionamento da indústria

Armbrust, Ghodsi, Xin e Zaharia (2021) argumentam que a arquitetura de *Data Warehouse* como a conhecemos hoje tende a perder importância nos próximos anos, e deverá ser substituída por um novo padrão arquitetônico, o *Data Lakehouse* que:

- Será baseado em formatos de dados abertos de acesso direto, como Apache Parquet;
- Terá suporte de primeira classe para aprendizado de máquina e ciência de dados;

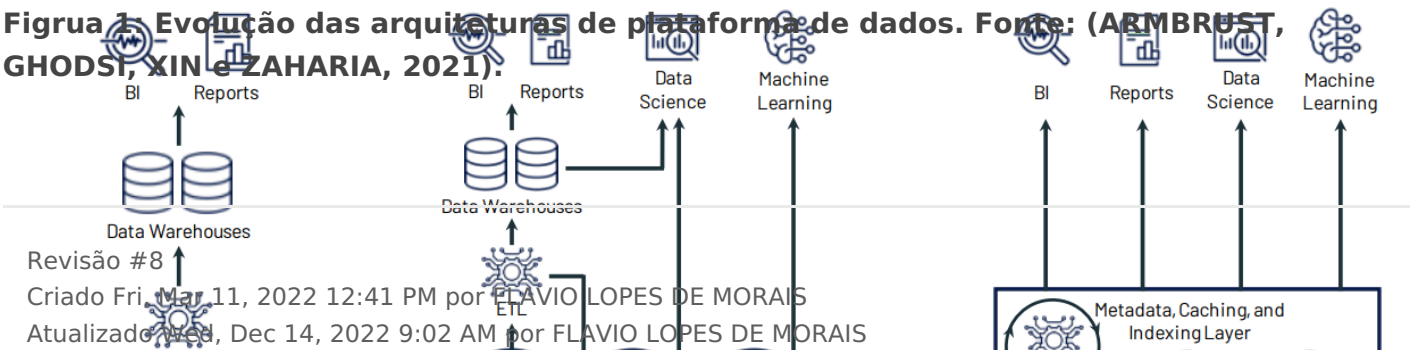
- Oferecerá desempenho de última geração.

De acordo com Armbrust, Ghodsi, Xin e Zaharia (2021), *Data Lakehouses* podem ajudar a enfrentar vários desafios importantes enfrentados atualmente com Data Lakes e *Data Warehouses* , incluindo:

- Dados desatualizados;
- Confiabilidade;
- Custo total de propriedade;
- Aprisionamento de dados;
- Suporte limitado a casos de uso específicos.

Armbrust, Ghodsi, Xin e Zaharia (2021) concluem que a indústria tende a convergir para a adoção de *Data Lakehouses* devido à vasta quantidade de dados já depositados atualmente em *Data Lakes*, e devido à grande oportunidade de simplificação das suas arquiteturas de dados.

Figrua 1: Evolução das arquiteturas de plataforma de dados. Fonte: (ARMBRUST, GHODSI, XIN e ZAHARIA, 2021).



Revisão #8
Criado Fri, Mar 11, 2022 12:41 PM por FLAVIO LOPES DE MORAIS
Atualizado Wed, Dec 14, 2022 9:02 AM por FLAVIO LOPES DE MORAIS