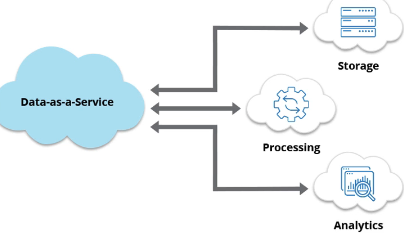
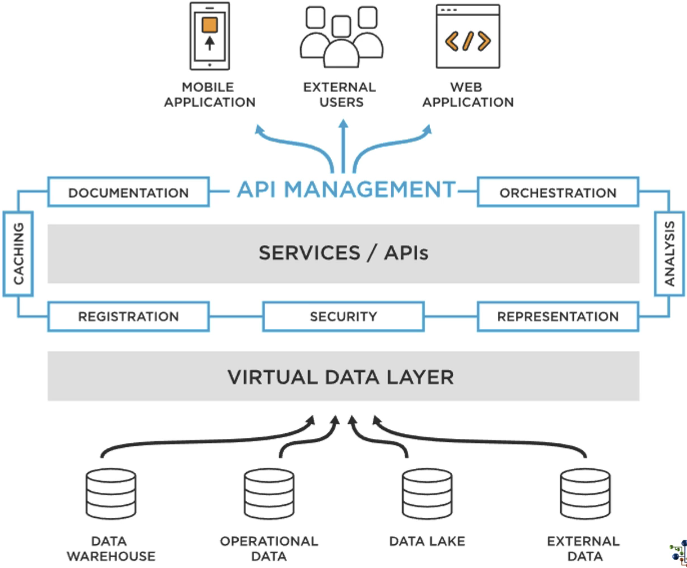
**DATA AS SERVICE (DaaS)**



* DaaS é uma estratégia de gerenciamento de dados que visa alavancar os dados como um ativo de negócios para maior agilidade no processo de análise
* Faz parte das ofertas “as a service” que se tornaram cada vez mais populares desde a expansão da internet nos anos 1990, que começou com a introdução do Software as a Service (SaaS)
* Semelhante a outros modelos “como serviço”, o DaaS fornece uma maneira de gerenciar as grandes quantidades de dados que as organizações geram todos os dias e fornecer essas informações valiosas em todas a empresa para a tomada de decisões baseadas em dados

**ARQUITETURA DaaS**

* A arquitetura DaaS (Data as a Service) se concentra no provisionamento de dados de uma variedade de fontes sob demanda por meio do uso de APIs
* Projetado para simplificar o acesso aos dados, o DaaS oferece conjuntos de dados já tratados ou fluxos de dados para serem consumidos em uma variedade de formatos, geralmente unificados usando virtualização de dados
* Na verdade, uma arquitetura DaaS pode incluir uma variedade de tecnologias de gerenciamento de dados, incluindo virtualização de dados, serviços de dados, análise de autoatendimento (Self-Service Analytics) e catalogação de dados

**PRINCIPAIS BENEFÍCIOS DE DaaS**

* Monetização de dados
* Redução de Custos
* Caminho mais rápido para inovação
* Agilidade no processo de decisão baseado em dados
* Menor risco no uso de dados
* Criação de uma cultura Data-Driven

**Arquiteturas Modernas de Big Data**

À medida que Big Data amadurece e ganha cada vez mais espaço nas empresas, novas

arquiteturas surgem com o objetivo de tornar o acesso aos dados mais fácil, eficiente, seguro e amplamente disponível.

Como vimos nas aulas anteriores neste capítulo, uma tendência é tornar os dados

disponíveis como serviço e algumas arquiteturas modernas

de Big Data estão sendo propostas para este fim.

As duas últimas tendências em arquiteturas de plataforma de dados emergentes são o

Data Lakehouse e o Data Mesh

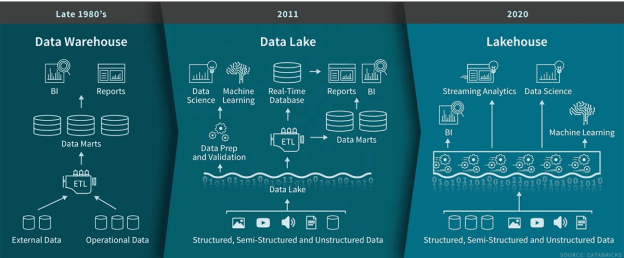
.

**Data LakeHouse**

Um Data Lakehouse é uma nova arquitetura de gerenciamento de dados que combina a flexibilidade, economia e escala de Data Lakes com o gerenciamento de dados e transações ACID (Atomicidade, Consistência, Isolamento, Durabilidade) de Data Warehouses, permitindo Business Intelligence (BI) e Machine Learning (ML) em todos dados armazenados em um único repositório.

Data Lakehouses são habilitados por um novo design de sistema aberto: implementação de estruturas de dados e recursos de gerenciamento de dados semelhantes aos de um Data Warehouse, diretamente no tipo de armazenamento de baixo custo usado para Data Lakes. Mesclá-los em um único sistema significa que as equipes de dados podem se mover mais rapidamente, pois podem usar os dados sem a necessidade de acessar vários sistemas.

Os Data Lakehouses também garantem que as equipes tenham os dados mais completos e atualizados disponíveis para projetos de ciência de dados, aprendizado de máquina e análise de negócios.

****

O termo Data Lakehouse foi definido pela Databricks (plataforma que fornece infraestrutura para processamento de grandes volumes de dados).

Abaixo o paper de pesquisa que originou o termo: Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

<https://databricks.com/wp-content/uploads/2020/12/cidr_lakehouse.pdf>

**Data Mesh**

Conforme definido pela primeira vez por Zhamak Dehghani, consultora da ThoughtWorks e a arquiteta original do termo, Data Mesh é um tipo de arquitetura de plataforma de dados que abrange a onipresença dos dados na empresa, permitindo um design orientado ao domínio e de autoatendimento. Data Mesh é amplamente considerado a próxima grande mudança arquitetônica em dados.

Data Mesh é uma nova abordagem para projetar e desenvolver arquiteturas de dados. Ao contrário de uma arquitetura centralizada e monolítica baseada em um Data Warehouse e/ou um Data Lake, Data Mesh é uma arquitetura de dados altamente descentralizada.

Data Mesh tenta resolver três desafios quando temos um Data Lake/Warehouse centralizado:

• Falta de propriedade. Quem é o proprietário dos dados - a equipe da fonte de dados ou a equipe de infraestrutura?

• Falta de qualidade. A equipe de infraestrutura é responsável pela qualidade, mas não conhece bem os dados.

• Escalonamento organizacional. O armazenamento central torna-se o gargalo, como no caso de um Data Lake/Warehouse empresarial.

O objetivo com Data Mesh é tratar os dados como um produto, com cada fonte tendo seu próprio gerente/proprietário de produto de dados (que fazem parte de uma equipe multifuncional de Engenheiros de Dados) e sendo seu próprio domínio claramente focado e com uma oferta autônoma, tornando-se os blocos de construção fundamentais de uma malha (Mesh), levando a uma arquitetura distribuída orientada por domínio.

Observe que, por motivos de desempenho, você pode ter um domínio que agrega dados de várias fontes. Cada domínio deve ser detectável, endereçável, autoexplicativo, seguro (governado por controle de acesso global), confiável e interoperável (governado por um padrão aberto). Cada domínio armazenará seus dados em um Data Lake e, em muitos casos, também terá uma cópia de alguns dos dados em um banco de dados relacional.

Outro componente do Data Mesh é a infraestrutura de dados como plataforma, que fornece armazenamento, pipeline, catálogo de dados e controle de acesso aos domínios. A ideia principal é evitar a duplicação de esforços. Isso permitirá que cada equipe de produto de dados crie seus produtos de dados rapidamente.

Vale ressaltar que Data Mesh ainda é uma tendência e sua implementação tem diversos desafios técnicos. Mas toda tecnologia de sucesso um dia foi uma tendência. Como dizia Bill Gates, fundador da Microsoft: O segredo do sucesso em tecnologia está em reconhecer o potencial de uma tendência no futuro e chegar lá primeiro.

**Data Mesh Como Paradigma de Arquitetura de Dados**

Data Mesh é um paradigma arquitetônico e organizacional que desafia a antiga suposição de que devemos centralizar os dados para usá-los, ter todos os dados em um só lugar ou ter os dados gerenciados por uma equipe de dados centralizada para agregar valor.

Para o Big Data fomentar a inovação, sua propriedade deve ser federada entre os proprietários de dados que são responsáveis por fornecer seus dados como produtos (com o suporte de uma plataforma de dados de autoatendimento para abstrair a complexidade técnica envolvida em servir produtos de dados). Também devemos adotar uma nova forma de governança federada por meio da automação para permitir a interoperabilidade de produtos de dados orientados a domínio. A descentralização, junto com a interoperabilidade e o foco na experiência dos consumidores de dados, são fundamentais para a democratização da inovação usando dados.

Se uma organização tem muitos domínios com vários sistemas e equipes gerando dados ou um conjunto diversificado de casos de uso e padrões de acesso orientados a dados, o uso de Data Mesh pode ser uma opção viável. A implementação de Data Mesh requer investimento na construção de uma plataforma de dados de autoatendimento e adoção de uma mudança organizacional para domínios a fim assumir a propriedade de longo prazo dos produtos de dados, bem como uma estrutura de incentivos que recompensa domínios que servem e utilizam dados como um produto.

Data Mesh marca uma mudança bem-vinda de paradigma arquitetônico e organizacional em como gerenciamos Big Data. O paradigma é baseado em quatro princípios:

(1) Descentralização orientada ao domínio da propriedade e arquitetura de dados;

(2) Dados orientados ao domínio servidos como um produto;

(3) Infraestrutura de dados de autoatendimento como uma plataforma para habilitar equipes de dados autônomas e orientadas para o domínio;

(4) Governança federada para permitir ecossistemas e interoperabilidade.

Há ainda uma grande lacuna de ferramentas comerciais para acelerar a implementação de Data Mesh e hoje o que vemos é a implementação de blocos isolados que então podem ser conectados criando assim uma grande malha de dados e tornando esse tipo de arquitetura uma realidade. Teremos ainda muita evolução e novas ferramentas surgindo nos próximos anos.

Arquitetos de Dados, Engenheiros de Dados e Engenheiros DataOps são os profissionais que provavelmente participarão de projetos para implementar esse tipo de arquitetura.

**Soluções Comerciais**

Vemos atualmente ocorrer com as arquiteturas modernas de Big Data, o mesmo fenômeno que ocorreu cerca de 10 anos atrás com o próprio Big Data: uma explosão de soluções comerciais que facilitam a implementação das arquiteturas. Aqui vão algumas dicas.

Para o Data Lakehouse, a solução mais comum é o Databricks. A DSA foi pioneira no Brasil ao trazer projetos de Databricks em nossos treinamentos e o último curso da Formação Engenheiro de Dados terá diversos novos projetos sobre a plataforma. Conheça o Databricks: <https://databricks.com/>

Data Mesh é um conceito de arquitetura e sua implementação pode envolver diversas tecnologias. Considerando o ambiente de Cloud Computing do Microsoft Azure, o Azure Purview seria seu ponto de partida para descobrir dados. Se você precisar fazer consultas entre domínios, também chamadas de consultas federadas, use o Synapse sem servidor com o Azure Virtual Network Peering se estiver consultando dados de contas de armazenamento (vinculando as contas de armazenamento em cada espaço de trabalho do Synapse). Se consultar dados de pools dedicados relacionais do Synapse, isso exigiria atualmente trabalho extra, como usar notebooks Synapse Spark, Databricks, Power BI ou fluxos de dados do Azure Data Factory para chamar vários bancos de dados hospedados em pools dedicados separados (mas há soluções mais fáceis).

As Formações Engenheiro de Dados e Arquiteto de Dados fornecem o conhecimento necessário para essas implementações.

REFERÊNCIAS 👍

Data Warehouses Meet Data Lakes

<https://databricks.com/p/whitepaper/data-warehouses-meet-data-lakes>

Data Lakehouse

<https://databricks.com/glossary/data-lakehouse>

Data Mesh

<https://www.thoughtworks.com/radar/techniques/data-mesh>

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

<https://databricks.com/wp-content/uploads/2020/12/cidr_lakehouse.pdf>

How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh

<https://martinfowler.com/articles/data-monolith-to-mesh.html>