PUC Minas Virtual

Arquiteturas e Serviços de Data Lakes e Data Warehousing

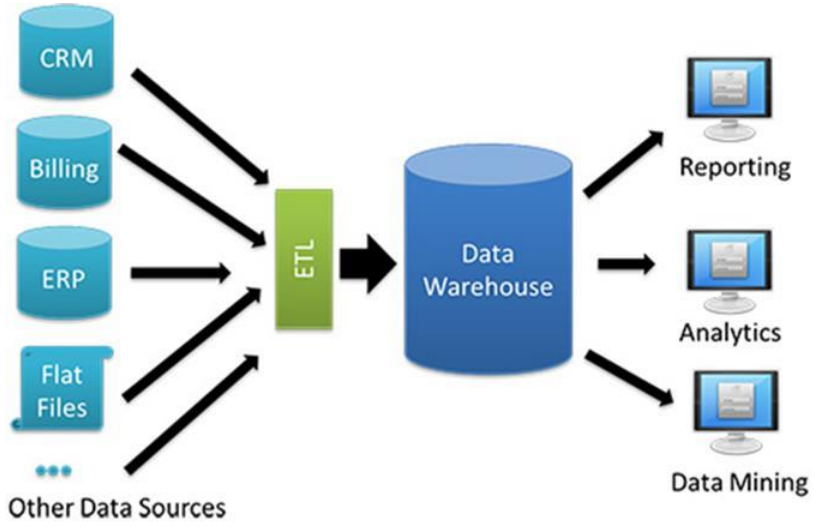
PUC Minas
Virtual

Unidade 3 – Data Lake

O que é um Data Lake

Data Warehouse

- Um Data Warehouse é um repositório central de informações que podem ser analisadas para tomar decisões mais informadas.
- Os dados fluem para um Data Warehouse a partir de sistemas transacionais, bancos de dados relacionais e outras fontes, normalmente em uma cadência regular.
- Analistas de negócios, engenheiros de dados, cientistas de dados e tomadores de decisão acessam os dados por meio de ferramentas de Business Intelligence (BI), clientes SQL e outros aplicativos de análise.

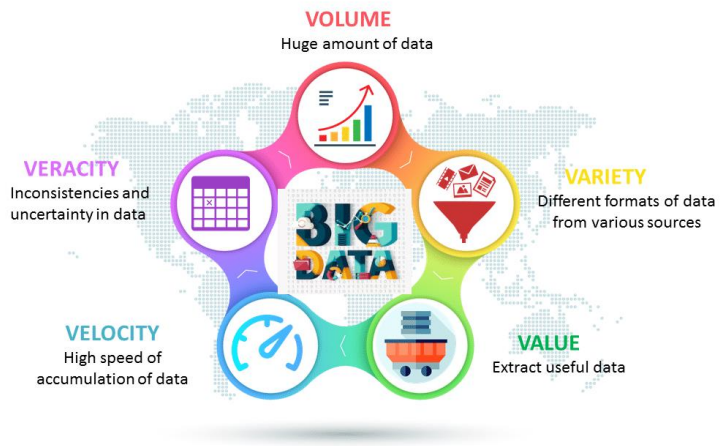


https://www.astera.com/type/blog/data-warehouse-definition/

Big Data

A Forbes relata que a cada minuto, os usuários assistem a 4,15 milhões de vídeos do YouTube, enviam 456.000 tweets no Twitter, postam 46.740 fotos no Instagram e há 510.000 comentários postados e 293.000 status atualizados no Facebook!

Imagine a enorme quantidade de dados que é produzida com essas atividades.



Fonte: https://www.edureka.co/blog/what-is-big-data/

Data Lake

Em 2010, James Dixon, CTO do Pentaho, apresentou um estudo no Hadoop World in New York.

- 80-90% das empresas estão lidando com dados estruturados ou semiestruturados (não estruturados).
- A fonte dos dados geralmente é um único aplicativo ou sistema.
- Os dados são tipicamente sub-transacionais ou não transacionais.
- Os dados são de uma escala ou volume diário tal que não se encaixam técnica e/ou economicamente em um RDBMS.

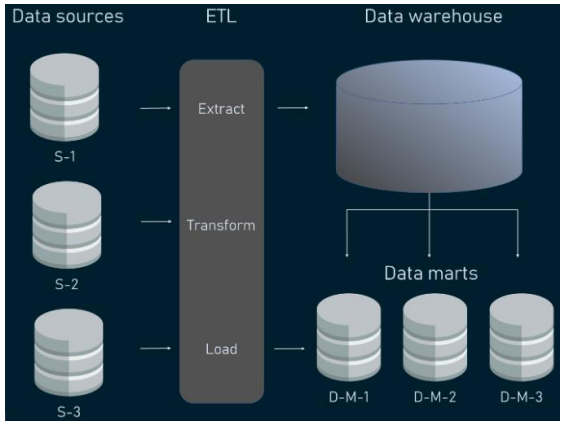
Fonte: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Data Lake

Ainda no Hadoop World, James Dixon disse o seguinte:

- No passado, a maneira padrão de lidar com relatórios e análises desses dados era identificar os atributos mais interessantes e agregá-los em um Data Mart.
- Existem vários problemas com esta abordagem:
  - Apenas um subconjunto dos atributos é examinado.
  - Os dados são agregados e a visibilidade nos níveis mais baixos se perde.

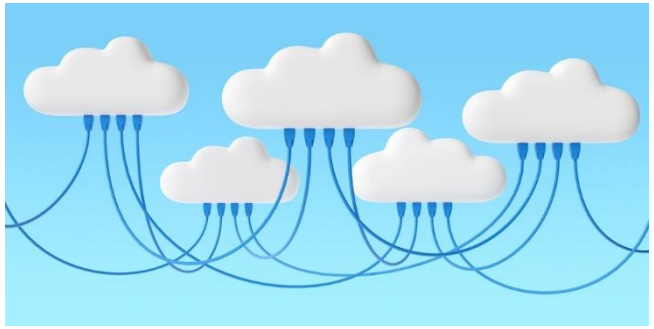
Com base nos requisitos acima e nos problemas das soluções tradicionais, **criamos um conceito chamado Data Lake para descrever uma solução ótima.**



Fonte: https://www.altexsoft.com/blog/what-is-data-mart/

Desafios

Nesse universo de Data Warehouses e Data Lakes, como obter vantagem com o que a nuvem oferece?

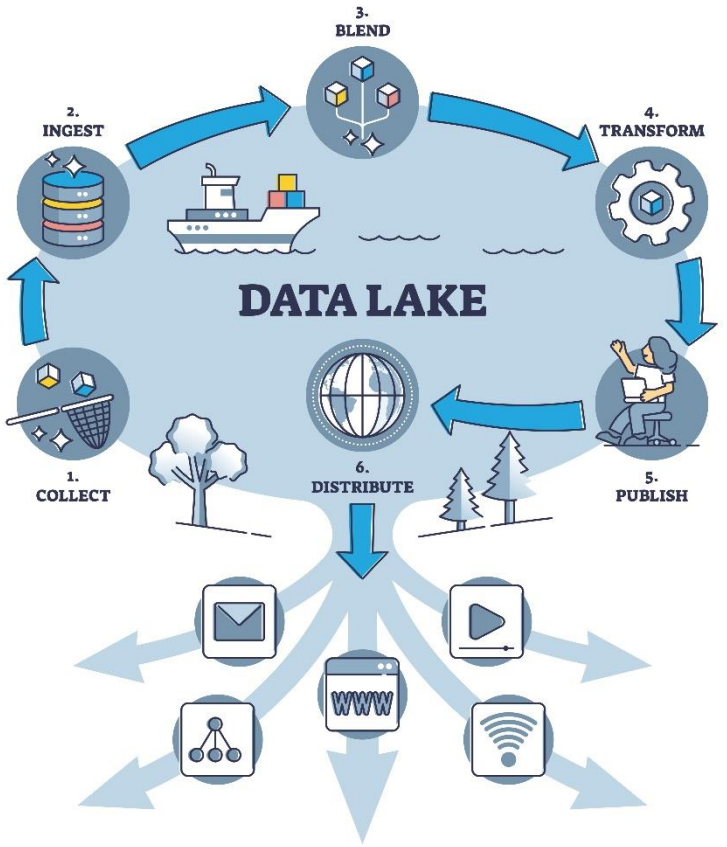


A arquitetura de dados evoluiu para um maior número de aplicações, mais simples e integradas, estimulada pelo uso da cloud, containers, microservices e service mesh.

Fonte: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Data Lake

Data lake ou lago de dados é um repositório utilizado para armazenar todos os dados estruturados e não estruturados. Ao armazená-los de forma não estruturada pode-se realizar diferentes tipos de análise, incluindo processamento de big data, análise em tempo real e machine learning, a fim de adquirir melhores decisões.



Data Lake

- ✓ Um Data Lake pode residir em ***Hadoop, NoSQL, Amazon Simple Storage Service, Banco de Dados Relacional, ou combinações diferentes deles.***
- ✓ Alimentado por fluxos de dados (Data Streams).
- ✓ Dados costumam ser não estruturados.

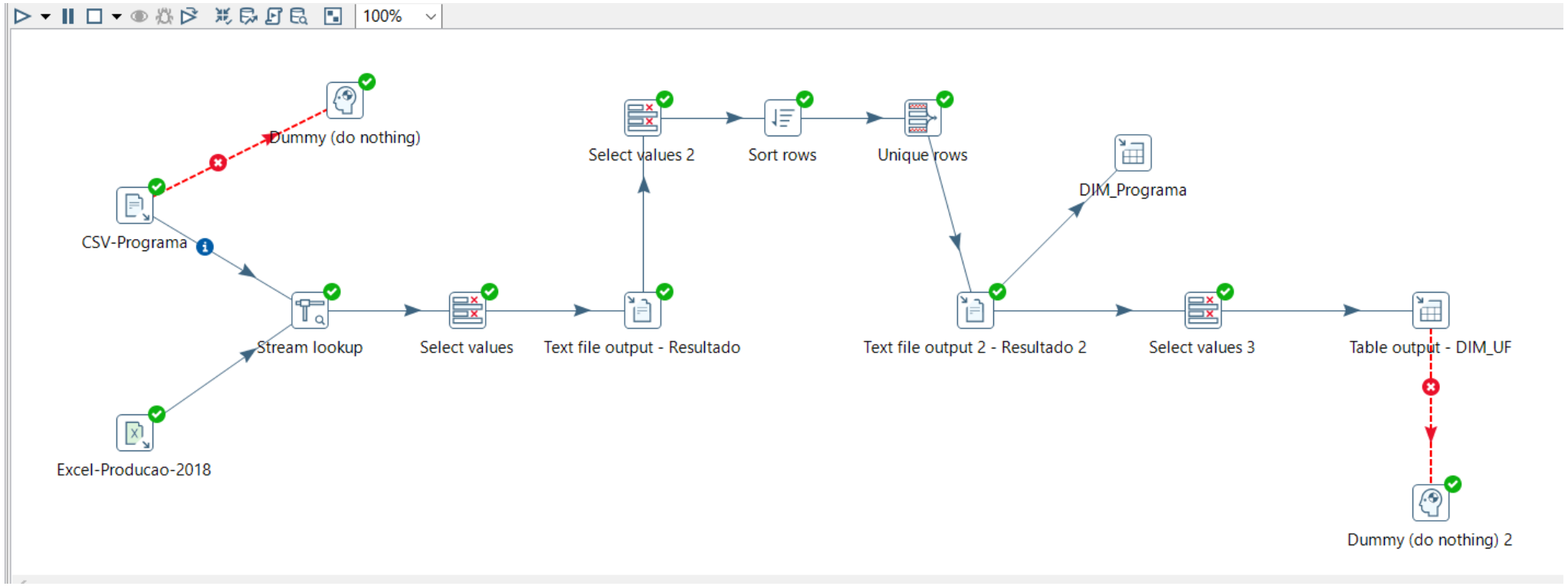
Data Lake

	Data Warehouse	Data Lake
Dados	· Estruturados · Processados	· Estruturados / Semi-estruturados / Não estruturados · Não processados (em estado bruto)
Processamento	· Esquema de dados gerado no momento da escrita	· Esquema de dados gerado no momento da leitura
Armazenamento	· Alto custo para alto volume de dados	· Criado para ser de baixo custo, independente do volume de dados
Agilidade	· Pouco ágil, configuração fixa	· Bastante ágil, pode ser configurado e reconfigurado conforme necessário
Segurança	· Estratégias de segurança bastante maduras	· Ainda precisa aperfeiçoar o modelo de segurança e acesso aos dados
Usuários	· Analistas de Negócios	· Cientistas e Analistas de Dados

ETL x ELT

ETL

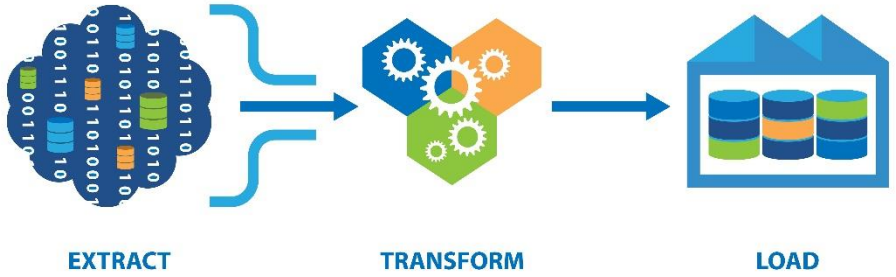
ETL significa "Extrair, Transformar e Carregar.



ELT

A função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim o carregamento dos dados geralmente para um Data Mart e/ou Data Warehouse.

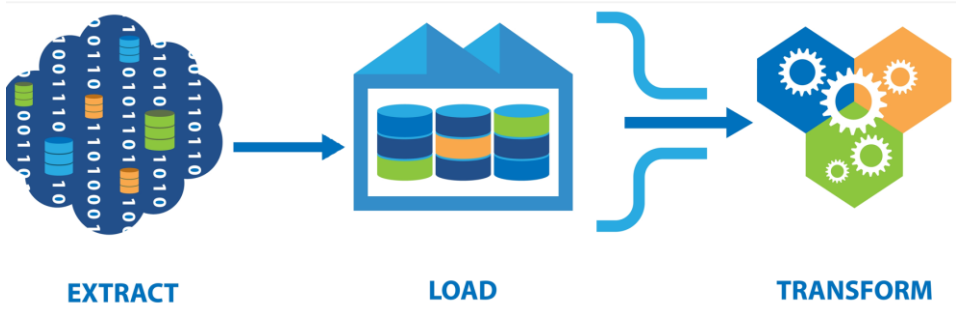
A extração e carregamento são obrigatórios para o processo, sendo a transformação/limpeza opcional, mas que são boas práticas, tendo em vista que os dados já foram encaminhados para o sistema de destino. É considerada uma das fases mais críticas do Data Warehouse e/ou Data Mart.



ELT

ELT significa "Extrair, Carregar e Transformar.

A transformação de dados ainda é necessária antes de analisar os dados com uma plataforma de inteligência de negócios. No entanto, a limpeza, enriquecimento e transformação de dados *ocorrem após o carregamento dos dados no Data Lake.*



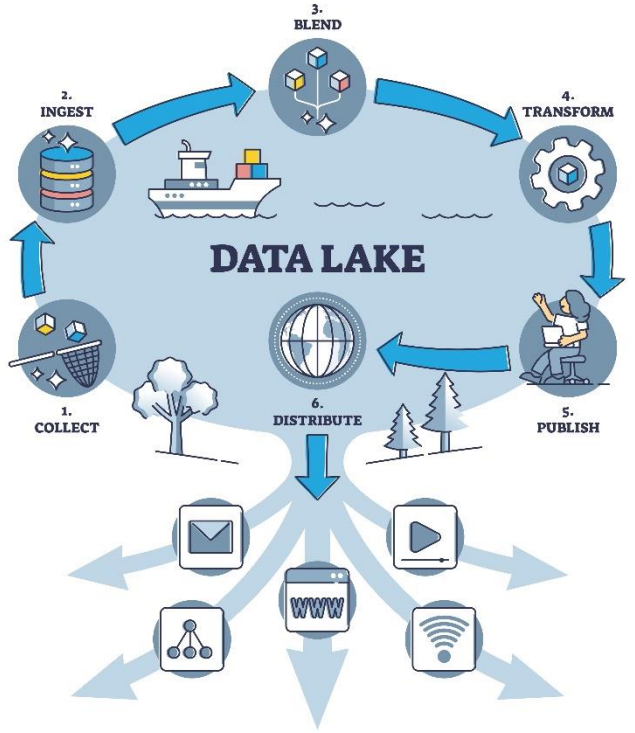
ETL x ELT

- Extract – Extrair.
- Transform – Transformar.
- Load – Carregar.
- _____
- Extract – Extrair.
- Load – Carregar.
- Transform – Transformar.
- O ***ELT*** é um processo de dados usado para replicar dados de uma fonte para um banco de dados de destino, sendo uma evolução ETL. Isso porque torna o processo de replicação de dados muito menos complexo, uma vez que o passo de transformação é realizado após os dados estarem no destino.

Características de um Data Lake

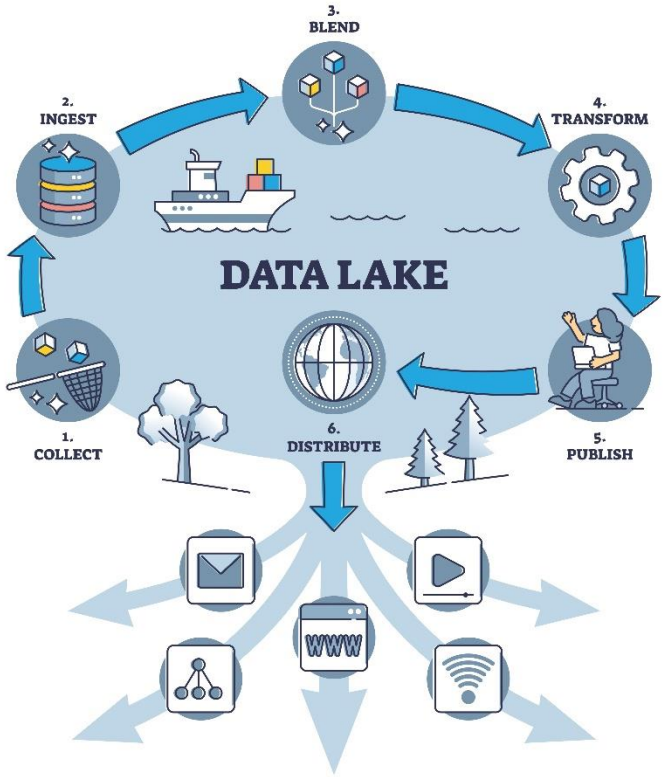
Características de um Data Lake

- Centralizar dados da organização num único local;
- Persiste dados estruturados, semi-estruturados e não-estruturados;
- Alta performance em escrita (ingestão) e em acesso (consumo);
- Baixo custo de armazenamento, se comparado a um Data Warehouse;
- Suporta regras de segurança e proteção de dados;
- Desacopla o armazenamento do processamento (ELT).



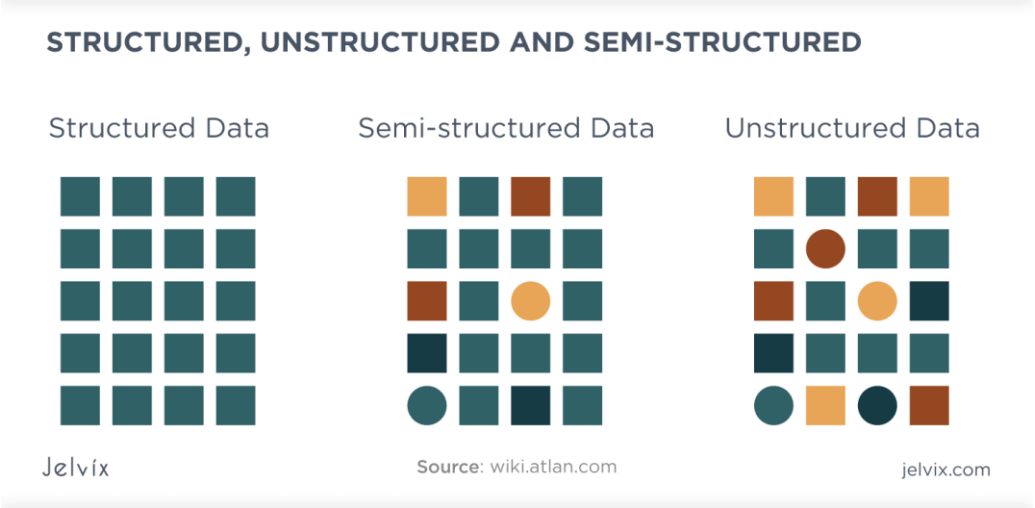
Recursos de um Data Lake

- Coleta e armazenamento de qualquer tipo de dado a baixo custo;
- Proteção de todos os dados armazenados no repositório central;
- Pesquisa e localização de dados relevantes no repositório central;
- Consulta aos dados, definindo a estrutura deles no momento do uso (esquema na leitura).



Armazenamento em um Data Lake

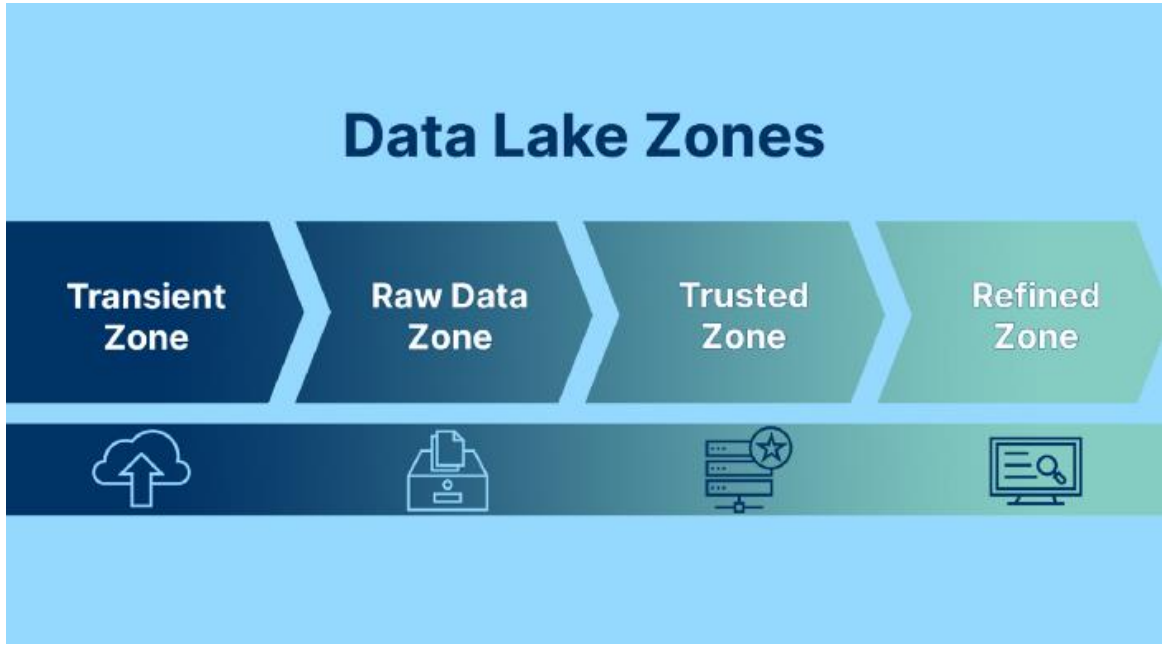
- **Dados estruturados**: banco de dados relacionais, Excel.
- **Dados semiestruturados:** arquivos HTML, XML, por exemplo
- **Dados não-estruturados:** arquivos de texto, imagens, vídeos e dados de redes sociais.



Organização de Camadas em um Data Lake

Amazenamento em um Data Lake

- Determinadas situações podem necessitar estruturações diferentes, o padrão é a divisão do Data Lake em 4 zonas:
  - Transient Zone;
  - Raw Data Zone;
  - Trusted Zone;
  - Refined Zone.



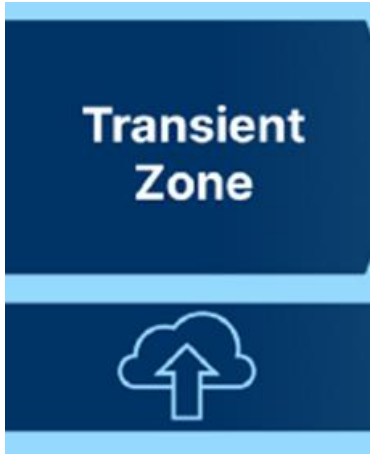
Fonte: Medium

Transiente Zone

A primeira zona é uma zona transitória, na qual os dados serão ingeridos pelo Data Lake.

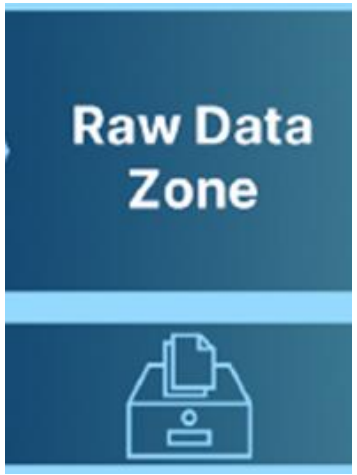
Aqui já pode-se iniciar o processo de governança com catalogação das origens e tipos de dados que estão entrando, e identificação do início das linhagens.

Depois que os dados ingeridos forem alocados na Raw Data Zone, os arquivos aqui são excluídos, tornando essa uma zona de arquivos temporários.



Raw Data Zone

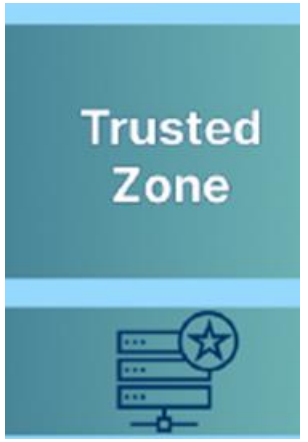
Um dos principais diferenciais de Data Lakes para Warehouses, já que nessa zona pode-se armazenar com agilidade todos os dados de fontes relevantes, independente de quanto será consumido de imediato.



Como estes dados são brutos, ainda não receberam os tratamentos necessários para serem consumidos em análises tradicionais, mas agregam muito valor por fornecerem aos cientistas de dados uma fonte crua, a partir da qual podem criar suas próprias modelagens para machine learning e AI.

Trusted Zone

Os Dados alocados nesta zona já foram tratados, sofreram as transformações necessárias para serem consumidos e já possuem garantias de Data Quality, podendo ser considerados exatos e confiáveis.



Refined Zone

Na Refined Zone é possível encontrar dados tratados e enriquecidos, estando prontos para serem consumidos por aplicações externas.

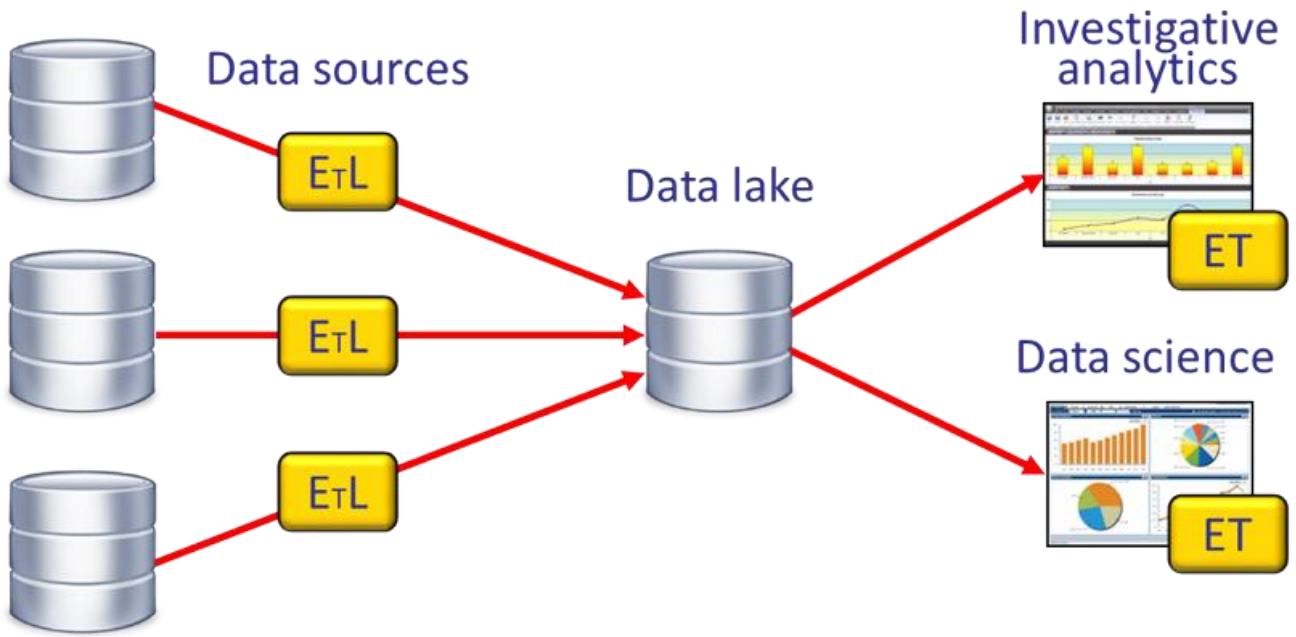
Justamente por esse uso, essa camada costuma ser construída com infraestrutura de bancos de dados relacionais.

Logical Data Lake

Data Lake

Um Data Lake é um repositório de armazenamento que contém uma grande quantidade de dados brutos em seu formato nativo até que seja necessário.

A Figura 1 contém uma arquitetura de alto nível representando um data lake.



Fonte: https://simplicitybi.com/data-virtualization/data-virtualization-and-the-logical-data-lake/

Data Lake

Infelizmente, existem complicações práticas que tornam o desenvolvimento de um armazenamento de dados centralizado contendo dados copiados difícil, impossível ou não permitido:

Complexo “T”: Todos os programadores de ETL concordam que passam a maior parte do tempo desenvolvendo o “T” e não tanto no “E” ou “L”.

Como os dados armazenados no data lake ainda estão em sua forma bruta, os cientistas de dados ainda precisam gastar tempo desenvolvendo o “T”.

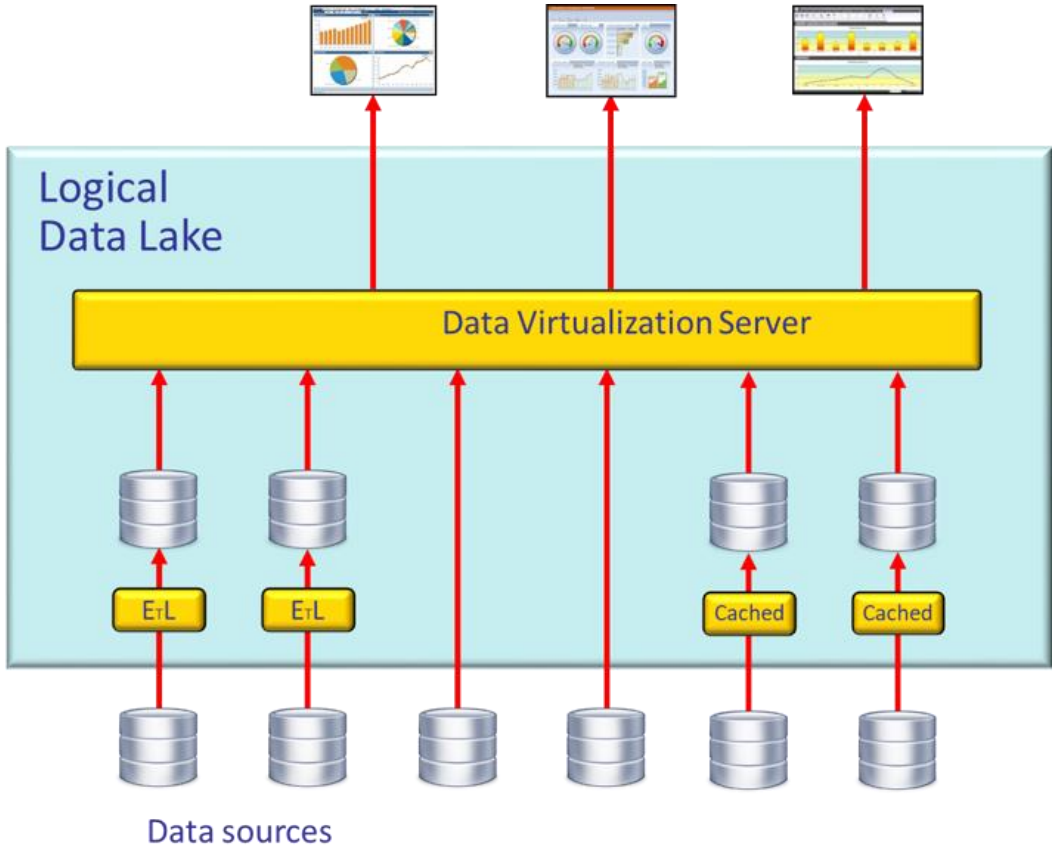
Big data muito grande para ser movido: em alguns ambientes, a grande quantidade de dados provenientes das fontes de dados pode ser muito grande para enviar e muito para copiar fisicamente. As limitações de largura de banda e ingestão de dados podem impossibilitar a cópia de uma fonte de big data para o data lake.

Data Lake

- Restringindo os regulamentos de privacidade e proteção de dados: Cada vez mais leis, regras e regulamentos proíbem o armazenamento de tipos específicos de dados juntos. Às vezes, os dados não têm permissão para sair de um país ou os regulamentos de privacidade e proteção de dados podem proibir que certos tipos de dados sejam armazenados centralmente em um data lake.
- Dados armazenados em sistema altamente seguro: Alguns sistemas de origem possuem um sistema altamente seguro para proteção contra o uso incorreto e fraudulento de dados. Os proprietários não podem permitir que seus dados sejam copiados fora do domínio de segurança original e no data lake menos seguro.

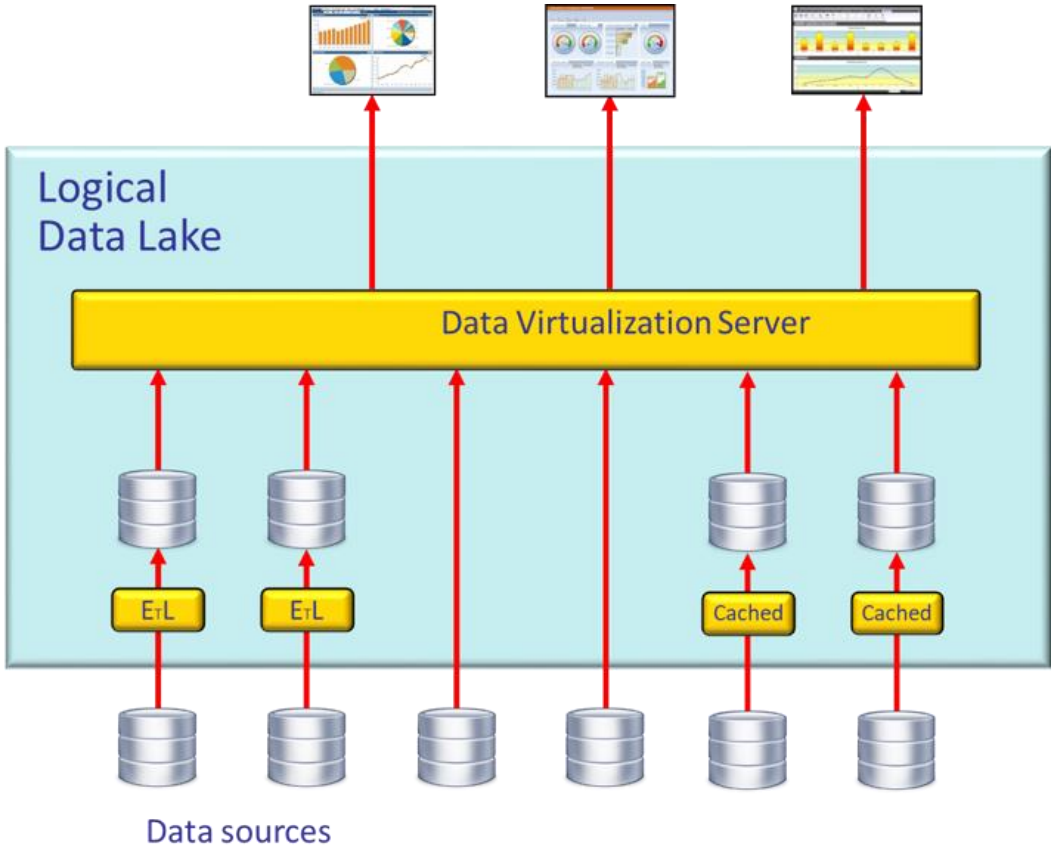
Logical Data Lake

- Todas essas são complicações realistas.
- Uma arquitetura de Data Lake alternativa que supera essas complicações é chamada de Data Lake Lógico.
- É baseado na tecnologia de virtualização de dados.



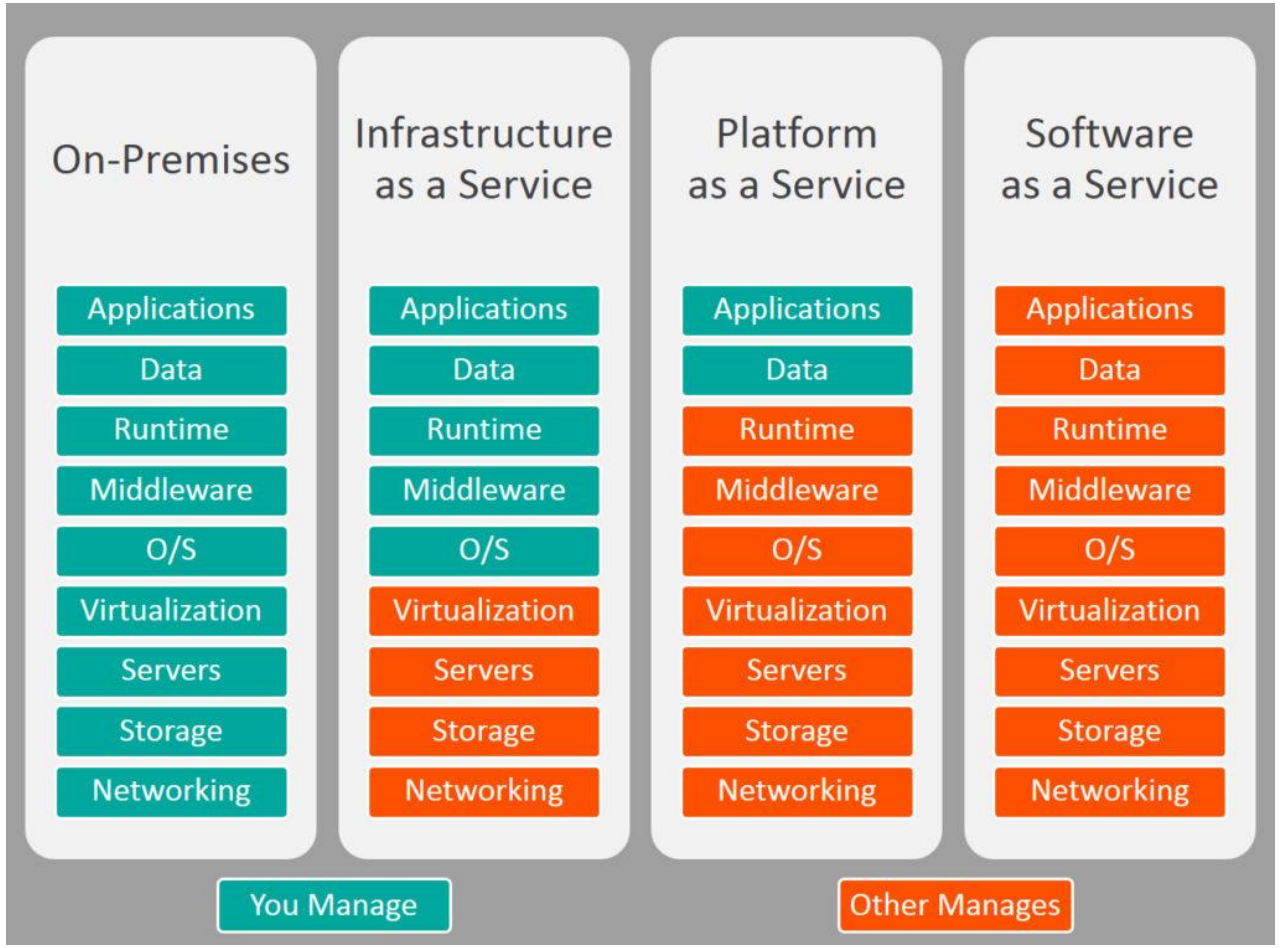
Logical Data Lake

- Em um Data Lake Lógico, os dados são apresentados aos cientistas de dados como se ainda estivessem armazenados centralmente em um repositório de armazenamento de dados.
- A virtualização de dados é uma camada de dados lógica que integra todos os dados corporativos isolados em sistemas distintos, gerencia os dados unificados para segurança e governança centralizados e os entrega aos usuários de negócios em tempo real.



Data Lake On-premise x Data Lake na Nuvem

TÍTULO



Fonte: AWS

Data Lake On Premisses

No Data Lake On premisses os softwares são instalados em máquinas próprias ou hospedadas em Data Center, geralmente se utiliza Hadoop para essa abordagem de Data Lake, visto que o Hadoop é a opção mais completa de software.

Atualmente o Data Lake on Premises deve ser bem avaliado, pois os custos de montagem da infra estrutura são bem pesados, devemos considerar : servidores, networking, instalação, update, e tudo que cerca a colocação de um servidor no ar : energia e resfriamento.

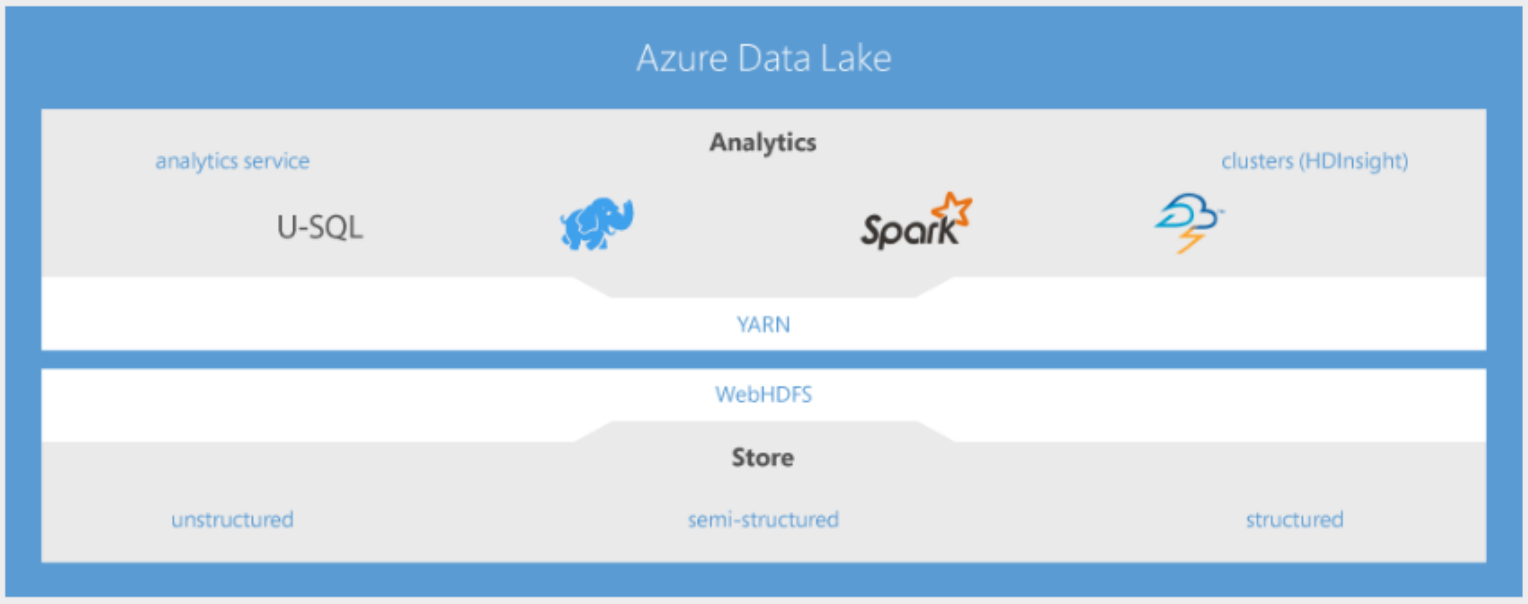
Data Lake On Premisses

Quando criado na nuvem, o Data Lake Cloud, tem sua infra estrutura fica sob responsabilidade de um provedor de Cloud (AWS, GCP, Azure, por exemplo), que é responsável por prover todos os softwares e serviços para o Lake funcionar.

Todos os provedores atualmente tem soluções para Data Lake, seja ela baseada em serviços próprios ou mesmo fazendo a hospedagem de máquinas com o Hadoop ou outra estrutura de Big Data.

Data Lake na Azure

O Azure Data Lake trabalha com investimentos existentes em TI para oferecer identidade, gerenciamento e segurança para que haja um controle e um gerenciamento de dados simplificados. Ele também tem integração direta com repositórios operacionais e data warehouses, de modo que você pode ampliar aplicativos de dados atuais.



Fonte: Microsoft

Arquitetura Corporativa - Enterprise Data Hub

Data Hub

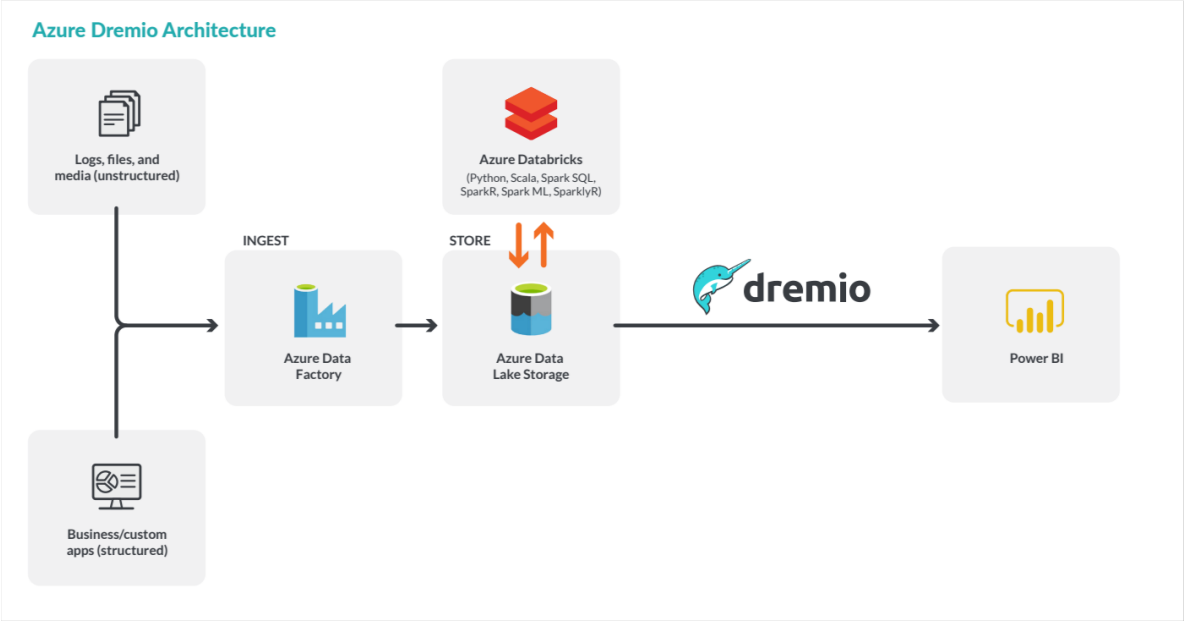
O Data Hub é o local para os dados principais de uma empresa.

Ele centraliza os dados da empresa que são críticos entre aplicativos e permite o compartilhamento contínuo de dados entre diversos setores, enquanto é a principal fonte de dados confiáveis para a iniciativa de governança de dados.

Eles são usados para conectar aplicativos de negócios a estruturas de análise, como Data Warehouses e Data Lakes.

Dremio

Dremio, projeto open-source que se descreve como The Data Lake Engine, é uma ferramenta que permite realizar a integração de dados provenientes das mais variadas fontes, sejam bancos de dados relacionais, bases NoSQL, colunares, indexadores e até mesmo o Hadoop sem nenhuma camada de abstração, como HIVE ou HBase.



Fonte: https://dremio.com

Dremio

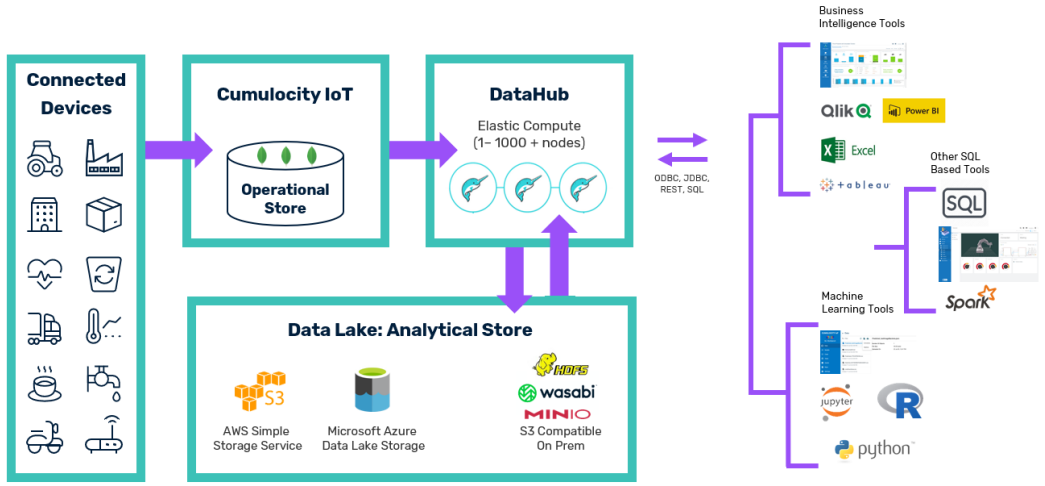
Fonte: Medium – O que é e como funciona o Dremio by Allan Sene

Cumulocity IoT

A plataforma Cumulocity IoT permite gerenciar e monitorar uma variedade de dispositivos.

Os dados emitidos por esses dispositivos são armazenados no Armazenamento Operacional do Cumulocity IoT, com os dados mais antigos sendo potencialmente removidos (com base nas configurações de retenção de dados).

Para executar uma consulta ad-hoc em relação aos dados recentes do dispositivo, o Cumulocity IoT oferece uma API REST.



Fonte: https://cumulocity.com/guides/datahub/datahub-overview/

Arquitetura Corporativa - Data Mesh

TÍTULO

- Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum.
- Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum.
- Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorem ipsum.

Data Mesh

O Data Mesh é uma nova abordagem para arquiteturas de dados.

Ao contrário de uma **arquitetura centralizada e monolítica** baseada em Data Warehouse e Data Lake, o Data Mesh é uma arquitetura de dados **descentralizada**.

O Data Mesh foi proposto por Zhamak Dehghani (diretora de tecnologia na ThoughtWorks) no artigo “How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh” como uma nova abordagem para projetar e desenvolver arquiteturas de dados.

Fonte: https://medium.com/data-hackers/data-mesh-indo-al%C3%A9m-do-data-lake-e-data-warehouse-465d57539d89

Princípios do Data Mesh

Arquitetura de dados descentralizada orientada ao domínio: a arquitetura deve ser modelada para organizar os dados analíticos por domínios.

Dados disponibilizados como produto: tratamento dos dados como um produto, onde cada fonte de dados tem seu próprio gerente/proprietário do produto de dados que fazem parte de uma equipe multifuncional de engenheiros de dados.

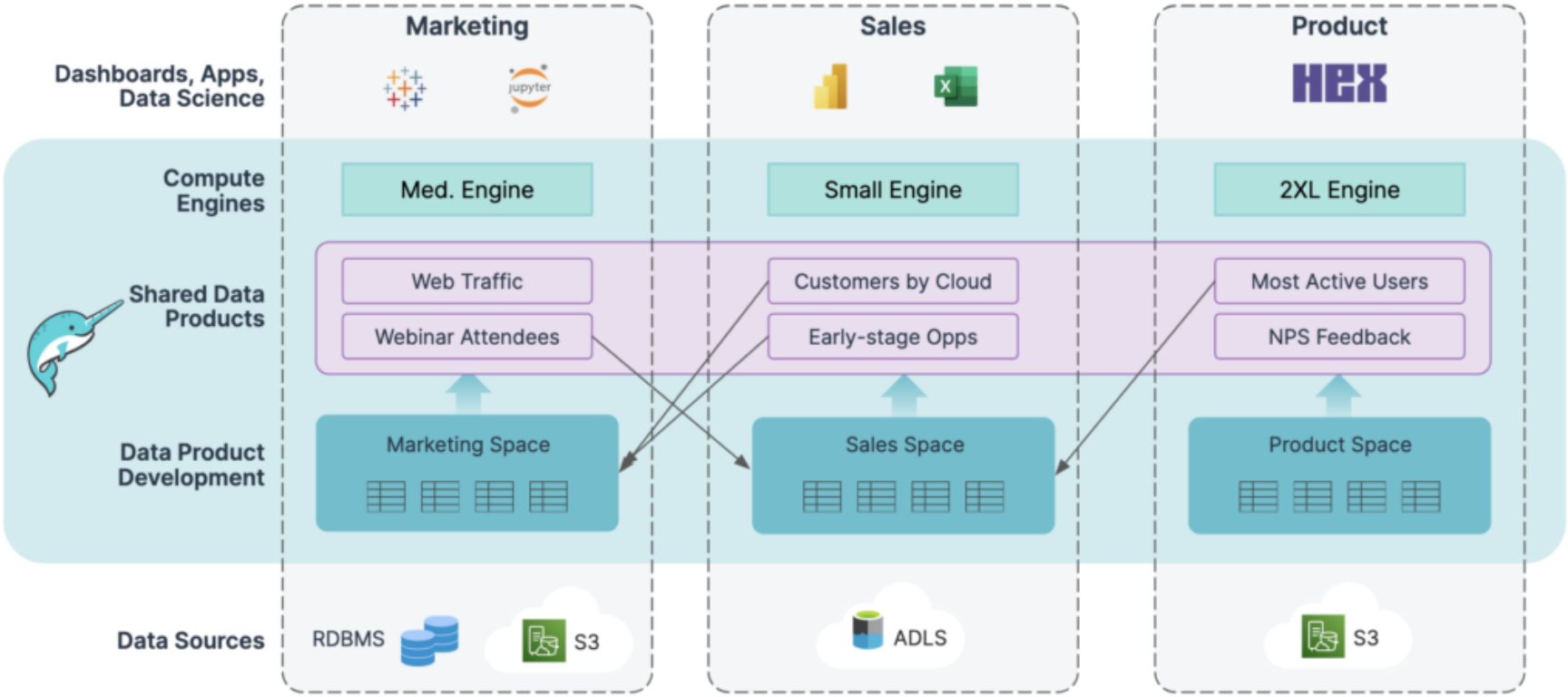
Plataforma de dados self-service: isso exige uma infraestrutura de dados self-service, como uma plataforma, para habilitar a autonomia do domínio.

Governança Federada: a implementação do data mesh requer um modelo de governança de dados que abrange a descentralização do domínio. Cada dono de um produto de dados tem autonomia e poder de decisão local de domínio, enquanto cria e adere a um conjunto de regras globais.

Fonte: https://medium.com/data-hackers/data-mesh-indo-al%C3%A9m-do-data-lake-e-data-warehouse-465d57539d89

Data Mesh com Dremio

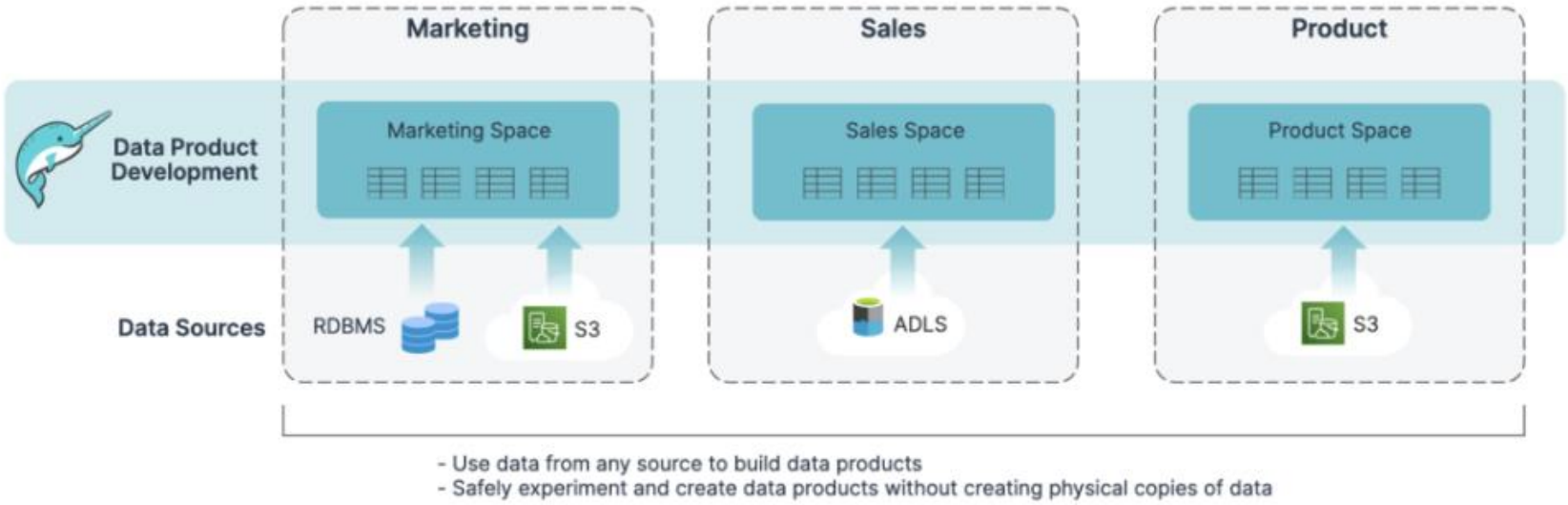
O Dremio é uma ferramenta Cloud que habilita o Data Mesh com um Open Lakehouse.



Fonte: https://www.dremio.com/blog/enabling-a-data-mesh-with-an-open-lakehouse/

Data Mesh com Dremio

Developing and managing data products



Fonte: https://www.dremio.com/blog/enabling-a-data-mesh-with-an-open-lakehouse/


PUC Minas
Virtual