



Camadas e Serviços de Consumo de Dados

Apresentação

Apresentação do Professor

Ricardo Brito Alves

Formação Acadêmica:

- Graduado em Ciência da Computação pela Pontifícia Universidade Católica de Minas Gerais, em 1994.
- MBA em Gestão Estratégica de Projetos pela Una, em 2009.
- Mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica de Minas Gerais, em 2018.
- Doutorando em Ciência da Computação pela Universidade Federal de Minas Gerais, em 2024.

Experiência Profissional:

- Desde 2002, atua com projetos de BI, Data mining.
- Atua há mais de 7 anos na área de Inteligência Artificial.



PUC Minas
Virtual

Unidade 1 – Produtores e Consumidores

Trabalhando com Dados

O que é ser Data Driven?

A gestão Data Driven é bem diferente dos modelos tradicionais, nos quais a tomada de decisão, geralmente, se baseava na intuição do dono ou nos “palpites” de especialistas, sem que houvesse dados reais para embasar essas atitudes.



Como Adotar o Data Driven?

O principal objetivo do Data Driven **é entregar respostas mais precisas e confiáveis por meio de dados.**

A ideia é **reduzir os “achismos” e entendam o valor dos dados.**

- **Transforme a sua cultura.**
- **Use boas soluções.**
- **Aprenda a entender os dados.**
- **Use os indicadores de performance.**



Data Driven

O Data Driven surgiu como extensão da Ciência de Dados utilizada para transformar dados estruturados e não estruturados em conhecimento.

Atualmente, no ambiente corporativo, isso é feito por meio de ferramentas como Big Data, Inteligência Artificial e Machine Learning, para obter insights a partir da coleta, cruzamento e interpretação de dados.

Produtores e Consumidores de Dados

Um **produtor de dados** é uma interface de usuário, sistema ou dispositivo que coleta dados relevantes para uma organização.

Um **consumidor de dados** é uma interface de usuário, sistema ou ferramenta que usa dados.

Um consumidor de dados pode ser qualquer usuário, aplicativo ou sistema que usa dados coletados por outro sistema ou armazenados em um repositório de dados. Dentro de uma organização ou empresa, muitos sistemas podem ser considerados produtores de dados, pois coletam ou geram dados que são armazenados para serem posteriormente utilizados pelos consumidores de dados.

Gerenciando os Dados

Data Management – Gestão de Dados

O gerenciamento de dados é uma tarefa complexa e demorada, exigindo muito tempo dos funcionários. As empresas querem ter certeza de que estão aproveitando seus dados para aproveitar ao máximo sua inteligência de negócios. Os escritórios tendem a ser divididos em dois campos separados: produção de dados versus consumo de dados. Sua empresa deveria consumir dados, não produzi-los. Além disso, você deseja uma solução para garantir que os dados funcionem para sua empresa e que sua empresa não precise trabalhar muito para minerar dados.



Lei Geral de Proteção de Dados – Lei 13.709/18

- Dispõe sobre o tratamento de dados pessoais – Nome, gênero, preferências de lazer, GPS, IP, etc.
- Conformidade / Compliance em agosto/2021.
- Regras para coletar, armazenar, tratar e compartilhar dados pessoais.
- Estende-se aos subcontratantes - fornecedores e parceiros.
- Multas podem chegar em 2% do faturamento / R\$ 50 milhões (por infração).

Com a LGPD vou ter que apagar toda a minha base de contatos?

Caso a base de contatos seja formada pela base de clientes da empresa, não será necessário apagar. Contudo, o tratamento dos dados deve ser para a finalidade específica que justifique o seu uso de acordo com a base legal da respectiva finalidade.

No entanto, se a base de dados foi adquirida ou repassada por terceiros, provavelmente terá que apagá-la, caso os titulares dos dados não saibam que você trata estes dados pessoais. Uma forma de remediar esta questão seria entrar em contato e informá-los de modo a pegar o seu consentimento.

E o que são os Dados?

O Que é Dado?

- Os **dados** são elementos que constituem a matéria-prima da informação.
- Podemos defini-los como o **conhecimento bruto**, que ainda não foi devidamente tratado de forma a prover insights para uma organização.
- **Dados, de forma isolada, não conseguem transmitir uma mensagem clara.**

O Que é Dado?

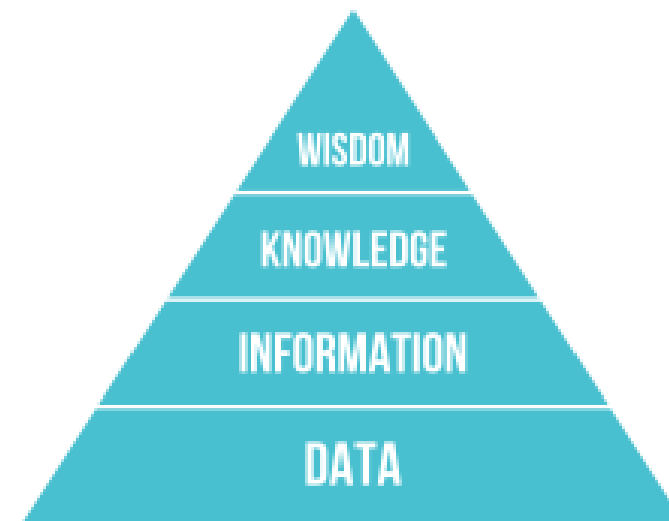
- Exemplos: Manga, Roupa, Sapato.
- Definimos dado como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado. De fato, as letras são símbolos quantificados, já que o alfabeto por si só constitui uma base numérica.

O Que é Dado?

- Também são dados as **imagens, sons e animação**, pois todos podem ser quantificados a ponto de alguém que entra em contato com eles ter eventualmente dificuldade de distinguir a sua reprodução, a partir da representação quantificada, com o original.
- É muito **importante notar-se que qualquer texto constitui um dado ou uma sequência de dados, mesmo que ele seja ininteligível** para o leitor.

Pirâmide DIKW

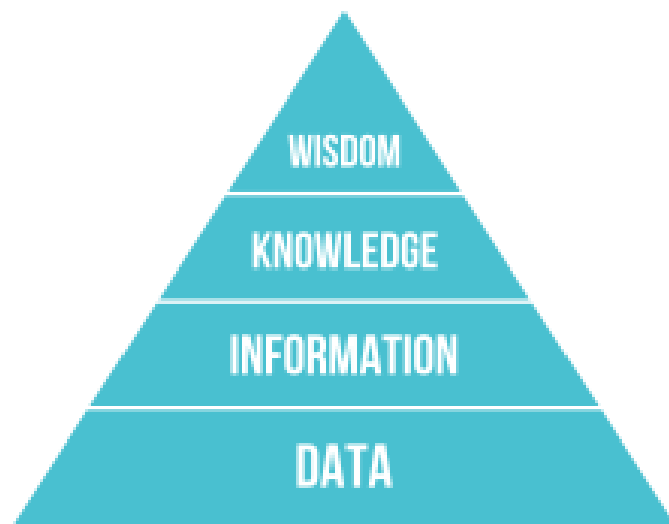
- A Hierarquia ou pirâmide **DIKW** é o modelo utilizado para discussão de **dados, informações, conhecimento, sabedoria** e suas **inter-relações**.
- Em **1955**, o economista e educador anglo-americano Kenneth Boulding apresentou uma variação na hierarquia que consiste em "sinais, mensagens, informações e conhecimento". No entanto, o primeiro autor a distinguir entre dados, informação e conhecimento e **também empregar o termo “gestão do conhecimento”** pode ter sido o educador americano Nicholas L. Henry, em um artigo de jornal de **1974**.



- Fonte: Wikimedia Commons

Pirâmide DIKW

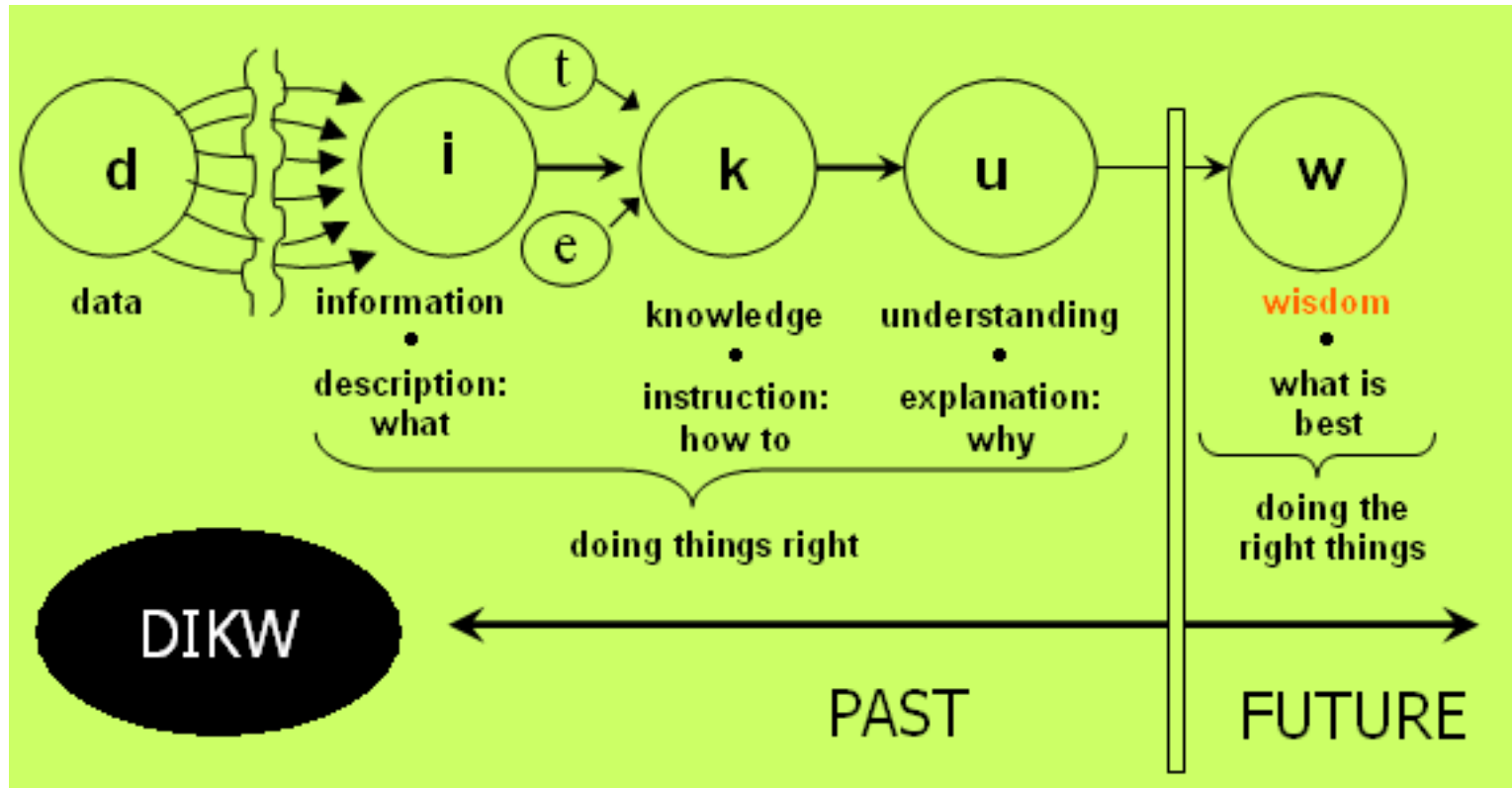
- A Hierarquia ou pirâmide DIKW é o modelo utilizado para discussão de dados, informações, conhecimento, sabedoria e suas inter-relações.



- Fonte: Wikimedia Commons

Gestão de Dados

DIKW – Diagrama de Fluxo



- Fonte: Creative Commons Attribution 4.0 International

O gerenciamento de informações combina processos de negócios, procedimentos e tecnologia para organizar, proteger e acessar os dados de uma organização independentemente do formato, incluindo dados digitais, documentos em papel e arquivos de áudio e vídeo.

- A informação consiste em dados que um usuário inseriu em um computador, processou e colocou em contexto.
- As empresas devem gerenciar essas informações com ferramentas como Microsoft SharePoint ou Alfresco; um sistema de gerenciamento de documentos, como Microsoft OneDrive ou Google Drive.

Gestão do Conhecimento - GC

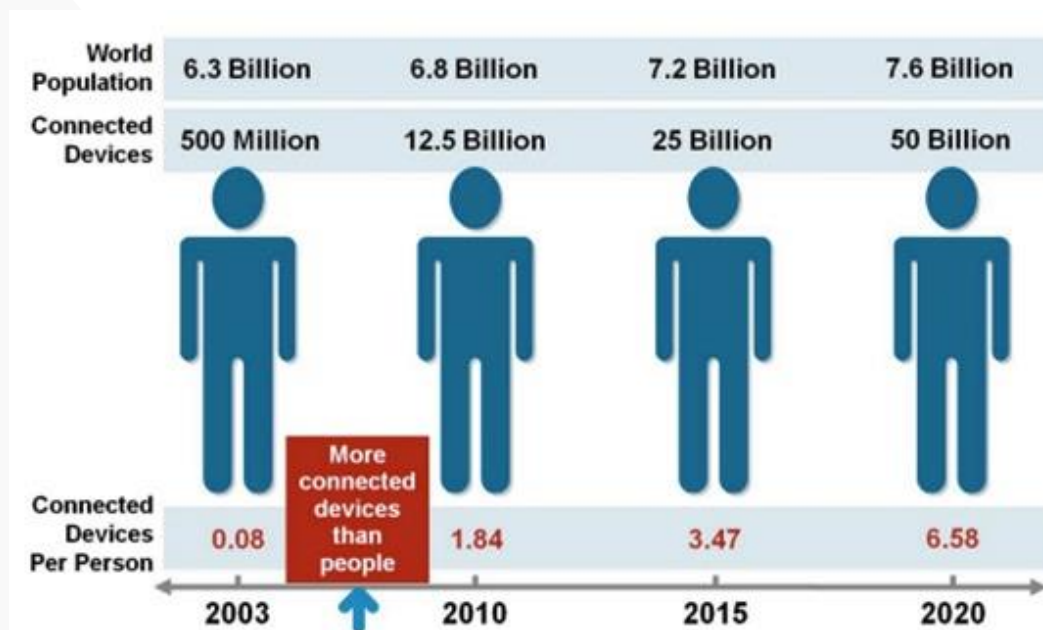
A gestão do conhecimento usa processos e ferramentas para transmitir sabedoria e compreensão de diferentes assuntos.

- Quando a informação é colocada no contexto de ser usada para maior compreensão de um assunto, ela se torna conhecimento.
- Esse conhecimento ajuda os funcionários a realizarem seus trabalhos, muitas vezes tornando-os mais eficientes. Também pode beneficiar os clientes de uma organização.
- **A gestão do conhecimento envolve a coleta, organização e compartilhamento do conhecimento.**
- Esse conhecimento pode estar na forma de documentos, vídeos e outros recursos destinados a ensinar as pessoas sobre um assunto específico.

Big Data

Big Data

- Qual é o tamanho da web?

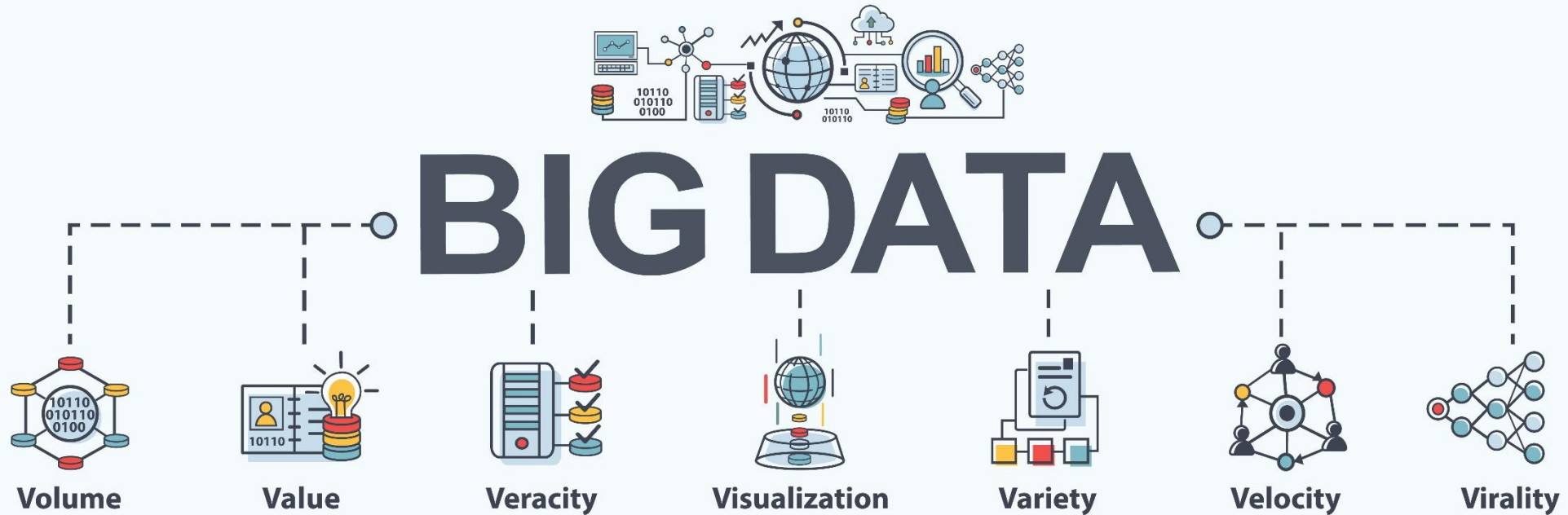


Fonte: https://www.researchgate.net/publication/316994150_THE_INTERNET_OF_THINGS_EVOLUTION

- Razões para usar o Big Data:
 - Entender padrões;
 - Prever situações;
 - Criar fronteiras;
 - Informar coleções de dados;
 - Estimar parâmetros escondidos;
 - Calibrar.

Big Data

- Os 3Vs
- Os 5Vs
- Os 7Vs



Os Vs do Big Data

- ❑ **Volume:** são os dados gerados a cada segundo. O volume define a quantidade de dados que temos – o que costumávamos medir em Gigabytes agora é medido em Zettabytes (ZB) ou mesmo Yottabytes (YB). A Internet das Coisas (IoT) cria um crescimento exponencial de dados. As projeções mostram o volume de dados mudando significativamente nos próximos anos.

Os Vs do Big Data

- ❑ **Velocidade:** um dos grandes desafios do Big Data. A velocidade representa a velocidade com que os dados são processados e se tornam acessíveis. Hoje, se a entrega não for em tempo real, geralmente não é rápida o suficiente.

Os Vs do Big Data

- ❑ **Variedade:** quanto mais dados e fontes, maior a complexidade, mas também é maior a chance de gerar informações úteis. A variedade descreve um dos maiores desafios do big data. Os insights podem vir sem estrutura. O ativo total pode incluir muitos tipos de dados, de XML a vídeo e SMS. Organizar os dados de maneira significativa não é tarefa simples quando os próprios dados mudam rapidamente.

Os Vs do Big Data

- ❑ **Variabilidade:** a variabilidade é diferente de variedade. Uma cafeteria pode oferecer seis misturas diferentes de café, mas se você obtém a mesma mistura todos os dias e tem um sabor diferente a cada dia, isso é variabilidade. O mesmo acontece com os dados. Se o significado mudar constantemente, isso pode afetar significativamente a homogeneização de seus dados.

Os Vs do Big Data

- ❑ **Veracidade:** o quanto a informação é verdadeira. A veracidade garante que os dados sejam precisos, o que requer processos para evitar que dados insuficientes se acumulem em seus sistemas. O exemplo mais simples é quando os contatos entram em seu sistema de automação de marketing com nomes falsos e informações de contato imprecisas. Quantas vezes você viu o Saci Pererê ou Mickey Mouse em seu banco de dados? É o clássico desafio de “entrar lixo, sair lixo”.

Os Vs do Big Data

- ❑ **Visualização:** a visualização é fundamental no mundo de hoje. Usar tabelas e gráficos para visualizar grandes quantidades de dados complexos é muito mais eficaz em transmitir significado do que planilhas e relatórios repletos de números e fórmulas.

Os Vs do Big Data

- ❑ **Valor:** deve-se gerar informações com valor. Depois de abordar volume, velocidade, variedade, variabilidade, veracidade e visualização – o que leva muito tempo, esforço e recursos –, você quer ter certeza de que sua organização está obtendo valor dos dados.

IoT

IoT - Internet das Coisas



IoT – Internet of Things

Internet das Coisas ou, pela sigla “**IoT**” (**Internet of Things**), traz um conceito atual e transformador sobre a conexão entre objetos físicos utilizando sensores, chips e softwares.

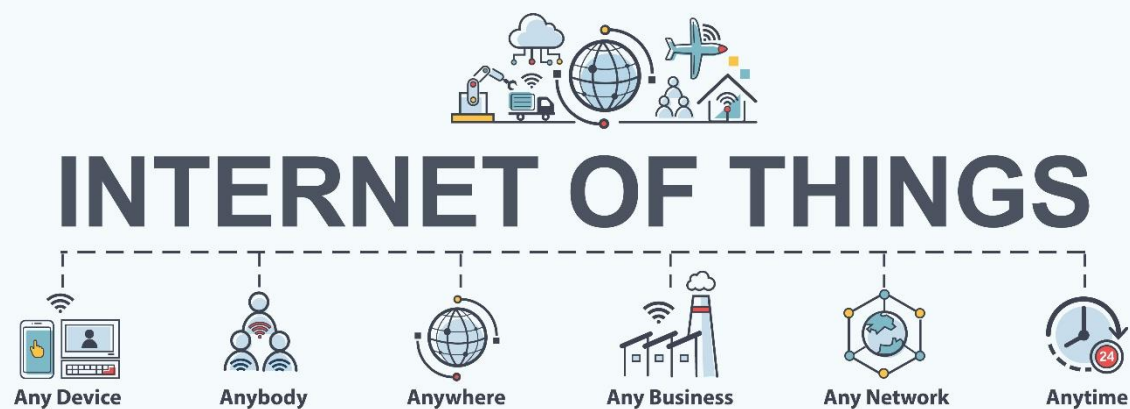
Quando falamos de “coisas” em IoT, estamos nos referindo a qualquer objeto que teve a implementação de sensores e outros sistemas digitais para funcionar de forma mais inteligente por meio da troca de informações com pessoas e outros objetos, não necessariamente utilizando conexões de internet.

IoT - Internet das Coisas

- O termo foi utilizado pela primeira vez em 1999 para descrever um sistema onde os objetos poderiam ser conectados à internet.
- Por causa do avanço do IoT, houve a necessidade de muito mais endereços de IP(IPv4 para IPv6 – 128 bits).
- O fato de ter tantos dispositivos conectados à rede literalmente tem revolucionado o modo como vivemos.

IoT - Internet das Coisas

- 32 bilhões de coisas vão estar conectadas a internet.
- 10% de todos os dados serão gerados por sistemas embarcados (versus 2% atualmente).
- 21% dos mais valiosos dados serão gerados por sistemas embarcados (versus 8% atualmente).
- Dados de telemetria - semi-estruturados e contínuos - representam um desafio para bancos de dados relacionais, que exigem um esquema fixo e dados estruturados.



Cidades Inteligentes

- Poluição Sonora.
- Otimizar coleta de lixo.
- Controle de tráfego.
- Controle de distribuição de energia elétrica.
- Segurança pública.



Casas Inteligentes



Produtores de Dados

O que é Produção de Dados?

Se sua empresa produz mais dados do que você pode gerenciar com eficiência, você é conhecido como produtor de dados. Os produtores de dados tendem a ter repositórios cheios de arquivos duplicados e arquivos enormes. Nesses ambientes, os dados não são escaláveis nem flexíveis. Os funcionários perdem horas tentando recuperar dados e entendê-los.

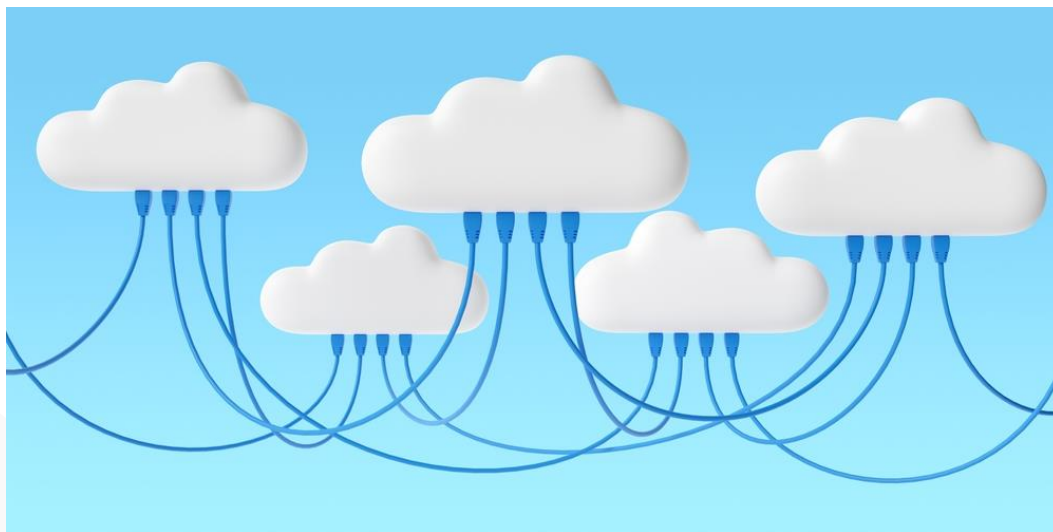


O que é Produção de Dados?

É muito caro ser um produtor de dados.

A produção de dados consome o tempo dos funcionários e os recursos da empresa.

E o desafio é aliar a produção de dados com um uso otimizado de cloud.



Na governança de dados, o papel do **consumidor de dados** traz consigo responsabilidades para garantir que os dados sejam confiáveis e protegidos. Essas responsabilidades se concentram em como os dados são protegidos, como serão usados, por quanto tempo podem ser retidos e com quem podem ser compartilhados.

Os administradores de dados conseguem controlar e distribuir os dados.

Por outro lado, **os proprietários dos dados** são responsáveis por definir o conteúdo, fonte e classificação dos dados, validá-los e autorizar o acesso.

Consumidores de Dados

O que é Consumo de Dados?

O consumo de dados levará a um novo fluxo de receita para sua empresa. Equipe seu time com dados rápidos e precisos para uma estratégia mais inteligente e uma tomada de decisão aprimorada. Sua empresa economizará tempo e dinheiro gastos anteriormente na produção de dados.

A análise de dados aprimorada fornecerá à sua equipe uma melhor compreensão das tendências do mercado e dos processos de negócios. Você pode se concentrar menos em gerenciar seus dados e mais em usá-los para criar estratégias com eficiência.

Consumidor de Dados

Um consumidor de dados pode ser qualquer usuário, aplicativo ou sistema que use dados coletados por outro sistema ou armazenados em um repositório de dados.

Dentro de uma organização ou empresa, muitos sistemas podem ser considerados produtores de dados, pois coletam ou geram dados que são armazenados para serem usados posteriormente pelos consumidores de dados.



Consumidor de Dados

Nesse contexto, o consumidor de dados é qualquer ator que pode acessar legitimamente os dados com um motivo aparente para fazê-lo. Eles podem usá-lo para criar novos dados, dos quais se tornam proprietários de dados com as devidas responsabilidades. Os consumidores de dados devem garantir que os dados tenham valor comercial e qualidade de dados clara. Eles também precisam garantir que ele seja usado exclusivamente para a finalidade para a qual foi coletado.

Data Domains e Data Products

Produtores de dados x consumidores de dados

- A maioria das grandes empresas Data Driven estão interessadas ou planejando sua próxima geração de uma plataforma de dados moderna.
- Nesta transição, a arquitetura de Data Mesh está ganhando muita força.
- Ao mesmo tempo, há preocupações sobre a interpretação desse conceito: falta orientação prática.

Produtores de dados x consumidores de dados

- **O Data Mesh defende o tratamento de dados como um produto.**
- A delegação de responsabilidades entre conjuntos de dados específicos **para domínios de dados significa que as empresas são sendo divididas em mini empresas**, cada uma assumindo a responsabilidade de construir abstrações, servir dados, manter metadados associados, melhorar a qualidade dos dados, aplicar gerenciamento de ciclo de vida, realizar controle de código e assim por diante.

Produção de dados x consumo de dados

- A transição de produtor de dados para consumidor de dados aprimorará seus processos, fornecendo uma estrutura para estratégias de negócios bem-sucedidas.
- O consumo aprimorado de dados impulsionará seus negócios com dados inteligentes e precisos.
- Agora é hora de pensar em domínios e tratar os dados como produto.

Arquitetura Corporativa - Enterprise Data Hub

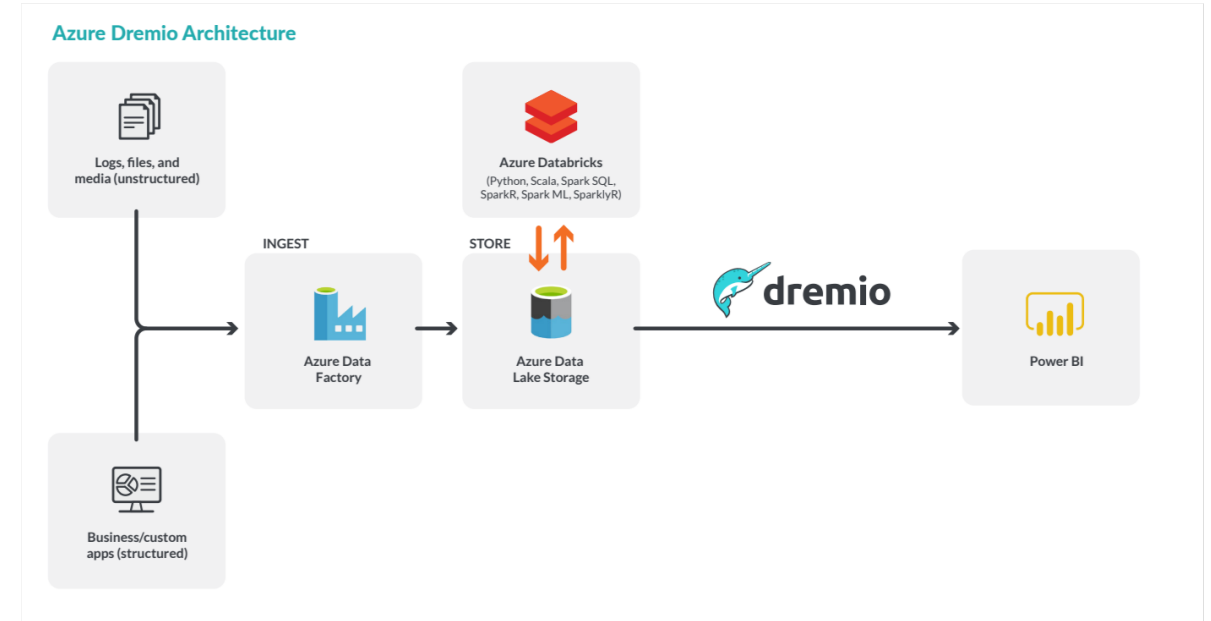
O Data Hub é o local para os dados principais de uma empresa.

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.

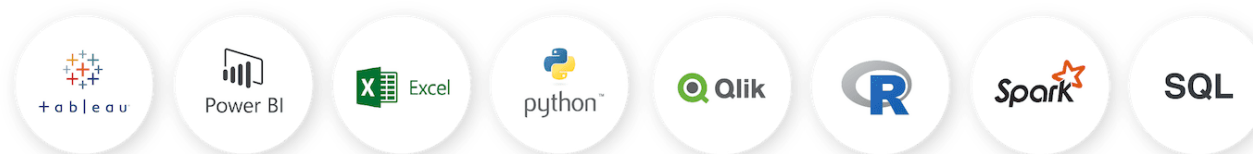
Eles são usados para conectar aplicativos de negócios a estruturas de análise, como Data Warehouses e Data Lakes.

Dremio

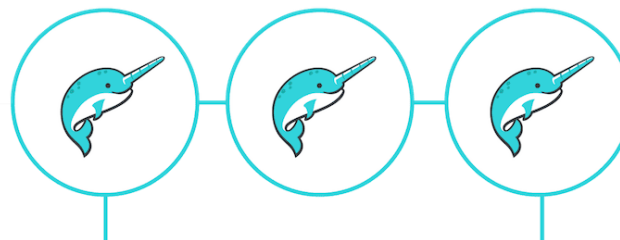
Dremio, projeto open-source que se descreve como The Data Lake Engine, é uma ferramenta que permite realizar a integração de dados provenientes das mais variadas fontes, sejam bancos de dados relacionais, bases NoSQL, colunares, indexadores e até mesmo o Hadoop sem nenhuma camada de abstração, como HIVE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



Dremio



Elastic Compute
(1 - 1000+ nodes)



Reflection Store
(S3, HDFS, ADLS)

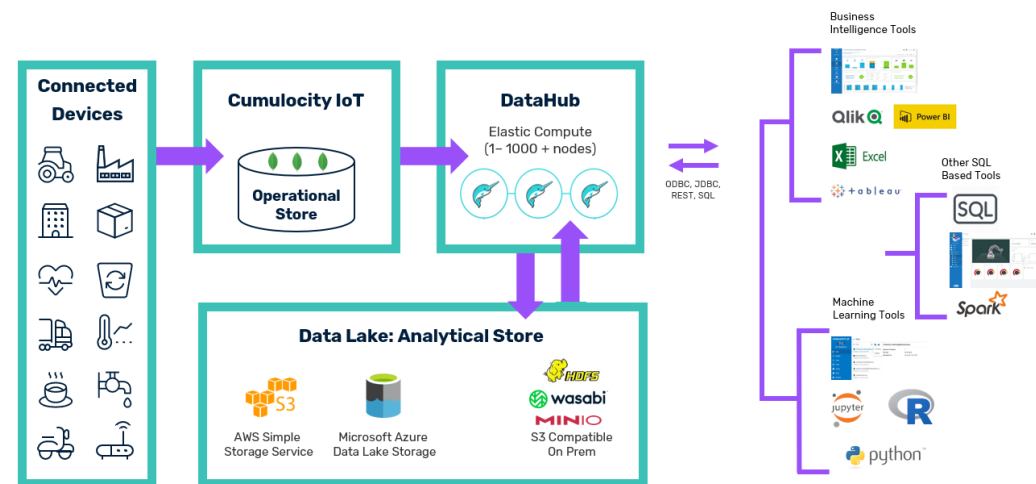


Cumulocity IoT

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.



Arquitetura Corporativa - Data Mesh

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-into-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.

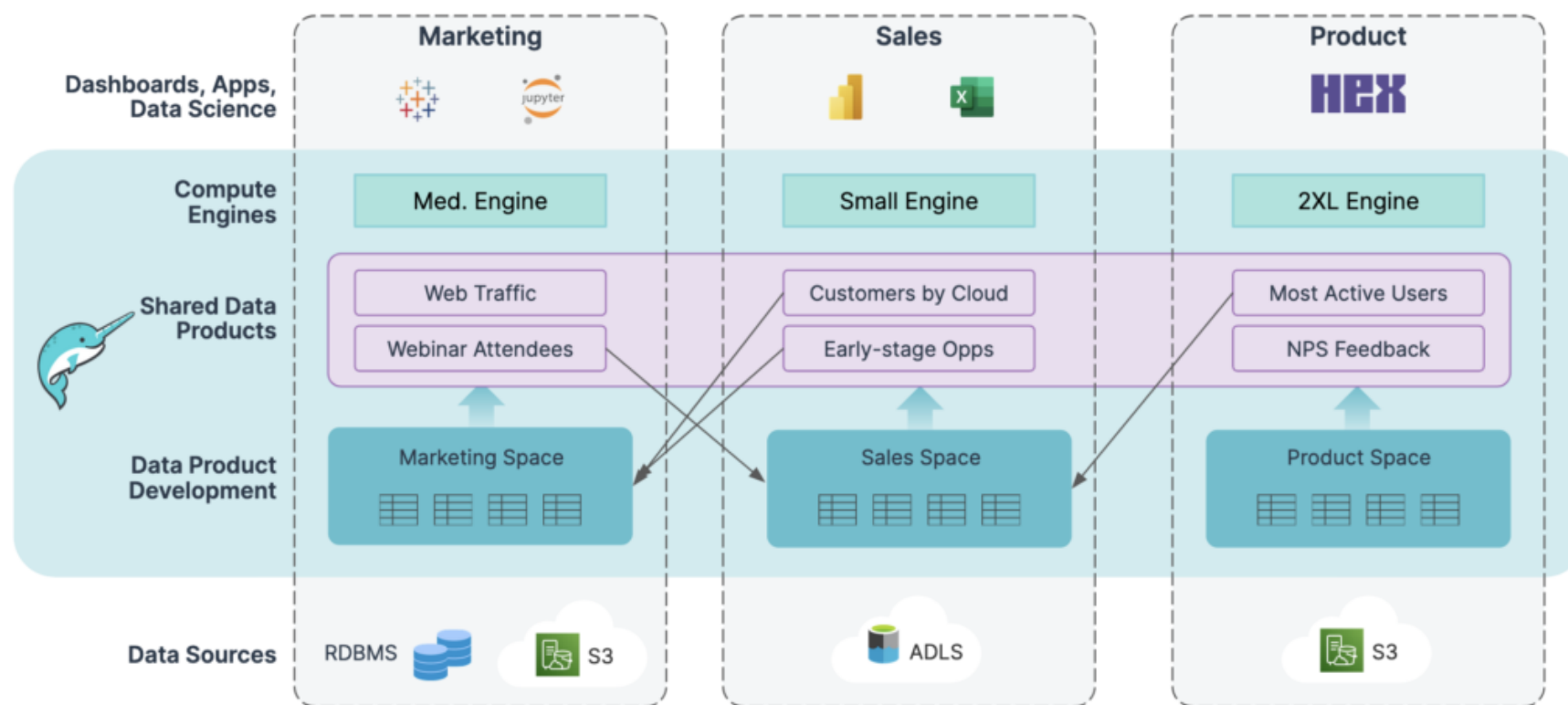
Dados disponibilizados como produto: tratamento dos dados como um produto, onde cada fonte de dados tem seu próprio gerente/proprietário do produto de dados que fazem parte de uma equipe multifuncional de engenheiros de dados.

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.

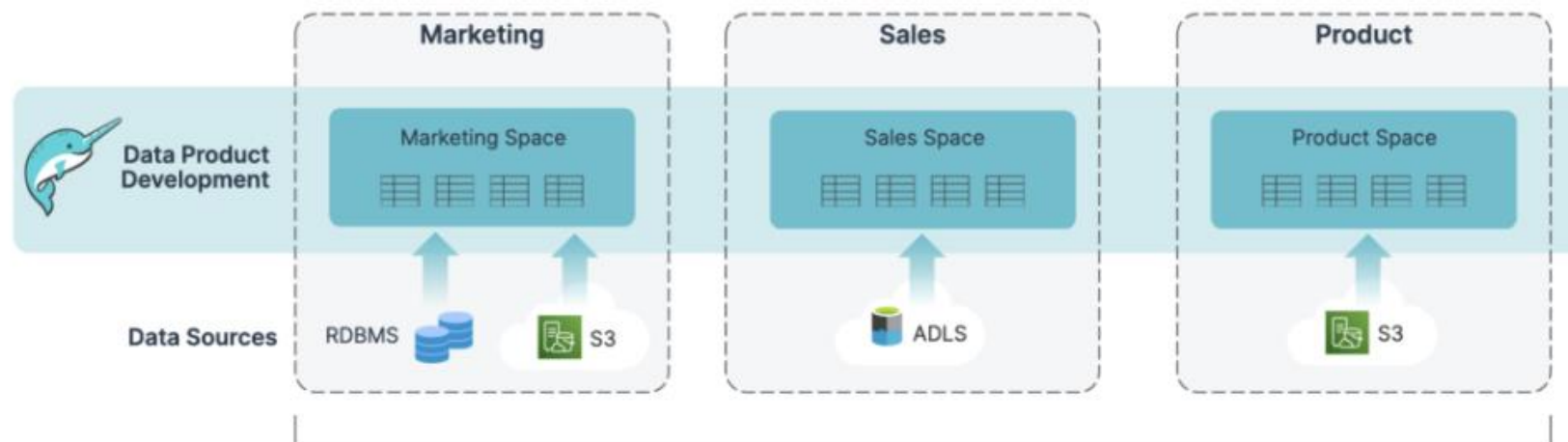
Data Mesh com Dremio

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



Data Mesh com Dremio

Developing and managing data products



- Use data from any source to build data products
- Safely experiment and create data products without creating physical copies of data



PUC Minas
Virtual