Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO Uni-FACEF

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis

Turma 2021-2022

Big Data com Python

André Morais

andre.morais@luizalabs.com

09/2022

Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis
Módulo: BIG DATA COM PYTHON - Professor: André Morais
Turma 2021-2022

Fonte Ilustração:
https://disabilityinbusiness.wordpress.com/2012/12/24/spy-uk/word-cloud-big-data/

“Dados representam o novo petróleo”

Para o diretor-geral da Intel Brasil, Maurício Ruiz, os dados crescem exponencialmente nos dias de hoje. “O grande diferencial é que agora temos tecnologia para processar tudo isso”, destaca

Por: Léia Machado, 27/11/2017 às 16h22 - Atualizado em 27/11/2017 às 16h22

CIO FROM IDG
Estratégias de negócios e TI para líderes corporativos
DIGITAL NETWORKIBRASILEIF
Login Registro Newsletter

Recursos/White Papers

Home Notícias Gestão Opinião Tecnologia Carreira Cursos Onlin

Opinião

Na nova sociedade digital, os dados são o novo petróleo, mas o motor é a IA

Os efeitos da IA ainda serão amplificados exponencialmente a partir da próxima década

Cezar Taurion *
Publicada em 17 de dezembro de 2017 às 10h36

“Os dados são importantes ativos para as empresas!”

Conceitos Big Data

Big Data
Tratar grandes volumes de dados

Gestão
Tomada de decisão

Ciência de Dados
Aplicar análise e ciência sobre os dados

Cultura Data Driven
Tomar decisões baseadas em dados

Machine Learning / IA
Treinar algoritmos inteligentes

Conceitos

Big Data é um termo em tecnologia amplamente utilizado na atualidade para nomear conjuntos de dados muito grandes ou complexos, que os aplicativos de processamento de dados tradicionais ainda não conseguem lidar. O conceito do Big Data se iniciou com 3 Vs : Velocidade, Volume e Variedade.

Conceitos

Ciência de Dados é um conjunto de métodos e técnicas, para aplicação de conceitos matemáticos e estatísticos, modelagem preditiva e aprendizagem de máquina, para analisar, interpretar e extrair conhecimento de grandes volumes de dados.”

Conceitos

MACHINE
LEARNING

Machine Learning ou aprendizado de máquina está relacionado ao uso de algoritmos inteligentes capazes de aprender através dos dados, gerando previsões ou respostas esperadas por meio da relação que existe nessa massa de dados. É uma das técnicas da Ciência de Dados e uma das áreas de conhecimento da Inteligência Artificial.

Conceitos

Cultura Data-Driven É a capacidade de ser orientado a dados, através da extração de informações para transformá-la em inteligência, que permite melhores tomadas de decisões, de forma pró-ativa, e oferecer melhores produtos e serviços.

Conceitos Os V's do Big Data

Volume
Geração de grandes volumes de dados

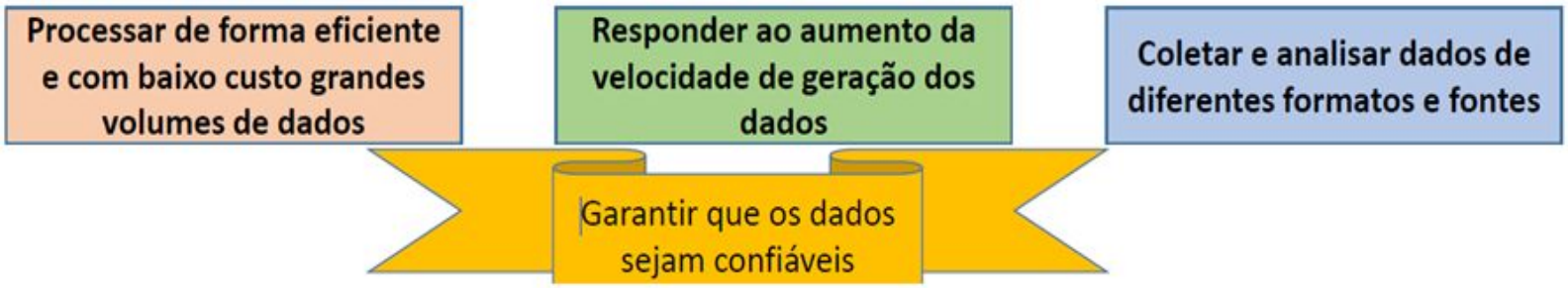
Velocidade
Agilidade com que os dados são gerados

Variedade
Os formatos em que os dados podem ser gerados

Veracidade
Garantia de que as informações são verdadeiras

Volatilidade
Sensíveis ao longo do tempo e periodicidade

Visualização
Apresentação de forma acessível e legível



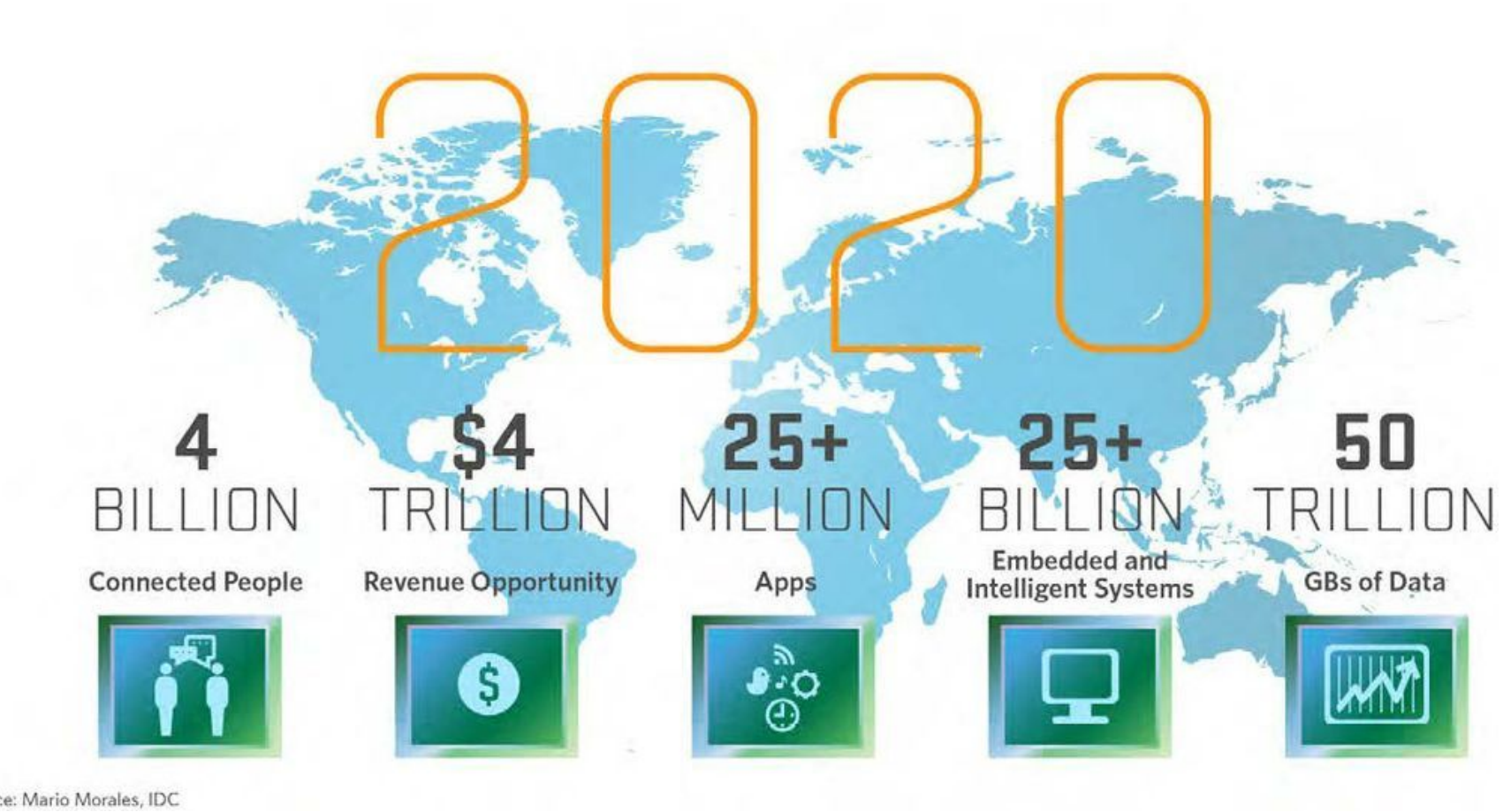
Valor
Análise precisa dos dados, exposição e o resultado para o negócio

Conceitos Crescimento exponencial

Segundo levantamento do IDC até 2020, foram gerados cerca de **40 trilhões de gigabytes**, o que dá uma média de 2,2 milhões de terabytes por dia.

De 2021 a 2024, a previsão é que se crie mais informações do que **nos últimos 30 anos somados.**

Conceitos Crescimento exponencial



Conceitos Variedade e formatos

	Estruturados	Semi-estruturados	Não-estruturados
Internos	 • Resultados de pesquisas • Registros de vendas • Medidas de controle de processos • Bancos de dados de sistemas internos (ERP, CRM)		 • E-mails, cartas, mensagens de texto • Legendas de vídeos • Comentários de clientes • Mensagens de voz • Imagens / ilustrações • Avaliação de funcionários
Externos	• Likes do Facebook, retweets • Horário de publicação de posts, tweets, updates • Pontuação em sites de classificação		• Conteúdo publicado em redes sociais • Comentários em foruns online • Imagens • Vídeos de câmeras de segurança

Conceitos Como gerar valor a partir do dados?

Em um mercado que exige decisões cada vez mais rápidas (velocidade) e assertivas (veracidade), o sucesso está no uso de tecnologias como Big Data Analytics:

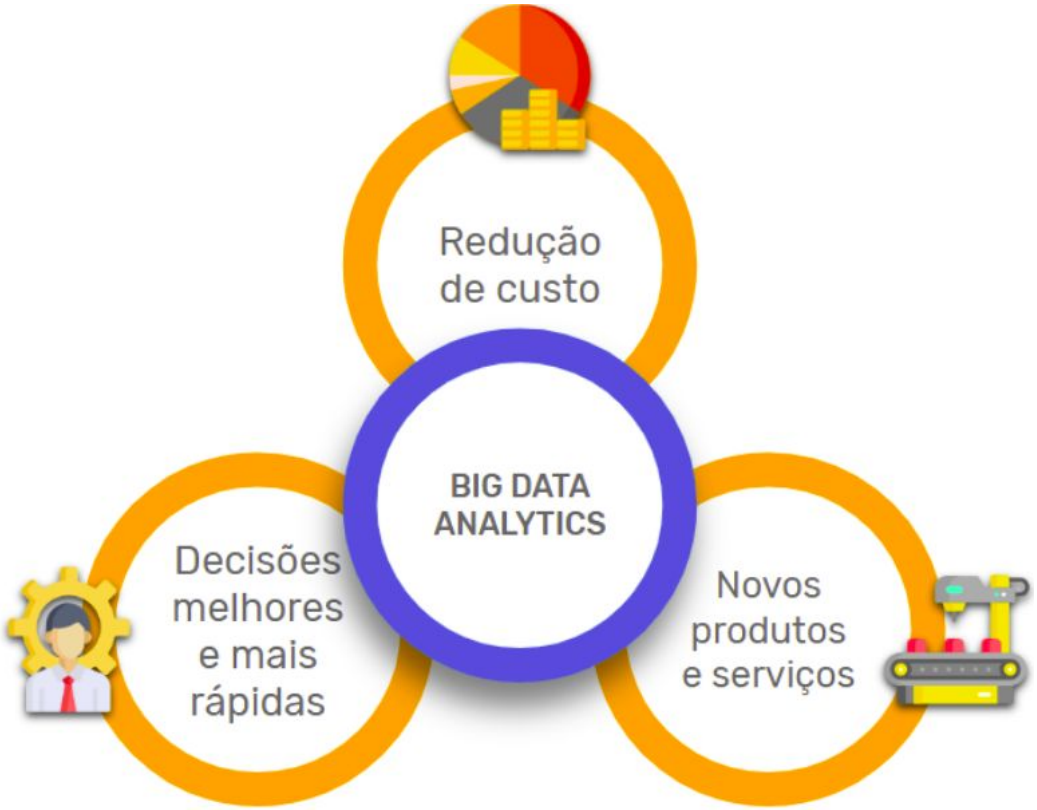
- Para prover melhor integração dos dados
- Para tomar decisões mais assertivas e baseadas em dados (cultura data-driven)
- Para dar respostas estratégicas para o negócio

Valor

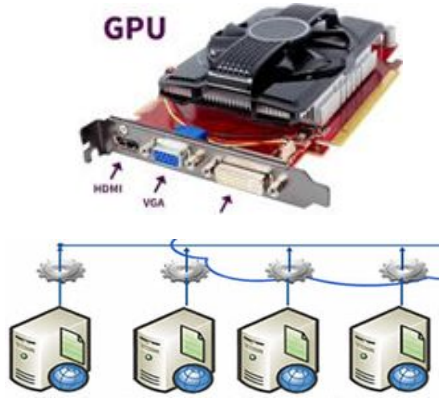
Conceitos

O Big Data Analytics pode ser aplicado em diversos setores

- E-commerce
- Entretenimento
- Marketing digital
- Mídias sociais
- Serviços financeiros
- Energia
- Saúde
- Astronomia
- Segurança da Informação



Crescimento Exponencial na geração de dados



Maior poder de Processamento

TECNOLOGIA

“Se dados são o insumo para construção de inteligência, a **tecnologia** é o instrumento.”

Menor Custo de Storage (Cloud Computing)

Ciclo da construção de inteligência

Coleta e Ingestão de Dados

Estruturação e processamento dos de Dados

Organização e democratização dos Dados

Datasets

BI, Análise e modelos estatísticos

Machine Learning

Relatórios Ad Hoc

Visualização e Dashboards

Aplicativos e APP´s orientados a Dados

Data Engineering **Analytics Engineering** **Data Science**

Conceitos Tipos de análise com dados

BI Tradicional

Ciência de Dados

BIG DATA + CIÊNCIA DE DADOS = BIG DATA ANALYTICS

Fonte Ilustração:
https://medium.com/@edselferri/an%C3%A1lise-preditiva-ou-prescritiva-sua-empresa-precisa-de-ambos-63b49caf09cf

Conceitos

Evolução das Arquiteturas de Dados Analíticos

- **Primeira Geração**

  Arquitetura de Data Warehouse - 163

- **Segunda Geração**

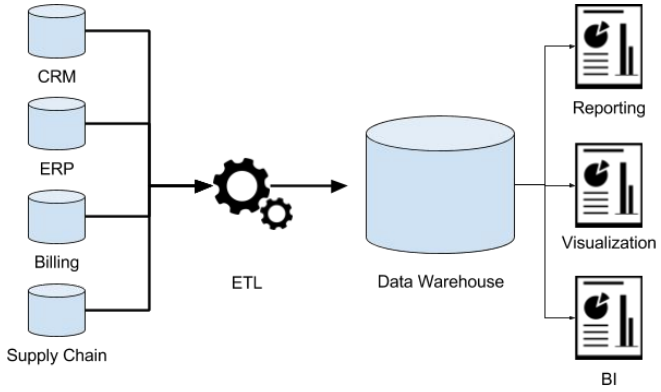
  Arquitetura Data Lake

- **Terceira Geração**

  Arquitetura de Nuvem Multimodal

Conceitos Primeira Geração

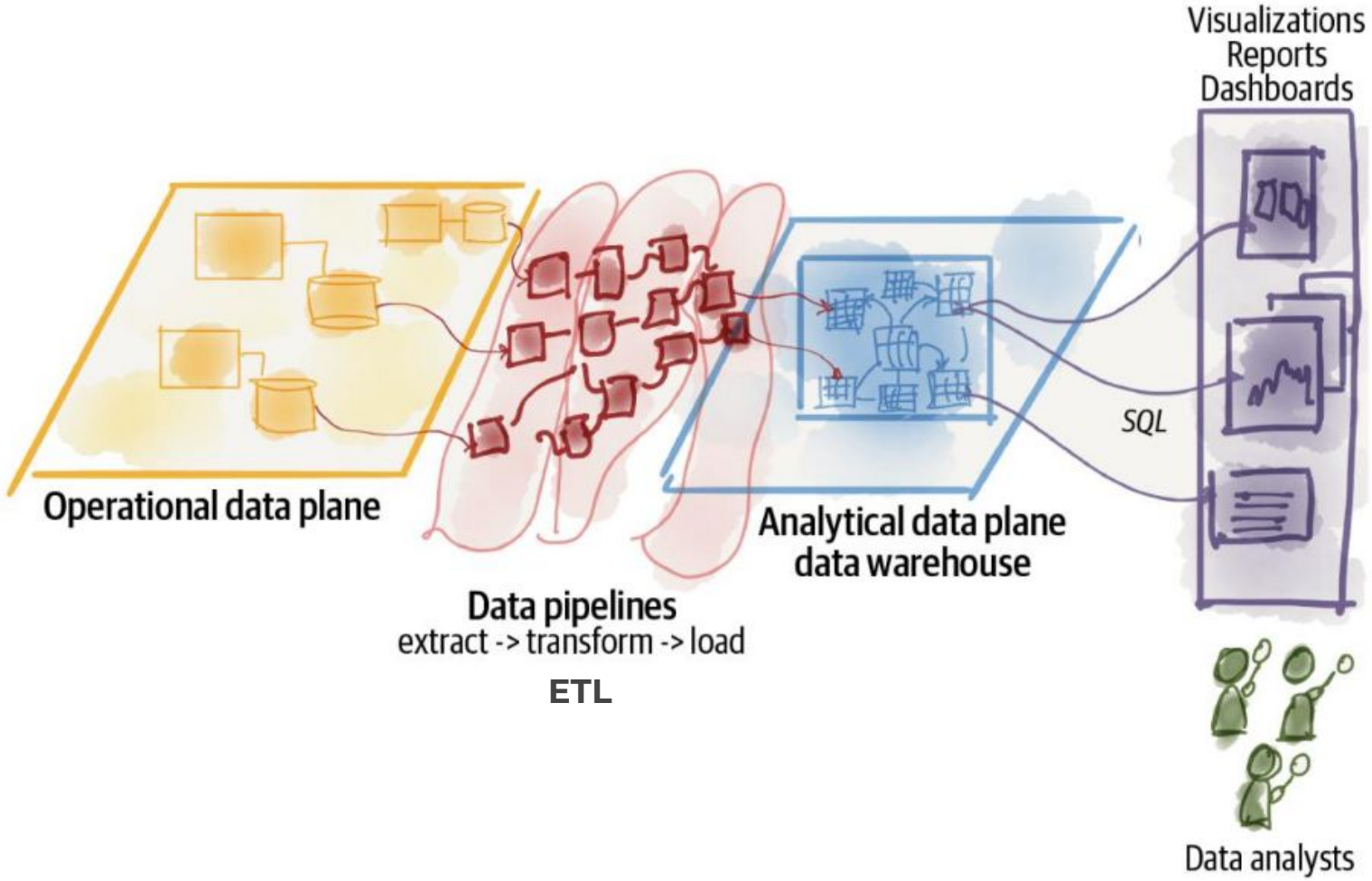
Arquitetura de Data Warehouse



- Extração de dados de sistemas operacionais para armazenamento em sistemas de Business intelligence (BI)
- Armazenamento estruturado em modelagens multidimensionais como fato e dimensões
- Alimentado por uma camada de ETL onde são transformados para atender a modelagem pré-definida
- Acessados por meio de consultas analíticas
- Dificuldade de armazenamento e processamento em alta escala, se tornando um ambiente de custo elevado.

Conceitos Primeira Geração

Arquitetura de Data Warehouse



Fonte Ilustração:
Livro O'Reilly Data Mesh Delivering Data-Driven Value at Scale - Autora Zhamak Dehghani

Conceitos Segunda Geração

Arquitetura de Data Lake

- Data Lake é um conceito, um repositório de armazenamento que contém uma grande quantidade de dados, sendo eles estruturados, semi-estruturados e não estruturados.
- Os dados não necessariamente precisam de uma modelagem prévia.
- A premissa primária é coletar e armazenar os dados em **DFS** (Distributed File System), e sob demanda, definir e realizar o processamento e análise a ser feita sobre os dados.

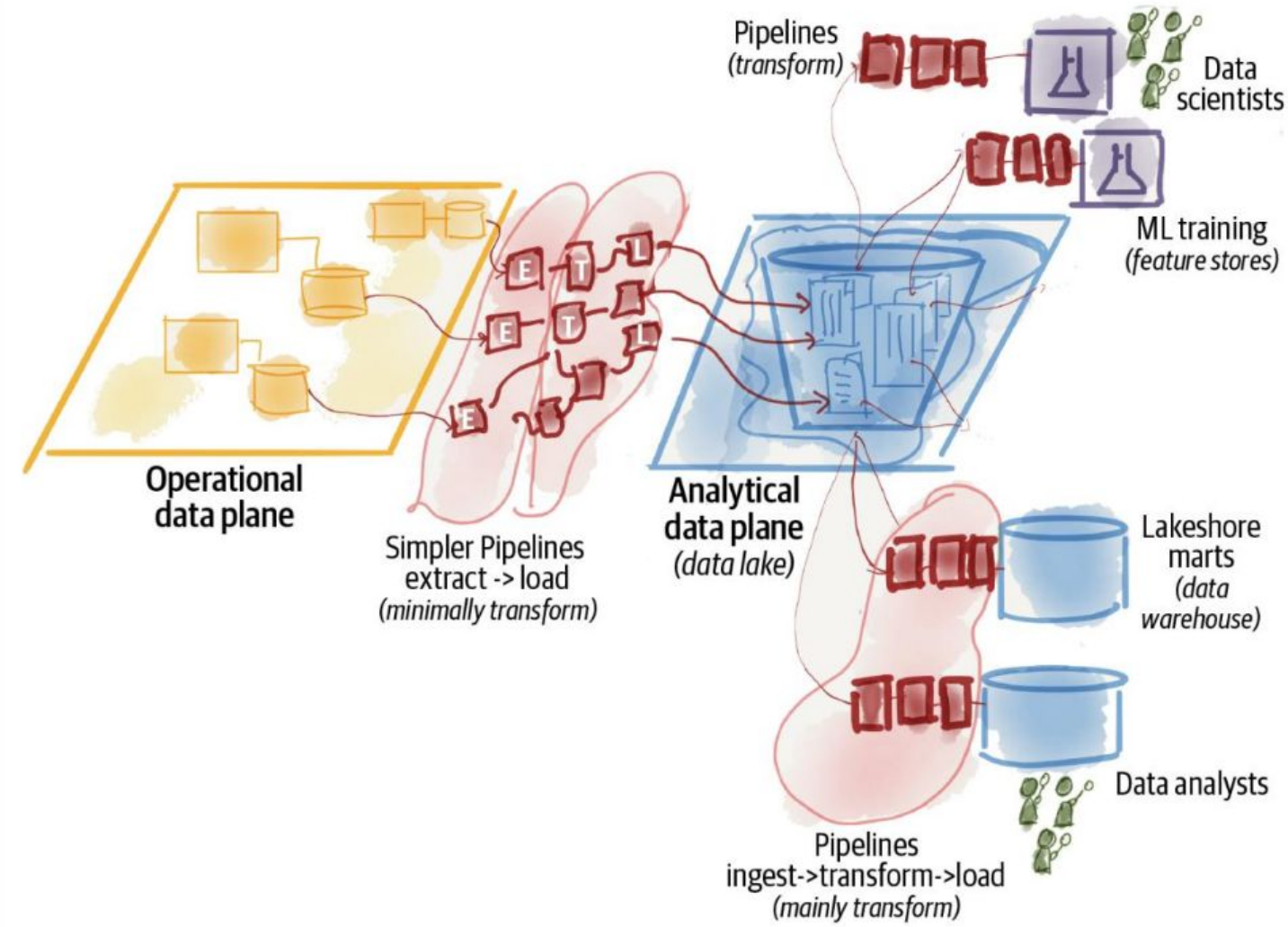
Conceitos Segunda Geração

Arquitetura de Data Lake

- Tecnologias relacionadas a clusters de computadores, storages e frameworks para processamento paralelo e distribuídos.
- Um ambiente que possibilite escalabilidade, alta disponibilidade e flexibilidade.

Conceitos Segunda Geração

Arquitetura de Data Lake



Fonte Ilustração:
Livro O'Reilly Data Mesh Delivering Data-Driven Value at Scale - Autora Zhamak Dehghani

Conceitos Data Warehouse x Data Lake

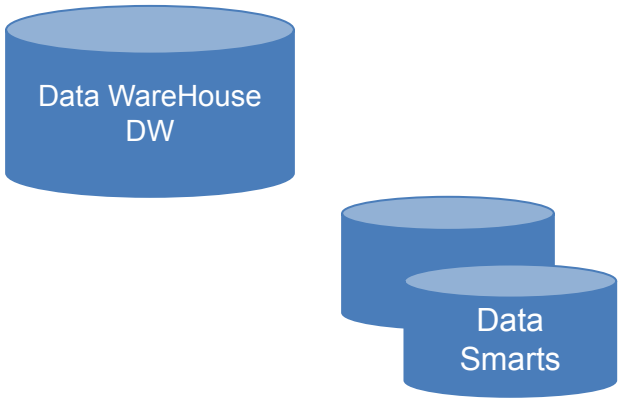
Principais Diferenças

DATA WAREHOUSE		DATALAKE
Estruturados	**Dados**	Estruturados, semi estruturados e não estruturados
Schema Pré-definido	**Modelagem**	Sem schema inicial e estruturação sob demanda
Conceito ETL - Tradicional	**Processamento**	Conceito ELT – Processamento Paralelo e Distribuído
SGBDs (Bancos de Dados relacionais)	**Armazenamento**	File Systems Distribuídos (Hadoop HDFS, Aws S3, Google GCS)
Alto custo para altos volumes	**custo**	Baixo custo e altos volumes
Consolidada	**Segurança**	Em evolução

Conceitos Data Warehouse x Data Lake

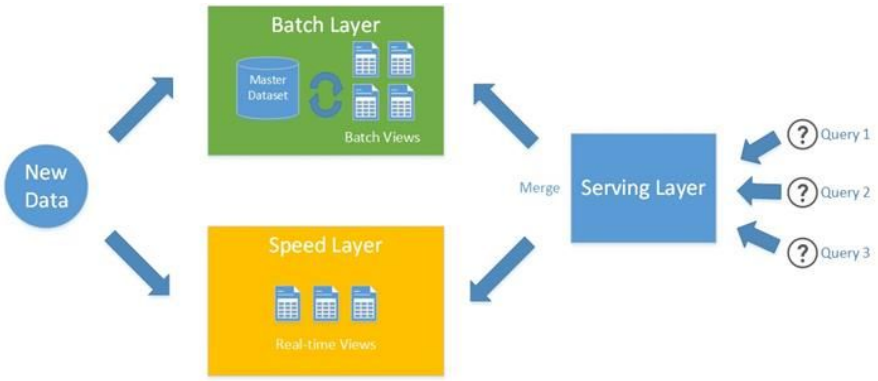
"Dentro de um contexto de plataforma de dados, o Data Lake venha a contribuir com o Data Warehouse. Onde o Data Lake será apenas uma das peças da arquitetura."

Dados estruturados e refinados (Analytical Storages)



Conceitos Terceira Geração

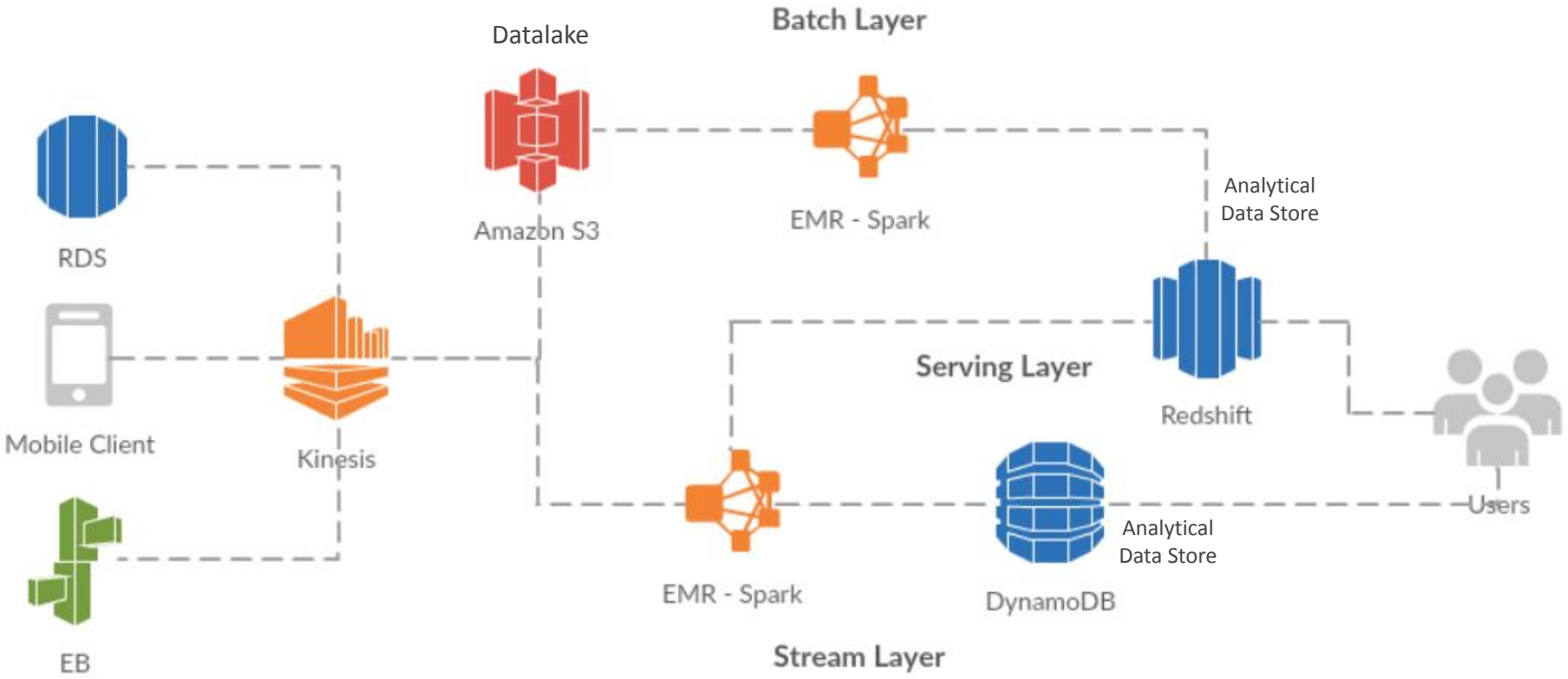
Arquitetura de Nuvem Multimodal



- Mais ou menos semelhante às gerações anteriores, com arquiteturas em nuvem e a modernidade dessas tecnologias
- Suporte ao streaming e análise em tempo real, com arquiteturas utilizando kafka e spark streaming por exemplo
- Aproveitamento da elasticidade da computação em nuvem e controles de custo de infraestrutura de Big Data.
- Convergência do warehouse e o data lake em uma tecnologia, incorporando treinamentos de ML, integridade e governança.

Conceitos Terceira Geração

Arquitetura de Nuvem Multimodal



Exemplo de Arquitetura Lambda em Nuvem usando AWS

Fonte Ilustração:
https://medium.com/@carleto/arquitetura-lambda-para-big-data-usando-a-aws-36d4c9eb968

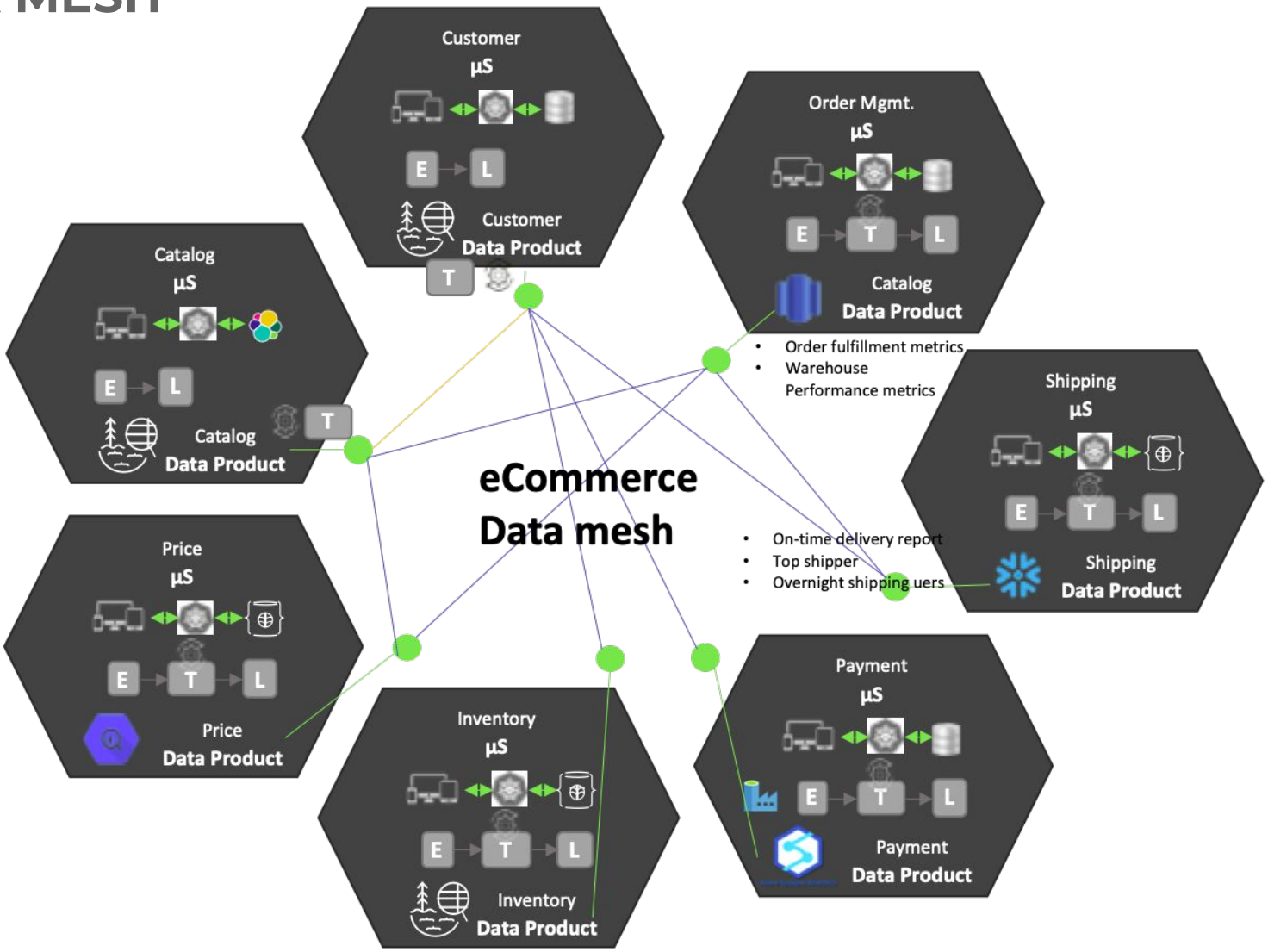
Conceitos Uma nova Geração?

DATA MESH

- Arquitetura descentralizada e baseada em “domínios”
- Pensar em dados como Produtos criados pelos seus respectivos domínios e consumidos por outros
- Plataforma Self Service para autonomia dos ownerships de domínios na criação de seus produtos de dados.
- Governança computacional federada com políticas definidas e sistematizada na plataforma

Conceitos Uma nova Geração

DATA MESH



Fonte Ilustração:
https://www.peerislands.io/demystifying-data-mesh/

No próximo Capítulo veremos:

- Por que usar Python para Big Data
- Conhecendo a Linguagem Python
- Hands on

"Os dados são importantes ativos para as empresas, pois são a bases para construção de Conhecimento"