Modernização de Dados:
De data lake a modern data stack

ORACLE

Regina
Cantele

Trilha Data Cloud Modernization

Regina Cantele

linkedin.com/in/regina-cantele-86a169

Agenda

1. Evolução
2. Data Warehouse e Delta Lake
3. Modern Data Stack
4. Tecnologias Oracle

Era das Fraldas e Cerveja

Primeira Geração

+ Dados transacionais

Alguns dados semiestruturados

DataWarehouses

fatos

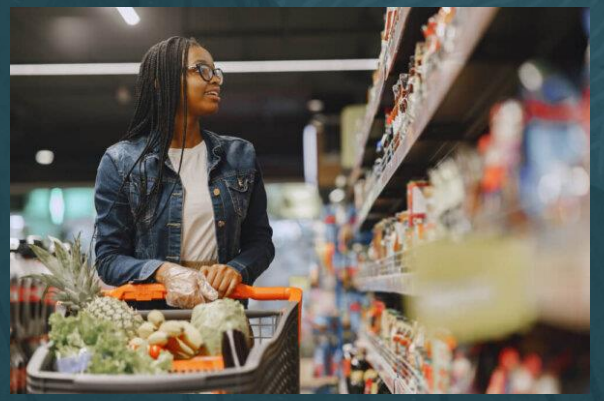
dimensões

Extract, Transform and Load (ETL)

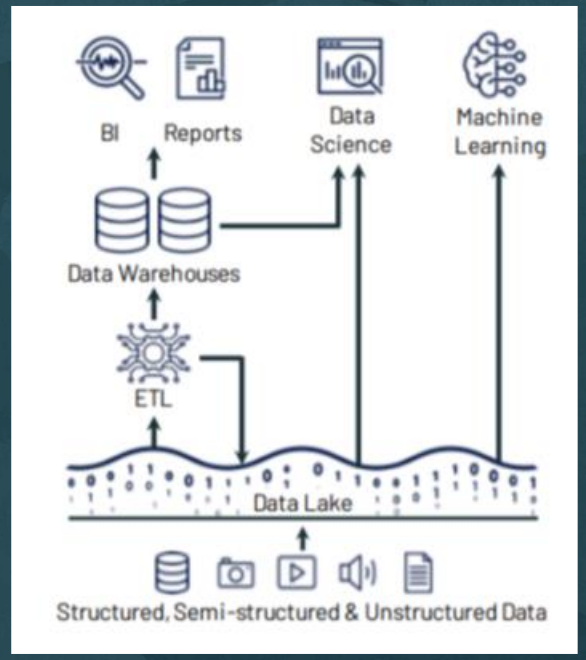
OLAP e cubos

Symmetric Multi-Processing (SMP)

Massive Parallel Processing (MPP).



Era Big Data



Segunda Geração

Volume, Variedade e Velocidade

Todos tipos de dados

Data Lake e NoSQL

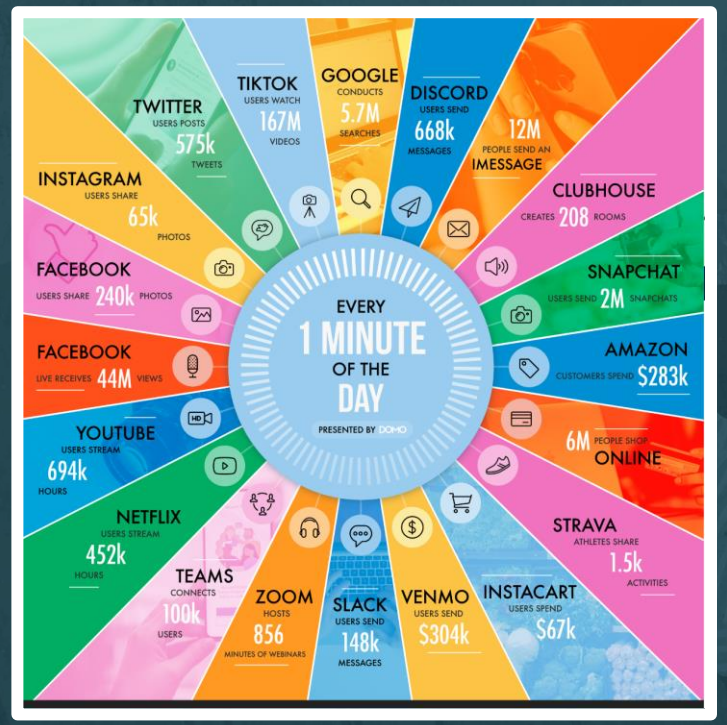
Extract Load Transform (ELT)

Elastic Parallel Processing (EPP)

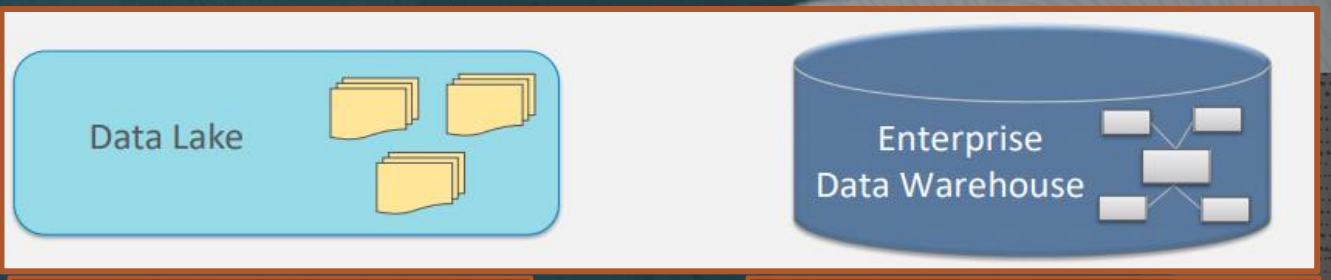
Lambda e Kappa

Armazenamento objetos

Ciência de Dados



Era Big Data



Vantagens:

- Agilidade
- Flexibilidade
- Entrega rápida
- Exploração

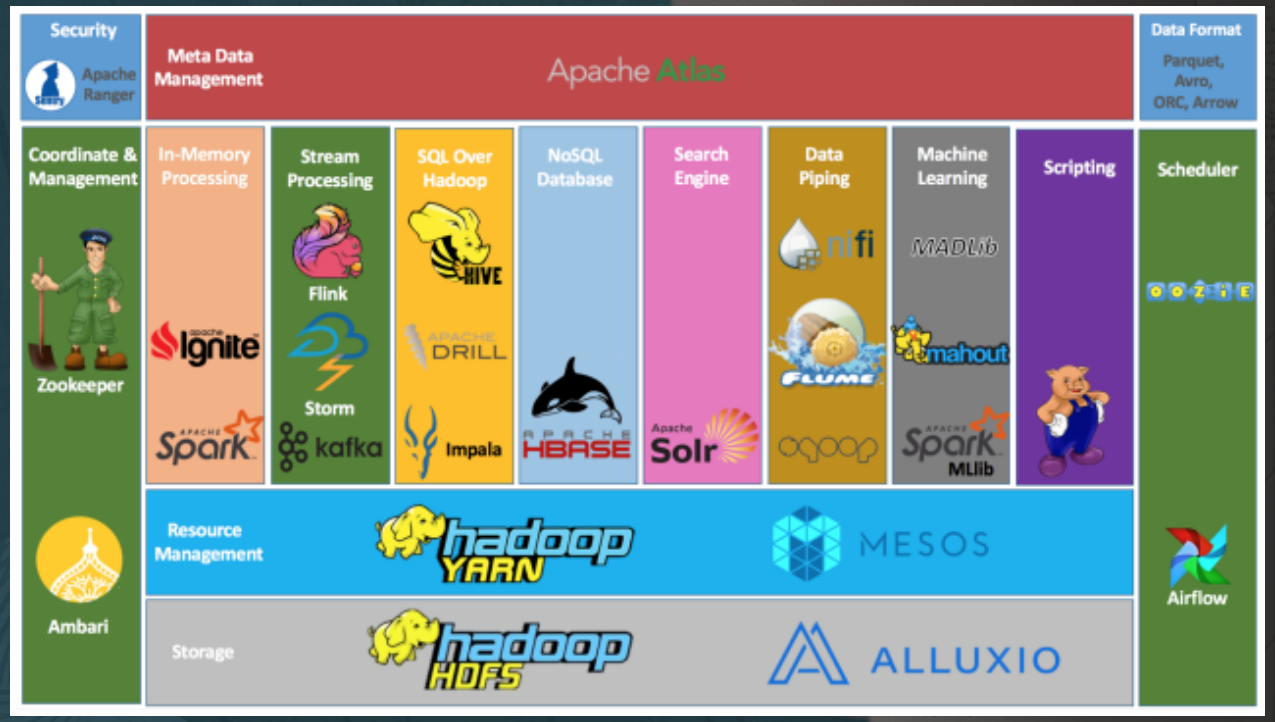
Vantagens:

- Governança
- Padrões
- Segurança
- Confiabilidade

Esforço Menor	**Aquisição dos dados**	Esforço Maior
Esforço Maior	**Recuperação dos dados**	Esforço Menor

Era Big Data

Soluções On Premise

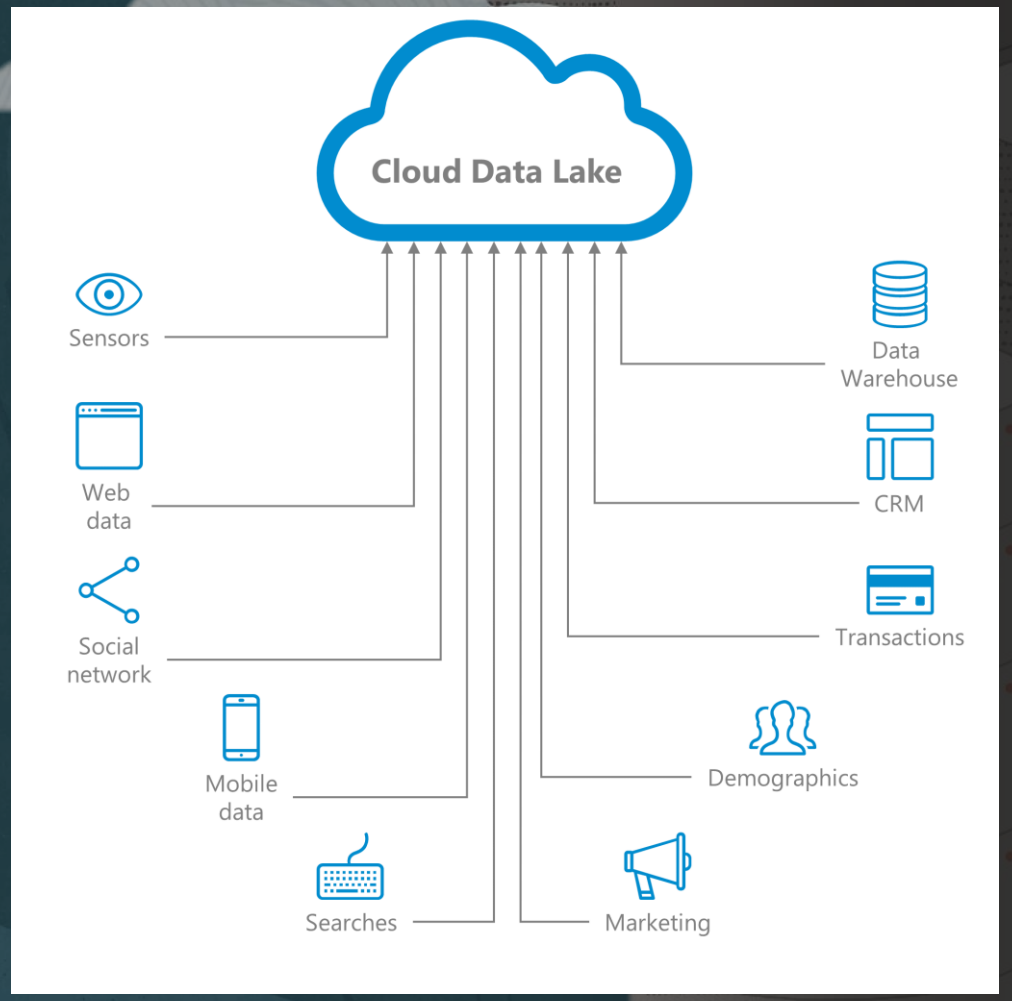


Era Big Data

Cloud Data Lake

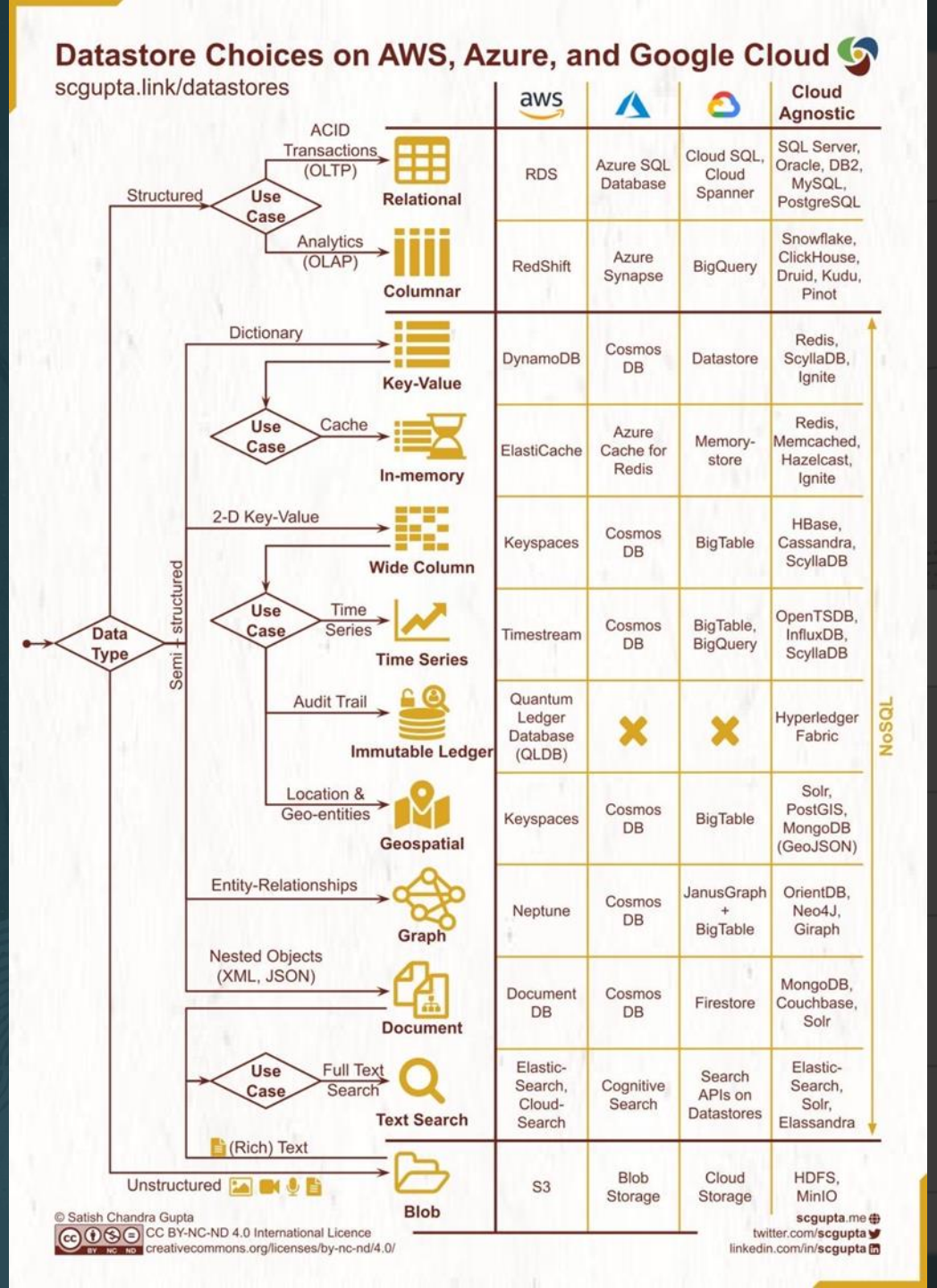
Flexibilidade e **escalabilidade**, separando a capacidade de **armazenamento e computação** em camadas fisicamente distintas, conectadas por conexões de rede rápidas.

Permite dimensionar sua capacidade de armazenamento à medida que o volume de dados aumenta e dimensione independentemente sua capacidade de processamento.

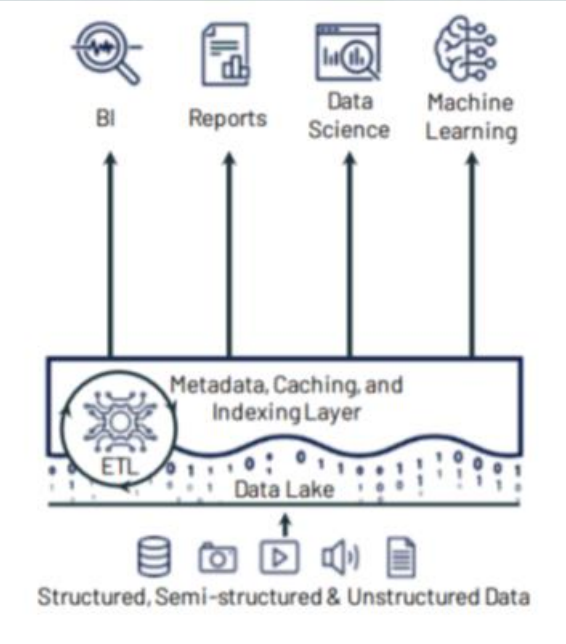


Era Big Data

Cloud Data Lake



Era Delta Lake Lakehouse

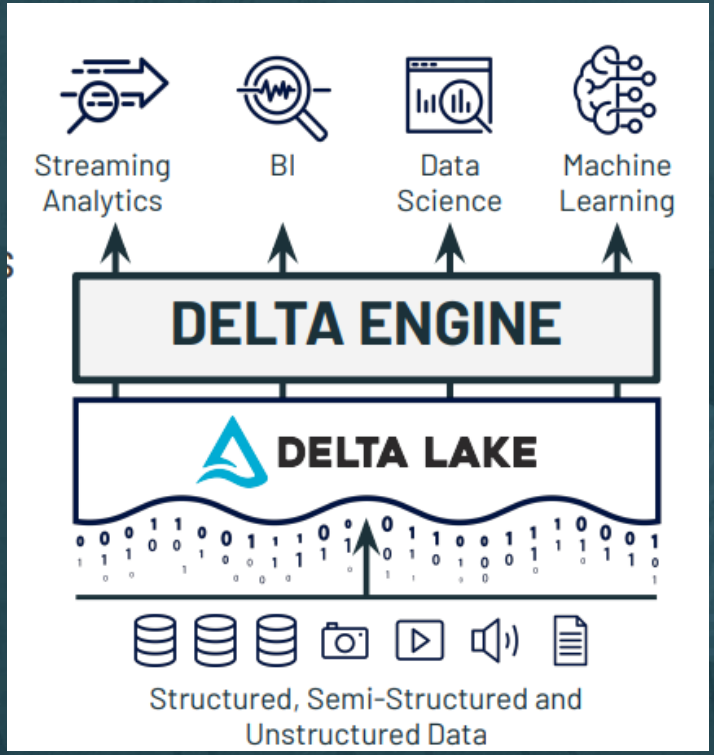


Terceira Geração

Simplicidade

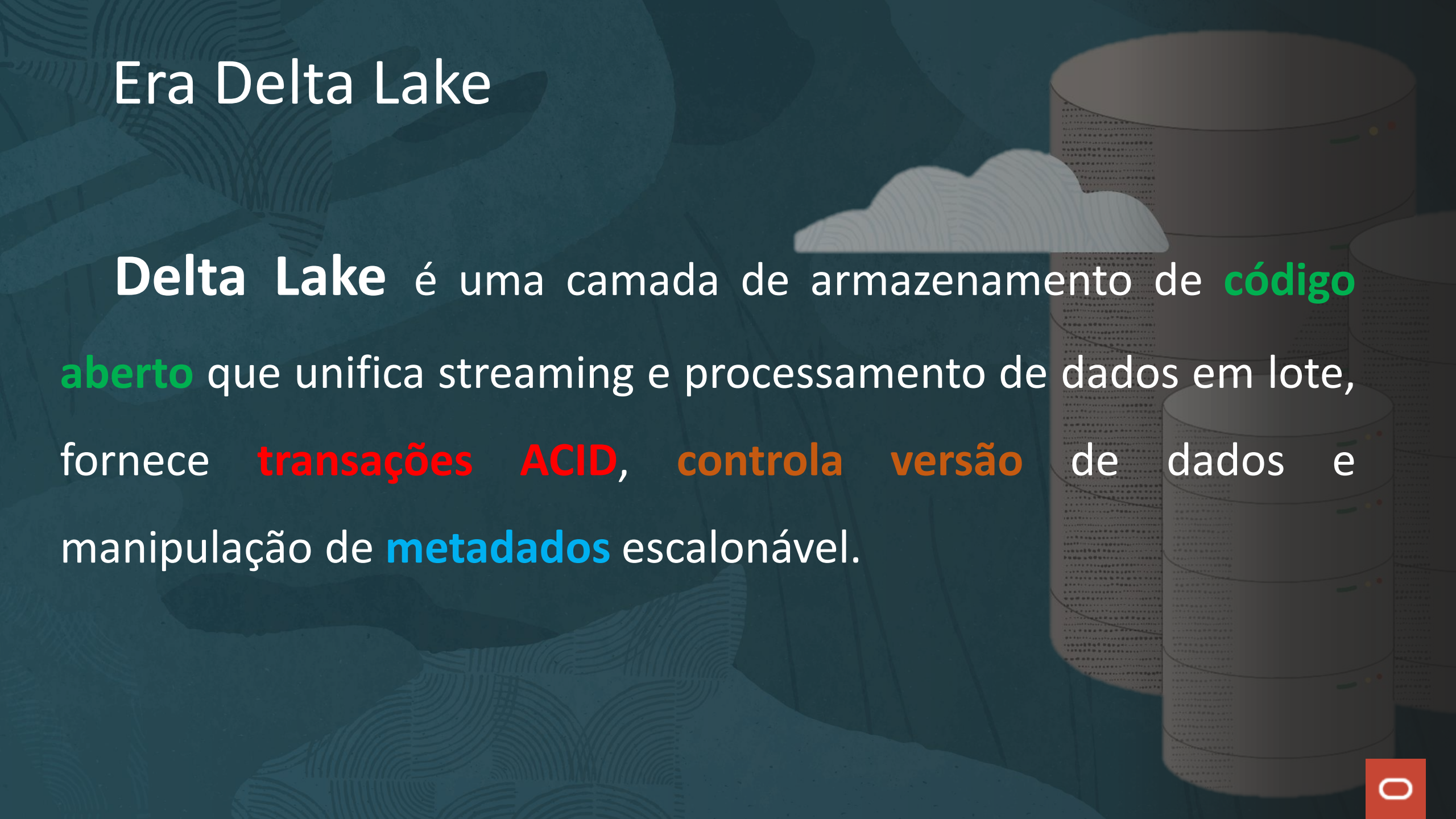
Escalabilidade

Custos operacionais mais baixos

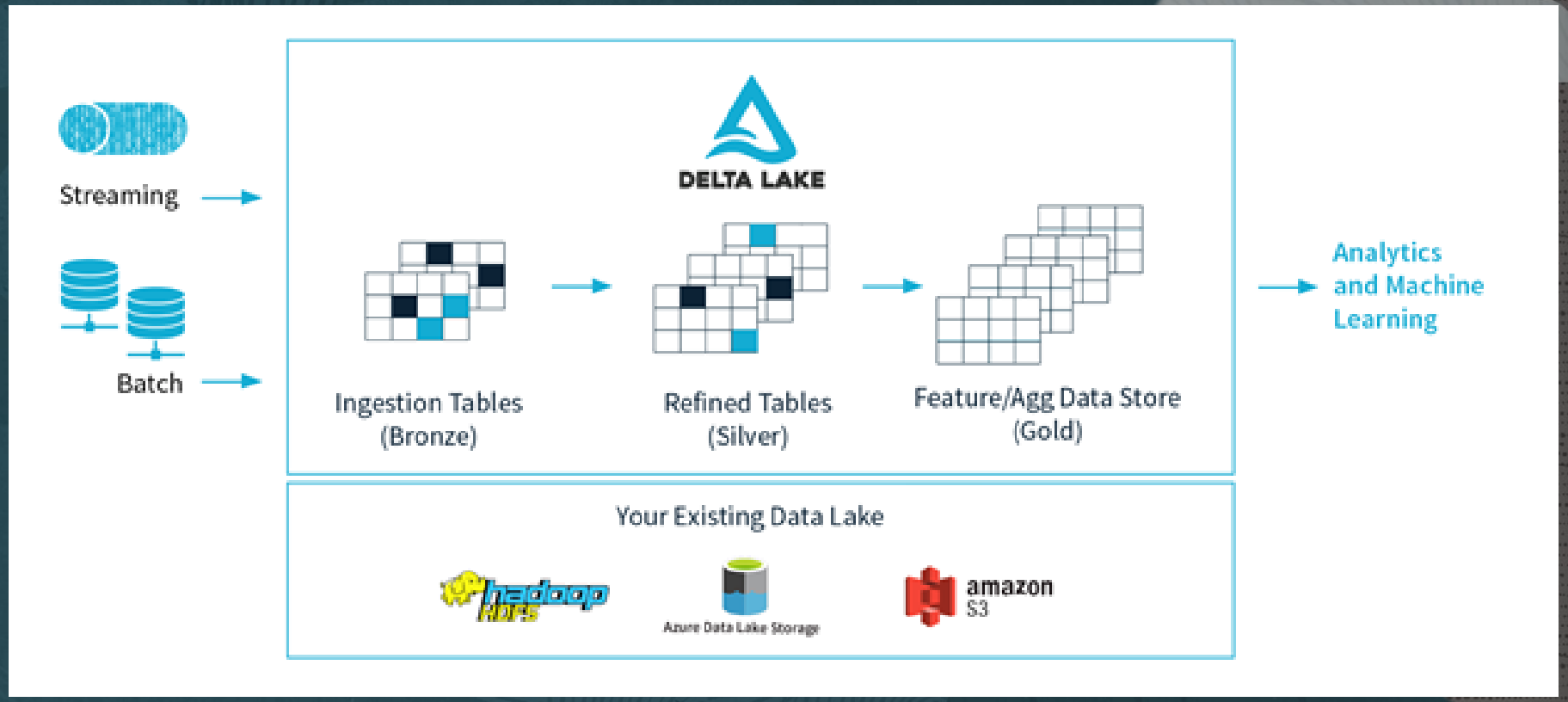


Era Delta Lake

Delta Lake é uma camada de armazenamento de **código aberto** que unifica streaming e processamento de dados em lote, fornece **transações ACID**, **controla versão** de dados e manipulação de **metadados** escalonável.

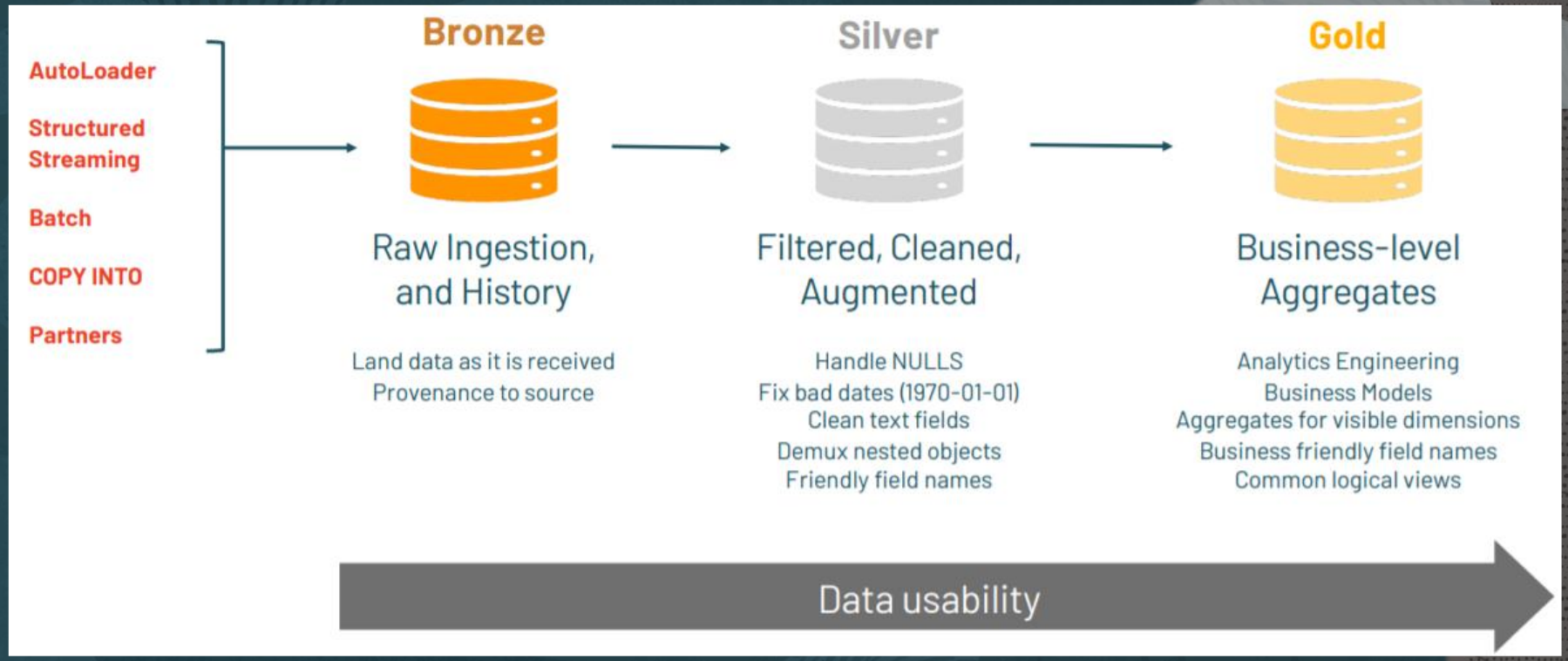


Era Delta Lake



Delta Lake roda em cima do data lake existente e é totalmente compatível com APIs Apache Spark.

Era Delta Lake



Era Delta Lake

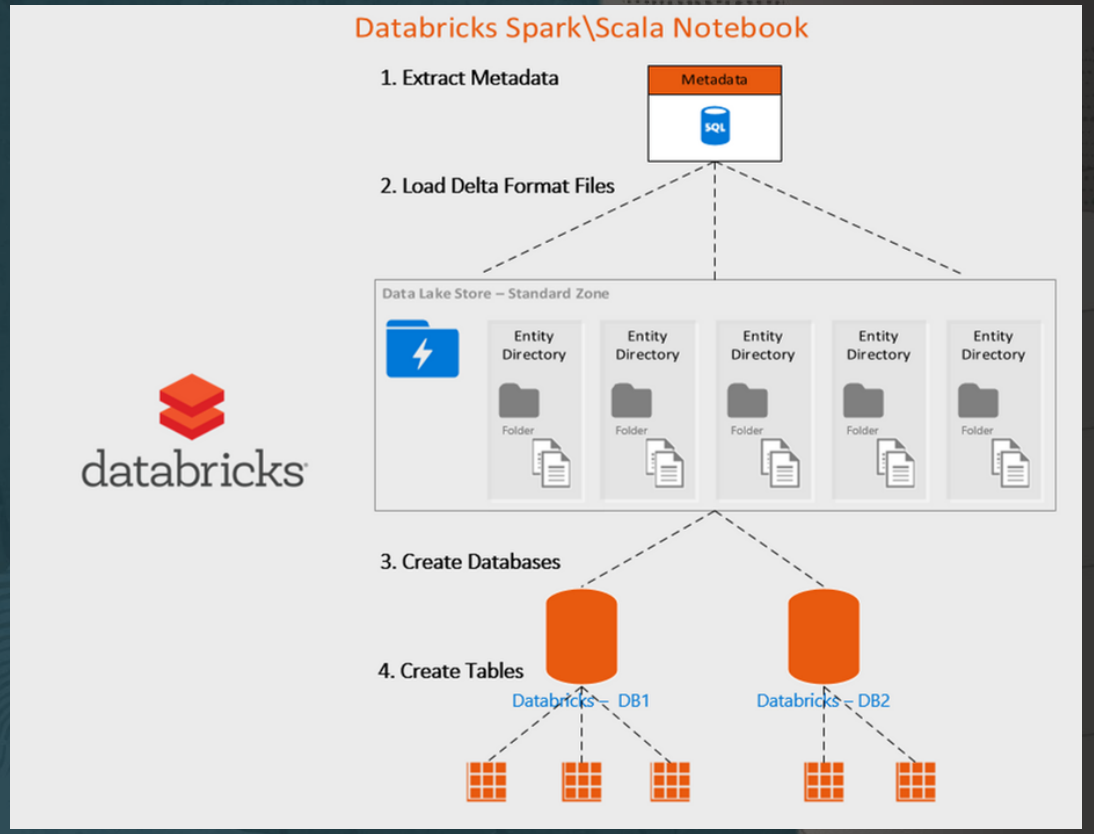
Databricks

- ✓ uma plataforma de análise baseada no Apache Spark.
- ✓ projetado com os fundadores do Apache Spark.
- ✓ permite fluxos de trabalho simplificados e um workspace interativo para colaboração entre os cientistas de dados, os engenheiros de dados e os analistas de negócios.

Era Delta Lake

Databricks

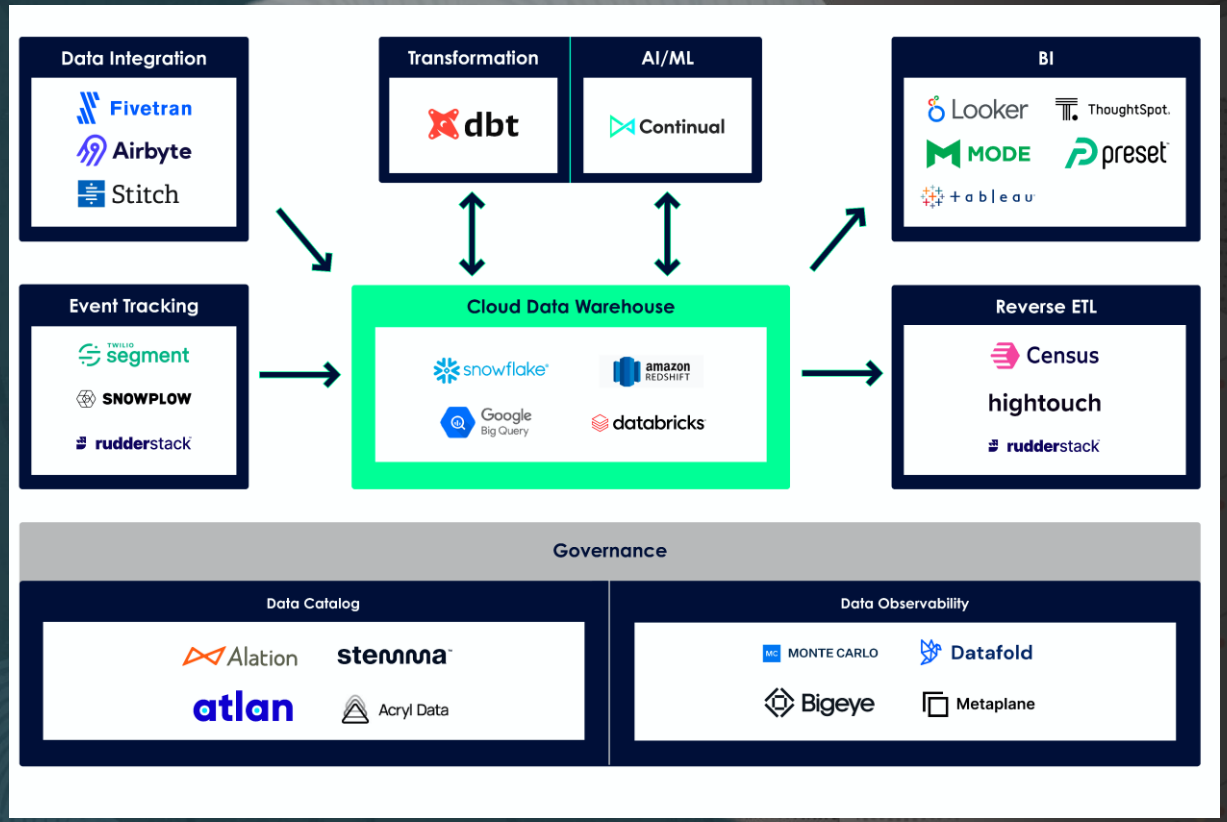
- Metadados
- Delta format
- Databases
- Tabelas



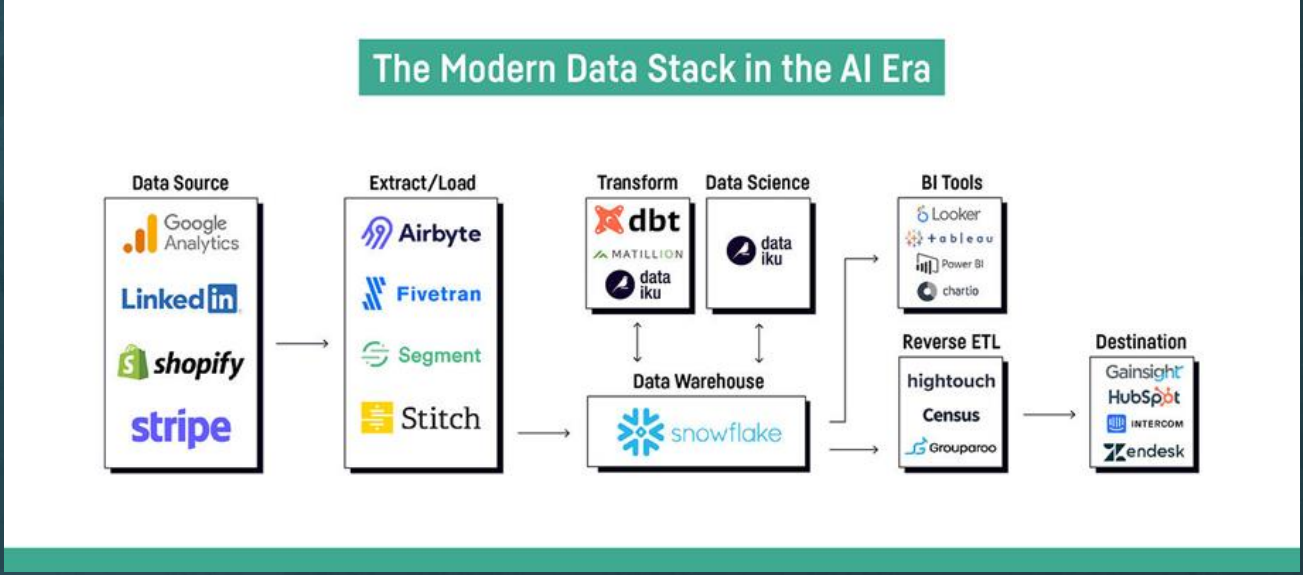
Modern Data Stack

Modern Data Stack se refere a uma coleção de tecnologias que compõem uma plataforma de dados nativa na nuvem com o objetivo de reduzir a complexidade na execução de uma plataforma de dados tradicional.

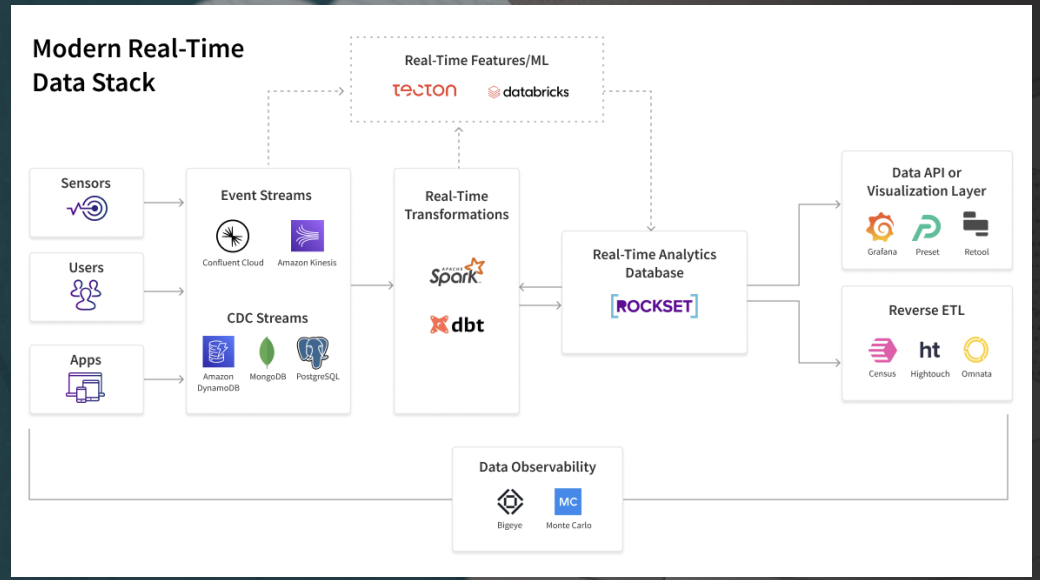
Desde 2010 com Big Query.



Modern Data Stack

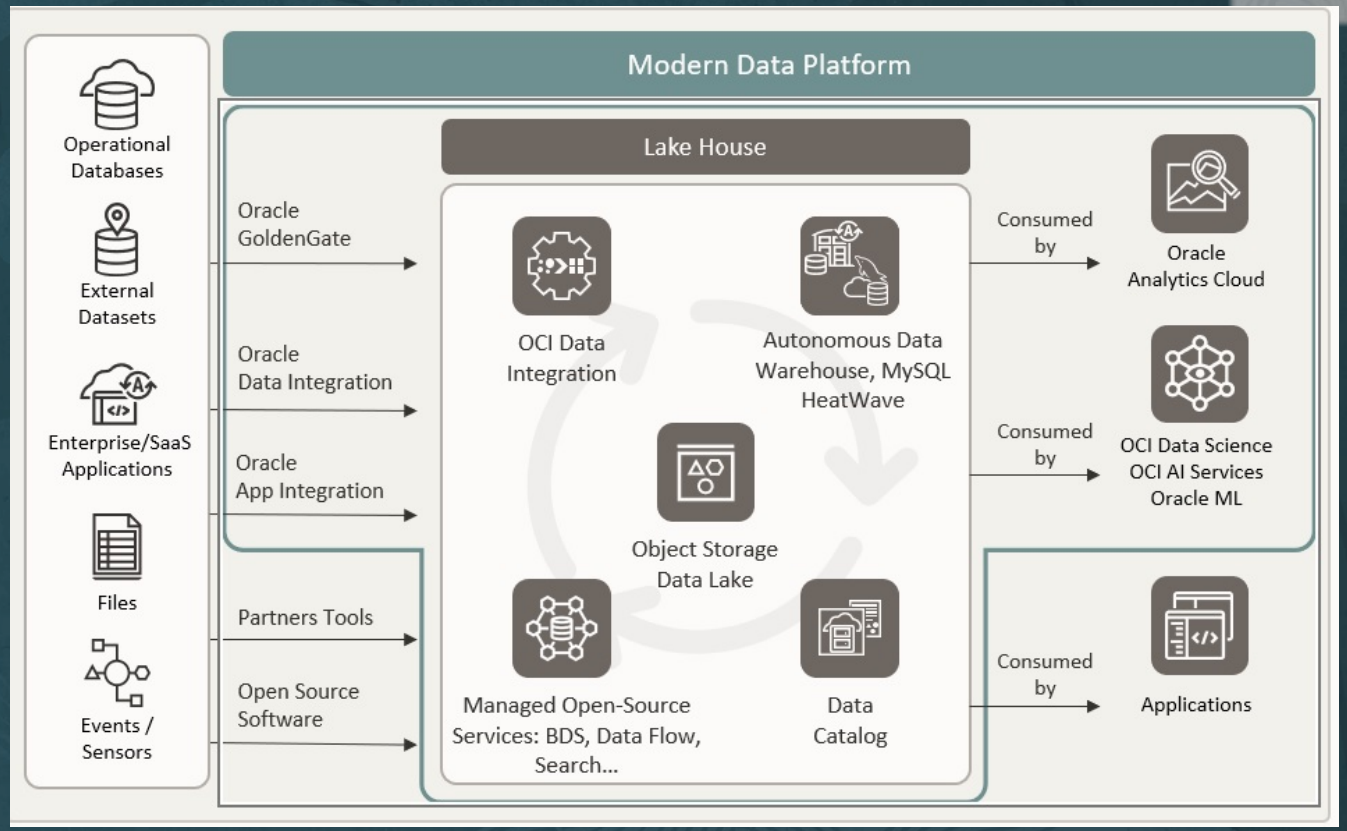


- Fluxos de captura de dados de eventos e alterações para ingestão
- ETL em tempo real (ou ELT) para transformações em tempo real
- Banco de dados de análise em tempo real para análises rápidas de dados atualizados



- API de dados ou camada de visualização
- ETL reverso para enviar insights para aplicativos de negócios
- Observabilidade de dados para garantir a qualidade dos dados em escala

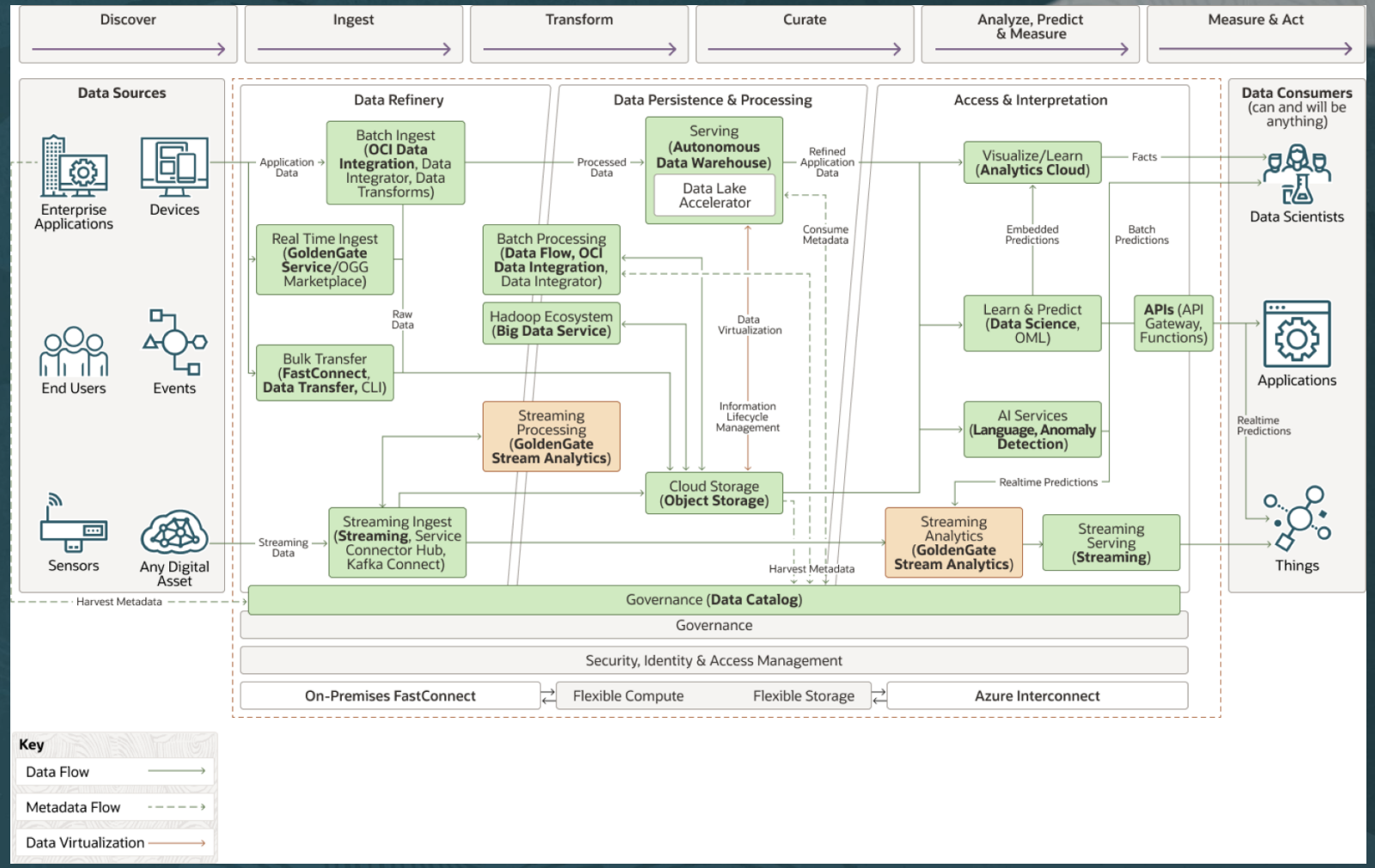
Tecnologias Oracle Modern Data Plataform



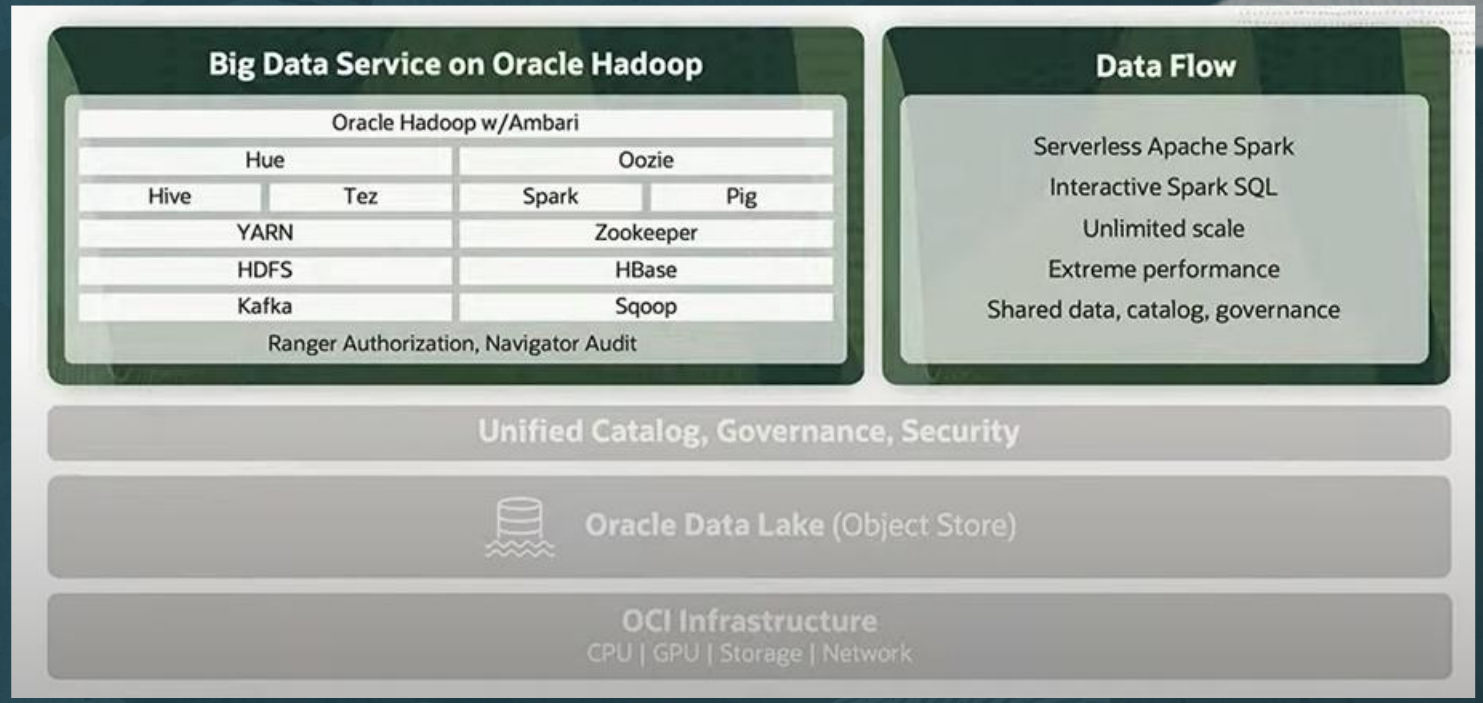
Os principais elementos do padrão do Oracle Lakehouse incluem

Integração de padrões de data warehouse e data lake.
Eliminação de silos de dados: fácil movimentação de dados entre armazém e lago, conforme necessário.
Metadados e governança unificados.
Suporte para ferramentas comerciais e de código aberto populares.
Suporte para uma ampla variedade de fontes de dados, formatos e tipos de dados (estruturados, semiestruturados e não estruturados)
Suporte a diversos consumidores e cargas de trabalho de dados, incluindo análise avançada de big data, SQL e BI, ciência de dados e aprendizado de máquina em todos os setores.

Tecnologias Oracle
Oracle Arquitetura de Referência

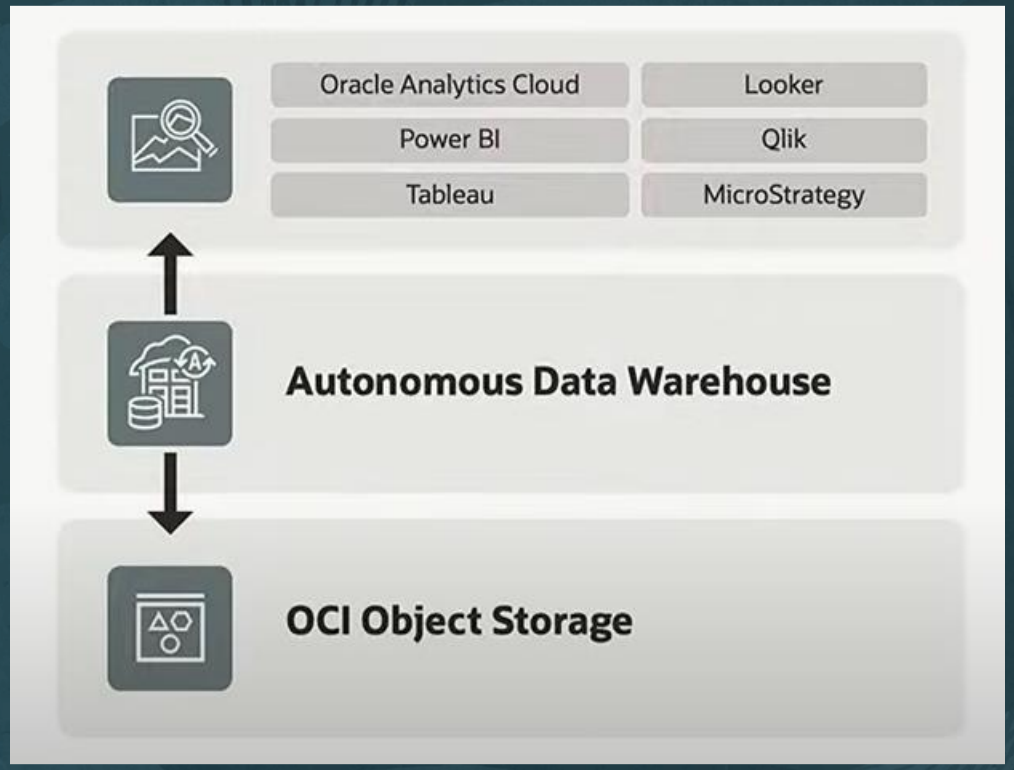


Tecnologias Oracle
Oracle Intelligent Lakehouse Plataform



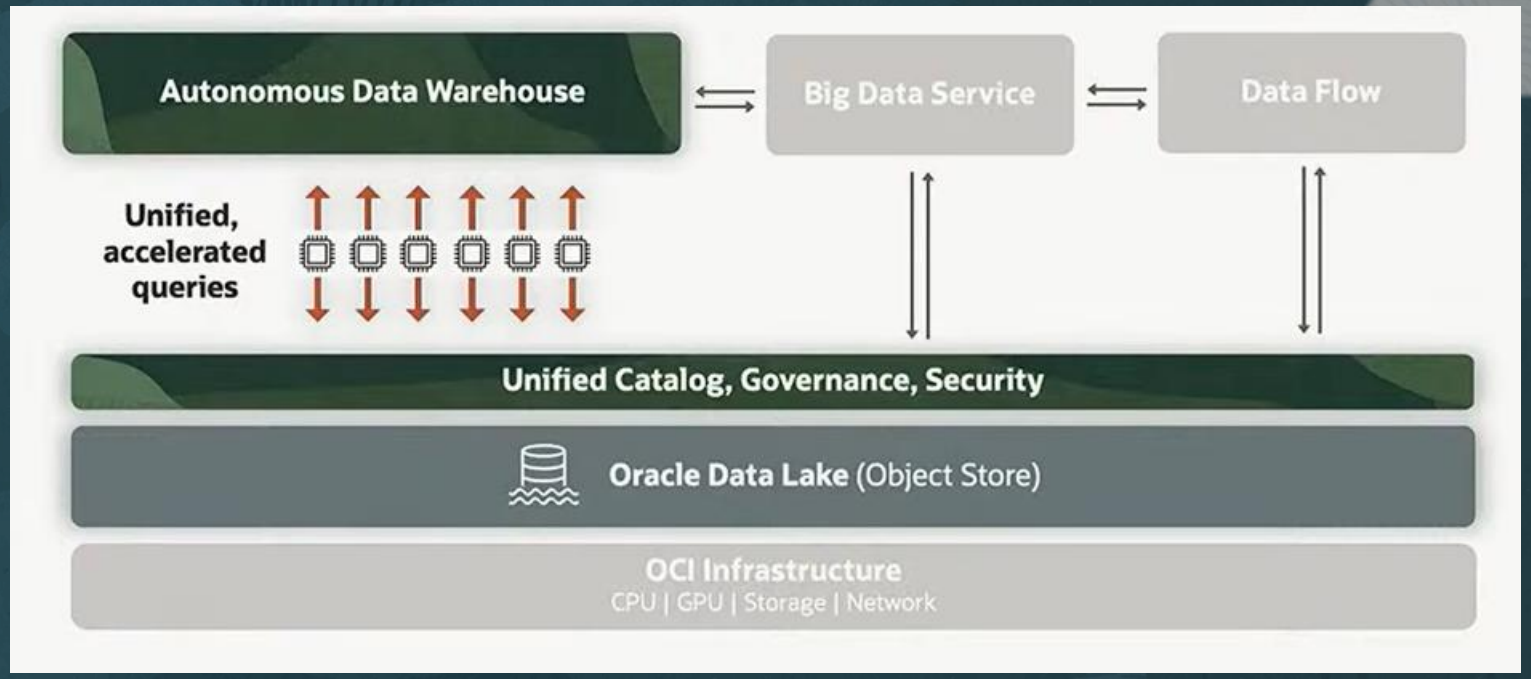
OCI Big Data Service e **OCI Data Flow** para gerenciar os serviços open-source com armazenamento e catálogos compartilhados.

Tecnologias Oracle Autonomous Data Warehouse



Analisar dados no data warehouse e data lake em escala com mesma query.

Tecnologias Oracle



Tecnologias Oracle

Oracle Big Data

fornece clusters com um ambiente Hadoop.

Data Catalog do Oracle Cloud Infrastructure

uma solução de governança e descoberta de dados;

um ambiente colaborativo único para gerenciar metadados técnicos, de negócios e operacionais.

Oracle Cloud Infrastructure Data Flow

um serviço gerenciado para executar aplicativos Apache Spark, suas dependências, parâmetros padrão e uma especificação de recurso de runtime padrão.

Oracle Autonomous Data Warehouse

um serviço de banco de dados independente, com autoproteção e autorreparo que é otimizado para cargas de trabalho de data warehouse.

Oracle Cloud Infrastructure Data Integration

um serviço de nuvem *server less* gerenciado para ingerir e transformar dados para simplificar processos complexos de ETL/E-LT em data lakes e warehouses.

Oracle Cloud Infrastructure Data Science

uma plataforma gerenciada, *server less* para cientistas de dados criarem, treinarem, implantarem e gerenciarem modelos de aprendizado de máquina e usarem a biblioteca ADS (Accelerated Data Science) aprimorada por **Oracle for Automated Machine Learning** (AutoML), avaliação de modelos e explicação de modelos.

Oracle Cloud
Modo Gratuito

https://bit.ly/TDCConnections2022

Uso Livre

(Always Free)

Serviços que você pode
usar por tempo ilimitado

\+

Avaliação Gratuita de 30 dias

US$ 500 em créditos gratuitos

O que está incluído no Oracle Cloud – Modo Gratuito (Free Tier)?

Uso Livre (Always Free)

Serviços em nuvem de Uso Livre:

- Dois Oracle Autoinomous Databases com ferramentas avançadas como Oracle APEX e Oracle SQL Developer
- Duas VMs de Computação AMD
- Até 4 instâncias em ARM Ampere A1 Compute
- Armazenamento de Bloco, Objetos e Arquivos; Balanceador de Carga e Saída de Dados; Monitoramento e Notificações

Avaliação Gratuita de 30 dias

US$ 500 em créditos gratuitos (exclusivo para os participantes do TDC Connections 2022)

- Acesso a uma ampla variedade de serviços da Oracle Cloud por 30 dias, incluindo Bancos de Dados, Análise Avançada, Computação e Container Engine for Kubernetes
- Até oito instâncias em todos os serviços disponíveis
- Até 5 TB de armazenamento

Mais detalhes em: https://www.oracle.com/br/cloud/free/

Muito Obrigada!!!