

MBA⁺

**Data Science &
Artificial Intelligence**



**MBA⁺****Data Architecture,
Integration and Ingestion**

Prof.: Ivan Gancev

Email: profivan.gancev@fiap.com.br



Ivan Gancev

- Formado em Processamento de Dados
- MBA em Gestão Estratégica de Projetos
- No mercado de tecnologia desde 2002
- Atuação em consultoria, gestão de projetos, governança de tecnologia
- Atuação nos segmentos de indústrias, varejo e financeiro
- Head de data engineering de plataformas de dados de ponta a ponta
- Responsável por jornada de Cloud (adoção, cultura, migração) e transformação digital



Data Architecture, Integration and Ingestion

(O que vamos explorar?)

Aula 1 – 12/abr (qua)

- Pilares de arquitetura: persistência, integração e consumo
- Estratégias de arquitetura
- Tipos de tratamentos e arquiteturas

Aula 2 – 19/abr (qua)

- Exemplos de Bancos, diferenças e usos:
 - Bancos Relacionais
 - Bancos Colunares

Aula 3 – 26/abr (qua)

- Exemplos de Bancos, diferenças e usos:
 - Bancos de documentos
 - Bancos chave-valor
 - Bancos de Grafos

Aula 4 – 03/mai (qua)

- Ingestão de dados, tratamentos e manipulações
- Pipeline de dados, governança e qualidade
- Integração de dados
 - Cargas batch, ETL, vantagens e desvantagens

Aula 5 – 10/mai (qua)

- Eventos, APIs, NRT e casos de uso
- Arquiteturas para analytics
- Boas práticas, recomendações e cuidados

Projeto Integrado DS&IA



Definição

Dado

Que caracteriza, que qualifica alguma coisa

Arquitetura de dados

Conjunto de regras, políticas e referências que governam como os dados são coletados, transformados e utilizados pelos sistemas de uma empresa

Banco de dados

Conjuntos de dados armazenados e organizados para serem acessados com facilidade e segurança.

Tipos de dados

Estruturados

Semi-estruturados

Não estruturados

Dados estruturados

Dados que possuem uma estrutura rígida e pré-definida. Seus formatos determinam tamanhos e capacidades.

Ex: código de barras, CPF, combo-box, QR Code

Dados semi estruturados

Eventualmente estruturado, podendo possuir uma parte com formatação rígida e pré-definida, juntamente com uma parte mutável.

Ex: Arquivos delimitados, XML, e-mails, HTML

Dados não estruturados

Não possui formatação ou estrutura padronizada

Ex: Fotos, vídeos, audios, emojis, sticker do whatsapp, gif

Questão

Quando um dado é armazenado ele passa a ser estruturado?

Pilares de arquitetura de dados

Persistência

Consiste no armazenamento de um dado em uma área não volátil e que possibilita que este dado possa ser recuperado futuramente no mesmo estado em que foi armazenado

Integração

É a consolidação de dados de origens distintas em um novo conjunto de dados, que atende a uma necessidade de negócio ou de novas informações

Consumo

Uso de dados para análises de cenários, históricos, previsões e qualquer outra informação utilizada para alguma tomada de decisão

Questão

Toda operação de um sistema de dados vai sempre utilizar um ou mais dos pilares de arquitetura?

Sistemas de Gerenciamento de Banco de Dados (SGBD)

Histórico e surgimento dos SGBD

FIAP

No início da década de 1960 surgiu o primeiro SGBD criado por Charles Bachman, chamado de integrated Data Store. Depois a IBM desenvolveu o Information Management System, um sistema hierárquico de dados utilizados por mainframes.

Na década de 1970 surgiram os primeiros sistemas relacionais que foram adotados por empresas no mundo todo e são amplamente usados até hoje.

Na década de 1990 surgiram os bancos NoSQL, que evoluíram muito desde então.

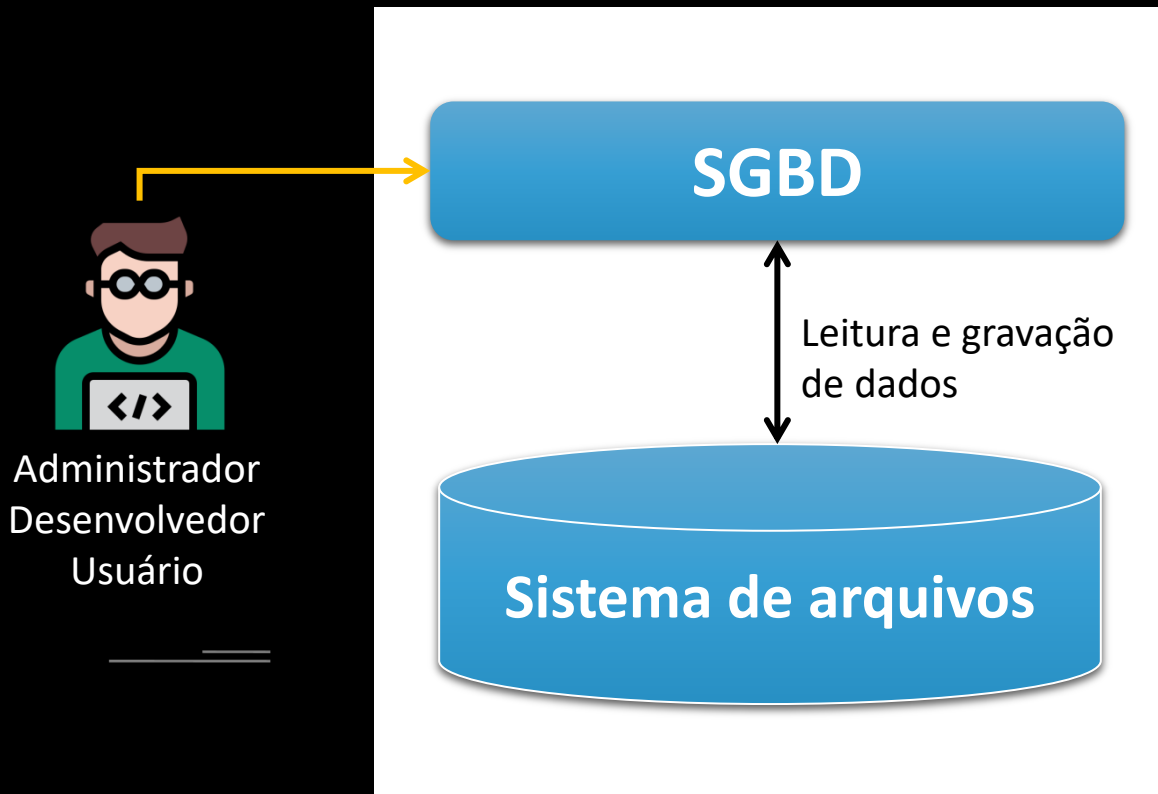
Definição de SGBD

Um Sistema de Gerenciamento de Banco de Dados (SGBD) é um software desenvolvido para armazenar, administrar e consultar dados. Possui regras e critérios sobre como os dados serão inseridos, lidos, alterados e removidos.

Além disso são capazes de gerenciar as permissões de cada ação sobre os dados. Fazendo o elo entre o usuário que manipula os dados e o armazenamento físico.

DBMS: Database Management System
RDBMS: Relational Database Management System

Funcionamento de um SGBD



- Composto por um conjunto de módulos que constituem o banco de dados
- Gerencia o armazenamento físico dos dados
- Controla permissões de acessos e operações
- Interpreta e valida os comandos solicitados
- Controla as sessões de usuários
- Efetiva as operações de manipulação dos dados
- Demais interações com o Sistema Operacional (SO)

Tipos de bancos de dados



Relacionais

- São os tipos de banco de dados mais populares
- As tabelas se relacionam através de chaves primárias e chaves estrangeiras
- Possuem propriedade ACID (Atomicidade, Consistência, Isolamento e Durabilidade)
- Escalabilidade vertical

Not Only SQL

- Capazes de armazenar seus dados como objetos
- Gerenciam grandes volumes de dados
- Consistência eventual
- Alta disponibilidade
- Escalabilidade horizontal

DB-ENGINES

Act in Time.
Build on InfluxDB.
The platform for building time series applications.

[Try for Free](#)

Knowledge Base of Relational and NoSQL Database Management Systems

[Home](#) [DB-Engines Ranking](#) [Systems](#) [Encyclopedia](#) [Blog](#) [Sponsors](#) [Search](#) [Vendor Login](#)

 Featured Products: [Neo4j](#) [Vertica](#) [Redis](#) [MariaDB](#) [Cassandra](#)
DB-Engines

DB-Engines is an initiative to collect and present information on database management systems (DBMS). In addition to established relational DBMS, systems and concepts of the growing NoSQL area are emphasized.

The [DB-Engines Ranking](#) is a list of DBMS ranked by their current popularity. The list is updated monthly.

The most important properties of numerous systems are shown in the overview of [database management systems](#). You can examine the properties for each system, and you can compare them side by side.

In the [database encyclopedia](#) terms and concepts on this topic are explained.

The latest news

[Snowflake is the DBMS of the Year 2022, defending the title from last year](#)

3 January 2023, Matthias Gelbmann, Paul Andlinger

[Turbocharge Your Application Development Using WebAssembly With SingleStoreDB](#)

17 October 2022, Akmal Chaudhri, SingleStore (sponsor)

[Ten years of DB-Engines.com](#)

13 October 2022, Matthias Gelbmann, Paul Andlinger

Featured Products

SkySQL, the ultimate MariaDB cloud, is here.

[Get started with SkySQL today!](#)



See for yourself how a graph database can make your life easier.

[Use Neo4j online for free.](#)

<https://db-engines.com/>

DB-Engines Ranking

The DB-Engines Ranking ranks database management systems according to their popularity. The ranking is updated monthly.

Read more about the [method](#) of calculating the scores.



410 systems in ranking, March 2023

Rank			DBMS	Database Model	Score		
Mar 2023	Feb 2023	Mar 2022			Mar 2023	Feb 2023	Mar 2022
1.	1.	1.	Oracle +	Relational, Multi-model i	1261.29	+13.77	+9.97
2.	2.	2.	MySQL +	Relational, Multi-model i	1182.79	-12.66	-15.45
3.	3.	3.	Microsoft SQL Server +	Relational, Multi-model i	922.01	-7.08	-11.77
4.	4.	4.	PostgreSQL +	Relational, Multi-model i	613.83	-2.67	-3.10
5.	5.	5.	MongoDB +	Document, Multi-model i	458.78	+6.02	-26.88
6.	6.	6.	Redis +	Key-value, Multi-model i	172.45	-1.39	-4.31
7.	7.	7.	IBM Db2	Relational, Multi-model i	142.92	-0.04	-19.22
8.	8.	8.	Elasticsearch	Search engine, Multi-model i	139.07	+0.47	-20.88
9.	9.	↑ 10.	SQLite +	Relational	133.82	+1.15	+1.64
10.	10.	↓ 9.	Microsoft Access	Relational	132.06	+1.03	-3.37
11.	↑ 12.	↑ 14.	Snowflake +	Relational	114.40	-1.26	+28.17
12.	↓ 11.	↓ 11.	Cassandra +	Wide column	113.79	-2.43	-8.35
13.	13.	↓ 12.	MariaDB +	Relational, Multi-model i	96.84	+0.03	-11.47
14.	14.	↓ 13.	Splunk	Search engine	87.97	+0.89	-7.39
15.	15.	↑ 16.	Amazon DynamoDB +	Multi-model i	80.77	+1.08	-1.03
16.	16.	↓ 15.	Microsoft Azure SQL Database	Relational, Multi-model i	77.44	-1.31	-7.23
17.	17.	17.	Hive	Relational	70.91	-1.21	-10.31
18.	18.	18.	Teradata	Relational, Multi-model i	63.74	+0.71	-5.11
19.	19.		Databricks	Multi-model i	60.86	+0.52	

É um ranking dos principais sistemas de gerenciamento de banco de dados ordenado por sua popularidade no mercado

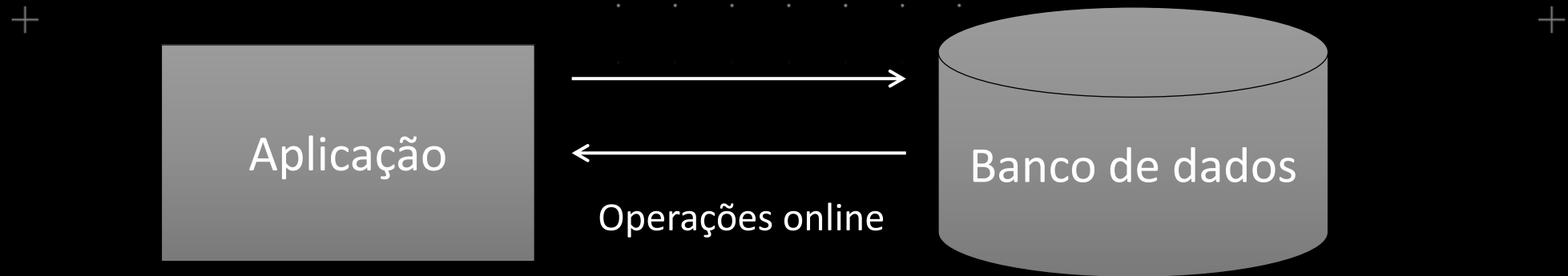
Serve como uma referência e também para buscar mais informações sobre algum SGBD

Vale sempre consultá-lo para manter-se atualizado com relação ao que o mercado está usando e o que surge de novo.

Tipos de arquitetura de dados

The background of the slide is a dark gradient, transitioning from a deep purple on the left to a black on the right. A large, faint, light-purple hexagonal shape is centered on the slide. Inside this hexagon, there is a complex pattern of dots and lines, resembling a stylized fingerprint or a data visualization. The dots are arranged in concentric, wavy lines, and there are some straight lines intersecting the pattern.

Arquitetura Transacional



As operações realizadas pela aplicação são refletidas instantaneamente no banco dados e sua confirmação é síncrona. Abordagem do tipo OLTP (Online Transaction Processing), com múltiplas transações simultâneas

— Aplicado em sistemas do time cliente-servidor, sistemas com cargas de trabalho sem elasticidade, sistemas com necessidade de consistência

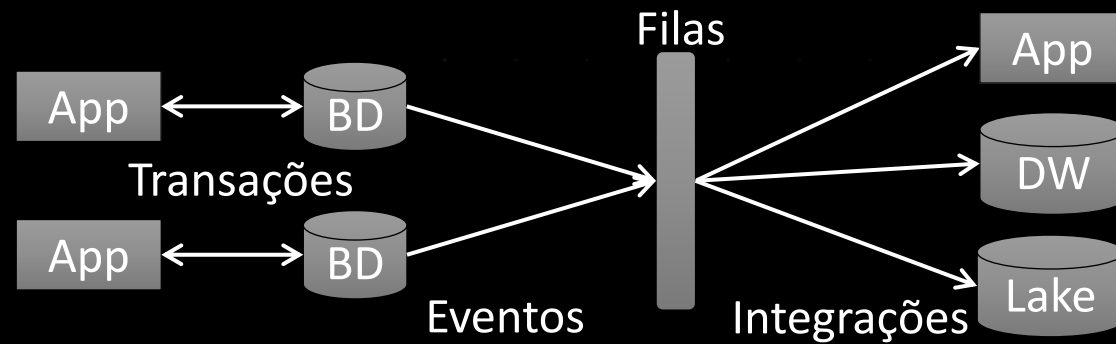
Arquitetura Analítica



Os dados gerados em operações são integrados em bases de dados construídas para análises complexas que combinam diferentes fontes de dados e seus históricos. Abordagem OLAP (Online Analytical Processing) para consultas múltiplas e robustas.

Aplicado em sistemas modelados para este propósito como data warehouse (DW) e Business Intelligence (BI).

Arquitetura aplicações digitais



Dados gerados por aplicações distribuídas e com cargas de trabalho elásticas, necessidade de integração em tempo real e por evento gerado

— Aplicado em sistemas que interagem com o cliente final ou que operem escalas muito grandes como um e-commerce, um centro de distribuição, aplicativos móveis.

Data warehouse

Surgimento do Data warehouse



Os Data warehouses surgiram com o crescimento de sistemas transacionais e volumes de dados gerados pelas empresas que trouxeram a necessidade de análises de dados mais robustas e que pudessem suportar decisões estratégicas de negócio.

Os sistemas transacionais (OLTP) não eram capazes de manter suas transações juntamente com o processamento de dados e geração de relatório (surgimento do mercado de BI)

Data Warehouse

Definição segundo Bill Inmon

“Um Data Warehouse é uma coleção de dados, orientado a assuntos, integrado, variável em relação ao tempo e não volátil, para suporte ao gerenciamento dos processos de tomada de Decisão”

Abordagem Top-down

Data Warehouse

Definição segundo Ralph Kimball

“Um Data Warehouse é conjunto de dados transacionais, especificamente estruturados para pesquisas e análises de negócio”

Abordagem Bottom-up

Características

- Orientado a assunto de negócio



Diversos sistemas transacionais combinados fornecem os dados para compor o assunto “cliente”

- Dados não voláteis



Sua persistência e duração é permanente

- Integração



Combinação de diferentes dados integrados formam um dado de negócio

- Dados variam em função do tempo

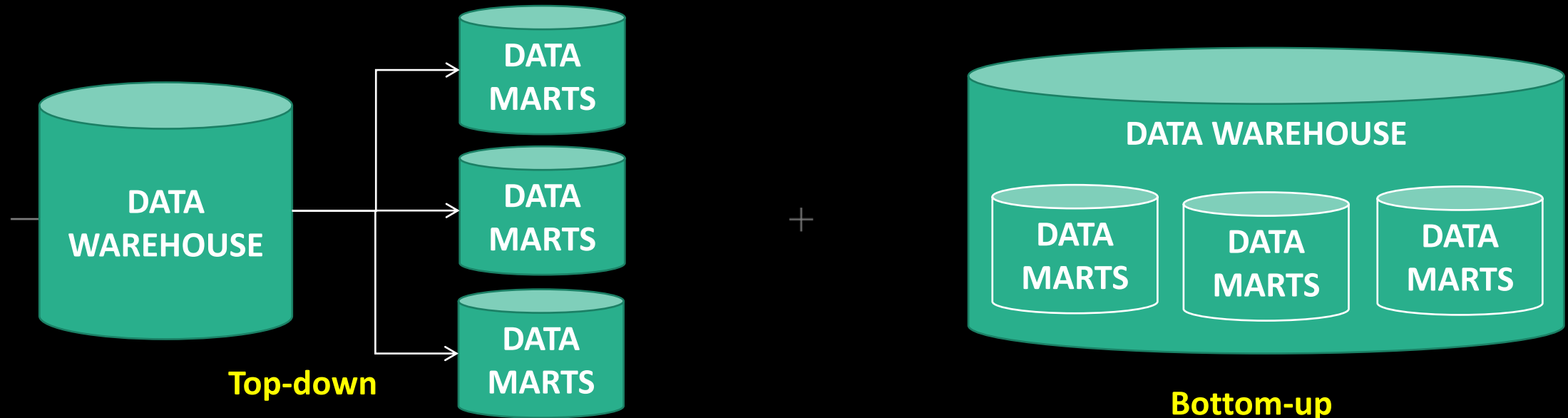


Dados armazenados em formato de série temporal

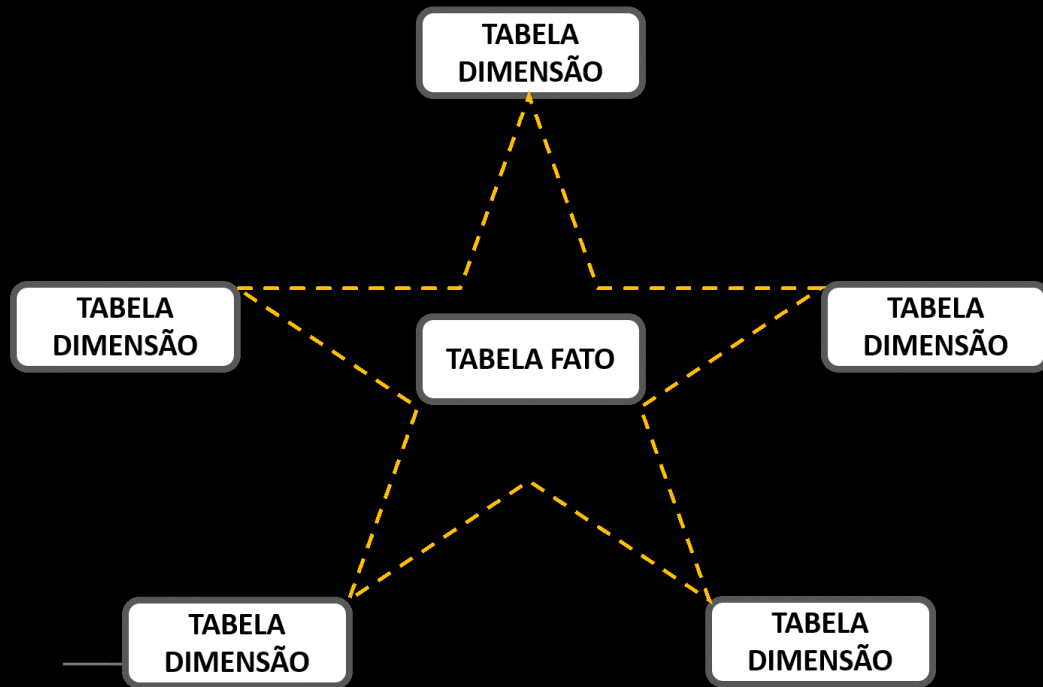
Data Marts

Data Marts são conjuntos de dados criados para atender especificamente uma necessidade com um conjunto de dados reduzido (ex: consolidação diária de uma informação)

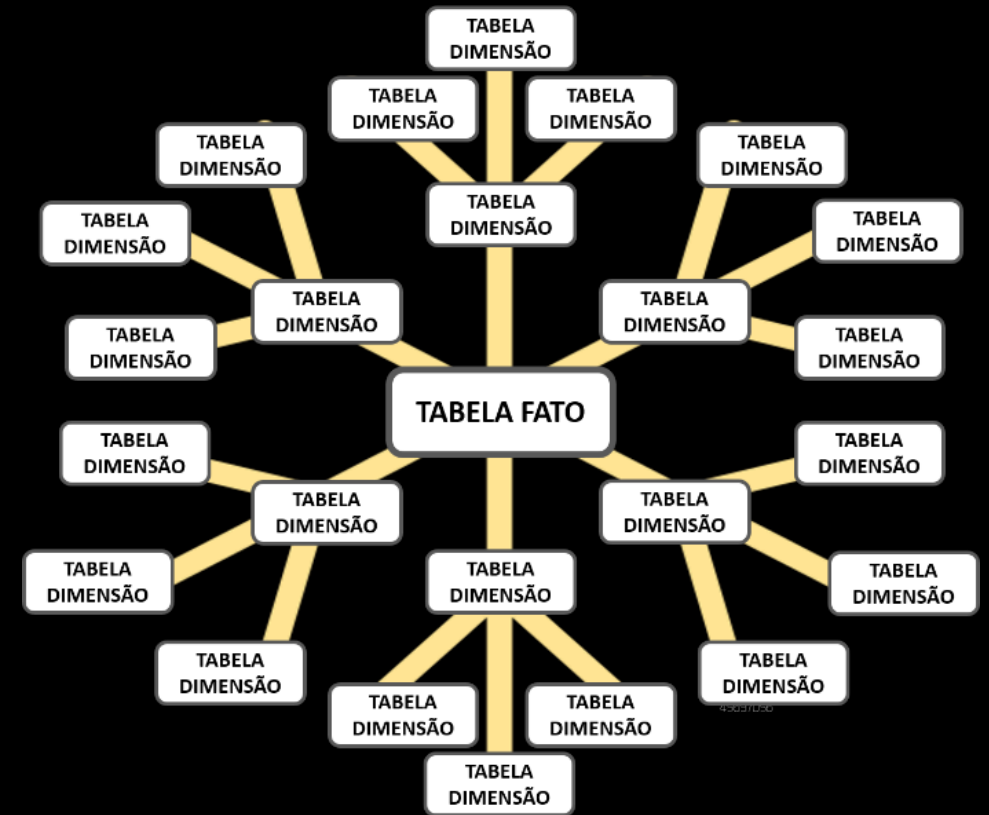
A combinação de Data Marts formam o Data Warehouse



Star schema



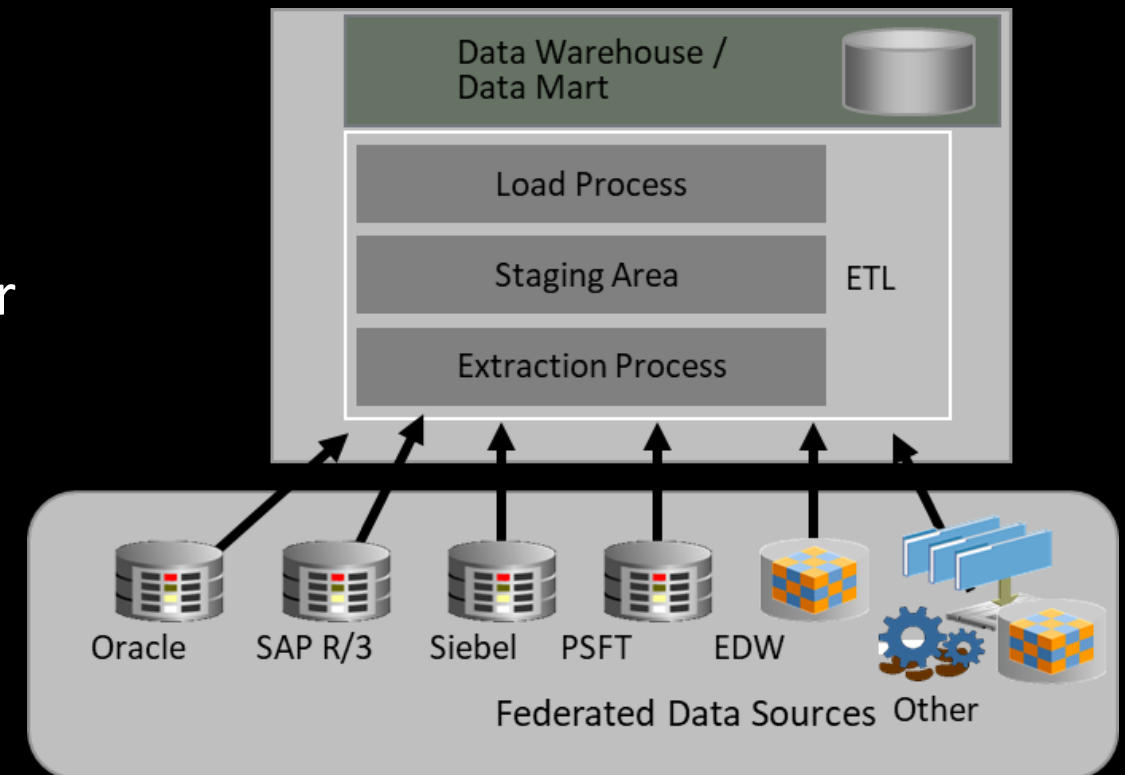
Snowflake schema



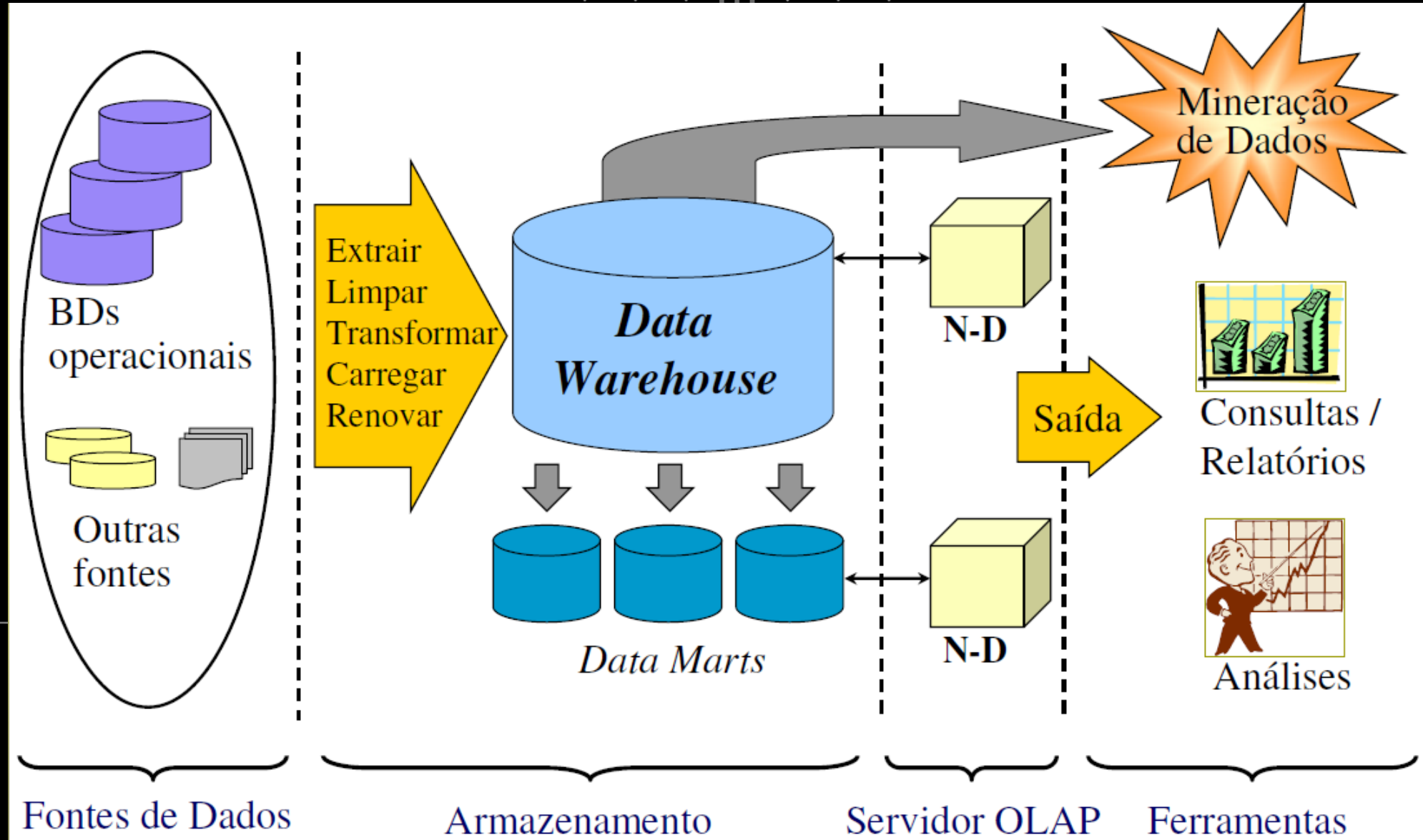
Extract, Transform, Load (ETL)

ETL é a camada da arquitetura para extrair, transformar e carregar os dados no DW / DM.




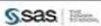
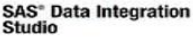










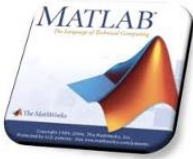








- Modular e isolada para cada sistema fonte
- Suporta diferentes tecnologias para atender cada sistema fonte
- Atua na transformação e carga entre fontes de dados e consumos



Visão Geral



Stack tecnológico DW

ETL	DATA BASES	ADVANCED ANALYTICS	DATA DISCOVERY OLAP
      	    	   	       

Evolução Data Warehouse

Fragilidades que surgiram nos DWs ao longo do tempo:

- Aumento no **volume de dados** com **mais dispositivos** conectados a internet
- Necessidade de análise de **dados sem estrutura** pré-definida
- **Alto custo** dos hardwares e software usados por ferramentas de DW
- Falta de **escalabilidade horizontal**
- Crescimento da necessidade de análise em **tempo real**, com escala global

Data Lake

Data Lake

- O que é um data lake?

É um repositório de dados centralizado capaz de armazenar, processar e disponibilizar grandes **volumes** de dados, com grande **velocidade** e nos mais **variados formatos**. Estes dados podem ser estruturados, semiestruturados ou não estruturados.

Tomadas de decisão: +rápidas +precisas

Os 3 V's - VOLUME

Capacidades de armazenamento
acima do suportado por DWs

Volume de informações digitais
estimada geradas pelo mundo em
2022 é de mais de **5 zettabytes**

1 byte (B) = 8 bits
1 kilobyte (KB) = 1.024 bytes
1 megabyte (MB) = 1.024 kilobytes
1 gigabyte (GB) = 1.024 megabytes
1 terabyte (TB) = 1.024 gigabytes
1 petabyte (PB) = 1.024 terabytes
1 exabyte (EB) = 1.024 petabytes
1 zettabyte (ZB) = 1.024 exabytes
1 yottabyte (YB) = 1.024 zettabytes

45697056

Os 3 V's - Velocidade

+

+

Para muitos negócios atualmente, a capacidade de processamento, análise e tomada de decisão baseada em dados precisa acontecer em tempo real.

Imagine se tivesse que pedir seu carro de aplicativo ou refeição com 1 dia de antecedência. Ou uma mensagem de texto demorasse 2 dias para ser entregue.

45697056

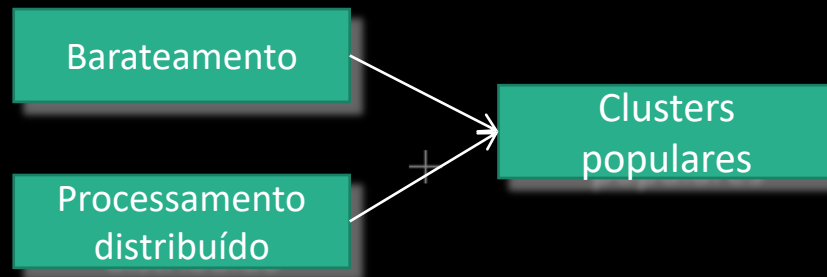
Os 3 V's - Variedade

A evolução de dispositivos e informações digitais trouxe o desafio de conseguir armazenar dados dos mais variados formatos

Dessa forma é possível fazer uma validação biométrica de uma imagem para autorizar ou não uma transação em tempo real

Como surgiram os Data Lakes?

O barateamento de recursos computacionais como armazenamento e processamento da década de 70 até 2000, unificado ao processamento distribuído de dados, fez as empresas adotarem Data Lakes alternativas mais baratas em relação aos Data Warehouses



45697056

Hadoop Distributed File System

HDFS

FIAP

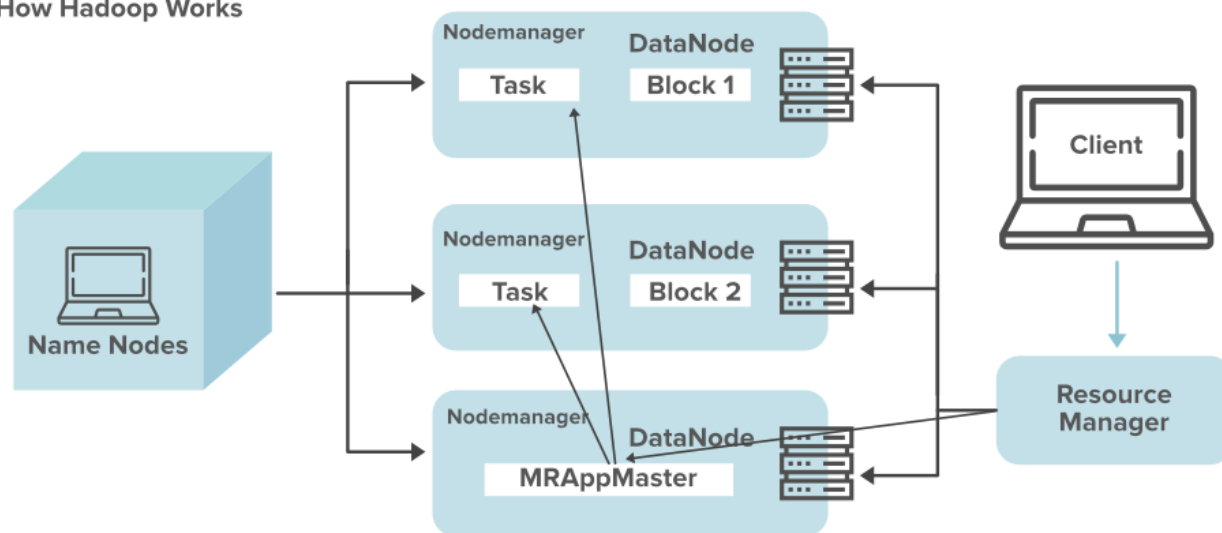


<https://hadoop.apache.org/>

O HDFS surgiu em 2006 possibilitando processamento distribuído com um modelo de processamento para larga escala (Map Reduce)

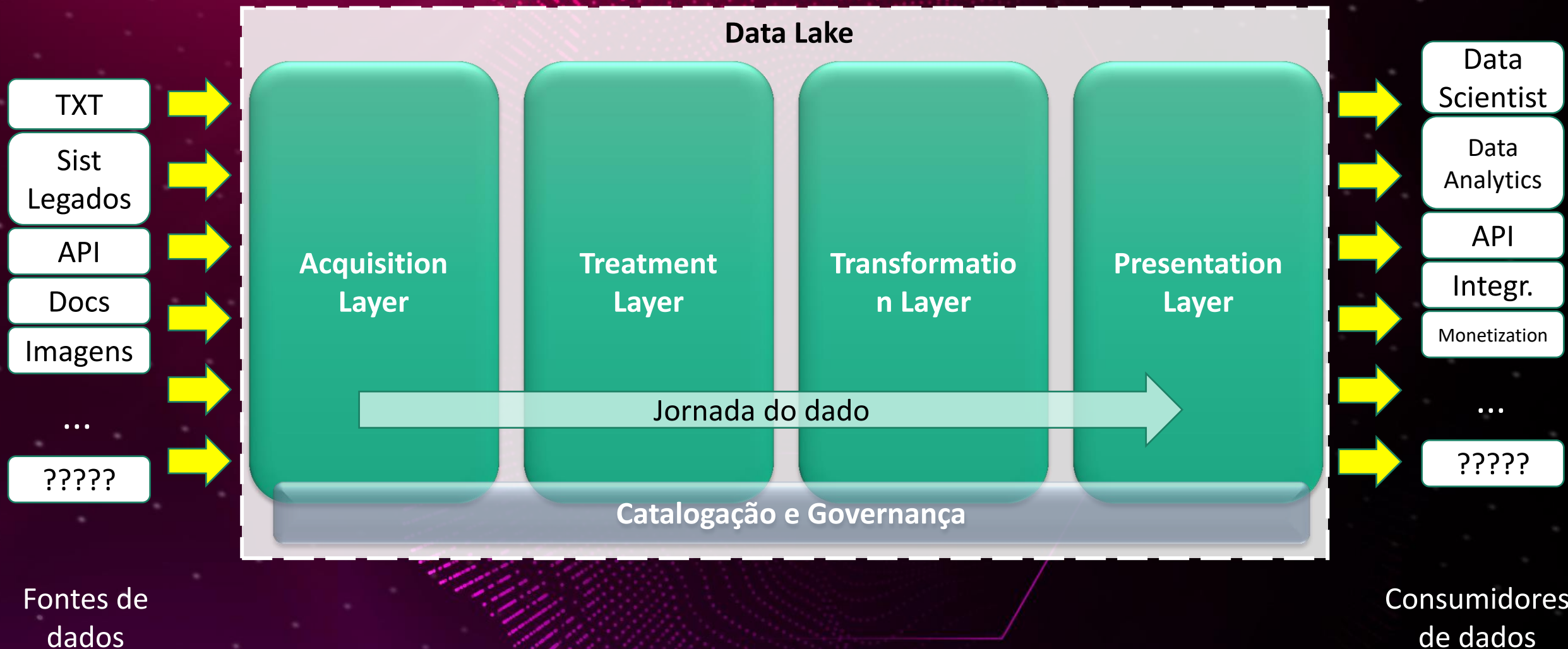
Acelerou o desenvolvimento de novos produtos de dados para processamento e armazenamento baseados neste sistema de arquivos. **Ecossistema Hadoop.**

How Hadoop Works



- Armazenamento distribuído
- Processamento paralelo
- Particionamento de dados

Camadas de dados



Catálogo de dados

A importância da catalogação de dados

É chave para manter a governança e um uso eficiente do data lake, além de prevenir que seu data lake se transforme em um **data swamp**.

Para cada camada de dados deve ser definida a profundidade e riqueza de detalhes que será necessária na catalogação de dados. Partindo de um visão macro na camada de aquisição até uma visão bem detalhada até a camada de exposição.

Atualmente muitas ferramentas possuem mecanismos robustos de descobrimento e linhagem de dados, automatizando boa parte da catalogação.

Sempre considere uma ferramenta altamente integrada a sua plataforma, mas que também flexibilize a inclusão de novo atributos quando necessário

Governança de dados



Políticas e regras de governança de dados

Data Lakes maduros e bem construídos possuem regras de governança de dados que estabelecem como são ingeridos, tratados e disponibilizados. Bem como indicadores de qualidade, propriedade de dados, ciclo de vida, entre outros.

Uma arquitetura padronizada e stack tecnológica ajudam a manter uma boa governança de uma plataforma de dados



45697056

Segurança de dados



Controle de acesso e rastreabilidade

Consiste nas regras de acesso aos dados como controles, aprovações e revisões que precisam fazer parte do dia a dia de um Data Lake.

Os acessos devem ser registrados e rastreados para qualquer necessidade de verificação de acesso realizada. A rastreabilidade garante a linhagem e melhora confiança nos dados.



45697056

Uso de Cloud

Surgimento das clouds

O uso de recursos computacionais em Cloud surgiu como alternativas aos datacenters on-premises, com alocação elástica e sob demanda de recursos provisionados de maneira automatizada e com pagamento sobre o consumo realizado

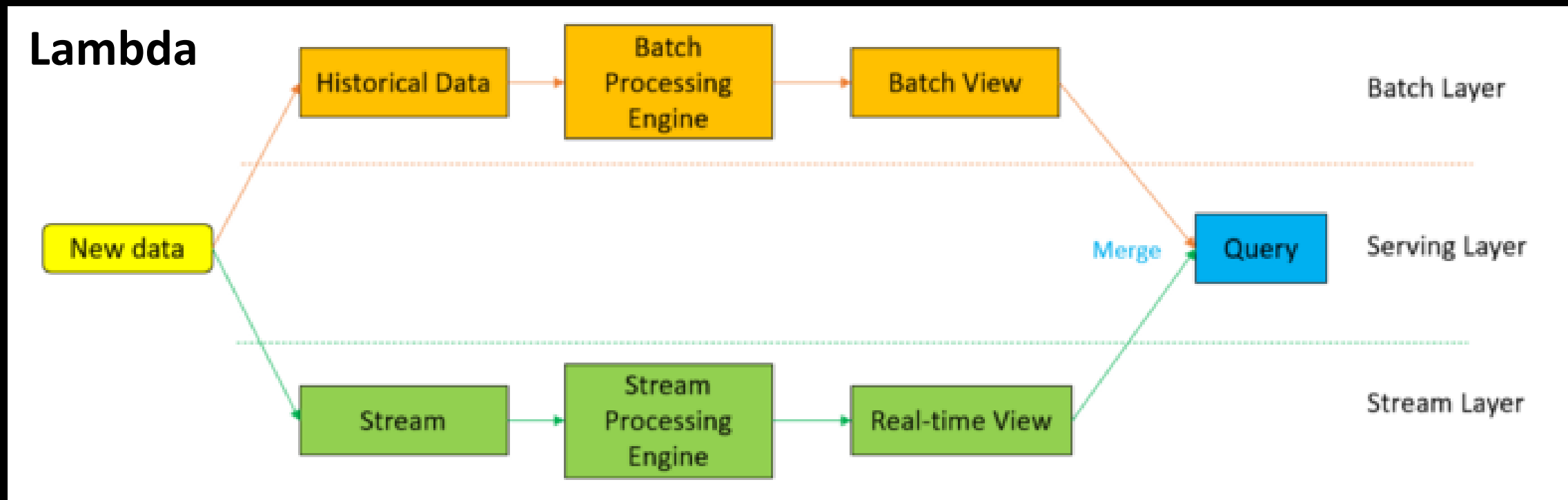
Evolução dos produtos de dados das clouds

O produtos de dados das clouds como os object storages e bancos de dados trouxeram grande capacidade de tratar os dados em nuvem mais rapidamente, com crescimento acelerado e eficiências financeiras.

Ex.: Um cluster Hadoop poderia ser escalado e utilizado de acordo com a demanda, com custo proporcional.

Arquiteturas de ingestão de dados

Arquitetura de Data Lakes



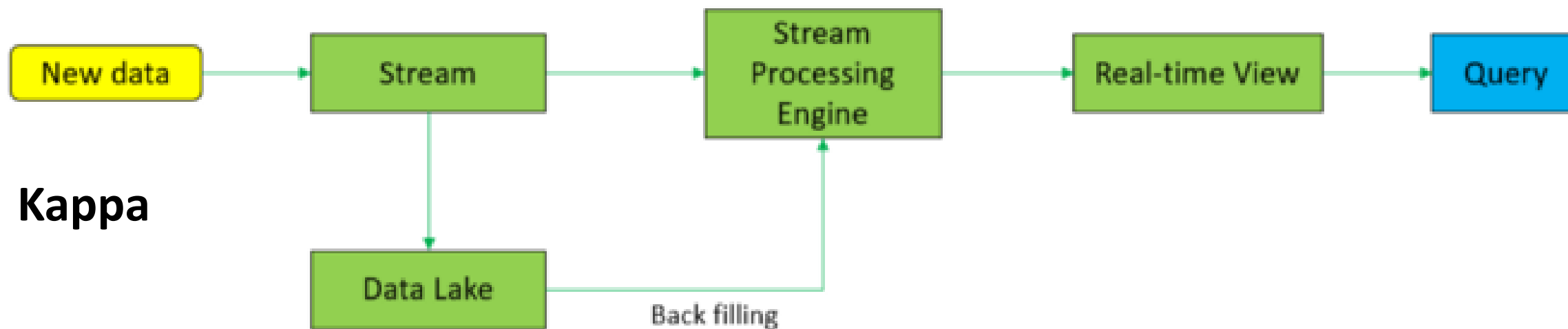
Prós

- Fluxos independentes e mais tolerantes a falhas
- Fluxos escaláveis separadamente
- Possibilidade de regras mais flexíveis

Contras

- Códigos duplicados terão 2x mais manutenção
- Disparidade de dados
- Mais consumo de processamento

Arquitetura de Data Lakes



Prós

- Menor esforço de escrita de código
- Sem duplicação de processamento ou regras
- Construção simplificada

Contras

- Necessidade de regras específicas para eventos
- + • Reprocessamento mais complexo 45697056
- Baixa tolerância a falhas e escalabilidade

Data Lakehouse

Data Warehouse + Data Lake = **Data Lakehouse**

Data Lakehouse

✓ • Data Warehouse

✓ • Data Lake

➔ • Data Lakehouse



- Transações ACID (Atomicity, Consistency, Isolation, Durability)
- Criação de relatórios para BI baseado em metadados pré-definidos
- Uso de dados não-estruturados e semi-estruturados
- Armazenamento barato
- Cargas de trabalho variadas para atendimento de Streaming
- Separação de armazenamento e processamento
- Rastreabilidade e linhagem dos dados

Data Lakehouse

Considerações:

- Não necessariamente toda base de dados precisa usar essa estrutura.
- Apesar de endereçar *real-time* analytics, a performance não deve ser um pré-requisito primordial
- Resolve fraquezas tanto do Data Warehouse quanto do Data Lake, por suportar dados não estruturados e também conseguir tratar transações concorrentes.
- Oferece robustez e escalabilidade mantendo a precisão para o negócio
- É uma evolução natural a organização de Data Lakes

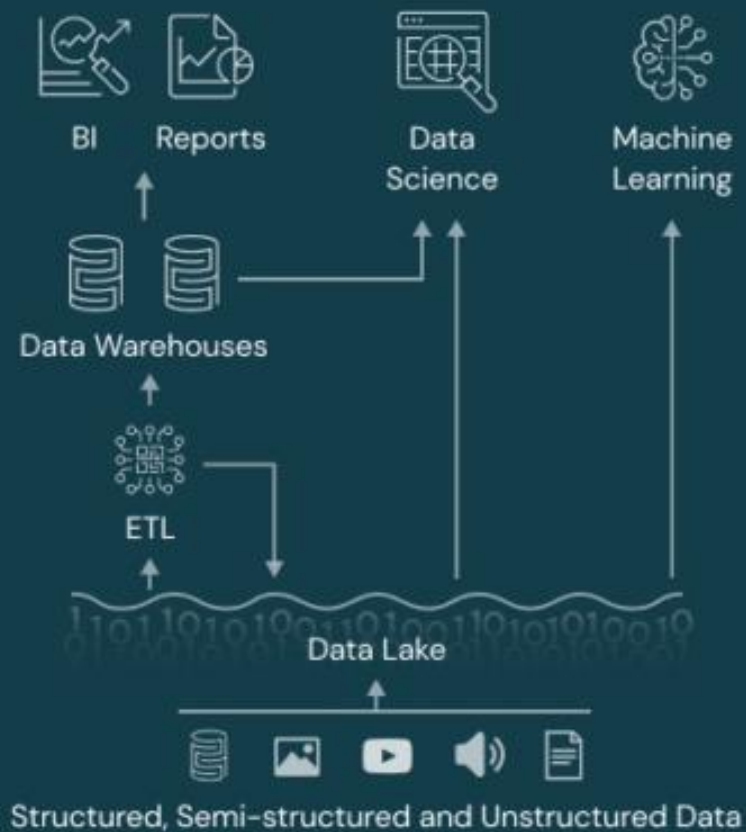
45697056

Data Lakehouse +

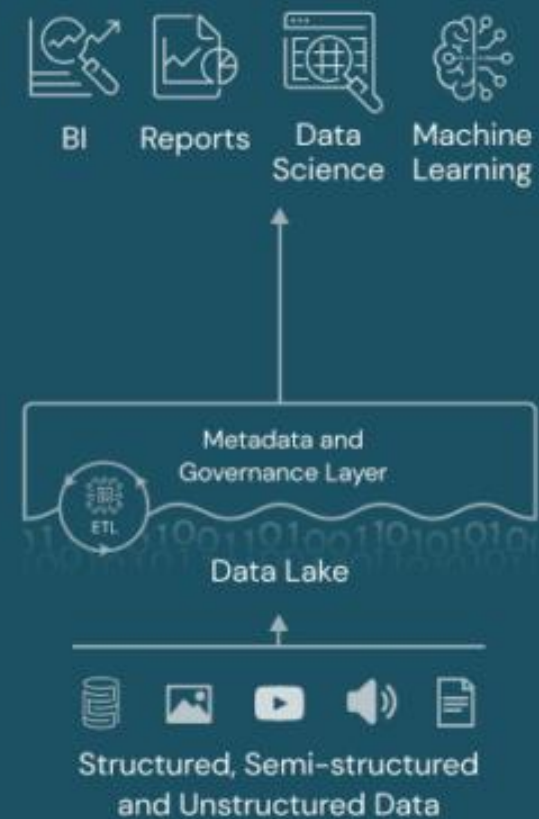
Data Warehouse



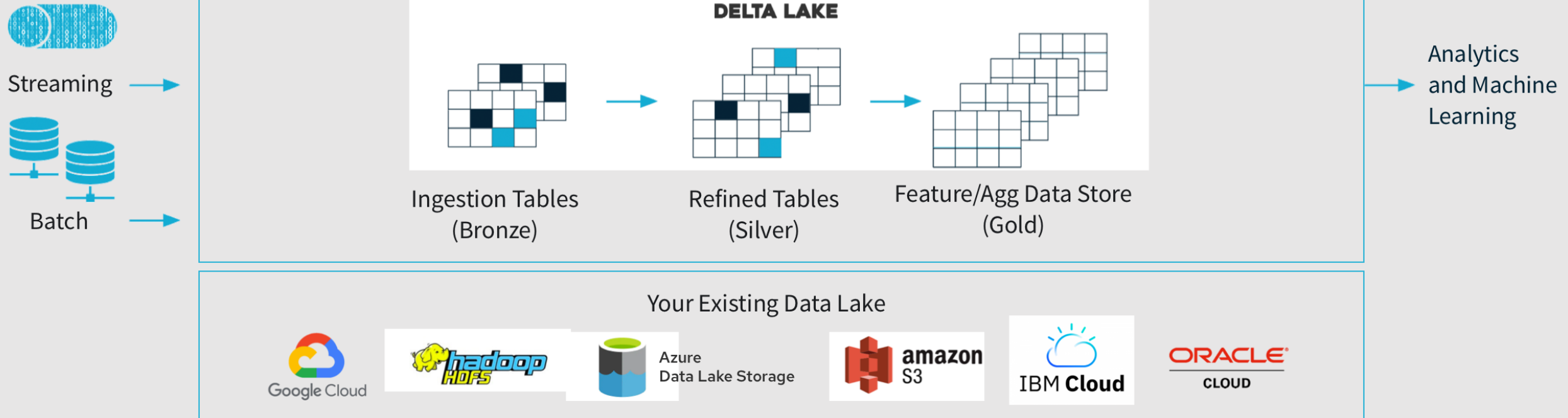
Data Lake



Data Lakehouse



Delta Lake



Delta Lake

Armazenamentos:

- Para conseguir garantir transações no uso do armazenamento o Delta Lake usa APIs nativas dos sistemas de armazenamentos (Log Store API) e implementa configurações específicas quando necessário (File System API).
- Os armazenamentos suportados são:
 - Amazon S3
 - Microsoft Azure storage
 - HDFS
 - Google Cloud Storage
 - Oracle Cloud Infrastructure
 - IBM Cloud Object Storage

Delta Lake

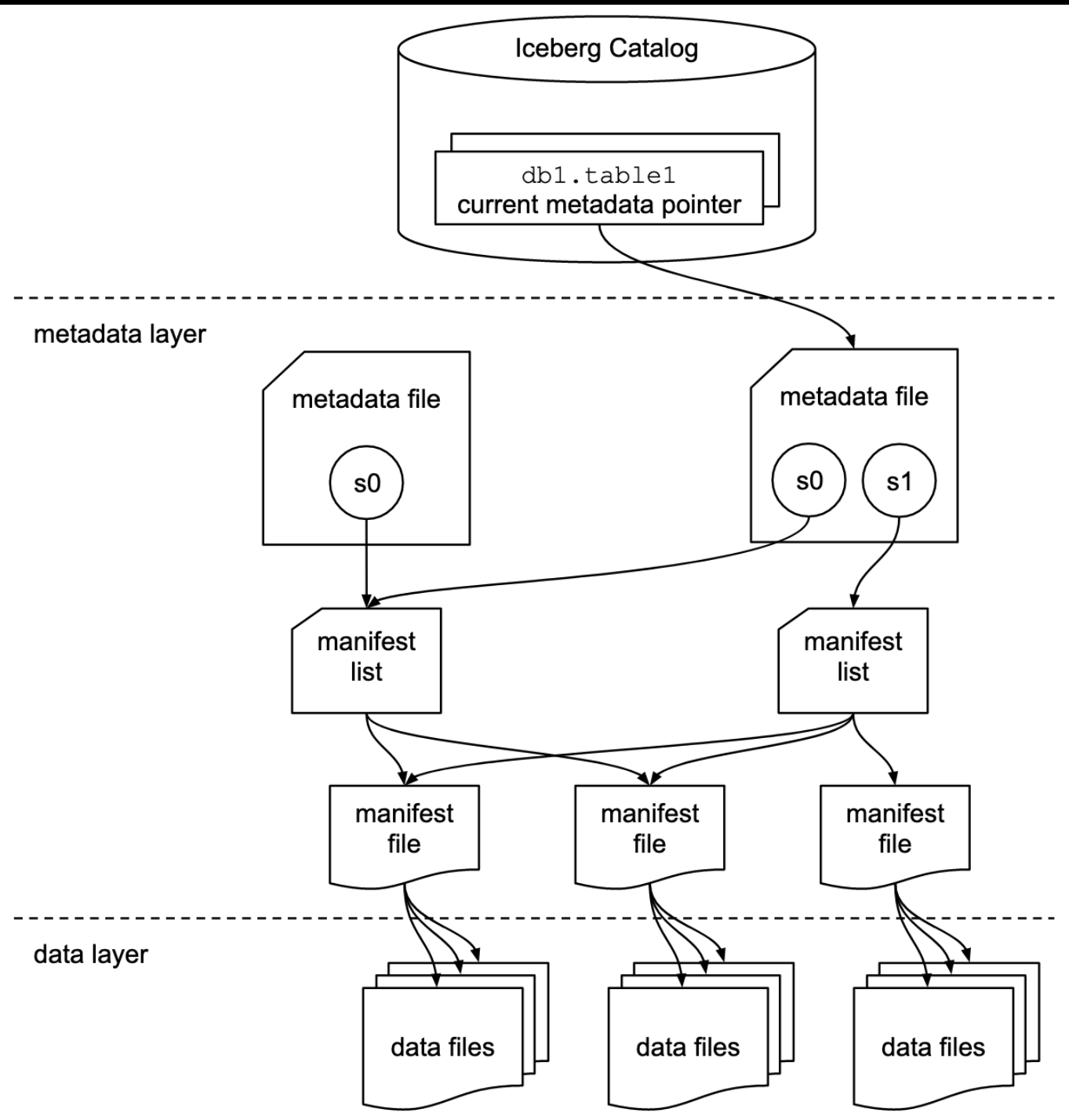
Melhores práticas:

- Escolha a coluna correta para particionamento
 - Não use uma coluna com muita granularidade. A coluna mais comum de se usar é a de *data*.
 - A recomendação é que cada partição tenha pelo menos 1 GB de dados.
- Compacte arquivos
 - Gravações pequenas e constante de dados podem gerar numerosos pequenos arquivos. Compacte-os em arquivos maiores a fim de obter melhor performance.
- Substituir conteúdo ou esquema de uma tabela pode ser feito através de uma operação de sobreposição para mudar um esquema, particionamento ou dados.
- Não use Spark Caching
 - A Databricks não recomenda o uso de Spark Caching pelo risco de perda de dados manipulados ou filtrados.

Apache Iceberg

Iceberg adiciona tabelas à ferramentas de processamento como Spark, Trino, PrestoDB, Flink, Hive e Impala, usando um formato de tabela de alta performance que funciona como uma tabela SQL

45697056



Apache Hudi

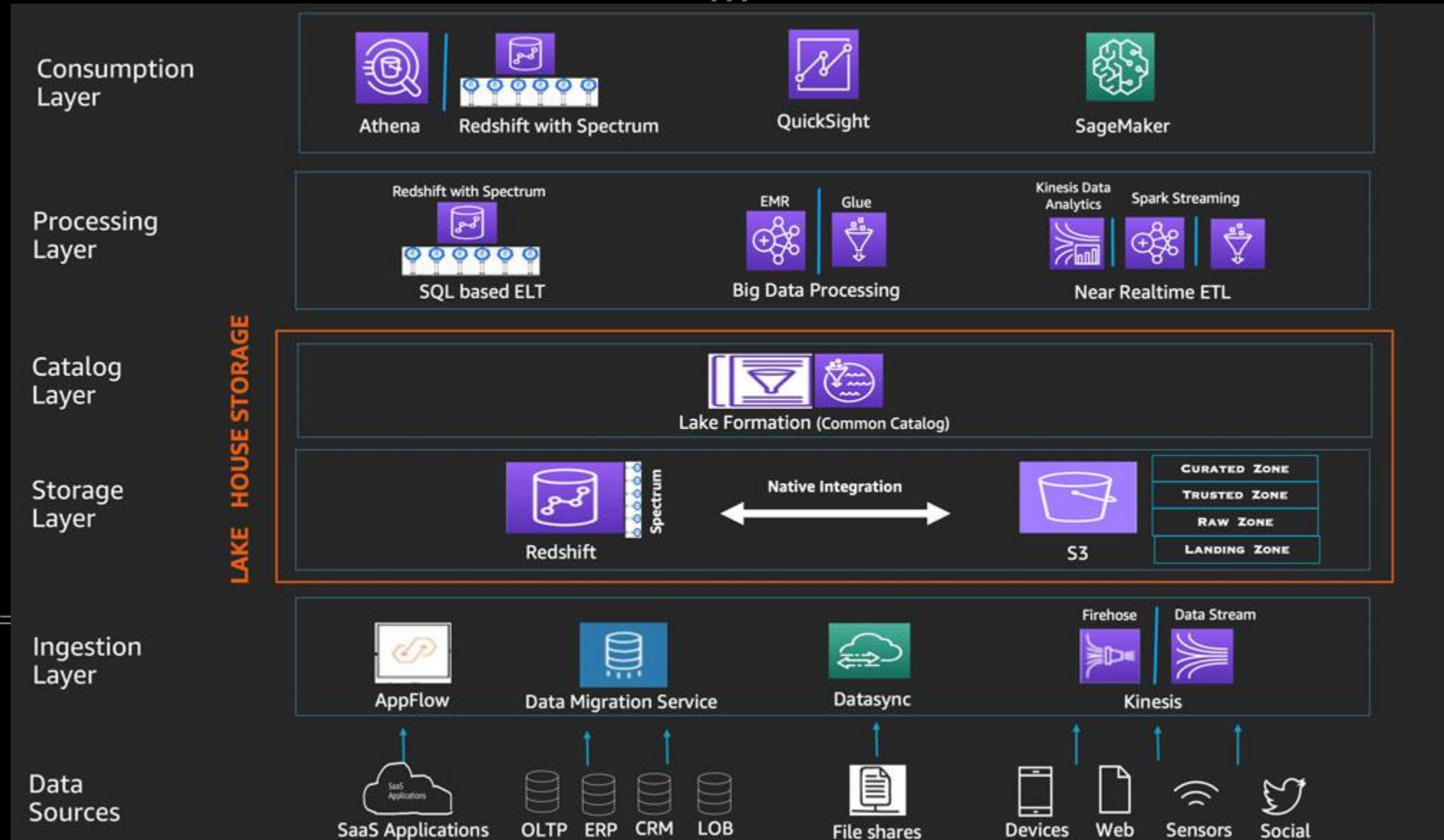


É uma plataforma de dados que traz capacidade de bancos de dados e data warehouse para um Data Lake, capaz de entregar processamentos robustos e analytics na casa de

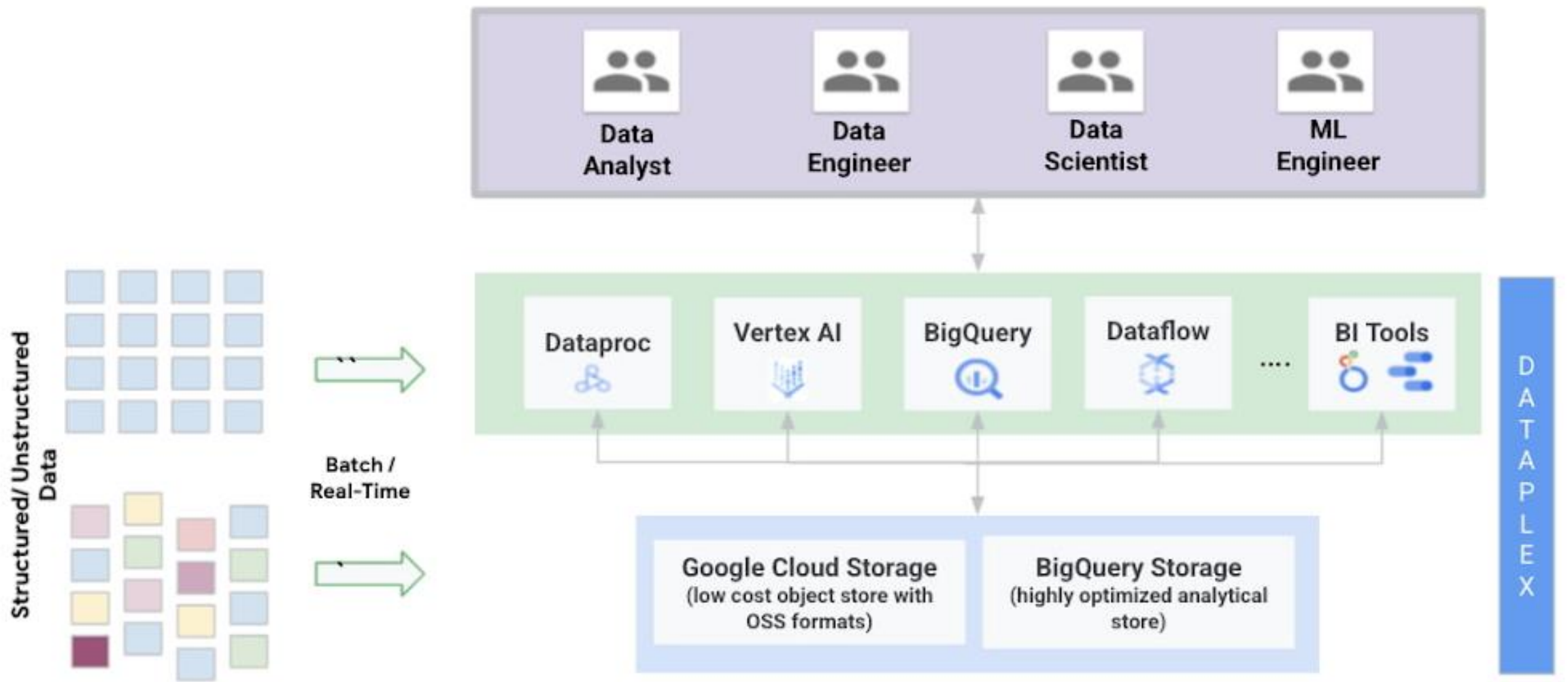
minutos

Ref.: <https://hudi.apache.org/>

AWS Data Lakehouse



Google Data Lakehouse



Data Fabric



Trata-se de uma abordagem de arquitetura de dados integrada a toda a empresa, adaptável, flexível e segura. É uma estratégia de armazenamento de dados que habilita o melhor uso de cloud, processamento interno e externo (como datacenters on-premises, IoT, mobile etc).

Com uma plataforma unificada é possível viabilizar a todos os departamentos da empresa um acesso e gerenciamento seguro.

Possibilita maior extração de valor dos dados disponíveis, criação de novos produtos e negócios baseados em dados

45697056

Data mesh

Dados e negócios orientados a domínios

Definição

Data Mesh é uma arquitetura descentralizada de dados focada em domínios de dados que facilitam a visão do dado como um produto de cada domínio. Esse modelo favorece a qualidade e acesso aos dados da companhia.

“Data Mesh is an analytical data architecture and operating model where data is treated as a product and owned by teams that most intimately know and consume the data” (*Zhamak Dehghani, 2019*)

Princípios

Propriedade sobre os domínios de dados

Encurtar a distância entre os consumidores de dados analíticos e os geradores de dados

Plataforma de dados self-service

Facilitar e automatizar o compartilhamento de dados pelos geradores e o acesso para os consumidores

Dados como produto

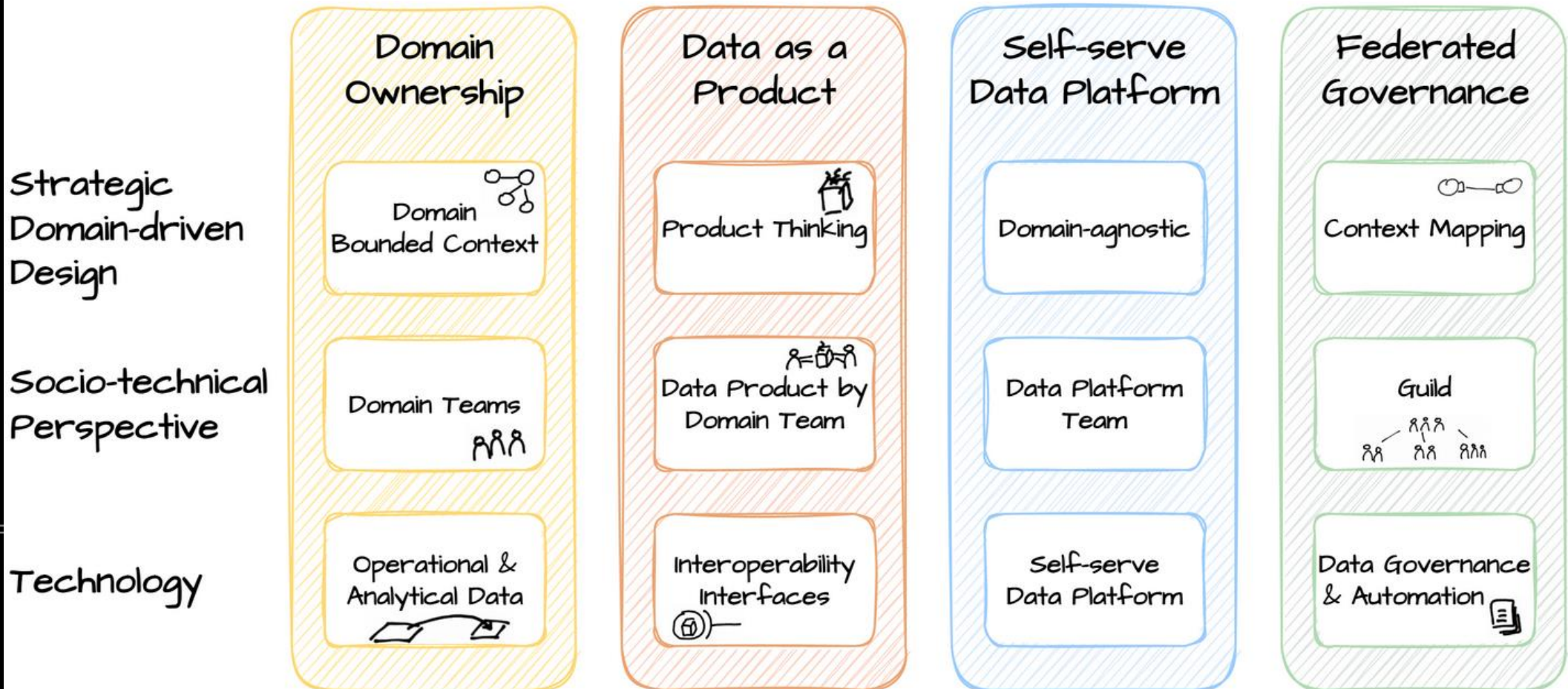
Produção de dados orientada para seus consumos

Governança federada

Políticas de dados globais e automatizadas que não centralizam responsabilidades sobre os dados

Data Mesh

What Is Data Mesh?



Benefícios

- **Democratização de dados** facilitada através da arquitetura e plataforma self-service
- **Eficiência Financeira** através da adoção de plataformas em cloud que dão melhor visão financeira para reduzir ineficiências
- **Menos débitos técnicos** devido a aproximação dos times que geram dados com os times que os consomem
- **Interoperabilidade** através da governança e padronização sobre os dados
- **Segurança e conformidade** por usar padrões de dados independentes de domínio, dando observabilidade e rastreabilidade.


MBA⁺

Copyright © **2023** Profs. Ivan Gancev e Leandro Mendes

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, dos Professores Ivan Gancev e Leandro Mendes

profivan.gancev@fiap.com.br

profleandro.mendes@fiap.com.br

