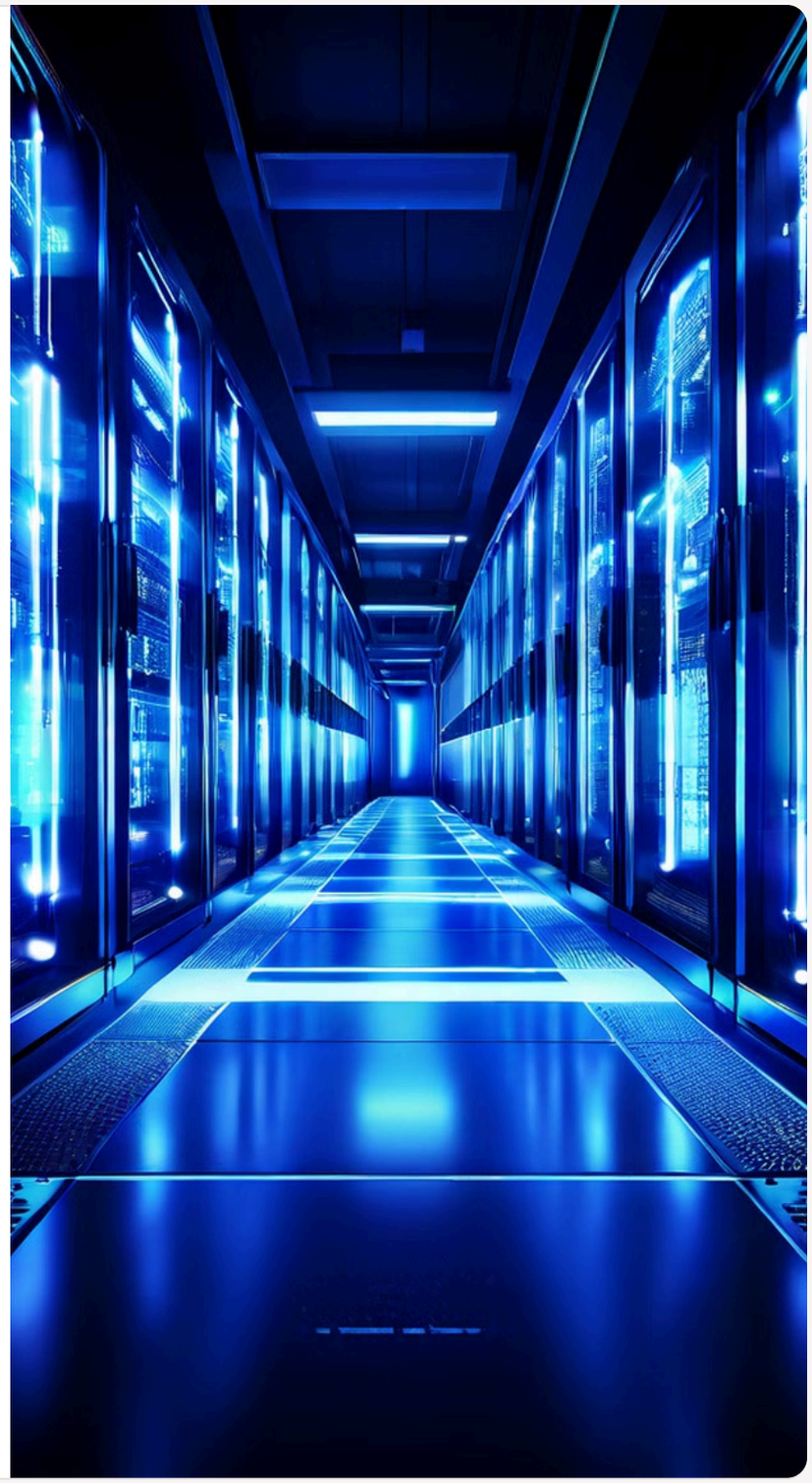
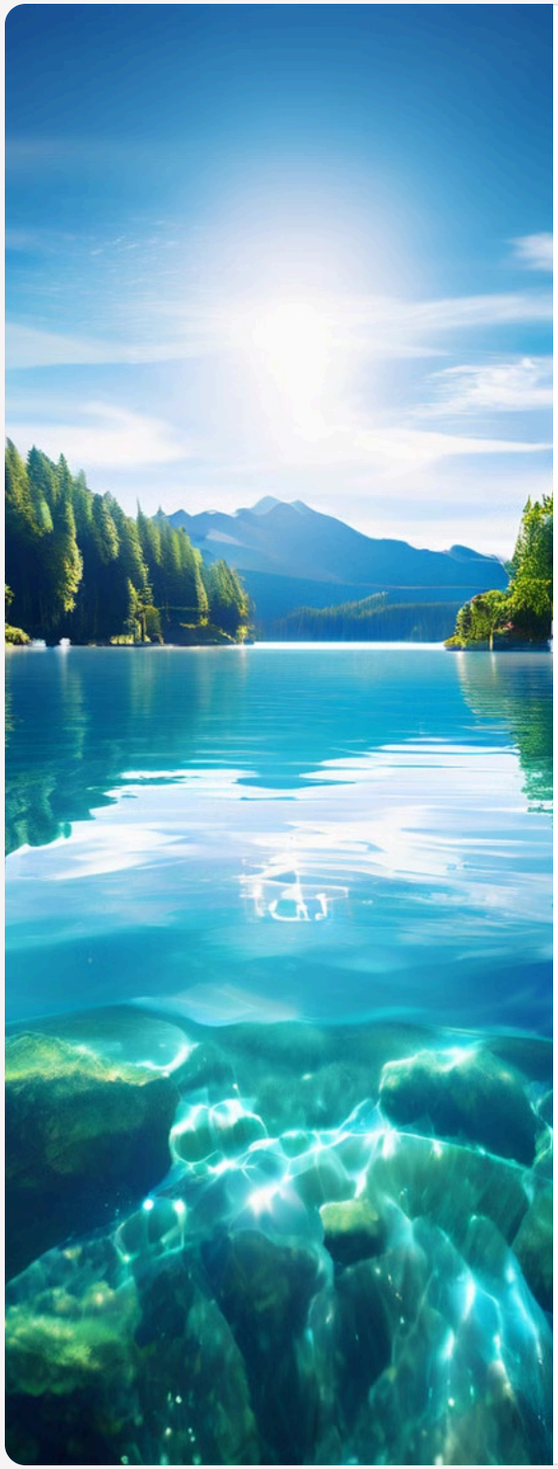
Construindo Datalakes Agnósticos de Baixo Custo com Hadoop e Outras Tecnologias

Nesta apresentação, vamos explorar como criar data lakes de forma econômica e eficiente, mantendo a independência em relação a plataformas de nuvem. Você aprenderá sobre o uso do Hadoop e tecnologias complementares, analisará casos de sucesso e comparará custos em diferentes cenários.

A **por Alessandro Binhara**





O que é um Data Lake?

Definição

Um data lake é um repositório centralizado que armazena grandes volumes de dados em seu formato bruto, prontos para serem processados e analisados conforme necessário. Diferente de um data warehouse, um data lake não impõe uma estrutura rígida de dados, permitindo maior flexibilidade e agilidade.

Benefícios

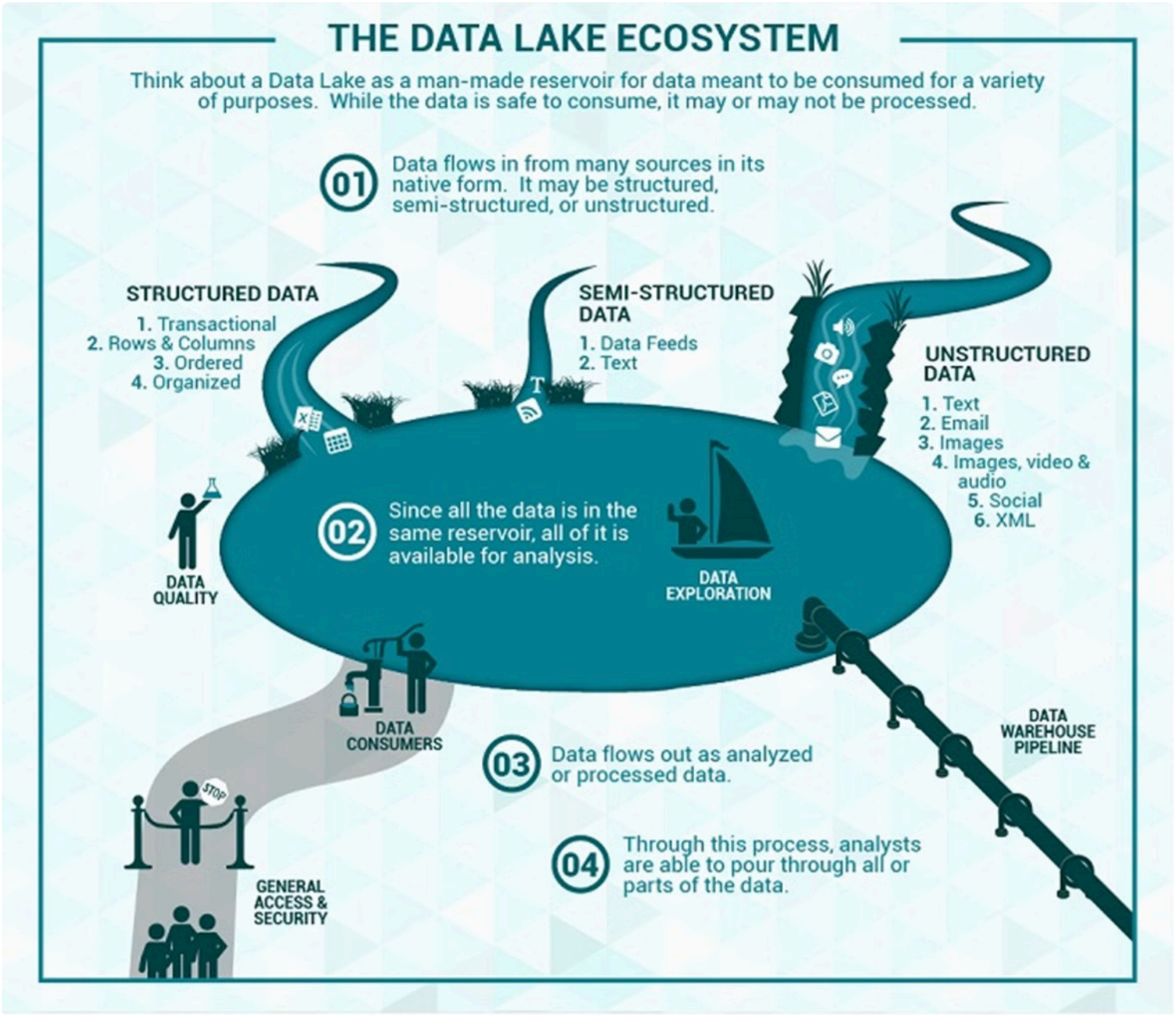
Os principais benefícios de um data lake incluem a capacidade de armazenar e processar uma ampla variedade de tipos de dados, a possibilidade de realizar análises avançadas e a redução de custos em comparação a soluções tradicionais de data warehousing.

Diferença para Data Warehouse

Enquanto um data warehouse é projetado para armazenar dados estruturados e otimizados para consultas analíticas, um data lake aceita dados em seu formato bruto, sejam eles estruturados, semi-estruturados ou não estruturados. Isso permite maior flexibilidade e capacidade de adaptação a requisitos futuros.

THE DATA LAKE ECOSYSTEM

Think about a Data Lake as a man-made reservoir for data meant to be consumed for a variety of purposes. While the data is safe to consume, it may or may not be processed.



1 Ingestão de Dados

O processo de ingestão utiliza pipelines de dados para coletar, centralizar e armazenar informações de diversas fontes, preparando-as para o processamento posterior.

2 Armazenamento

O armazenamento do data lake utiliza camadas flexíveis que suportam uma variedade de formatos de dados, escalando conforme a demanda por capacidade aumenta.

3 Processamento

O processamento dos dados é realizado por frameworks de big data, permitindo transformações, enriquecimento e modelagem dos dados de acordo com as necessidades analíticas.

4 Orquestração

A orquestração coordena e automatiza os diversos processos do data lake, desde a ingestão até a entrega dos dados, garantindo a integridade e disponibilidade.

5 Monitoramento

O monitoramento acompanha o estado geral do data lake, identificando potenciais gargalos e garantindo o desempenho ideal dos diversos componentes.

6 Ciência de Dados

A camada de ciência de dados permite a aplicação de técnicas avançadas de análise e a criação de modelos preditivos sobre os dados, gerando insights valiosos.

7 Entrega de Dados

A entrega de dados disponibiliza os resultados das análises de forma flexível e acessível para os usuários finais, permitindo a geração de relatórios e painéis personalizados.

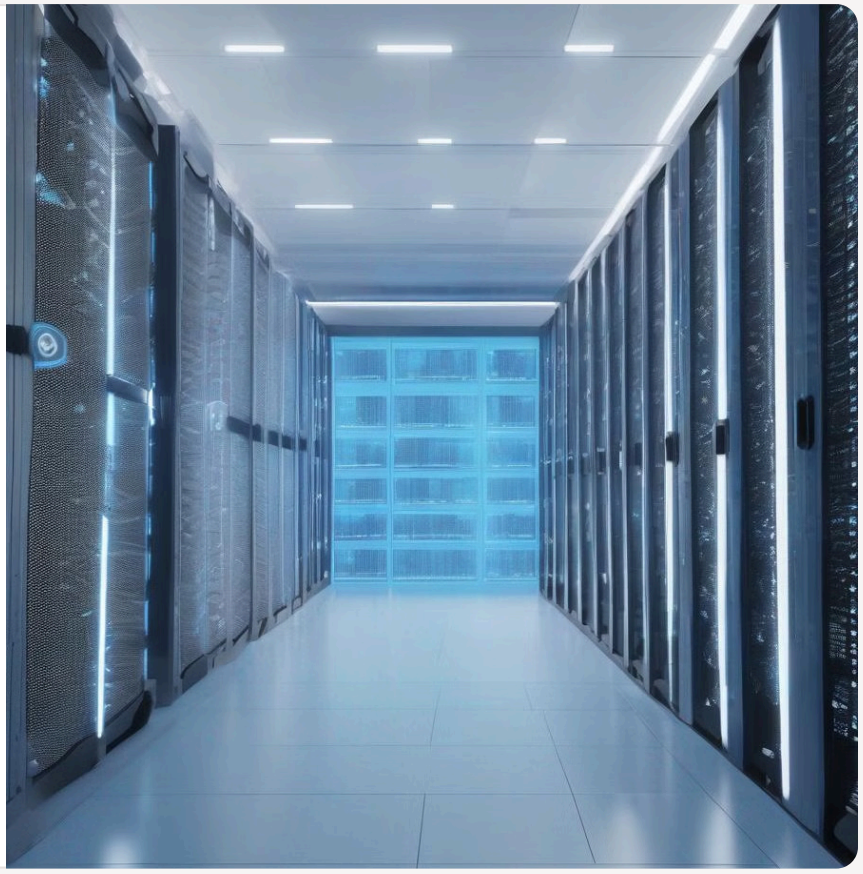
8 Backup(Expurgo) e Recuperação

A camada de backup e recuperação garante a proteção e a restauração dos dados críticos do data lake, evitando perdas e assegurando a continuidade operacional.

Pilha de Tecnologias

Data Lake Agnóstico

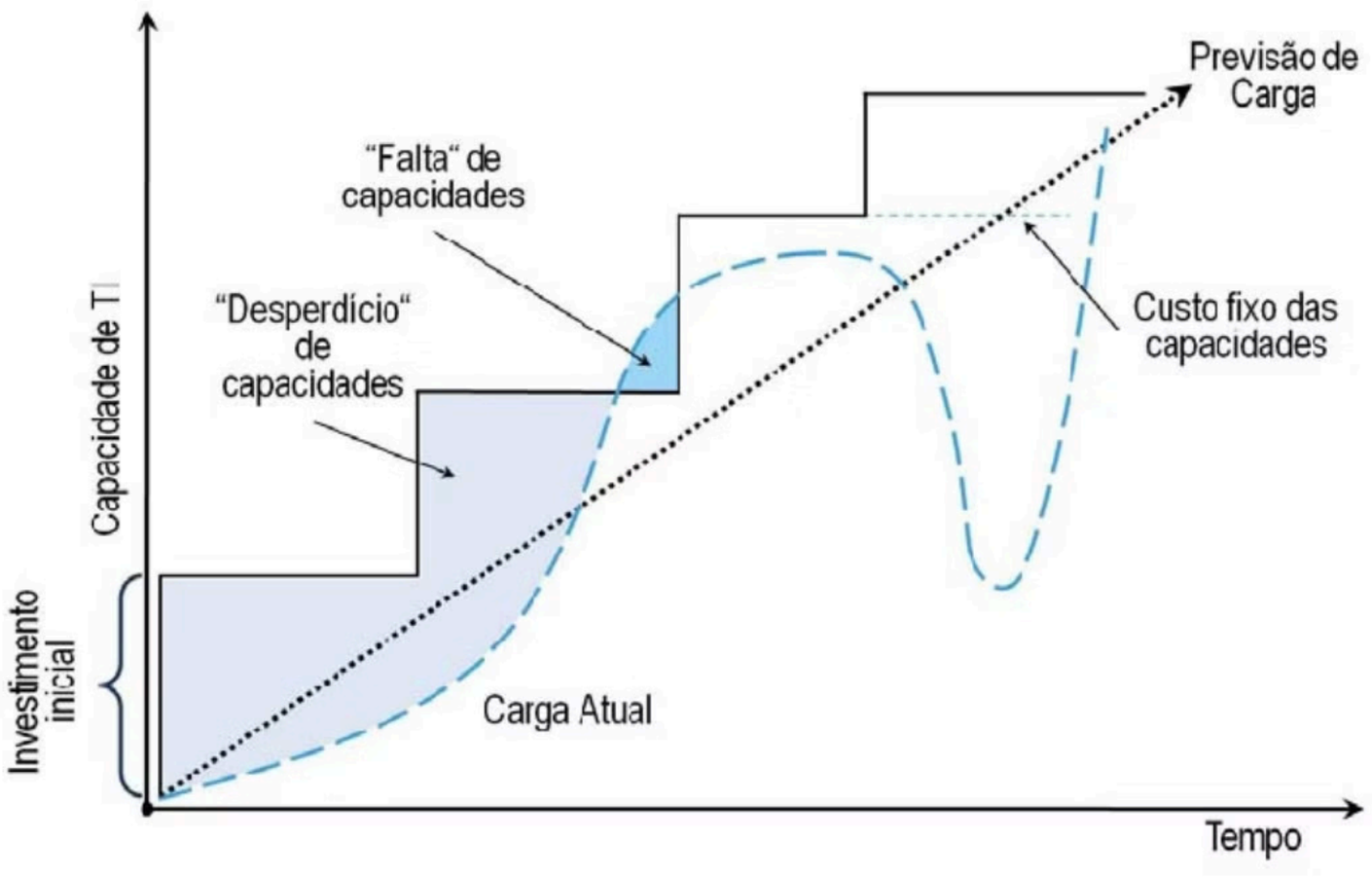
A ideia central é construir um sistema de armazenamento e processamento distribuído sem depender de ferramentas e serviços de terceiros, permitindo o uso de qualquer fornecedor de nuvem ou até mesmo a instalação do sistema em uma nuvem privada da própria empresa.



A Curva de uso do Cloud Computing

O uso do cloud computing tem crescido exponencialmente nos últimos anos, à medida que as empresas buscam maior flexibilidade, escalabilidade e redução de custos de infraestrutura de TI.

À medida que a adoção do cloud aumenta, é importante entender as tendências de utilização dessa tecnologia para aproveitar seus benefícios de maneira estratégica.



Tec_Mundo

Por que empresas estão movendo aplicações de volta para infraestru...

Entenda por que a nuvem híbrida está no centro das estratégias de TI e como a consultoria do Dell Expert Network pode te guiar nesse processo

Tendência Global Confirmada

A nuvem híbrida se consolidou como a preferência estratégica das empresas.

Gestão Unificada

A demanda é por soluções que permitam uma gestão unificada dos ambientes.

Desempenho e Segurança

Requisitos específicos de performance e segurança guiam as decisões.

Além da Escalabilidade

As vantagens da nuvem híbrida vão além da escalabilidade e economia.

publicado 03/01/2024 às 10:00

IGN Brasil

Adeus à nuvem: por que as empresas voltam a ser donas de sua infra...

Repatriação de dados tem sido cada vez mais comum

Esse é o nome desse fenômeno que já está em andamento há alguns anos e consiste em algo simples: **sair da nuvem e voltar a ter todos os serviços e dados em infraestrutura local**. Segundo relata o portal InfoWorld, um total de 25% das empresas contatadas em uma pesquisa no Reino Unido já fizeram um movimento parcial ou total nesse sentido.

Entre as razões apresentadas pelas empresas que participaram no inquérito e que repatriaram as suas infraestruturas estavam os problemas de segurança e as grandes expectativas que tinham com esta mudança (33% aludiram a esta causa), enquanto 24% explicaram que os seus objetivos e expectativas não foram alcançados. Publicado 7 de Março de 2024 às 09:00

Baguete

Serpro volta para seus data centers

Acordos com grandes players de nuvem parecem ter ido para o saco.

O Serpro, maior estatal federal de tecnologia, parece ter feito uma virada na sua estratégia dos últimos anos, deixando para trás os acordos fechados com gigantes de nuvem como AWS e Microsoft e voltando a centralizar a sua oferta para o governo nos próprios data centers.

Pelo menos, é o que se desprende de um comunicado da estatal divulgando a **"Nuvem de Governo"**, uma solução, que, de acordo com o Serpro, transforma o Brasil "na única nação com nuvem 100% soberana no hemisfério Sul".

"Os dados estarão em ambiente 100% controlado pelo Serpro, no ambiente de São Paulo e de Brasília", afirma o diretor-presidente do Serpro, Alexandre Amorim.

03/04/2024 05:02

Agnosticismo em Data Lakes

O que é Agnosticismo?

1 Ser agnóstico em relação à nuvem

O data lake não depende de uma plataforma de nuvem específica. Isso permite maior flexibilidade e independência.

2 Implantação em diferentes provedores

A solução pode ser implantada em diferentes provedores de nuvem ou mesmo em infraestrutura on-premises.

Vantagens

1 Redução da dependência de um único provedor de nuvem

2 Aproveitamento dos melhores recursos de cada plataforma

3 Mitigação de riscos relacionados a vendor lock-in

4 Solução robusta e adaptável às necessidades da empresa

Desafios

1 Arquitetura complexa

O agnosticismo em relação à nuvem requer uma arquitetura mais complexa para garantir a integração e o funcionamento adequado dos diferentes serviços e plataformas.

2 Tecnologias heterogêneas

A utilização de múltiplos provedores de nuvem implica na integração de tecnologias heterogêneas, o que pode exigir esforço adicional e conhecimento especializado.

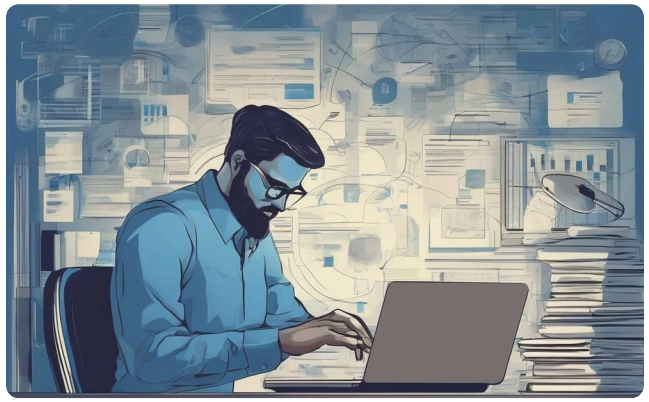
3 Gestão de segurança e compliance

Ao lidar com múltiplos ambientes de nuvem, é necessário garantir a segurança e a conformidade em cada um deles, o que pode ser desafiador.

Benefícios de um Data Lake Agnóstico

1. Controle de Custos
2. Domínio Tecnológico
3. Flexibilidade quanto a tecnologia
4. Vantagem Competitiva
5. Não depender de um único fornecedor
6. Poder mudar de servidor de cloud sem depende dele.
7. Ser um diferencial competitivo
8. Grande possibilidade inovações tecnológicas

Desafios de um Data Lake Agnóstico



Falta de Suporte dos Fornecedores

Ao optar por uma solução agnóstica, pode faltar o suporte técnico e de desenvolvimento dos principais fornecedores de nuvem, exigindo uma equipe interna mais robusta.

Maior Responsabilidade

A liberdade de escolher as tecnologias traz também a responsabilidade de projetar, implementar e manter todo o ecossistema do data lake.

Necessidade de Suporte Interno

Para garantir a estabilidade e o bom funcionamento do data lake, é essencial contar com uma equipe interna qualificada para operar e solucionar problemas.

Desafios de Manter um Data Lake Agnóstico

Equipe Qualificada

Manter uma equipe de profissionais qualificados e atualizados é fundamental para o sucesso do projeto.

Compatibilidade

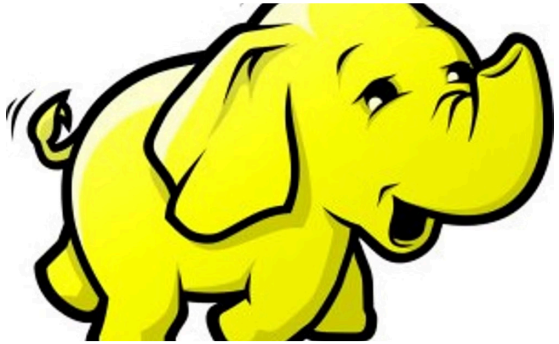
Garantir a compatibilidade de funcionalidades entre as diferentes ferramentas utilizadas é um desafio constante.

Instabilidades

Problemas de hardware fora do nosso controle podem causar instabilidades no data lake.

Atualização Constante

Manter uma stack de serviços atualizados e compatíveis entre si é um desafio constante.



O que é Hadoop?

Hadoop é um framework de código aberto projetado para processar e armazenar grandes volumes de dados de forma distribuída e escalável.

Componentes Principais

Os principais componentes do Hadoop incluem o HDFS (Hadoop Distributed File System)

, MapReduce e YARN.

Por que escolher Hadoop?

O Hadoop se destaca por sua capacidade de lidar com grandes volumes de dados, sua escalabilidade horizontal, seu baixo custo de implementação e sua robustez.

HADOOP ECOSYSTEM TIMELIENE

altexsoft

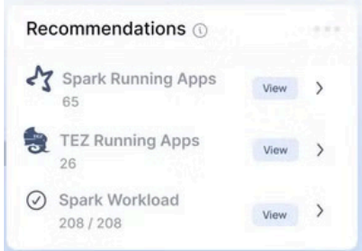
Hadoop Core	2008-2009	2010-2012	2013 and later
Hadoop MapReduce Processing engine	HBase NoSQL database	Apache Kafka Event streaming	Apache Drill SQL on Hadoop
Hadoop HDFS File system	ZooKeeper Coordination	Flume Log data ingesting	Apache Spark Real-time data analytics
Hadoop YARN Resource management	Apache Pig SQL-like scripting	Hive SQL-like scripting	Apache Phoenix SQL on HBase
	Mahout Machine learning	Oozie Scheduling	Apache Giraph Graph processing
	Sqoop RDBMS to HDFS data transfer	Flink Stream and batch processing	Apache Storm Stream processing

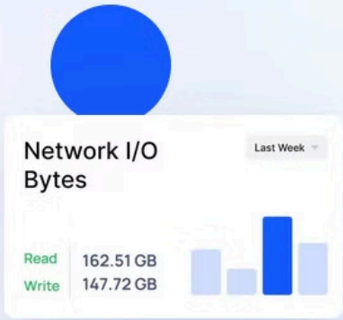

Security
Apache Ranger
Meta Data Management
Apache Atlas
Data Format
Parquet, Avro, ORC, Arrow
Coordinate & Management
Zookeeper
Ambari
In-Memory Processing
Stream Processing
Flink
Storm
SQL Over Hadoop
HIVE
APACHE DRILL
Impala
NoSQL Database
HBASE
Search Engine
Solr
Data Piping
nifi
FLUME
oqoop
Machine Learning
MADLib
mahout
Spark MLlib
Scripting
Apache Pig
Scheduler
OOZIE
Airflow
Spark
kafka
Resource Management
hadoop YARN
MESOS
Storage
hadoop HDFC
ALLUXIO

Embrace Openness with the Open Data Platform

Break Free from Vendor Lock-In and Get Seamless Access to the Latest Open Source Data Platform.

Request Demo





TRUSTED BY ENTERPRISE DATA TEAMS WORLDWIDE

HCSC Health Care Service Corporation · PhonePe · dun & bradstreet · HERSHEY THE HERSHEY COMPANY · MERKLE · PubMatic · true

Telkom Indonesia · Circana. · RD STATION · ACT FIBERNET

Existem outras possibilidades além do Hadoop ?

Outras ferramentas agnósticas de cloud

Databricks

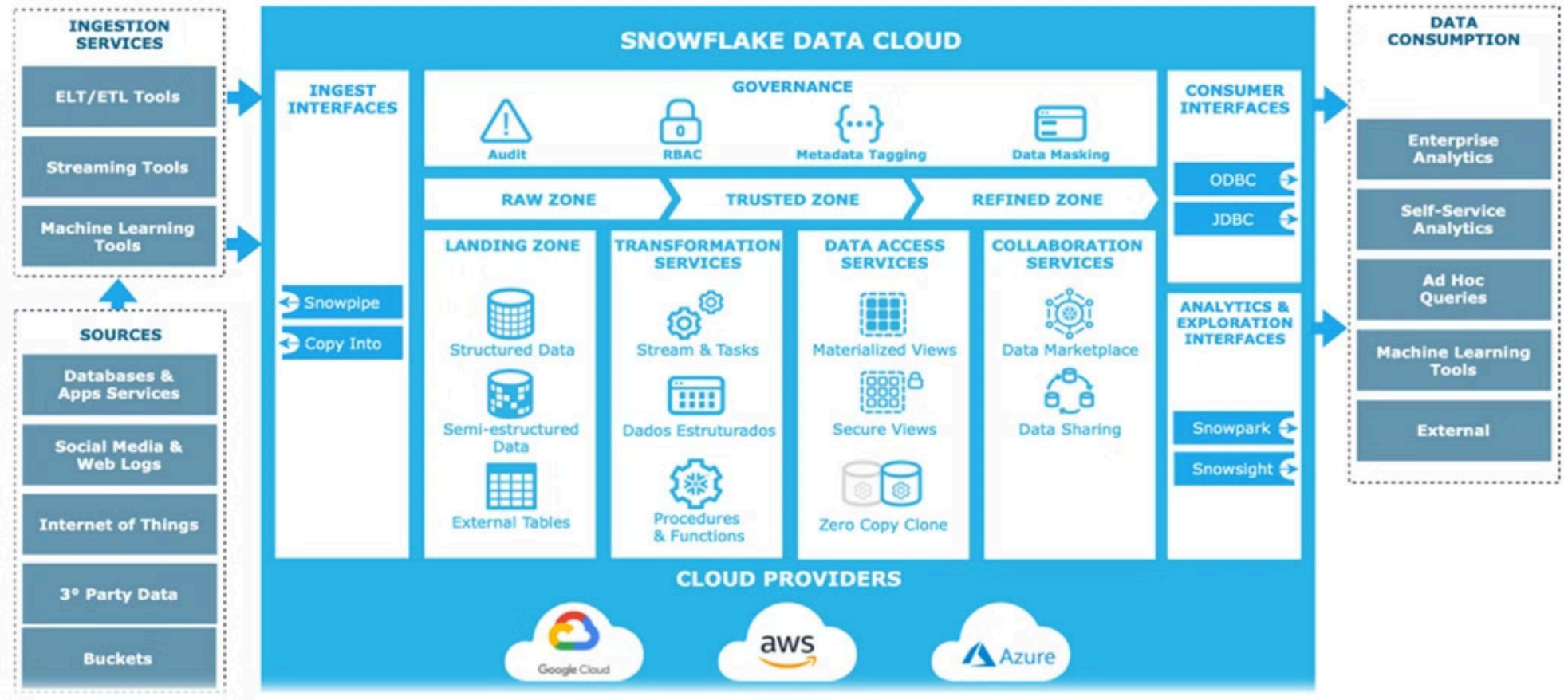
Plataforma de análise de dados escalável e unificada, permitindo processamento de dados em larga escala.

SnowFlake

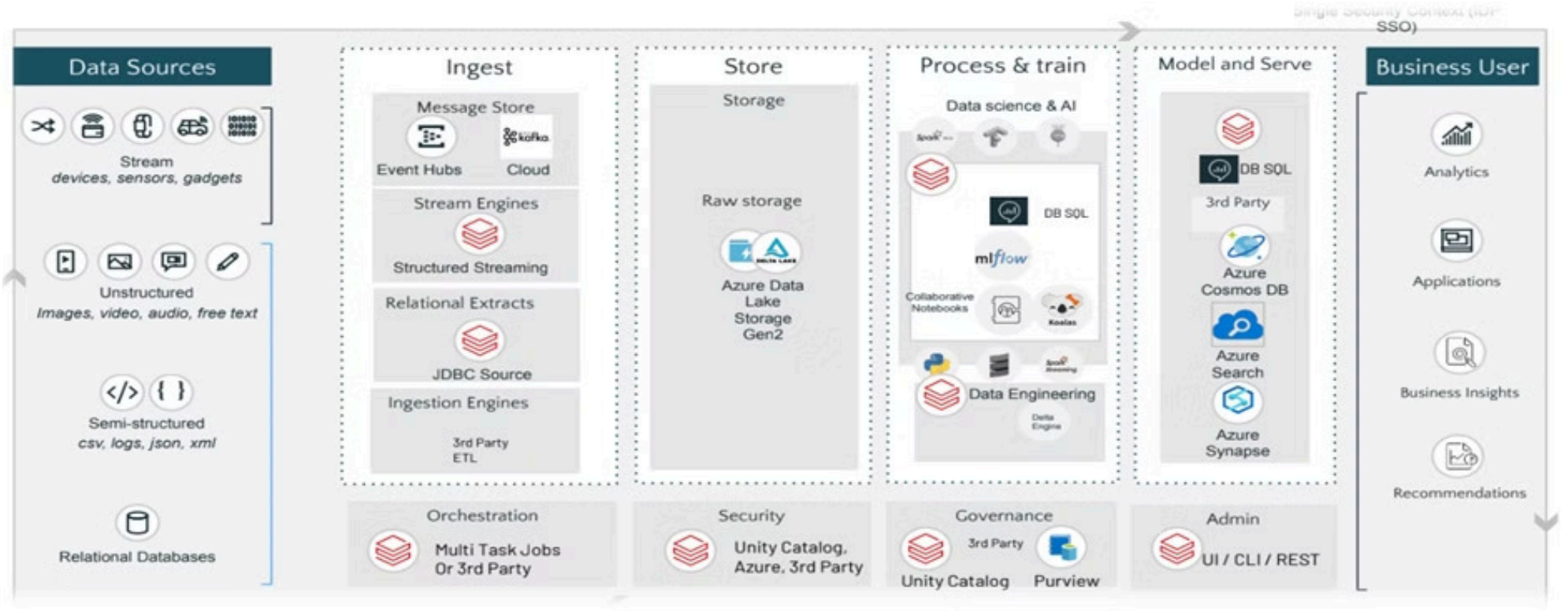
Data warehouse na nuvem com escalabilidade elástica e processamento de consultas em tempo real.

GaioDATA OS

Sistema operacional voltado para análise e processamento de grandes volumes de dados.

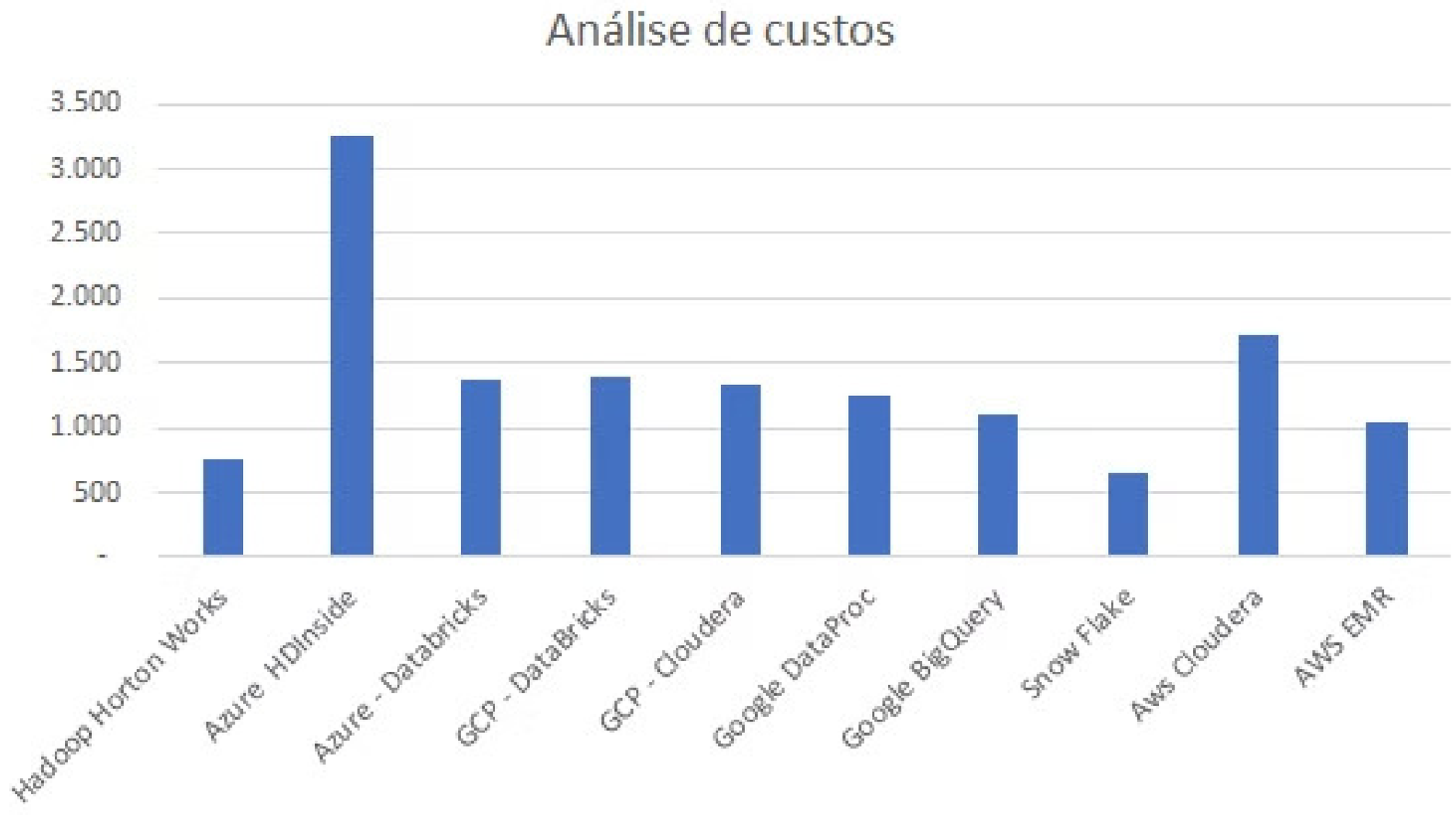


Arquitetura snowflake



Arquitetura databricks

Comparativo Geral entre Clouds



Comparativos de custos por camadas

Descritivo	Hadoop Horton	Azure HDInside	Azure - Databricks	GCP - DataBricks	GCP - Cloudera	Google DataProc	Google BigQuery	Snow Flake	Aws Cloudera	AWS EMR
Camada de Carga de dados	77	102	102	17	136	17	22	77	160	57
Camada de Armazenamento de Dados	258	328	328	67	67	120	78	466	100	59
Camada de Processamento	258	2.079	587	754	517	409	658	-	1.066	803
Camada Orquestração de serviço	26	94	102	39	6	39	39	26	53	56
Camada de Gestão e Monitoramento	2	5	17	6	105	6	6	-	53	-
Camada de entrega de dados	50	165	165	357	357	357	191	-	103	-
Camada de Ciência de Dados	60	481	65	63	63	63	63	-	124	73
Camada de Backup	9	-	-	40	40	119	18	38	28	-
Total em pesos	**740**	**3.254**	**1.366**	**1.343**	**1.292**	**1.130**	**1.077**	**607**	**1.689**	**1.049**

Startup Ebehavior: Uma Solução Escalável e Econômica

Aquisição pelo Buscapé

A Ebehavior, uma startup brasileira, foi adquirida pelo grupo Buscapé por cerca de R$10 milhões. Essa incorporação permitiu que a plataforma da Ebehavior se tornasse um padrão adotado pela Buscapé.

Escalabilidade Impressionante

O sistema da Ebehavior é capaz de processar incríveis 1,5 milhões de registros por segundo em um cluster de 70 máquinas.

Cada nó possui 48 CPUs, 56GB de RAM e 10TB de armazenamento, garantindo uma enorme capacidade de processamento.

RD STATION

Quem é a RDStation

Ajudar pequenas e médias empresas

a crescer de maneira previsível, escalável e sustentável, impactando a Economia dos países onde elas estão inseridas, através da Tecnologia.

RD Station Marketing e RD Station CRM

Nossos dois principais produtos.

+20.000 clientes em 1.700 agências

de Marketing Parceiras, além de +700 RDoers trabalhando para impactar este incrível ecossistema.

The Hadoop Stack

1 **Storage**

HDFS, WebHDFS

2 **Processing**

MapReduce, Tez, Spark2

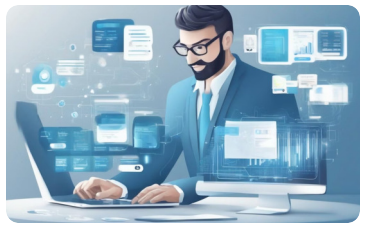
3 **Analysis**

Hive, Presto

4 **Ecosystem**

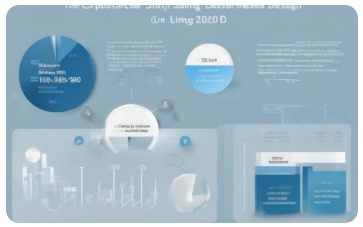
HBase, Solr, Knox, Ranger, Airflow, NiFi, Sqoop, Gateway NFS

Solução Personalizada e Econômica



Solução Personalizada

Uma plataforma flexível que se adapta às necessidades únicas de cada cliente.

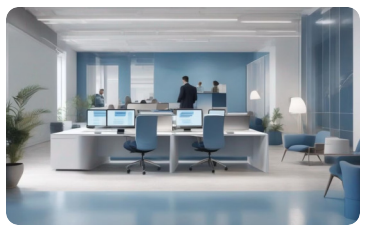


Economia Significativa

Redução de custos de infraestrutura em dezenas de milhares de reais por mês.

Custo Fixo

Uma estrutura de custos previsível, sem surpresas ou variações inesperadas.

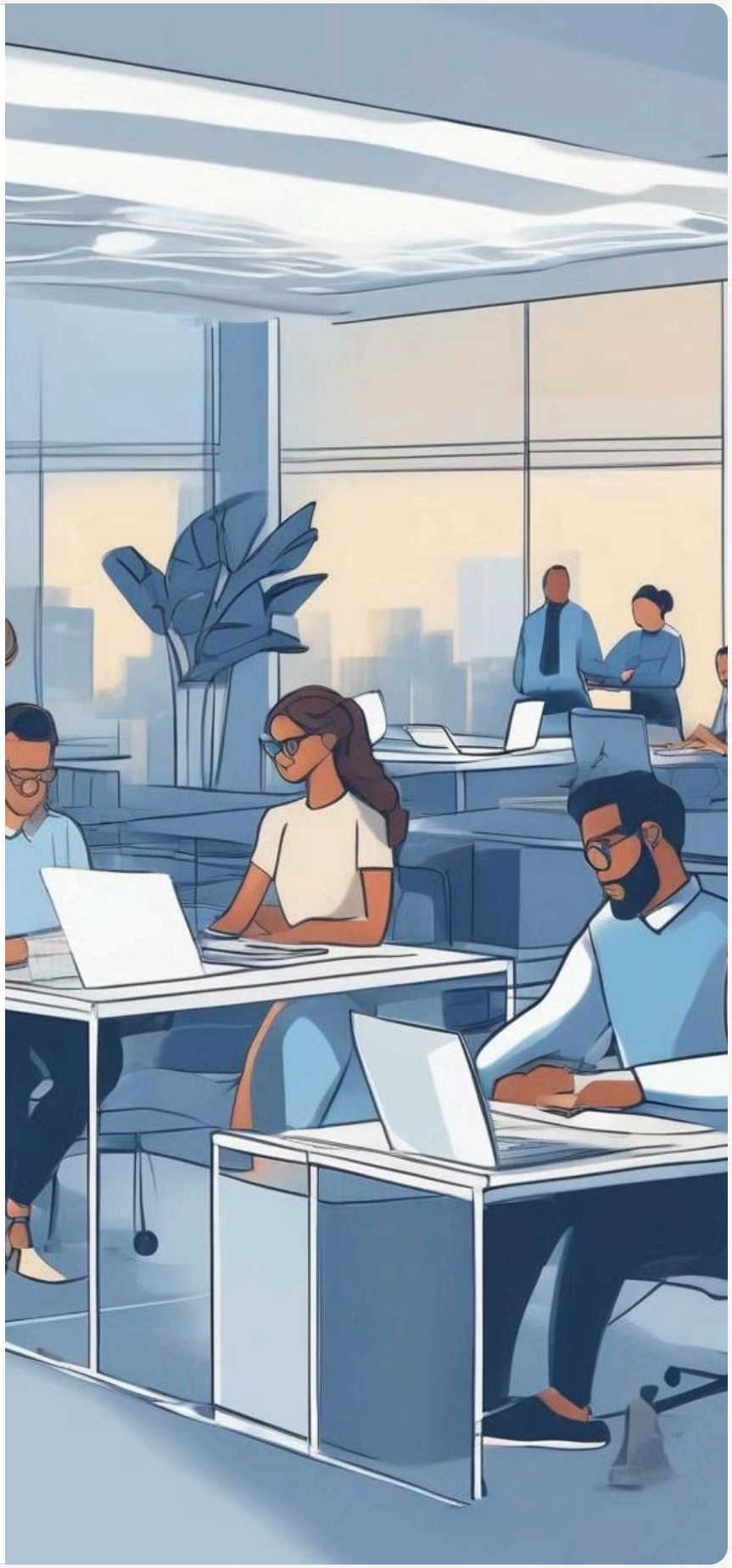


Autonomia para os usuários

Os usuários podem trabalhar com grande flexibildade.

Compatibilidade

Aderente a padrões de mercado.



Quer Entrar ou Sair do Hadoop?