Banco de Dados Tecnologias Atuais II

Prof. Dr. Vladimir Costa Alencar

LANA - UEPB

valencar.com

Histórico (Ano 2.000 – Dias Atuais)

Machine Learning – Deep Learning

Big Data

Data Science

NoSQL (SGBD não Relacionais)

Engenharia de Dados (Hadoop, Spark)

Big Data

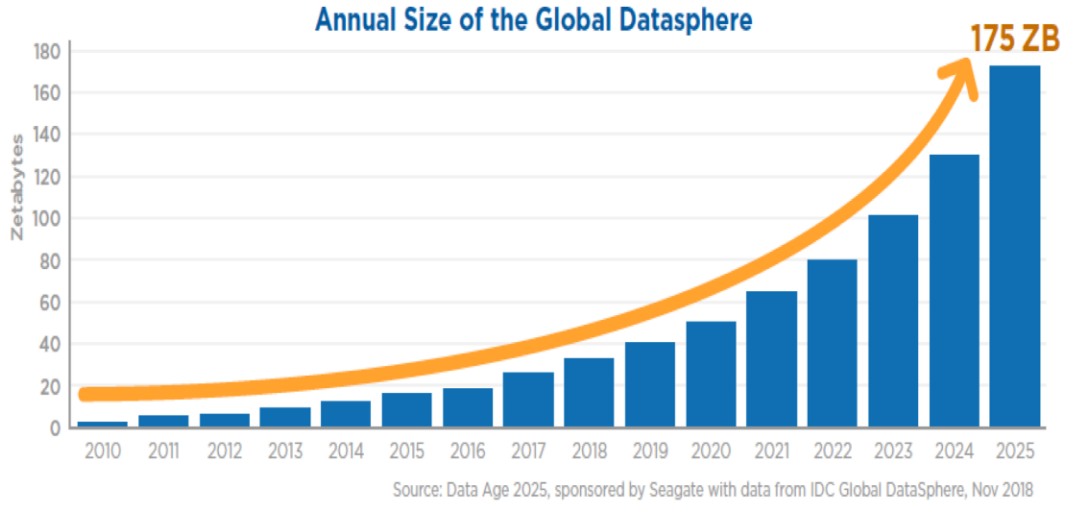
Todos os dispositivos conectados

Até meados de 2024, foram produzidos cerca de **149 trilhões de gigabytes**

90% dos dados no mundo foram criados nos últimos dois anos

Adesão das grandes empresas à internet, Redes sociais, dados dos GPS, dispositivos embutidos e móveis, sensores

Utiliza também posts de mídias sociais, vídeos, geolocalização e comportamento de clientes



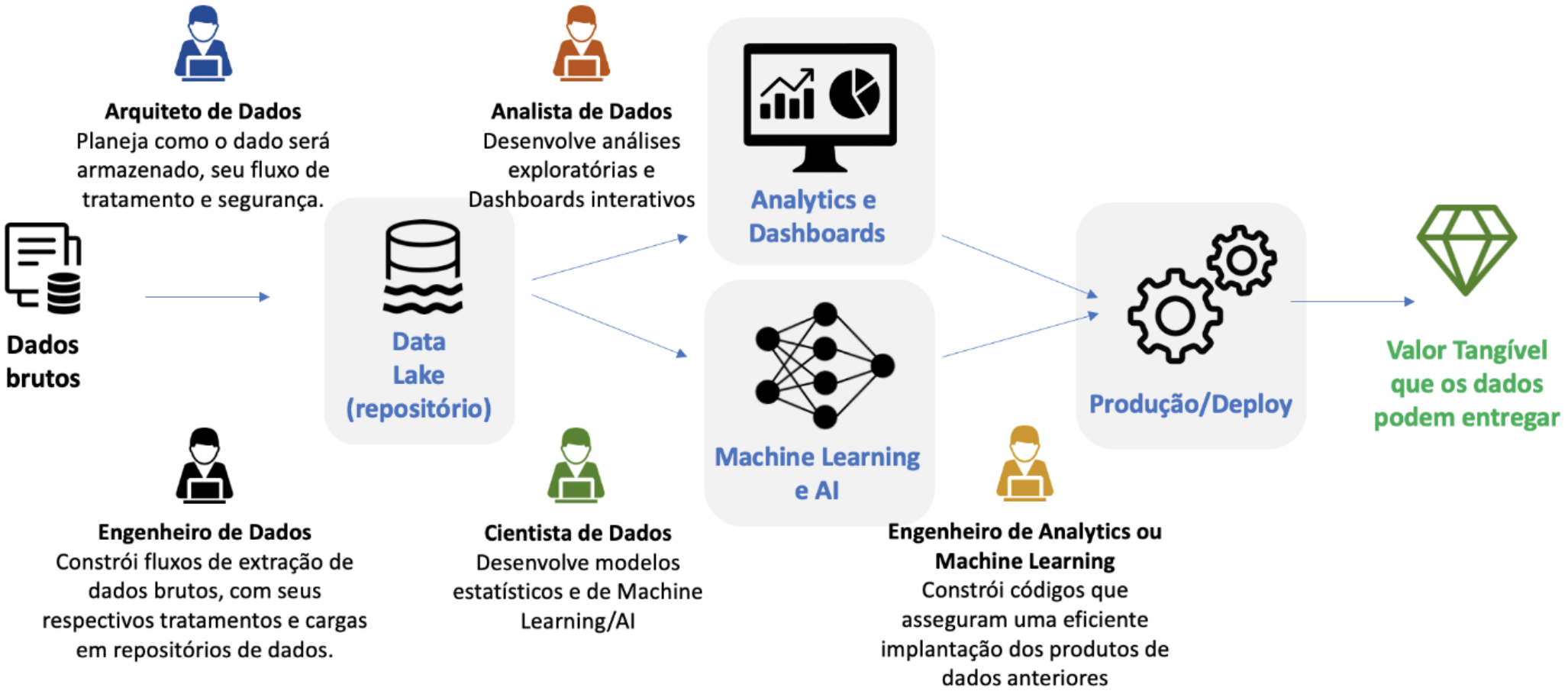
Category	Proportion of Internet Data Traffic
Video	53.72%
Social	12.69%
Gaming	9.86%
Web browsing	5.67%
Messaging	5.35%
Marketplace	4.54%


Nesse contexto, precisamos de profissionais para analisar dados, contruir modelos preditivos, infraestrutura, etc.

Jornada dos dados

As principais funções analíticas na jornada dos dados

Preditiva.ai



Ciência de Dados

Um cientista de dados é alguém que sabe como extrair "significados" de dados e interpretá-los, isso requer habilidade e conhecimento de ferramentas, métodos e técnicas oriundas da estatística, matemática, aprendizagem de máquina, Inteligência Artificial.

Ele passa uma boa parte do seu tempo num processo de coletar o dado, limpá-lo e transformá-lo, pois o dado nunca vem perfeito, ou limpo.

Analista de Dados

- O Analista de Dados é responsável por extrair, preparar e analisar grandes conjuntos de dados para resumir, identificar tendências, padrões e insights que possam ajudar uma empresa a tomar decisões baseadas em dados.

- Ele utiliza ambientes com muitos dados disponíveis, usa estatística e ferramentas como SQL, Power BI e Linguagem Python.

E o que são dados?

JUL 2024

ESSENTIAL DIGITAL HEADLINES

OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES

GLOBAL OVERVIEW

TOTAL POPULATION	UNIQUE MOBILE PHONE SUBSCRIBERS	INDIVIDUALS USING THE INTERNET	SOCIAL MEDIA USER IDENTITIES
8.12 BILLION	**5.68 BILLION**	**5.45 BILLION**	**5.17 BILLION**
URBANISATION	vs. POPULATION	vs. POPULATION	vs. POPULATION
57.9%	**70.0%**	**67.1%**	**63.7%**

we are social

Meltwater

KEPIOS

we

JUL 2024

DIGITAL GROWTH

CHANGE IN THE USE OF CONNECTED DEVICES AND SERVICES OVER TIME

GLOBAL OVERVIEW

TOTAL POPULATION	UNIQUE MOBILE PHONE SUBSCRIBERS	INDIVIDUALS USING THE INTERNET	SOCIAL MEDIA USER IDENTITIES
+0.9%	+2.3%	+3.2%	+5.8%
YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE
+74 MILLION	+126 MILLION	+167 MILLION	+282 MILLION

Meltwater

KEPIOS

we are social

we

JAN 2024

DAILY TIME SPENT WITH MEDIA

THE AVERAGE AMOUNT OF TIME EACH DAY THAT INTERNET USERS AGED 16 TO 64 SPEND WITH DIFFERENT KINDS OF MEDIA AND DEVICES

GLOBAL OVERVIEW

TIME SPENT USING THE INTERNET

GWI.

6H 40M

YEAR-ON-YEAR CHANGE

+0.8% (+3 MINS)

TIME SPENT WATCHING TELEVISION (BROADCAST AND STREAMING)

Meltwater

3H 06M

YEAR-ON-YEAR CHANGE

-8.2% (-17 MINS)

TIME SPENT USING SOCIAL MEDIA

GWI.

2H 23M

YEAR-ON-YEAR CHANGE

-5.5% (-8 MINS)

TIME SPENT READING PRESS MEDIA (ONLINE AND PHYSICAL PRINT)

1H 41M

YEAR-ON-YEAR CHANGE

-22.2% (-29 MINS)

TIME SPENT LISTENING TO MUSIC STREAMING SERVICES

we are social

1H 25M

YEAR-ON-YEAR CHANGE

-12.8% (-13 MINS)

TIME SPENT LISTENING TO BROADCAST RADIO

GWI.

0H 50M

YEAR-ON-YEAR CHANGE

-15.5% (-9 MINS)

TIME SPENT LISTENING TO PODCASTS

KEPIOS

0H 49M

YEAR-ON-YEAR CHANGE

-20.3% (-13 MINS)

TIME SPENT USING A GAMES CONSOLE

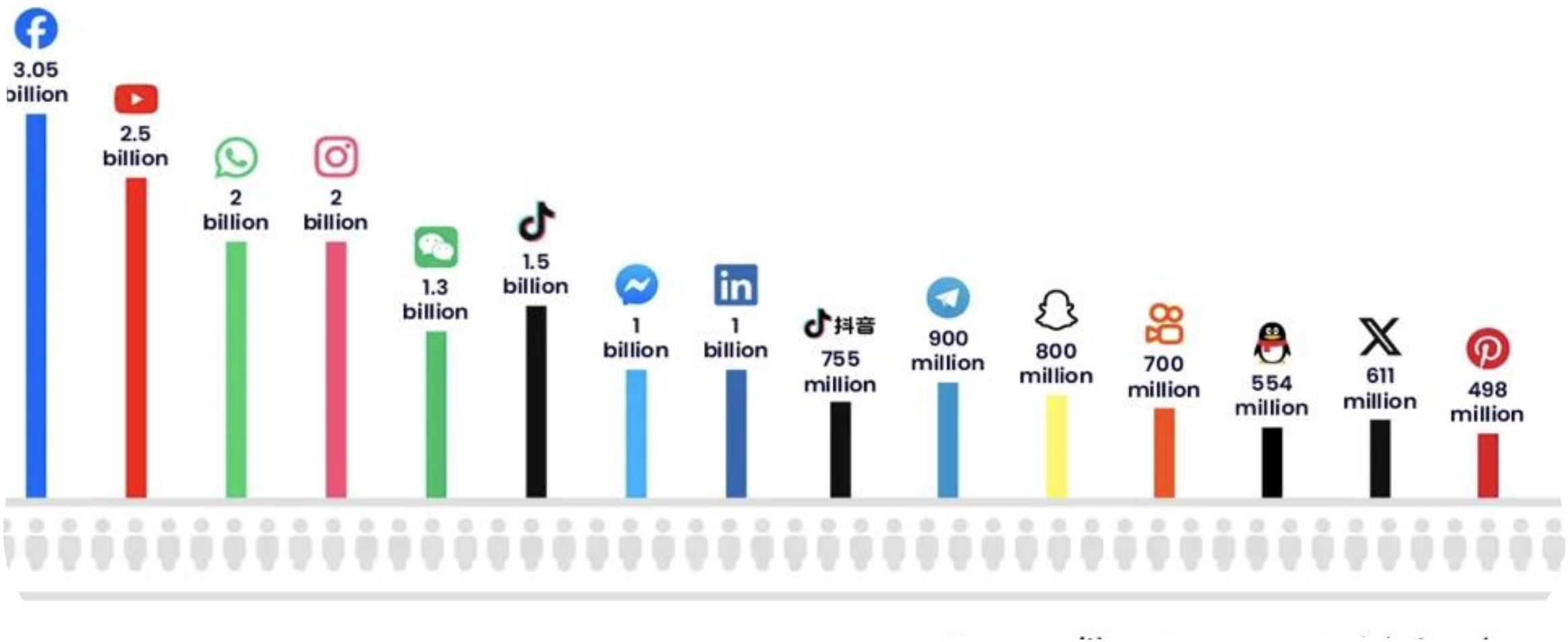
1H 02M

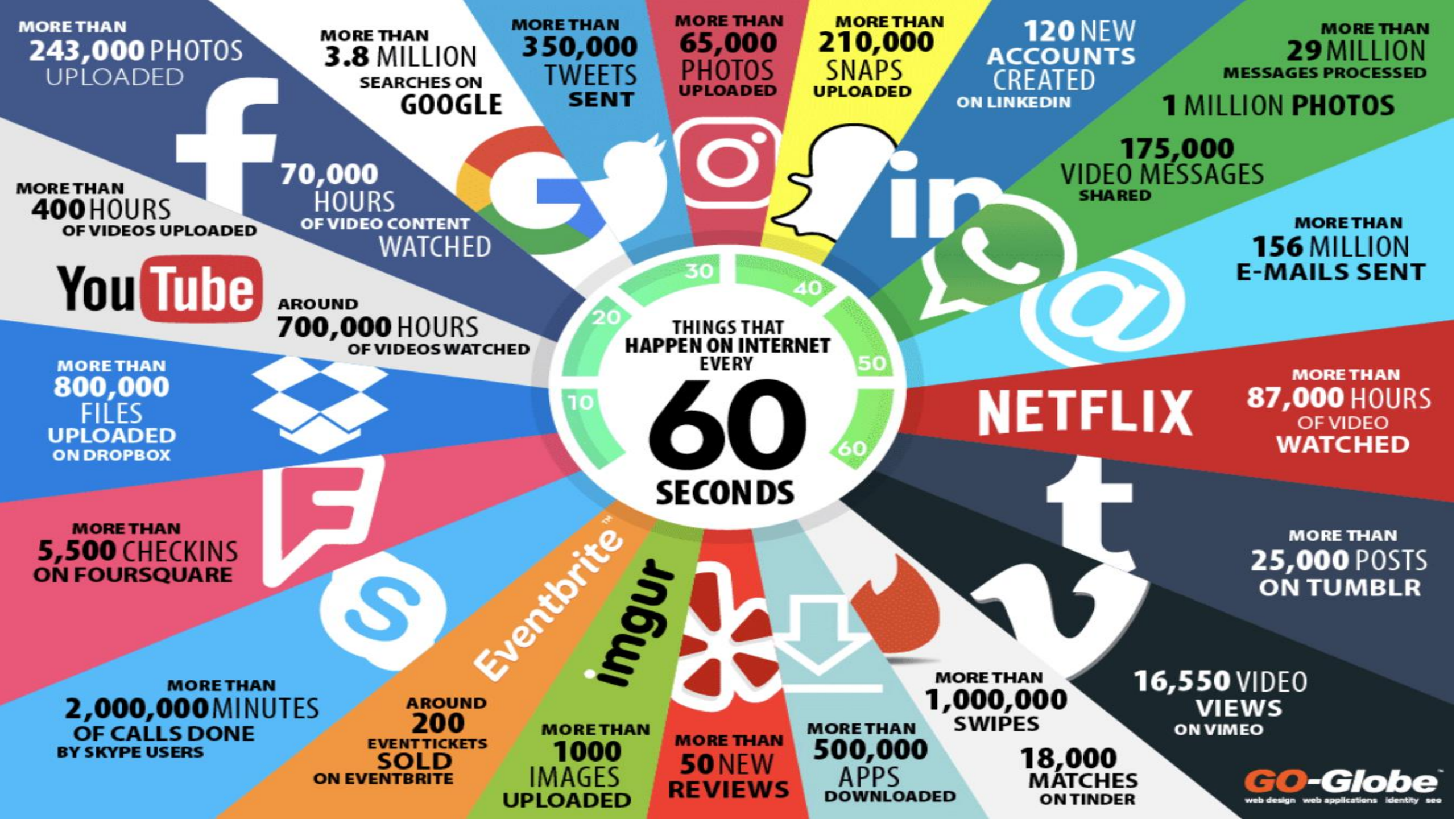
YEAR-ON-YEAR CHANGE

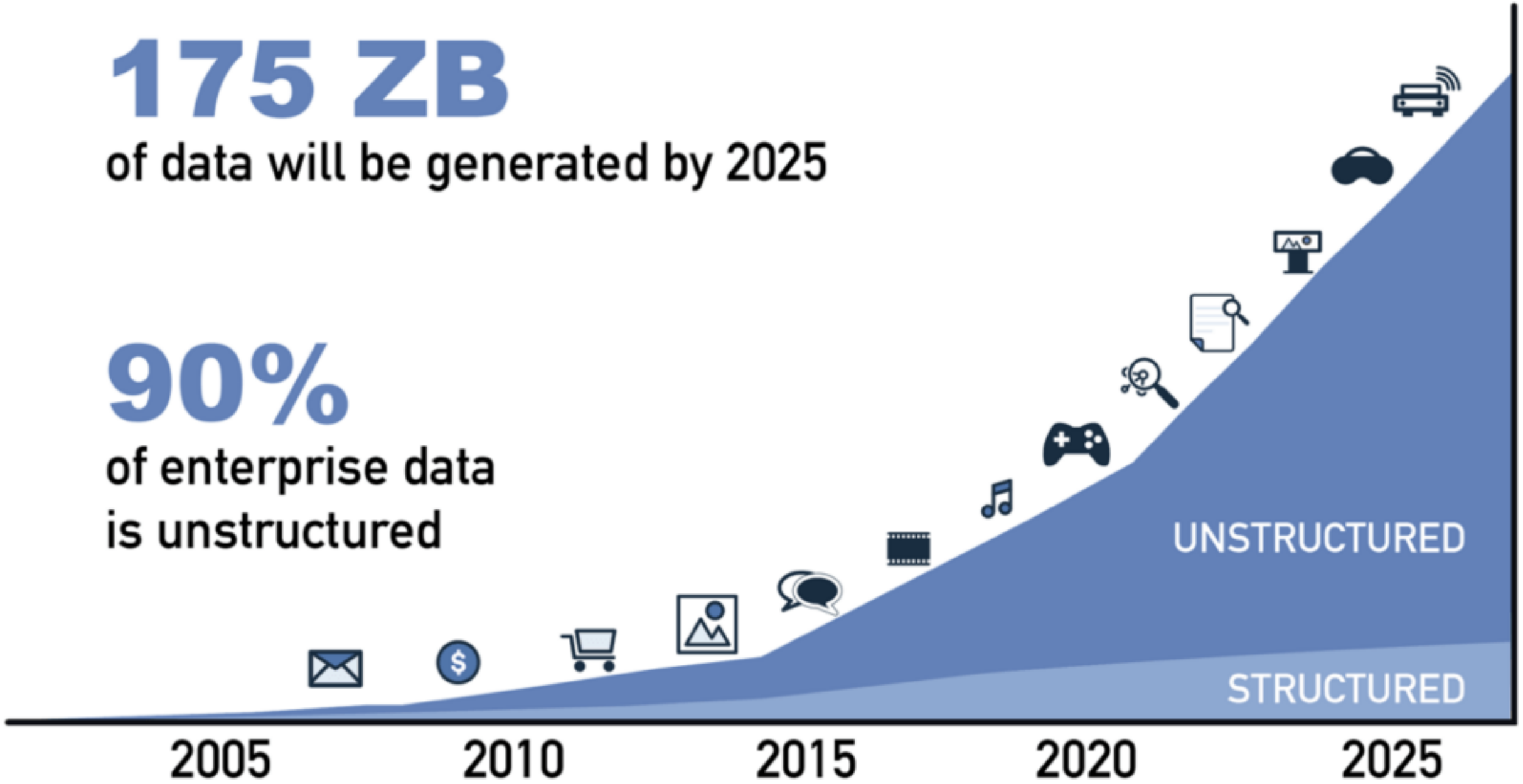
-16.7% (-12 MINS)

SOURCE: GWI (Q3 2023). FIGURES REPRESENT THE FINDINGS OF A BROAD SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GWI.COM. **NOTES:** PEOPLE MAY CONSUME DIFFERENT MEDIA CONCURRENTLY

Most Popular Social Networks 2024

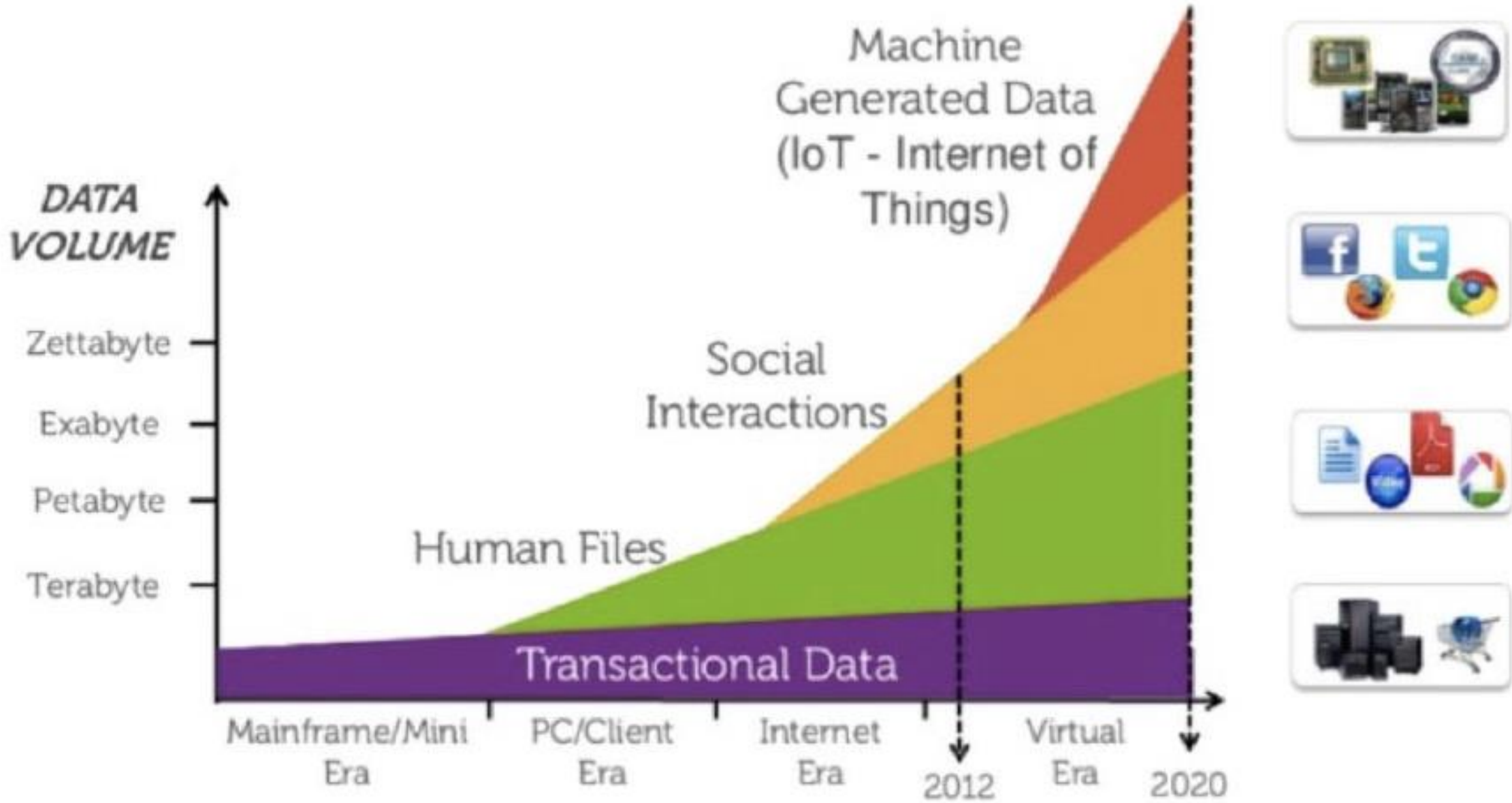



MORE THAN 243,000 PHOTOS UPLOADED
MORE THAN 3.8 MILLION SEARCHES ON GOOGLE
MORE THAN 350,000 TWEETS SENT
MORE THAN 65,000 PHOTOS UPLOADED
MORE THAN 210,000 SNAPS UPLOADED
120 NEW ACCOUNTS CREATED ON LINKEDIN
MORE THAN 29 MILLION MESSAGES PROCESSED
1 MILLION PHOTOS
175,000 VIDEO MESSAGES SHARED
70,000 HOURS OF VIDEO CONTENT WATCHED
MORE THAN 400 HOURS OF VIDEOS UPLOADED
MORE THAN 156 MILLION E-MAILS SENT
AROUND 700,000 HOURS OF VIDEOS WATCHED
THINGS THAT HAPPEN ON INTERNET EVERY 60 SECONDS
MORE THAN 800,000 FILES UPLOADED ON DROPBOX
MORE THAN 87,000 HOURS OF VIDEO WATCHED
MORE THAN 5,500 CHECKINS ON FOURSQUARE
MORE THAN 25,000 POSTS ON TUMBLR
MORE THAN 2,000,000 MINUTES OF CALLS DONE BY SKYPE USERS
AROUND 200 EVENT TICKETS SOLD ON EVENTBRITE
MORE THAN 1000 IMAGES UPLOADED
MORE THAN 50 NEW REVIEWS
MORE THAN 500,000 APPS DOWNLOADED
MORE THAN 1,000,000 SWIPES
16,550 VIDEO VIEWS ON VIMEO
18,000 MATCHES ON TINDER
GO-Globe
web design web applications identity seo


175 ZB
of data will be generated by 2025
90%
of enterprise data
is unstructured
UNSTRUCTURED
STRUCTURED
2005
2010
2015
2020
2025

SOURCE: IDC

The Explosion of Data



Artificial Intelligen...
Termo de pesquisa

Data Science
Termo de pesquisa

Big Data
Termo de pesquisa

+ Adicionar comparação

Mundo | 01/08/2010 – 18/08/2024 | Todas as categorias | Pesquisa na Web

Interesse ao longo do tempo

As marcas mais valiosas do mundo - 2024

1	2	3	4	5
+1	+2	0	-3	+1
	Microsoft	Google	amazon	SAMSUNG
$516.6 bn	$340.4 bn	$333.4 bn	$308.9 bn	$99.4 bn
+74%	+78%	+19%	+3%	-0%

6	7	8	9	10
-1	+3	+6	+2	-3
Walmart	TikTok	facebook	T	ICBC 中国工商银行
$96.8 bn	$84.2 bn	$75.7 bn	$73.3 bn	$71.8 bn
-15%	+28%	+28%	+17%	+3%

As marcas mais fortes do mundo - 2024

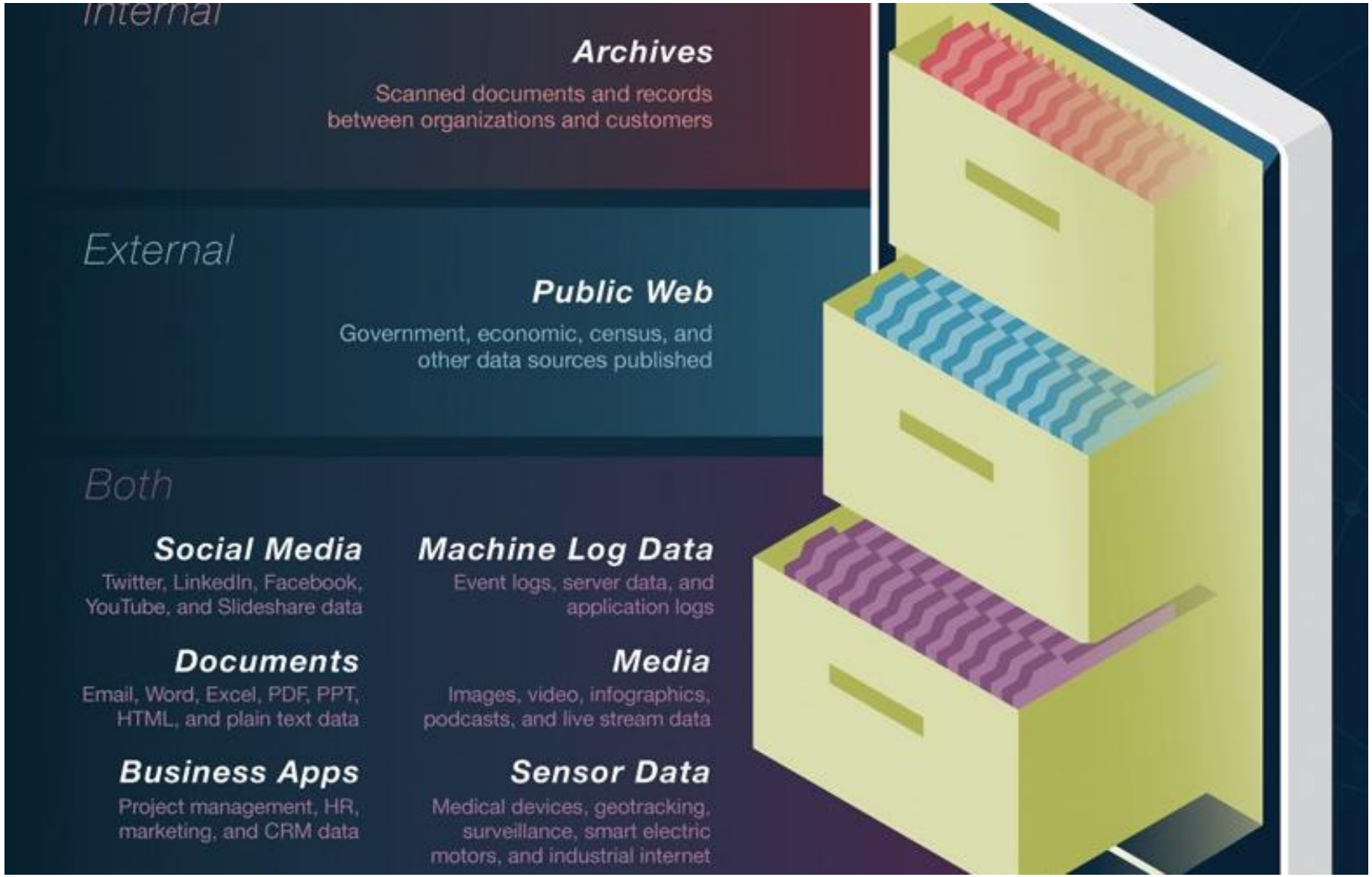
1	2	3	4	5
+3	0	-2	New	+6
	YouTube	Google	MARINA BAY Sands SINGAPORE	
94.3 +2.7	92.7 +0.0	92.5 -0.8	91.8 +5.0	90.7 +0.9

6	7	8	9	10
-1	+5	+40	+10	-2
Deloitte.	Coca-Cola	NETFLIX	ROLEX	
90.6 -0.7	90.4 +0.8	90.2 +4.3	90.2 +2.3	90.0 -0.7

Data Science e Big Data

- Big Data é uma coleção de conjuntos de dados, grandes e complexos,

Não podem ser processados por bancos de dados ou aplicações de processamento tradicionais

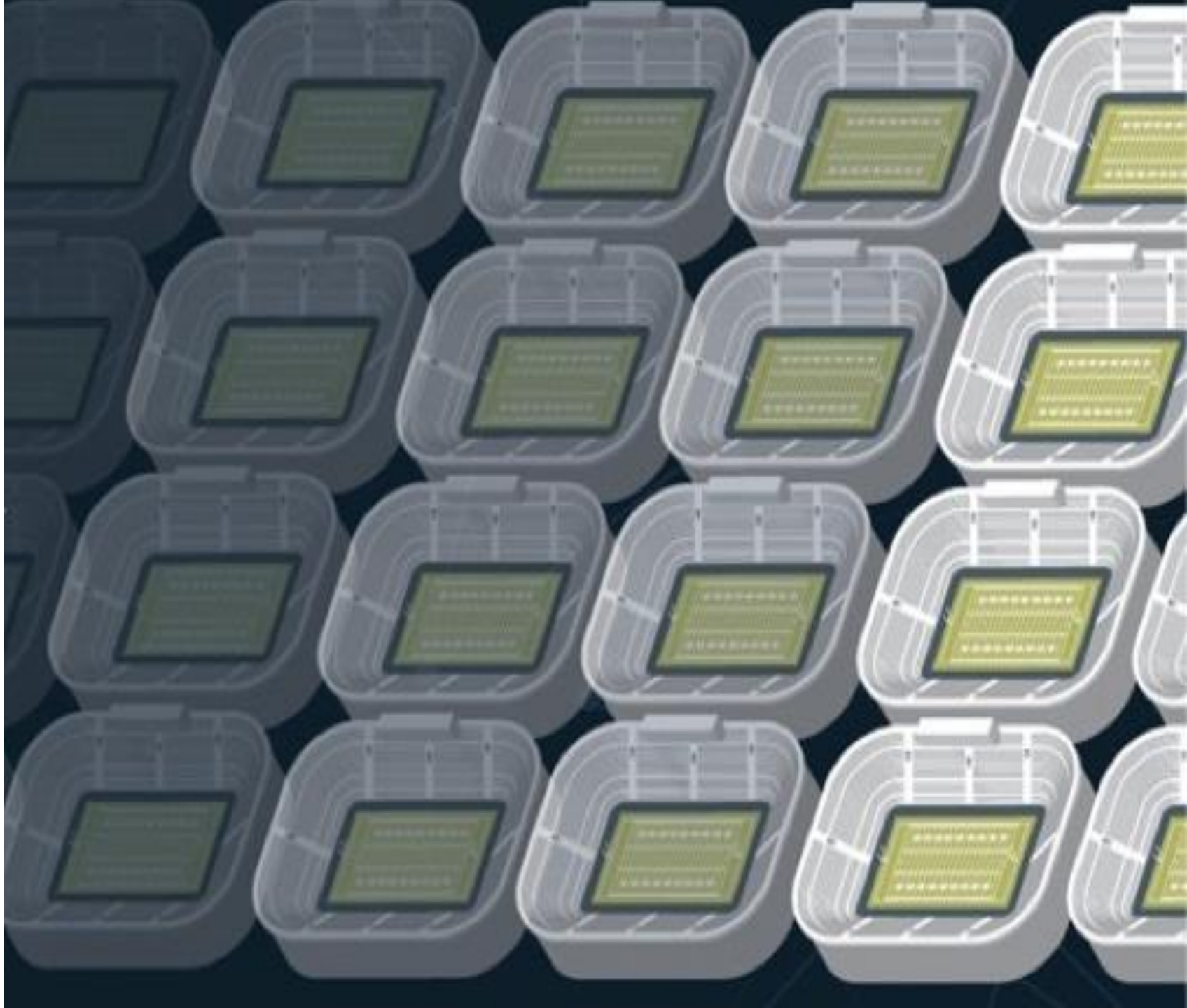

Internal
Archives
Scanned documents and records between organizations and customers
External
Public Web
Government, economic, census, and other data sources published
Both
Social Media
Twitter, LinkedIn, Facebook, YouTube, and Slideshare data
Machine Log Data
Event logs, server data, and application logs
Documents
Email, Word, Excel, PDF, PPT, HTML, and plain text data
Media
Images, video, infographics, podcasts, and live stream data
Business Apps
Project management, HR, marketing, and CRM data
Sensor Data
Medical devices, geotracking, surveillance, smart electric motors, and industrial internet

A quantidade total de dados digitais no mundo foi estimada em aproximadamente 149 zettabytes (ZB) ou **149 trilhões de gigabytes** (2024).

Para armazenar os dados atuais colocados em data centes:

1 Datacenter = 1 Bilhão de GB (25.000 m^2)

Os dados ocupariam cerca de 521.710 campos de futebol



Comparação de Medidas

Maior estádio do mundo

May Day, em Pyongyang, capital da Coréia do Norte.

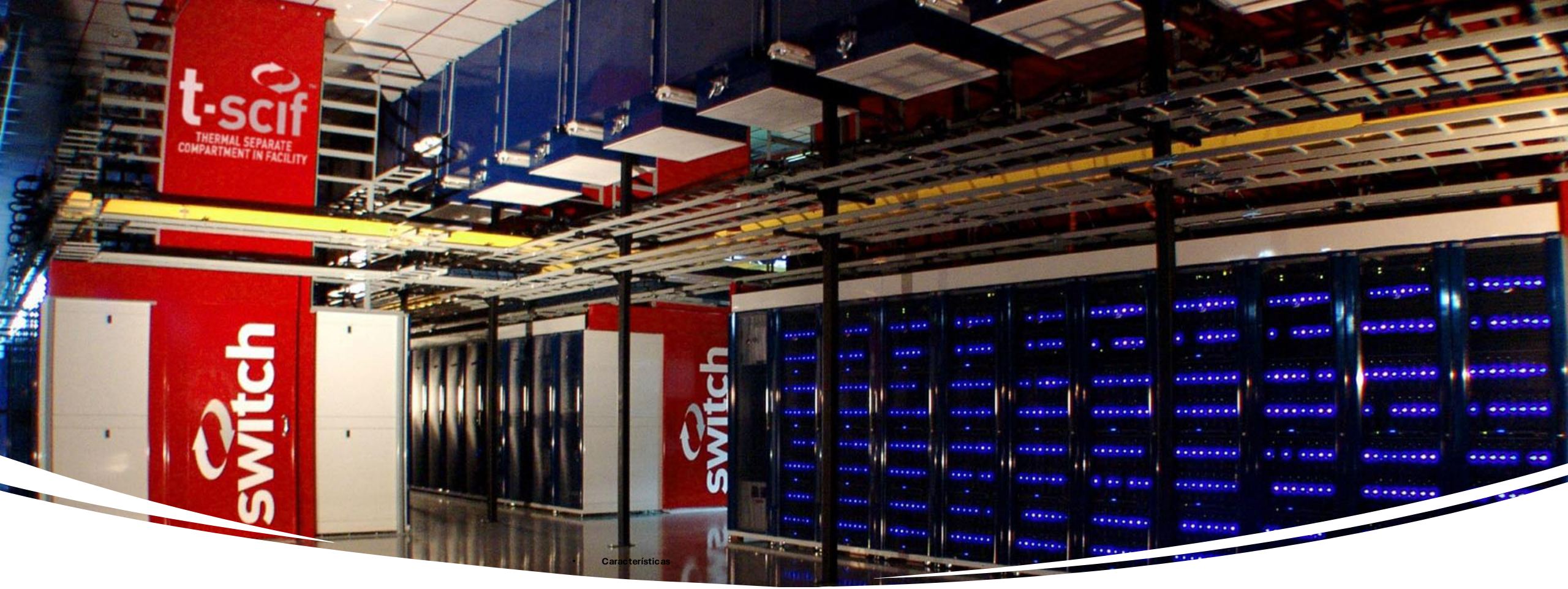
Ele tem impressionantes **207.000 metros quadrados**.

A arena tem capacidade para **110.000** espectadores

Switch SuperNAP (Las Vegas, Nevada) - Data Center

Ocupando uma área de **3.5 milhões de metros quadrados (Aproximadamente 17 Estádios de Futebol do May Day, Coréia do Norte)**

A escolha de sua localização em **Las Vegas, Nevada**, foi estratégica para protegê-lo de desastres naturais.

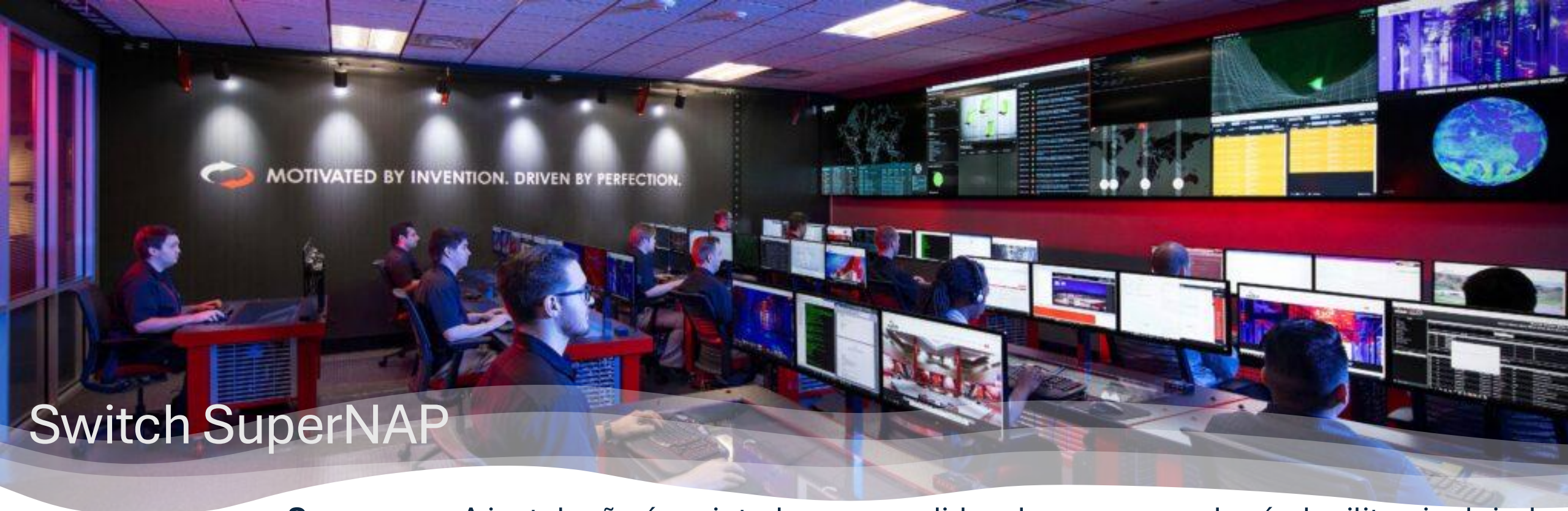


Características

Switch SuperNAP

- **Capacidade de Energia:** Projetado para suportar mais de **430 MW (megawatts)** de potência, oferecendo uma infraestrutura de energia robusta e redundante para acomodar ambientes de computação de alta densidade.
- 1 cidade = 97 MW por hora => Aproximadamente 4.4 cidades

- **Certificação Tier IV:** É o nível mais alto de certificação para a confiabilidade de data centers. Isso significa que eles possuem tolerância a falhas e redundância excepcionais.



Switch SuperNAP

- **Segurança:** A instalação é projetada com medidas de segurança de nível militar, incluindo segurança armada 24/7, controles de acesso biométrico e um sistema patenteado de contenção térmica em corredores para proteger o hardware e a integridade dos dados.
- **Sistema de Resfriamento:** A Switch utiliza um design proprietário de HVAC (aquecimento, ventilação e ar condicionado) conhecido como **T-SCIF (Thermal Separate Compartment in Facility)**, que mantém temperaturas ideais enquanto minimiza o consumo de energia.
- **Conectividade:** O hub de interconexão fornece acesso a mais de **60 operadoras nacionais e internacionais**. Está estrategicamente localizado para oferecer conectividade de baixa latência à costa oeste dos Estados Unidos e outros mercados globais

Switch SuperNAP

O Campus Citadel está localizado em um no Tahoe Reno Industrial Center, ao lado da Tesla Gigafactory, e é totalmente alimentado por energia renovável. Assim, a Switch é a única empresa com um Índice de Energia 100% Limpa.

- O Tahoe Reno Citadel Campus é cercado por um muro de concreto sólido de 6 metros de altura, garantindo segurança e confiabilidade.
- Com um custo de US$ 1 bilhão, 800 quilômetros de cabos de rede de fibra óptica conectarão São Francisco, Reno, Los Angeles e Las Vegas.
- As 50 milhões de pessoas que vivem nas quatro cidades poderão acessar dados em apenas 14 milissegundos.

Switch TAHOE RENO
The Citadel
Campus Nevada, USA

O campus de 8.094 hectares (80.940 metros quadrados), localizado no Tahoe Reno Industrial Center, ao lado da Tesla Gigafactory, é alimentado por energia 100% renovável, comprometendo-se com uma liderança no setor de energia verde.

SuperNap Las Vegas

Data Center da China Telecom (Hohhot, Mongólia Interior)

Data Center da China Telecom (Mongólia)

- **Tamanho:** Aproximadamente **1,2 milhão de metros quadrados** (120 hectares) , 7 estádios de futebol tipo May Day da Coréia do Norte.
- **Localização:** Hohhot, Mongólia Interior, China.
- **Finalidade:** Este data center serve como um hub para **serviços de nuvem, serviços de internet, e-commerce earmazenamento** para diversas empresas e agências governamentais.
- **Características:** A instalação utiliza tecnologias avançadas de resfriamento, incluindo resfriamento natural devido ao clima mais frio da Mongólia Interior, o que ajuda a reduzir o consumo de energia e os custos. Ela é projetada para lidar com **grandes quantidades de dados e necessidades de computação de alto desempenho**.

Um técnico demonstra um sistema de big data na zona de computação em nuvem da China Telecom, localizada na Mongólia Interior, em Hohhot, na Região Autônoma da Mongólia Interior, no norte da China.

Em 2016, a região designou as indústrias de big data e computação em nuvem como novos motores para o desenvolvimento local, prometendo que o valor de produção da indústria de big data regional superará 100 bilhões de yuan (14,5 bilhões de dólares dos EUA) em 2020.

As autoridades locais visam tornar a zona de big data da Mongólia Interior o maior centro de big data do norte da China. (Xinhua/Lian Zhen)

Big Data

Data Science

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.5 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

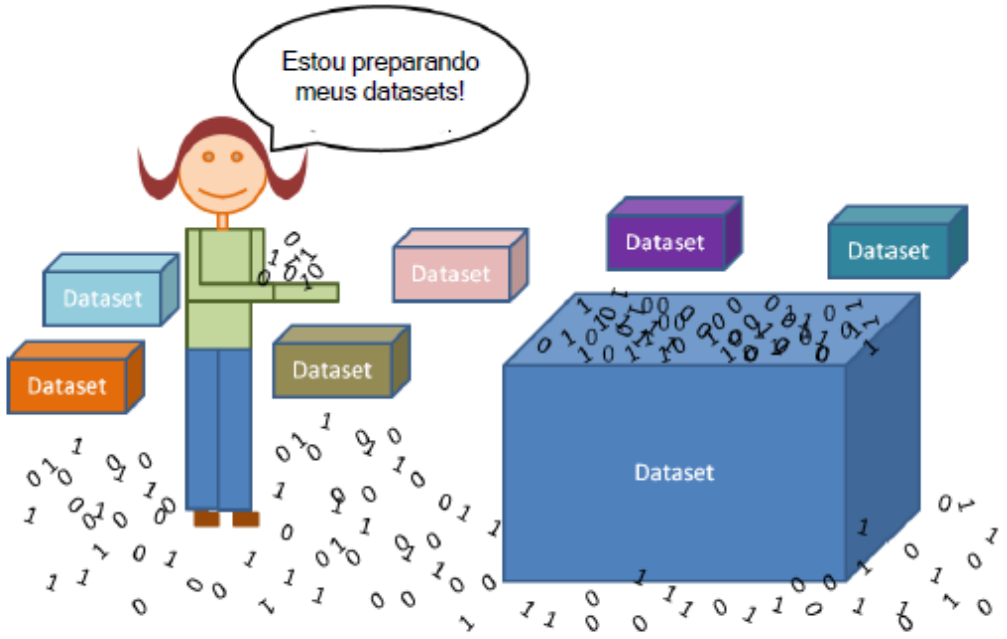
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

O que é um Dataset?

Coleção de observações

Cada observação é tipicamente chamada de registro

Cada registro tem um conjunto de atributos que apontam características, comportamentos ou resultados



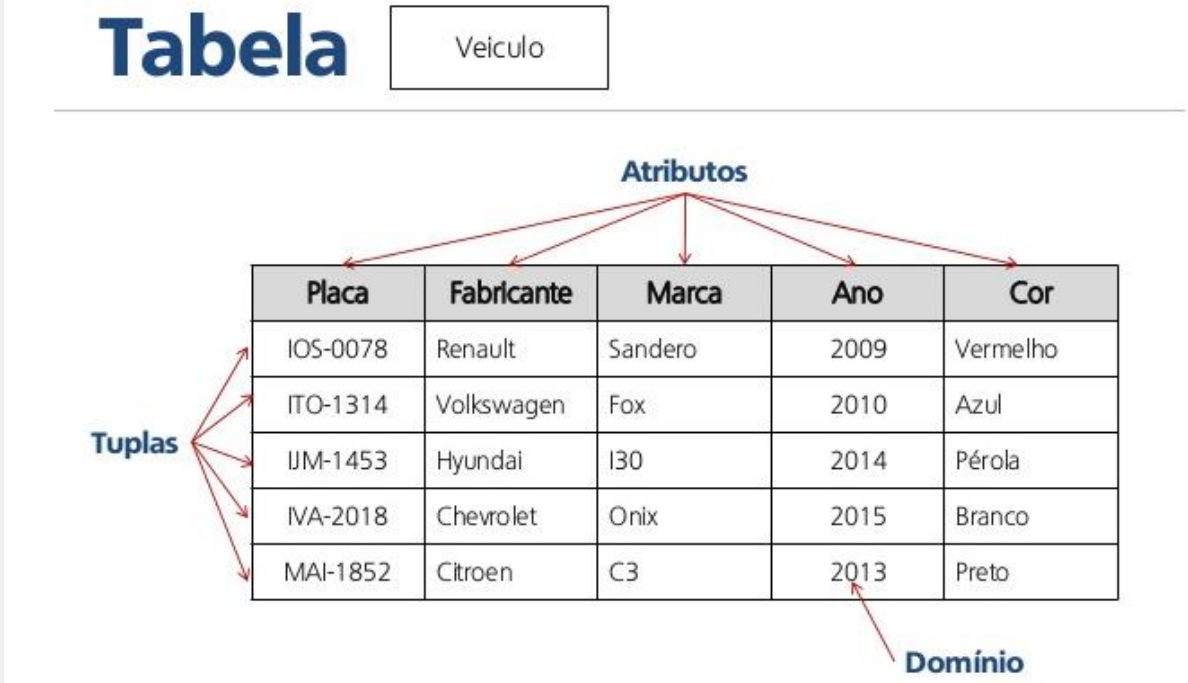
Tipos de Dados

Banco de Dados

XML, JSON

Twitter, Posts do Facebook, Fotos, Vídeos, Música.

Tipos de Dados: Estruturados

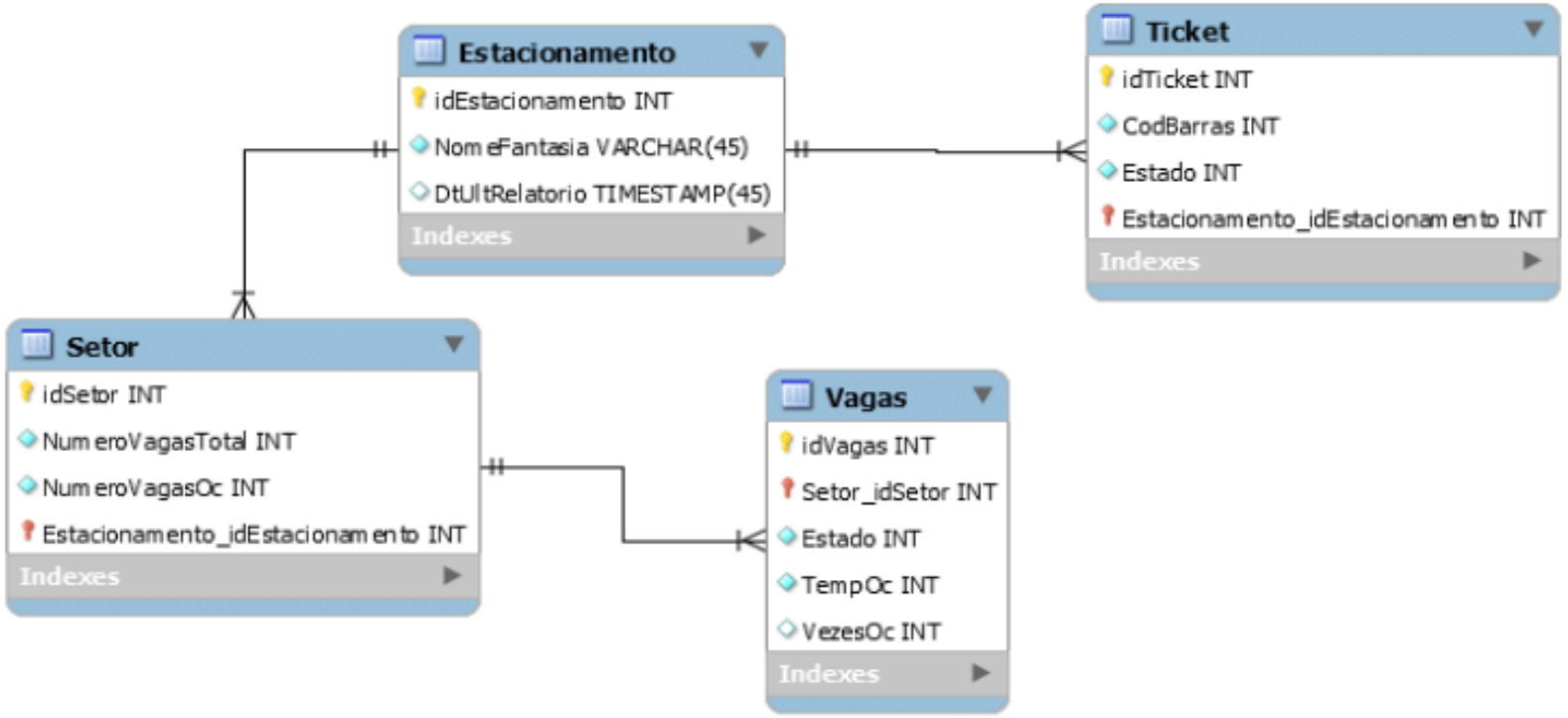


Banco de Dados Relacionais

- Armazena dados em tabelas
- Possuem atributo tipo chave
- implementam funcionalidades simples do tipo CRUD
  (Create, Read, Update e Delete)
- Controla o armazenamento, recuperação, exclusão, segurança e integridade dos dados

	Id_Cliente	Nome	Data_Nascimento	Salario
1	1	João	1981-05-14 00:00:00.000	4521
2	2	Marcos	1975-01-07 00:00:00.000	1478,58
3	3	André	1962-11-11 00:00:00.000	7151,45
4	4	Simão	1991-12-18 00:00:00.000	2584,97
5	5	Pedro	1986-11-20 00:00:00.000	987,52
6	6	Paulo	1974-08-04 00:00:00.000	6259,14
7	7	José	1979-09-01 00:00:00.000	5272,13

Tipos de Dados: Estruturados – Banco de Dados Relacionais




SQLite
ORACLE
PostgreSQL
db isam
TURBODB
Embedded
Microsoft
SQL Server
Microsoft
SQL Azure
elevatedb
MySQL

Tipos de Dados: Semi-Estruturados - XML

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
- <MUSICAS xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  - <MUSICA>
      <NOME>A Fórmula Do Amor</NOME>
      <CANTOR>Kid Abelha</CANTOR>
      <LETRA>Eu tenho gestos aptos</LETRA>
    </MUSICA>
  - <MUSICA>
      <NOME>A Viagem</NOME>
      <CANTOR>Roupa Nova</CANTOR>
      <LETRA>Há tanto tempo que eu deixei você</LETRA>
    </MUSICA>
  - <MUSICA>
      <NOME>Águas De Março</NOME>
      <CANTOR>Elis Regina</CANTOR>
      <LETRA>É pau é pedra</LETRA>
    </MUSICA>
</MUSICAS>
```

Semi-Estruturado – Arquivo .JSON (Java Script Object Notation)

{"id_str": "694877215301873666",

"text": "When Ebola performs duty in Sierra Leone,US UK "Warships" were sent 2 treat patients. Now Zica: sending 2 #Brazil",

"coordinates": 10000,20000,

"followers_count": 722,

"description": "Portfolio Manager BehindtheScenes Money Maker Management.

"friends_count": 165,

"location": "ATHENS",

"screen_name": "anthonysamaha",

"lang": "en",

"favourites_count": 4963,

"name": "Anthony Samaha",

"url": "http://t.co/o6uil766Ds",

"time_zone": "Athens",

"lang": "en",

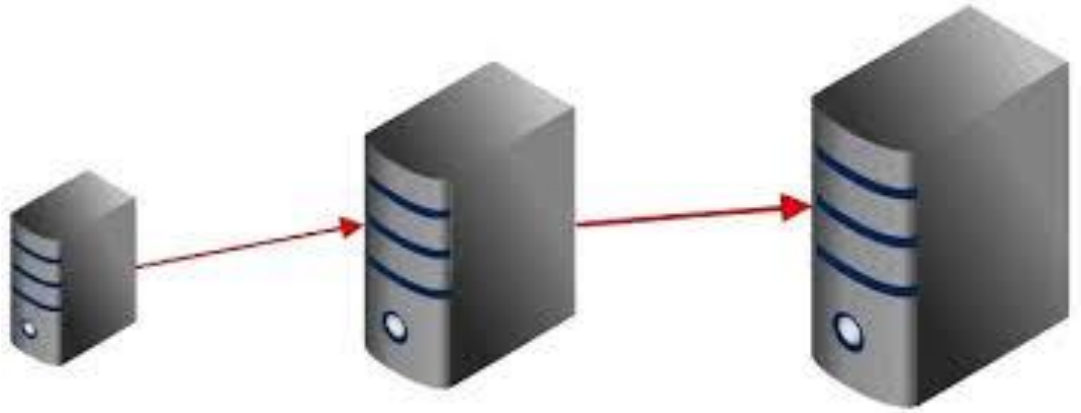
"created_at": "Wed Feb 03 13:36:52 +0000 2016"

"place": Athens,

}

Banco de Dados Não-Relacionais (NoSQL) Escalabilidade

- Volume de dados crescente
- Disponiblidade dos dados (tempo real)

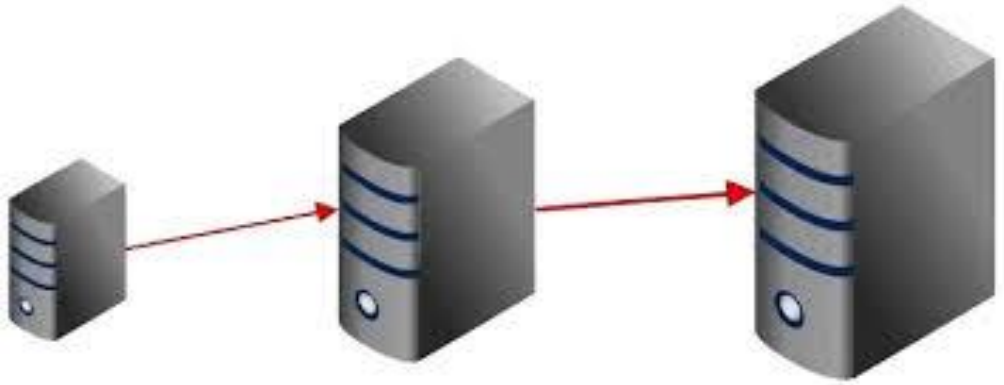


Escalabilidade

- Característica de **Aumentar a capacidade** do sistema

(seja processamento, armazenamento, E/S)

- Há um Aumento da **Performance** depois da adição de **Hardware ou Serviço.**

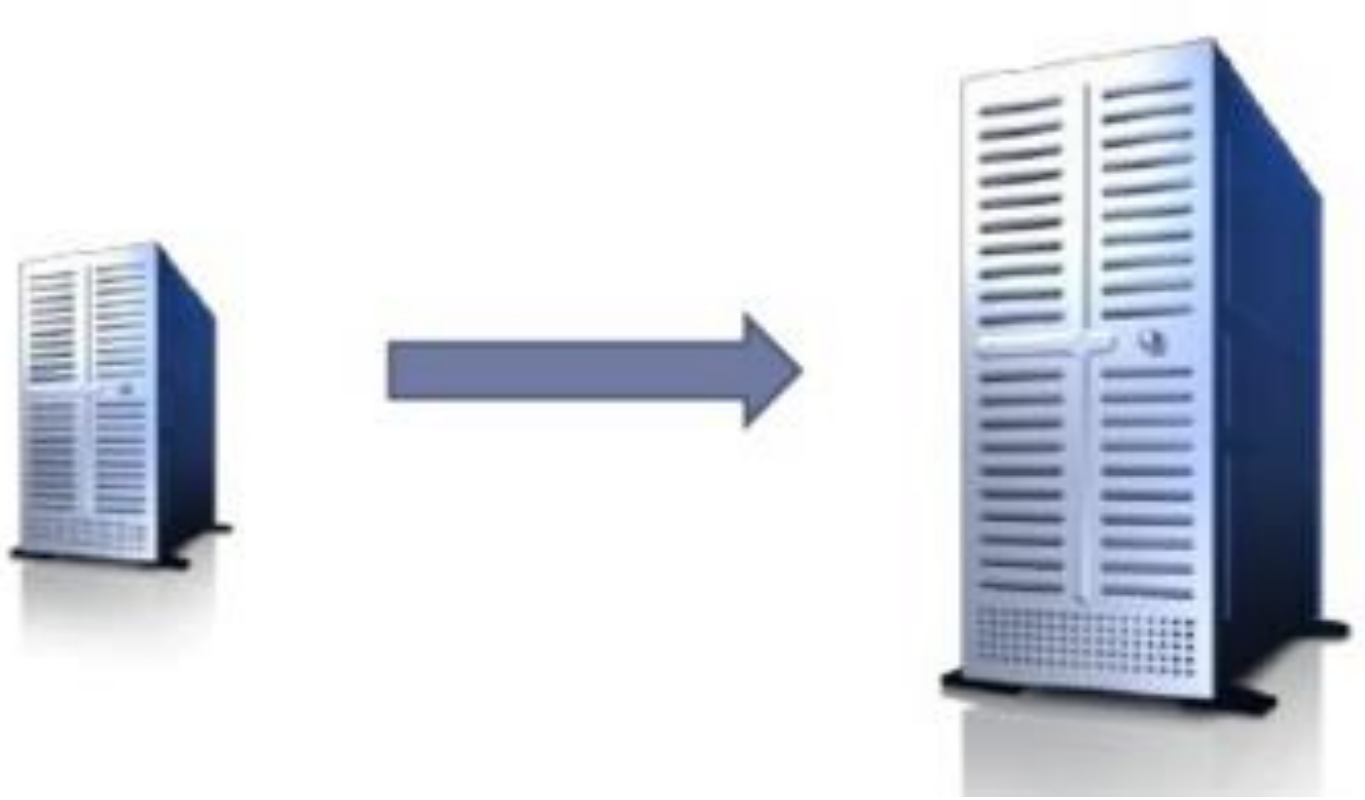


Escalabilidade Vertical

Adicionar novos componentes na máquina.

É quando você coloca mais memória, mais disco, mais CPU no seu servidor.

Geralmente requer desligar a máquina, adicionar recursos e ligar novamente.



Escalabilidade Horizontal

Adicionar máquinas em Paralelo

Coloca mais servidores para atender a demanda

Carga é balanceada entre os servidores
(ex. Cluster)

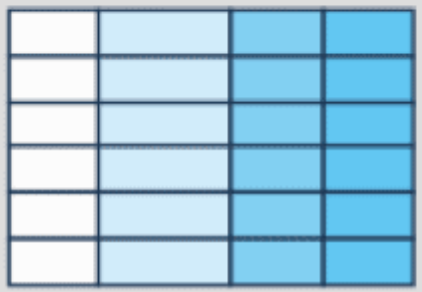
NoSQL - Escalabilidade

Escalabilidade Vertical
(Scale Up)

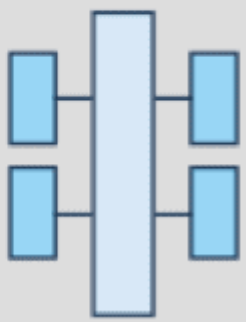
Escalabilidade Horizontal
(Scale Out)

SQL

Relational

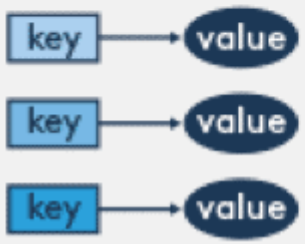


Analytical (OLAP)



NoSQL

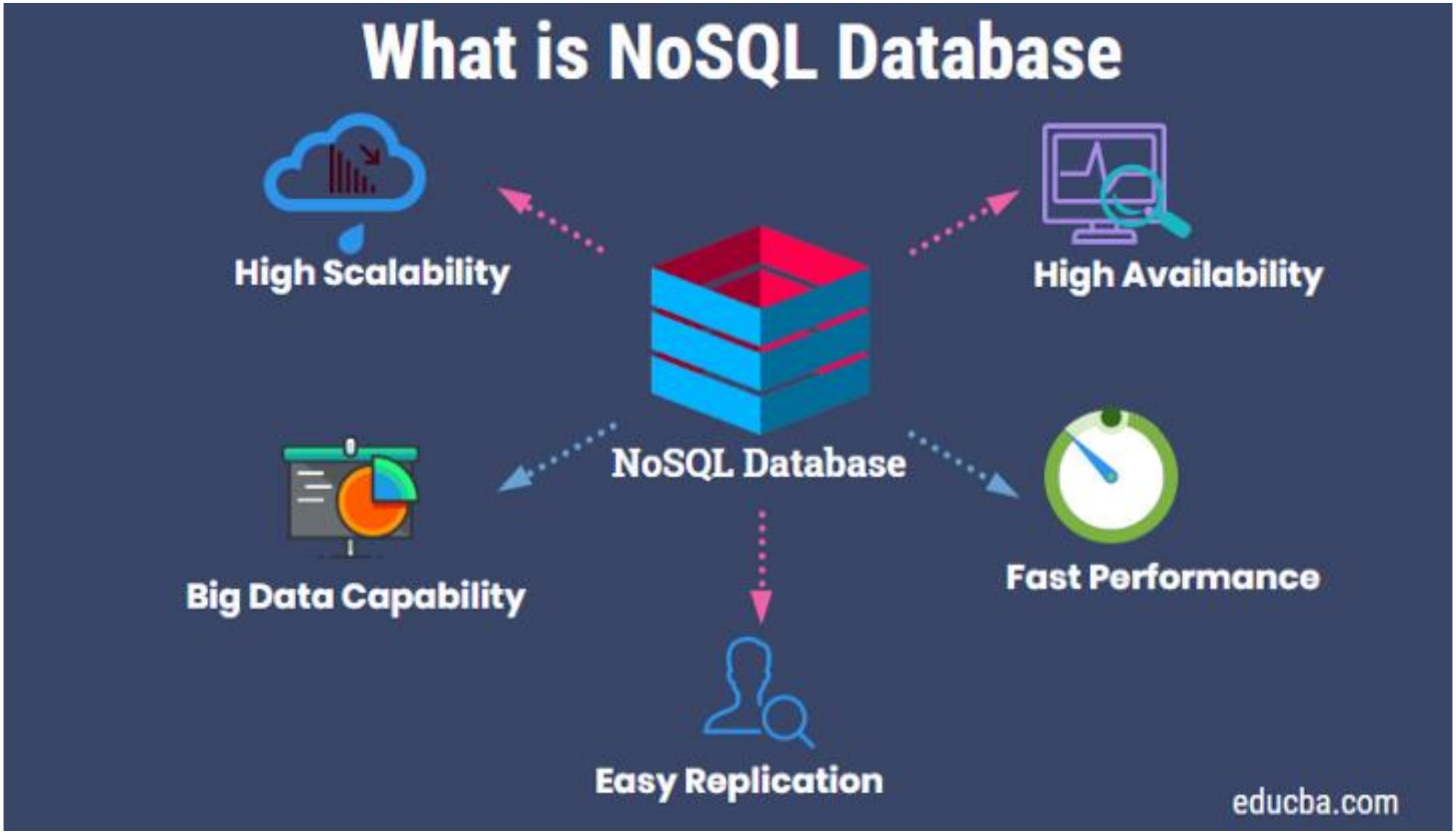
Key-Value



Column-Family

Graph

Document


What is NoSQL Database
High Scalability
High Availability
NoSQL Database
Big Data Capability
Fast Performance
Easy Replication
educba.com


mongoDB


Apache
CASSANDRA™


APACHE
HBASE


amazon
DynamoDB


neo4j


RAVENDB


CouchDB
relax


redis


HYPERTABLE INC


OrientDB

NoSQL – Não Relacionais

- São Facilmente escaláveis Horizontalmente

- Trabalham com quantidades maciças de dados (Big Data)

- Não utilizam o modelo relacional para suas estruturas de dados

- NoSql – Not Only Sql

Banco de Dados Não-Relacionais (NoSQL)

- Gerenciar os grandes volumes de dados
- Buscar alto desempenho e disponibilidade
- Permitem uma escalabilidade mais barata e menos trabalhosa
- Características de poder trabalhar com dados semi-estruturados ou crus vindos de diversas origens
- (arquivos de log, web-sites, arquivos multimídia, JSON, etc...)

- Modelos Baseados em Documentos, Colunas, Chave-Valor e Grafos.

Tipos (NoSQL)

Chave/Valor (Key/Value)

Orientados a Documentos

Orientados a Colunas

Grafos

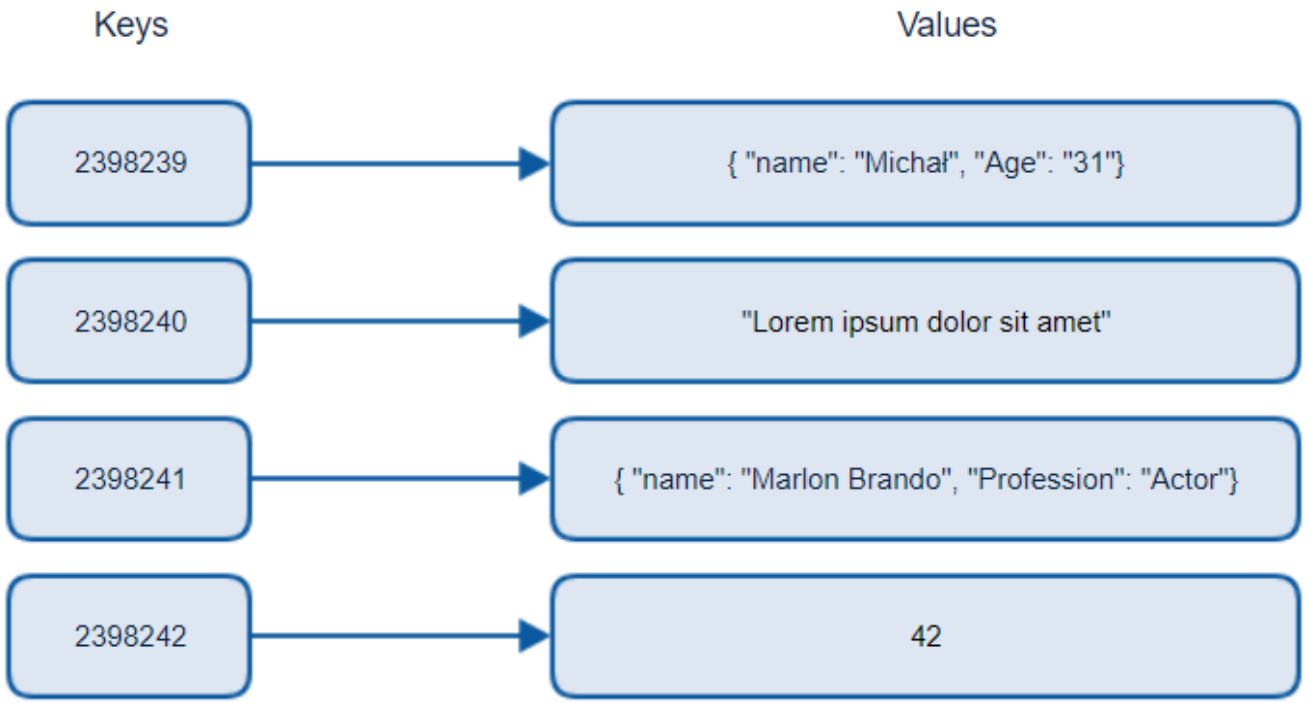
NoSQL – Chave/Valor

O modelo chave-valor é o mais simples e fácil de implementar

Usa uma tabela hash na qual há

uma chave única e um indicador

de um dado ou de um item em particular



NoSQL – Chave/Valor - Aplicações:

- Cache de conteúdo

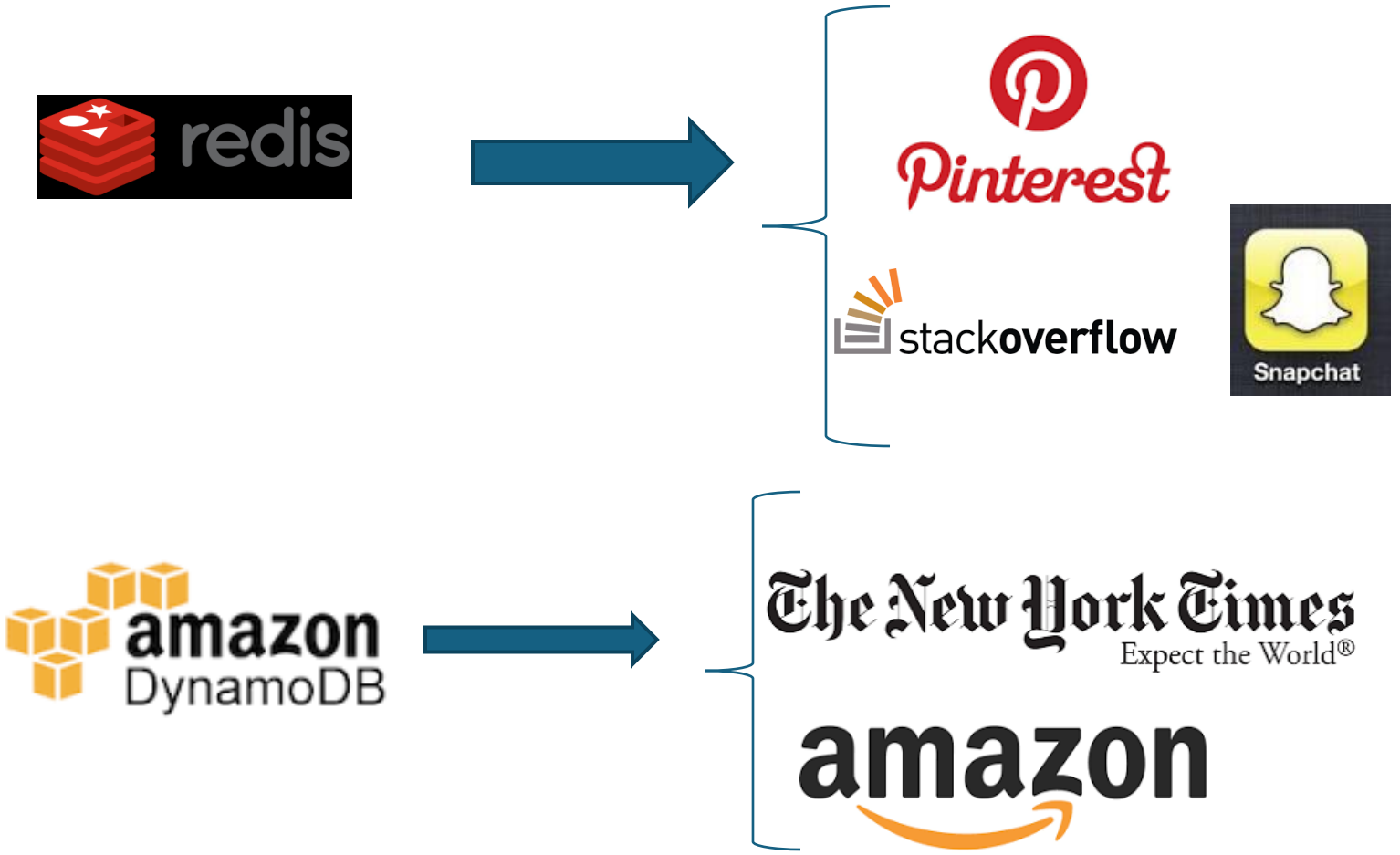
(grande quantidade de dados, carregamento massivo)

- Pesquisas rápidas

- Logging

(registro de eventos relevantes)

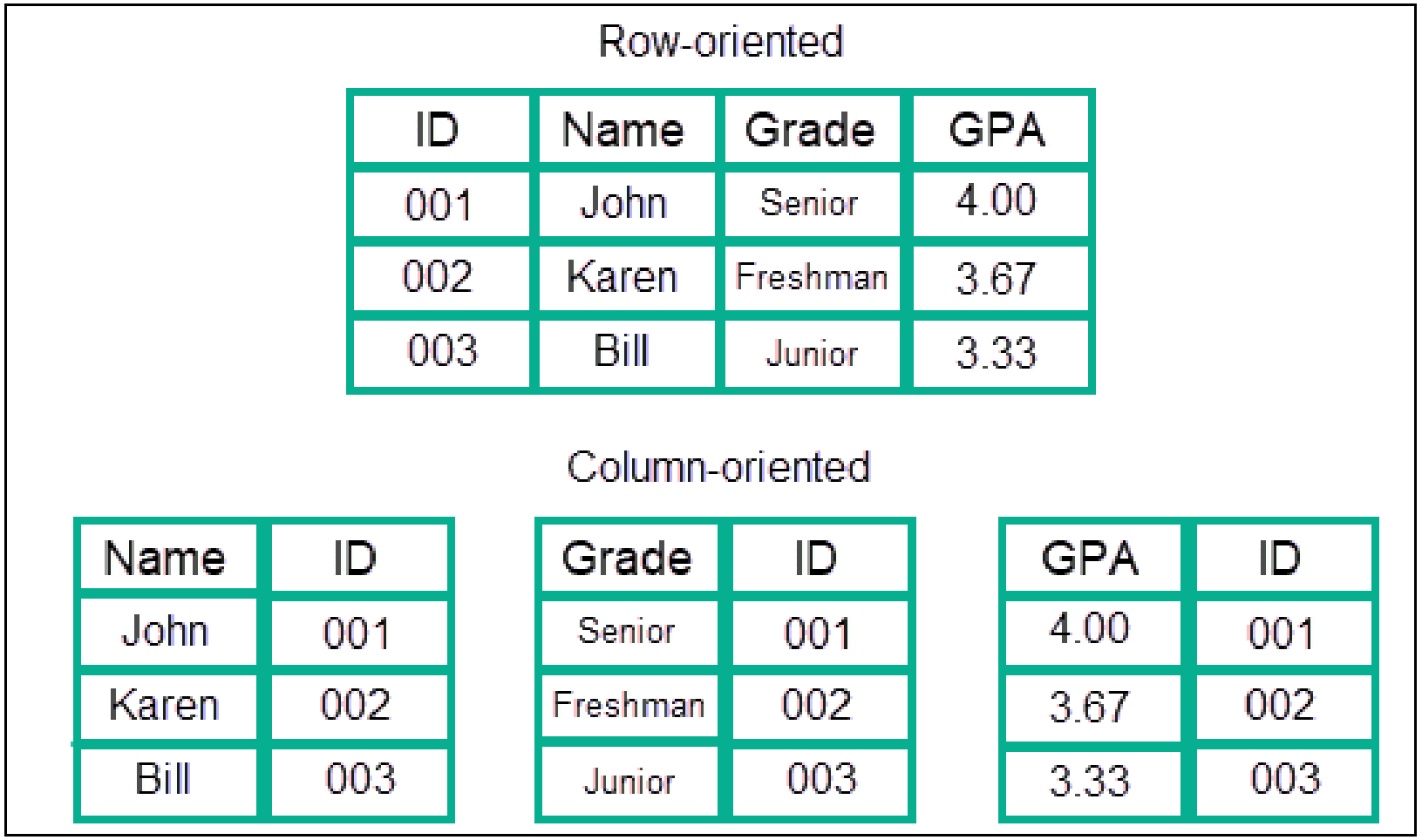
NoSQL – Chave/Valor - uso:



NoSQL – Orientados a Colunas

- Enquanto um banco de dados relacional é otimizado para armazenar linhas de dados, geralmente para aplicativos transacionais, um banco de dados colunar é otimizado para recuperação rápida de colunas de dados, normalmente em aplicativos analíticos
- Armazena e processa grandes quantidades de dados distribuídos em muitas máquinas.
- Reduz expressivamente os requisitos gerais de E/S de disco e diminui a quantidade de dados que você precisa carregar do disco.

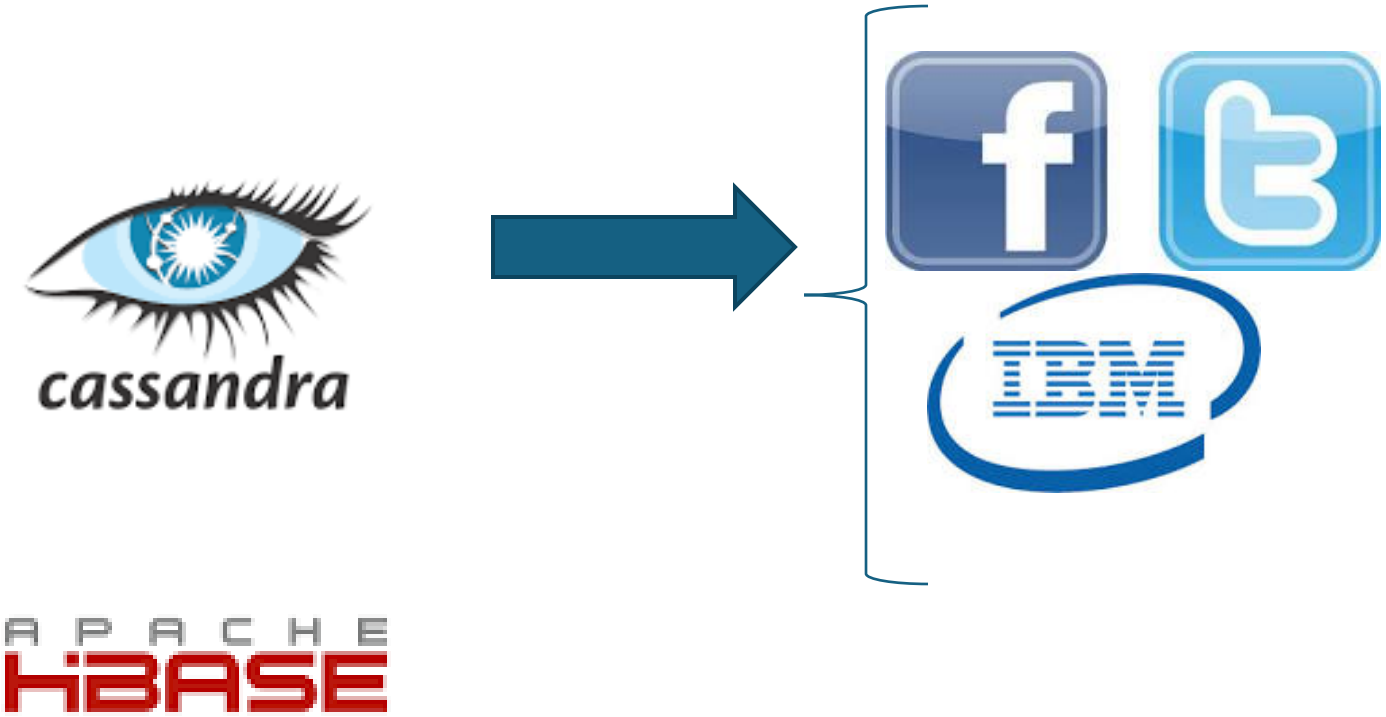
NoSQL – Orientados a Colunas



NoSQL – Orientados a Colunas - Aplicação

- Aplicações distribuídas com uso intensivo de dados
- Ex. Facebook

NoSQL – Orientados a Colunas - Uso



NoSQL – Orientados a Documentos

- Um documento é:
- -um objeto
- -tem um identificador único
- - tem um conjunto de campos, que podem ser strings, listas ou documentos aninhados

- Em cada documento temos um conjunto de campos (chaves) e o valor deste campo

NoSQL – Orientados a Documentos

Documento JSON

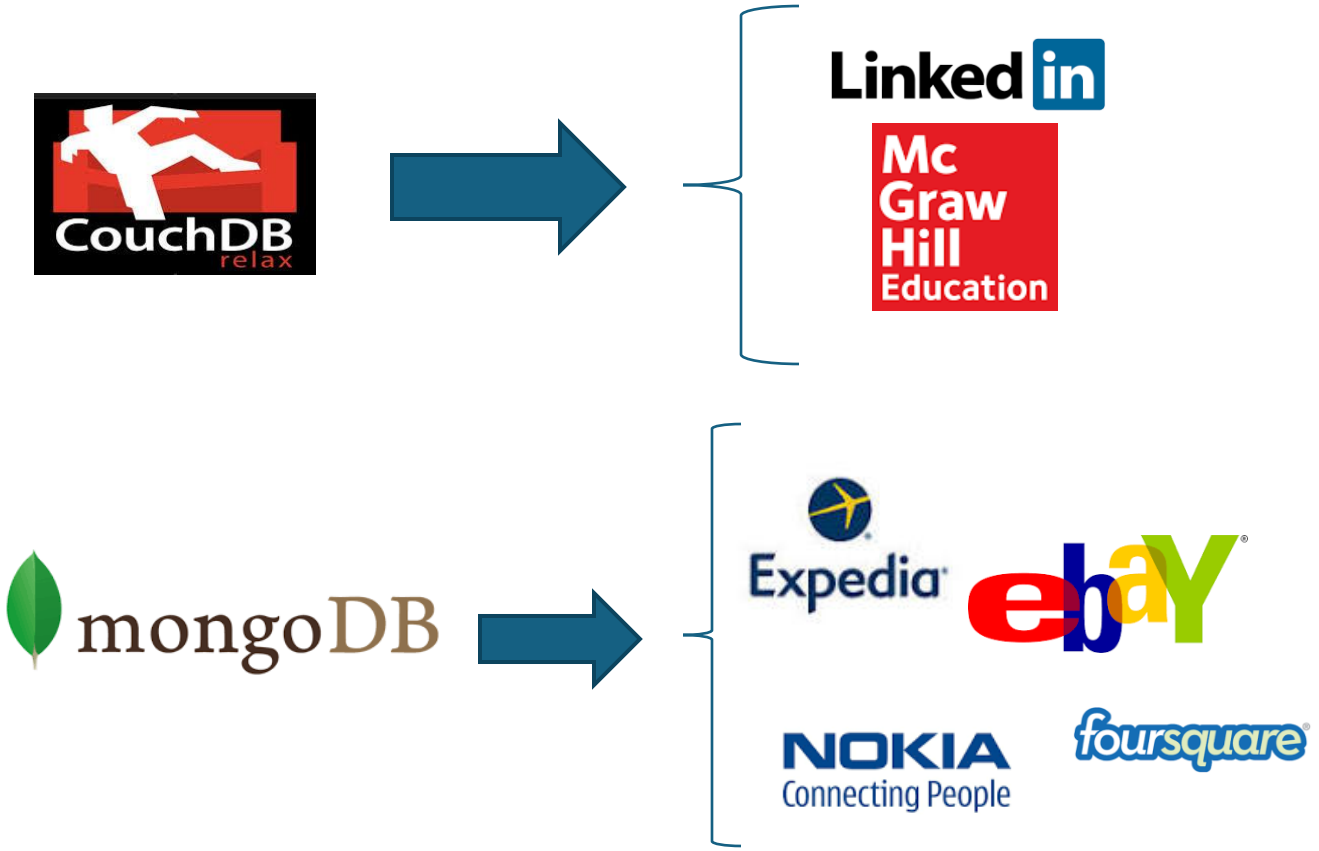
{ “screen name”: “@infomagazine”
"text": “The 8 most stressful jobs in tech. Is yours one of them?”
"url": "https://t.co/qOfPJJrfcw",
“screen name”: @infomagazine
"lang": “En",
"created_at": ”Tue Feb 05 02:31:38 +0000 2019",
"place": “United States“, …
}
{ “Nome”: Pedro,
“Endereco”: { “rua”: “AVenida Canal”, “numero”: 290 }
“Idade”:20
}

NoSQL – Orientados a Documentos – Aplicação

- Aplicações Web
- (Similar ao armazenamento chave-valor)

- Tolerante a dados incompletos

NoSQL – Orientados a Documentos - Uso



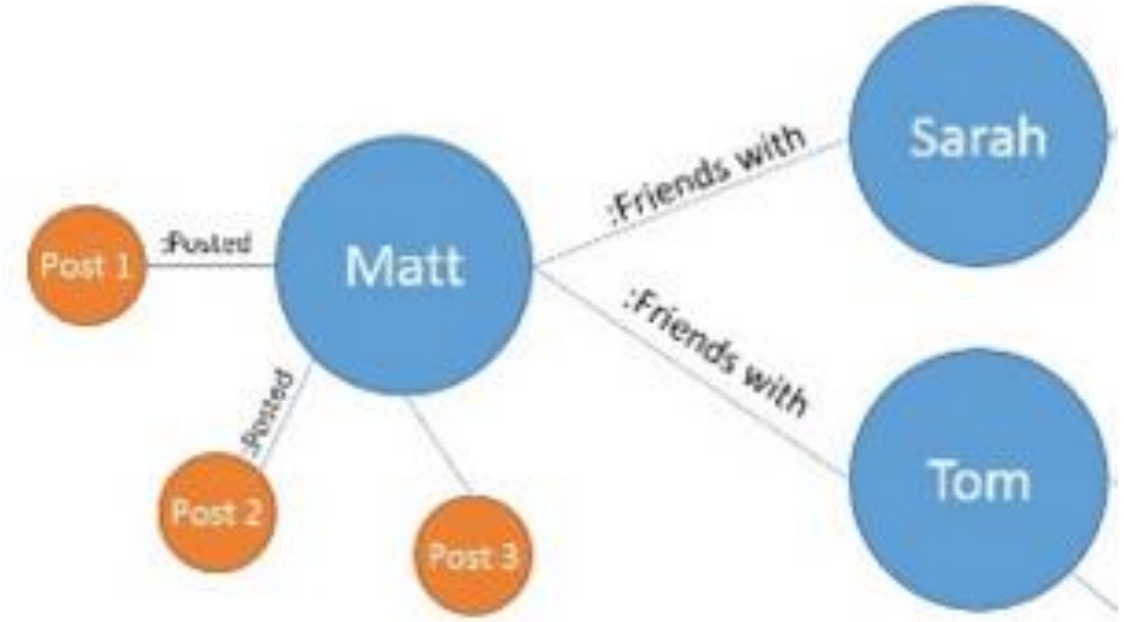
NoSQL – Grafos

Possuem:

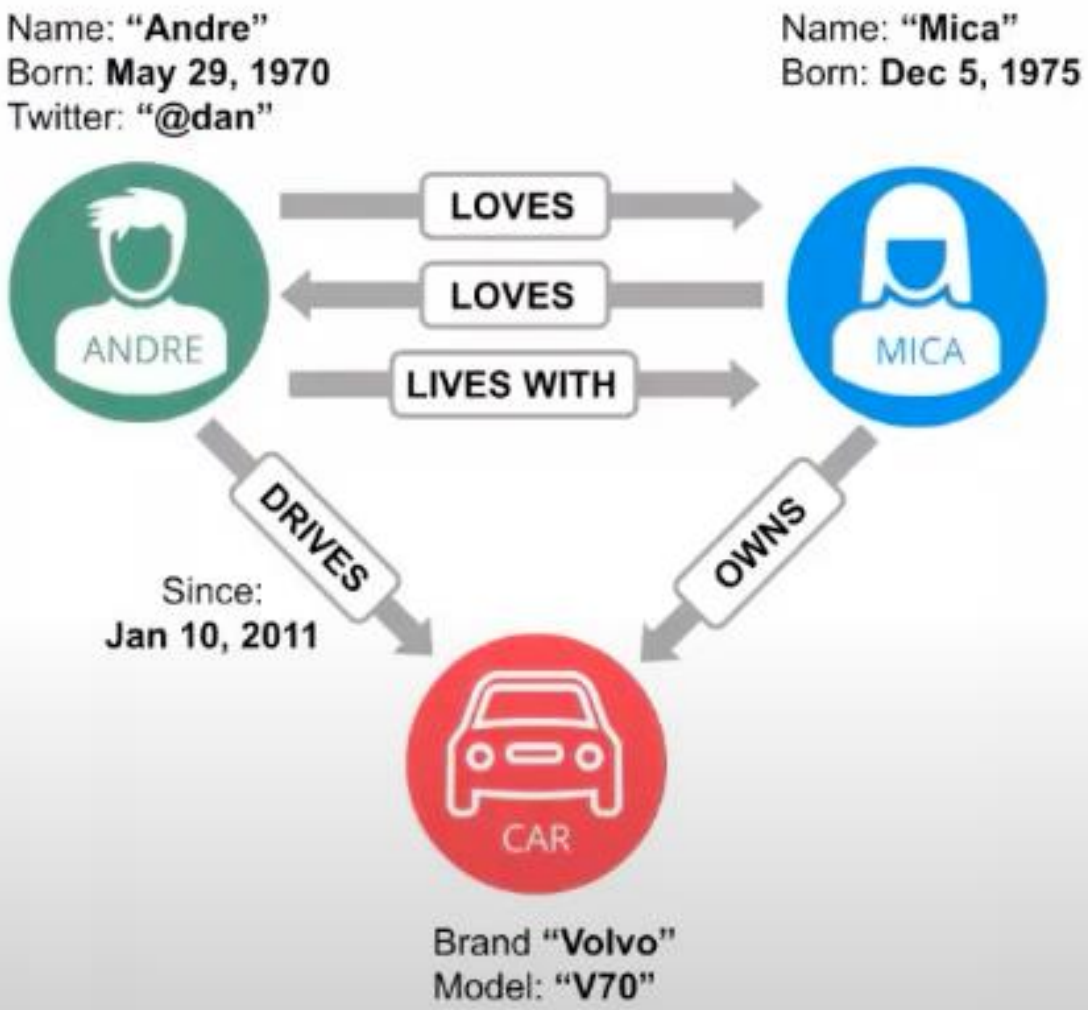
os nós (são os vértices do grafo)

os relacionamentos (são as arestas)

e as propriedades (ou atributos) dos nós e relacionamentos



What's a graph?



Node

Represents an entity in the graph

Relationship

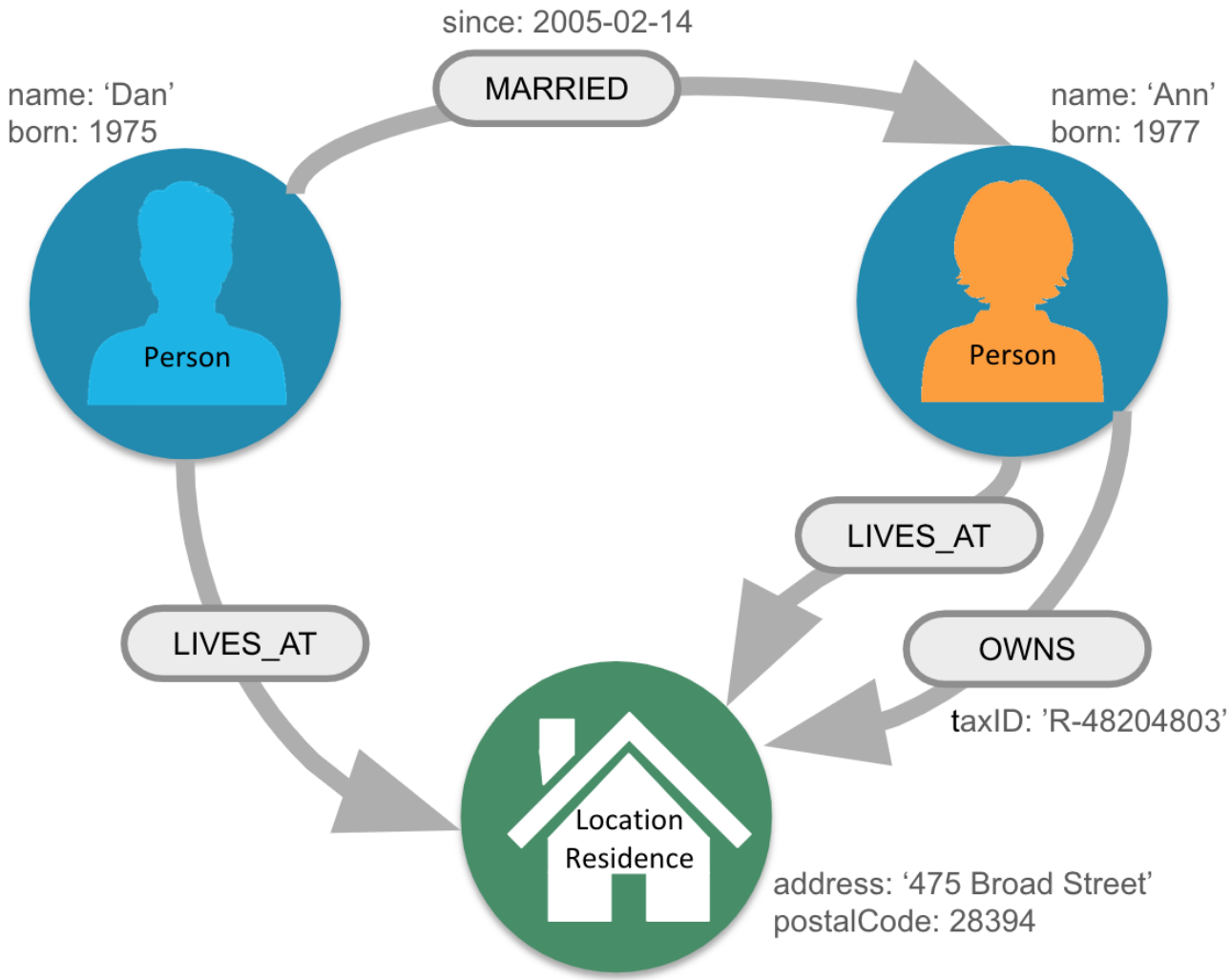
Connect nodes to each other

Property

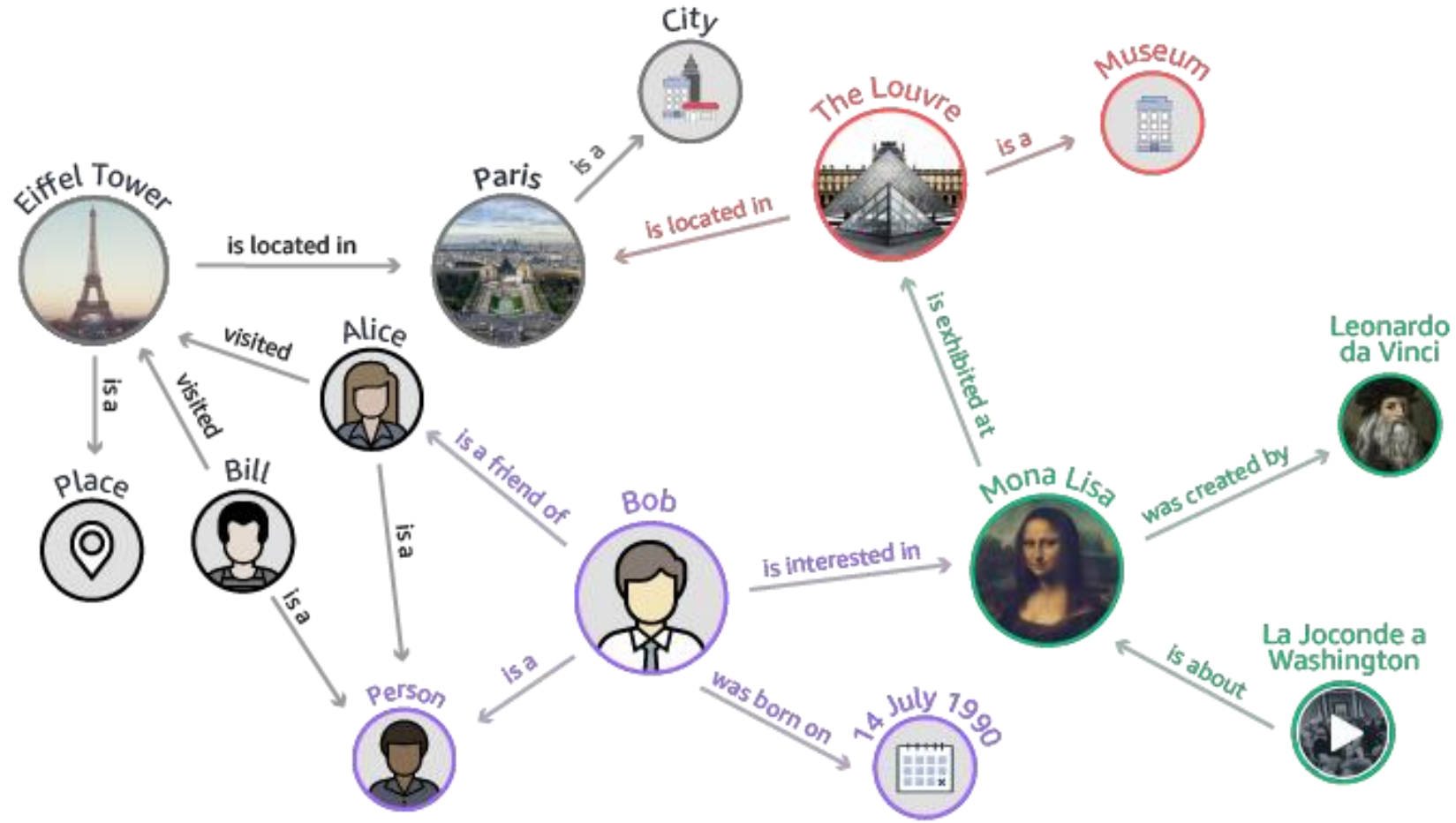
Describes a node or relationship:
e.g. name, age, weight etc

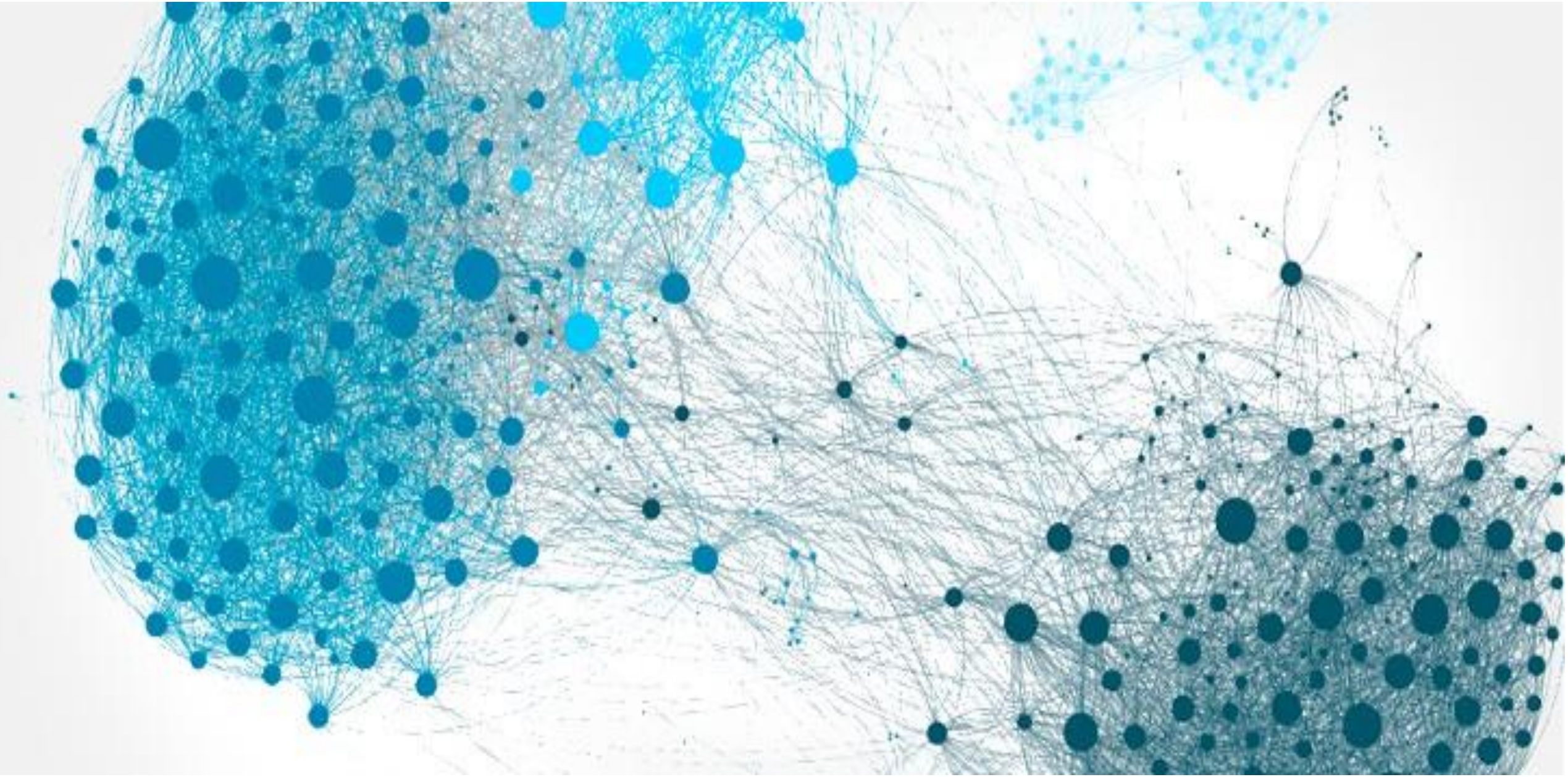
neo4j

NoSQL – Grafos



NoSQL – Grafos





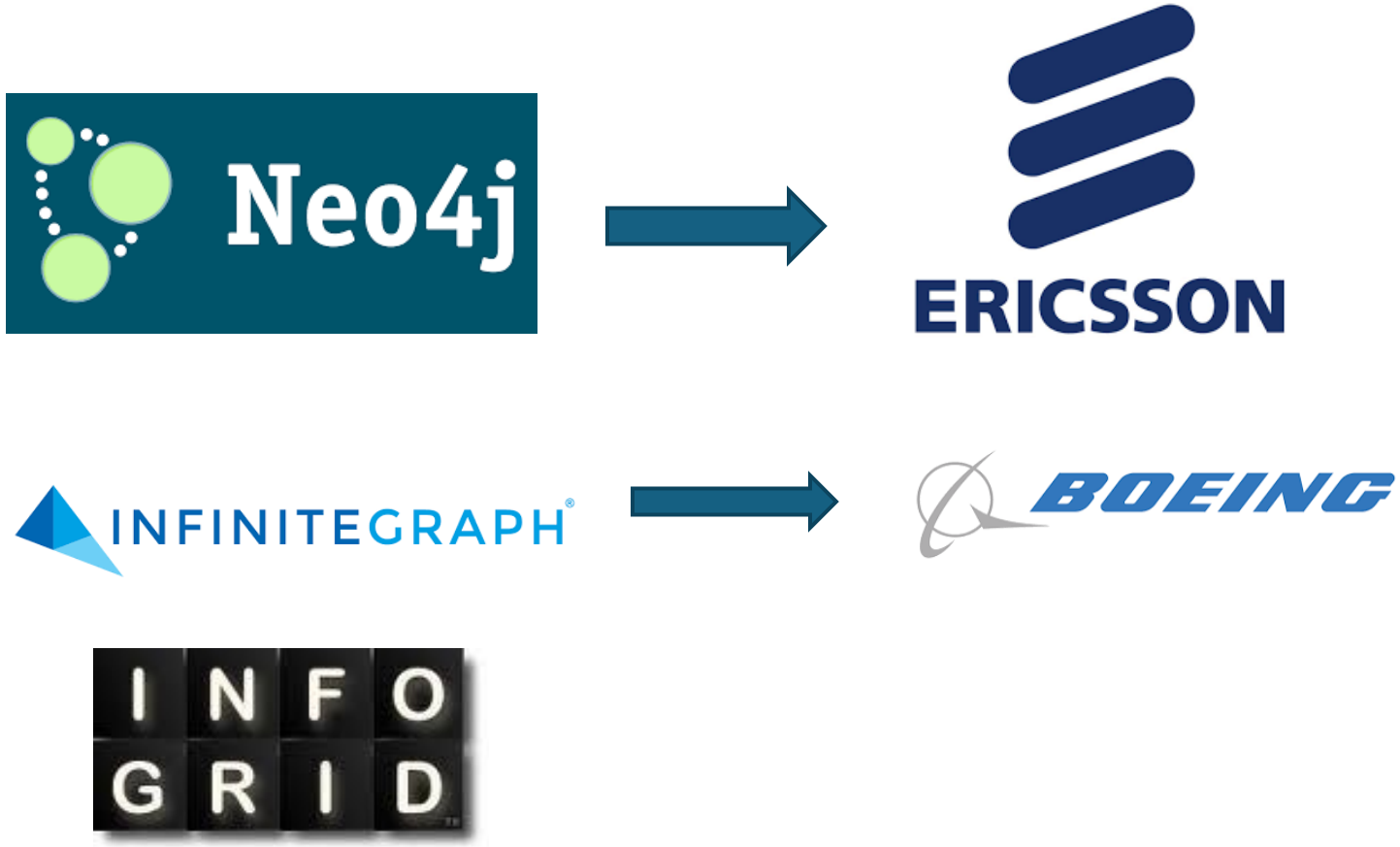
NoSQL – Grafos - Aplicação

- Manipulação de dados estatísticos, frequentemente escritos mas raramente lidos
- (como um contador de hits na Web)

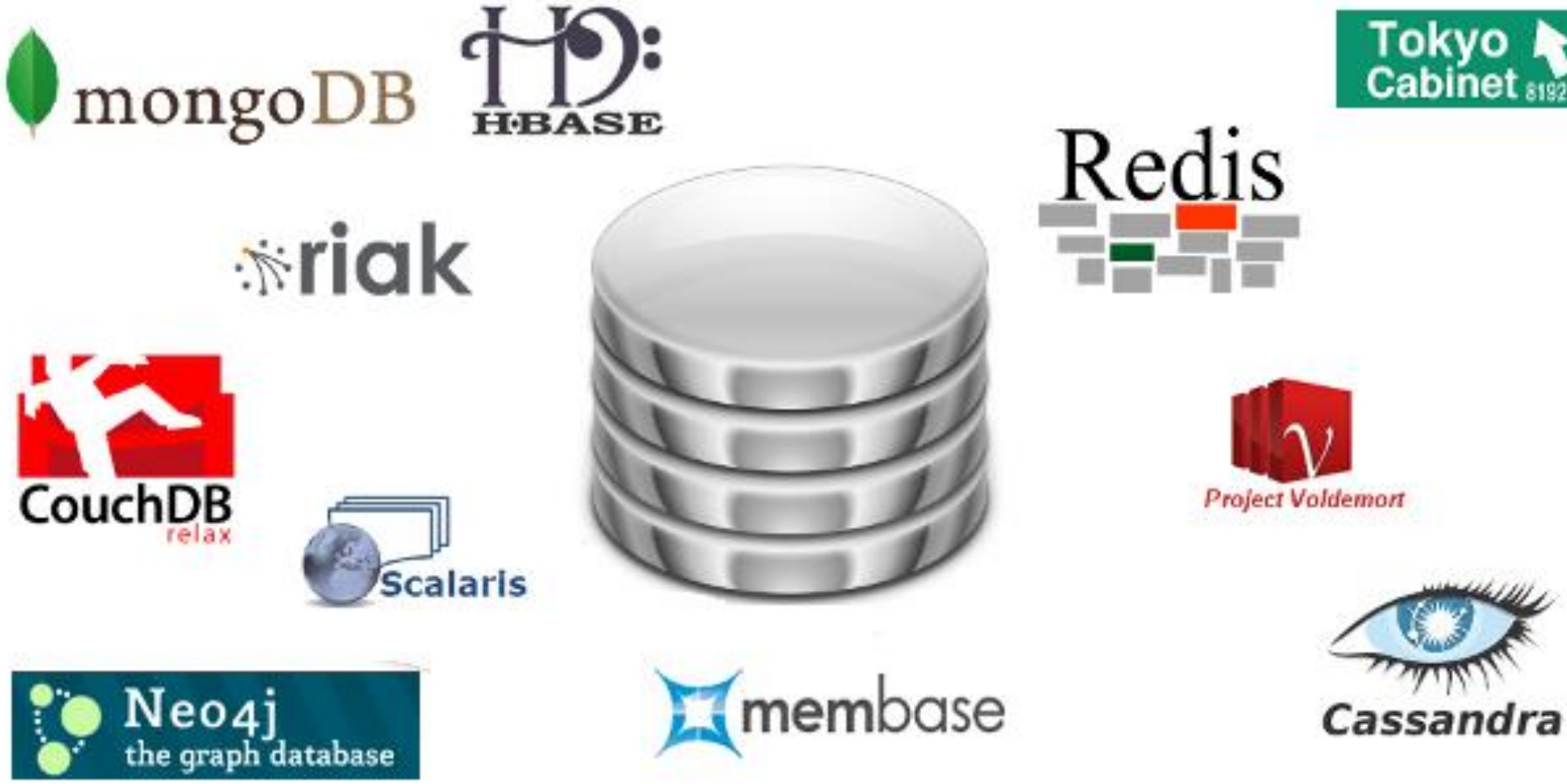
- Aplicações que exigem alto desempenho em consultas com muitas junções (Tabelas)

- Redes Sociais, Recomendações (Foco em modelar a estrutura dos dados – interconectividade)

NoSQL – Grafos - Uso



SGBD’s NoSQL


mongoDB
HBASE
Tokyo Cabinet
Redis
riak
CouchDB
relax
Scalaris
Project Voldemort
Neo4j
the graph database
membase
Cassandra

50 GB de dados para armazenar

- Mysql (Relacional)
- 300 ms – Write
- 350 ms – Read

- Cassandra (NoSQL)
- 0.12 ms – Write
- 15 ms – Read

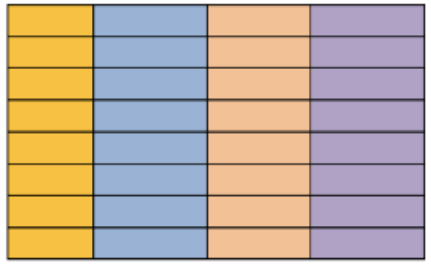
- Diferença:
  - Write – 2.500 vezes mais rápido o NoSQL
  - Read – 23 vezes mais rápido o NoSQL

Quando Usar?

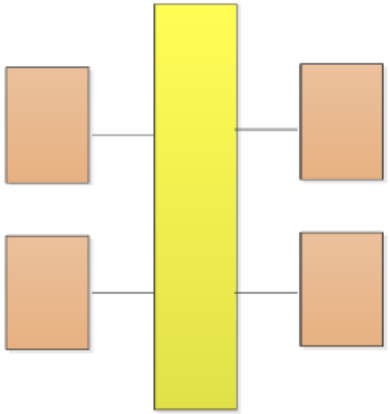
SGDB-Relacional	**NoSQL**
	O armazenamento deve ser capaz de lidar com **carregamentos pesados**
Armazenamento é esperado para apresentar carregamento pesado também, mas consiste principalmente na **leitura** de operações	Você executa muitas operações de **escrita** no armazenamento
Você **prefere performance** a uma estrutura de dados sofisticada	Você quer um armazenamento que seja **escalável horizontalmente**
Você precisa de uma linguagem **SQL query poderosa**	Simplicidade é bom, como em uma **linguagem query bem simples** (sem joins)

SQL Databases

Relational

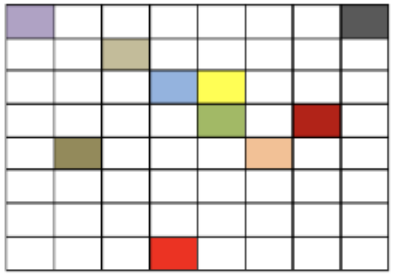


Analytical (OLAP)

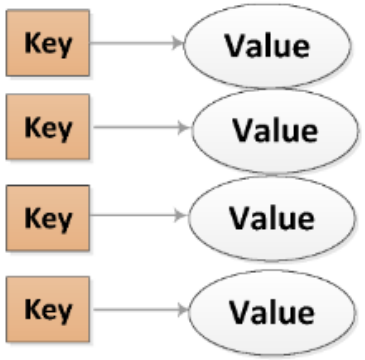


NoSQL Databases

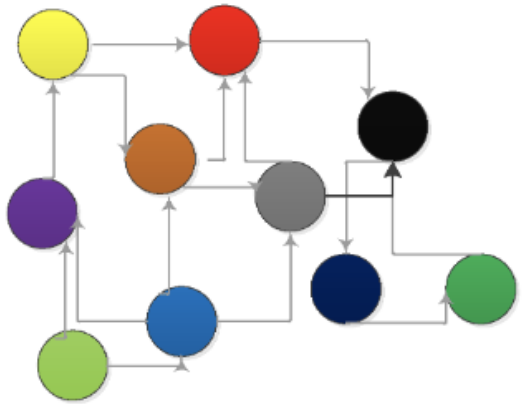
Column Family



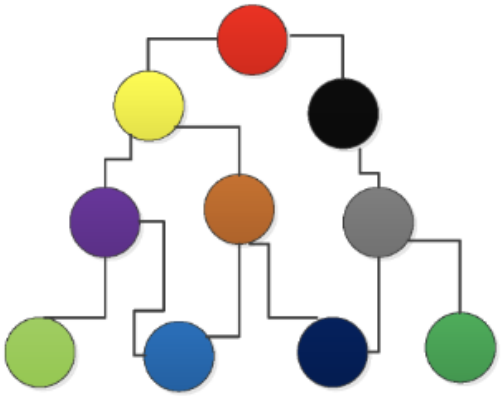
Key-Value



Graph



Document

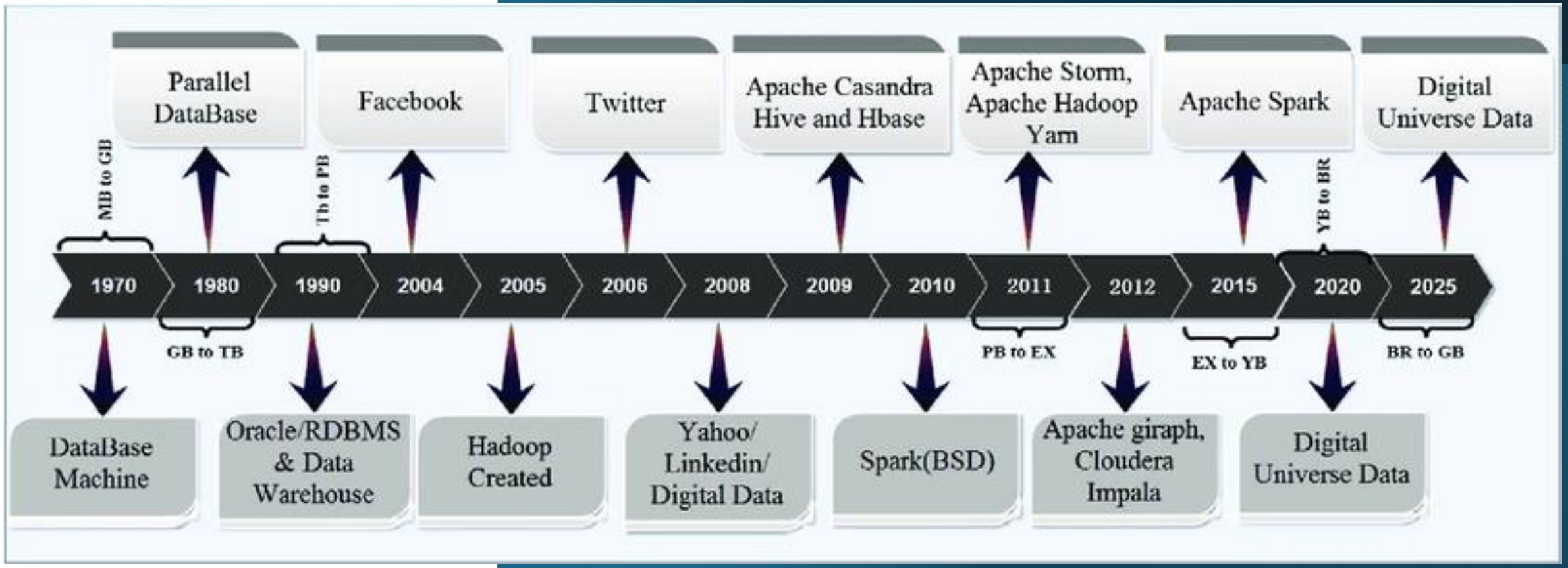


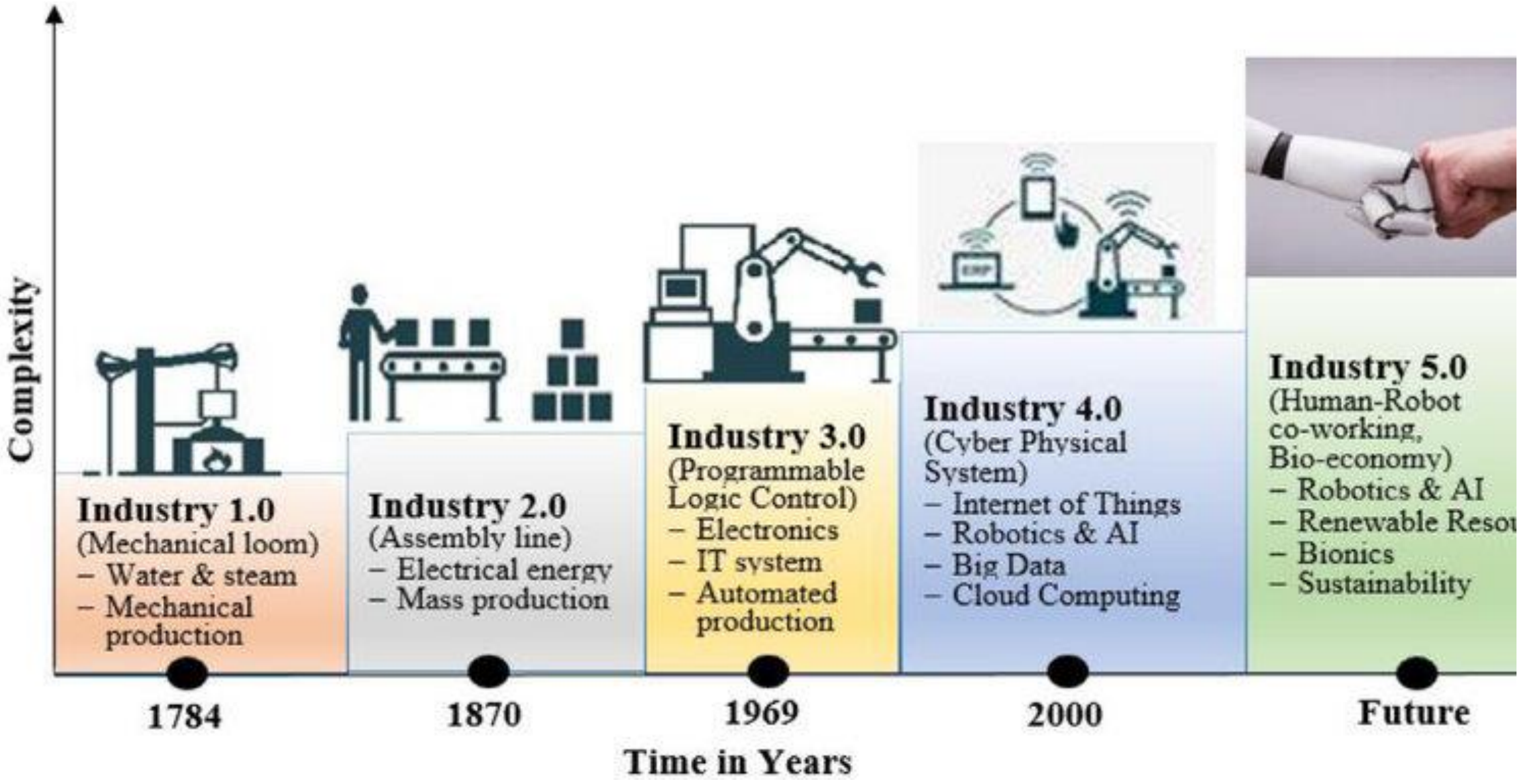
423 systems in ranking, September 2024

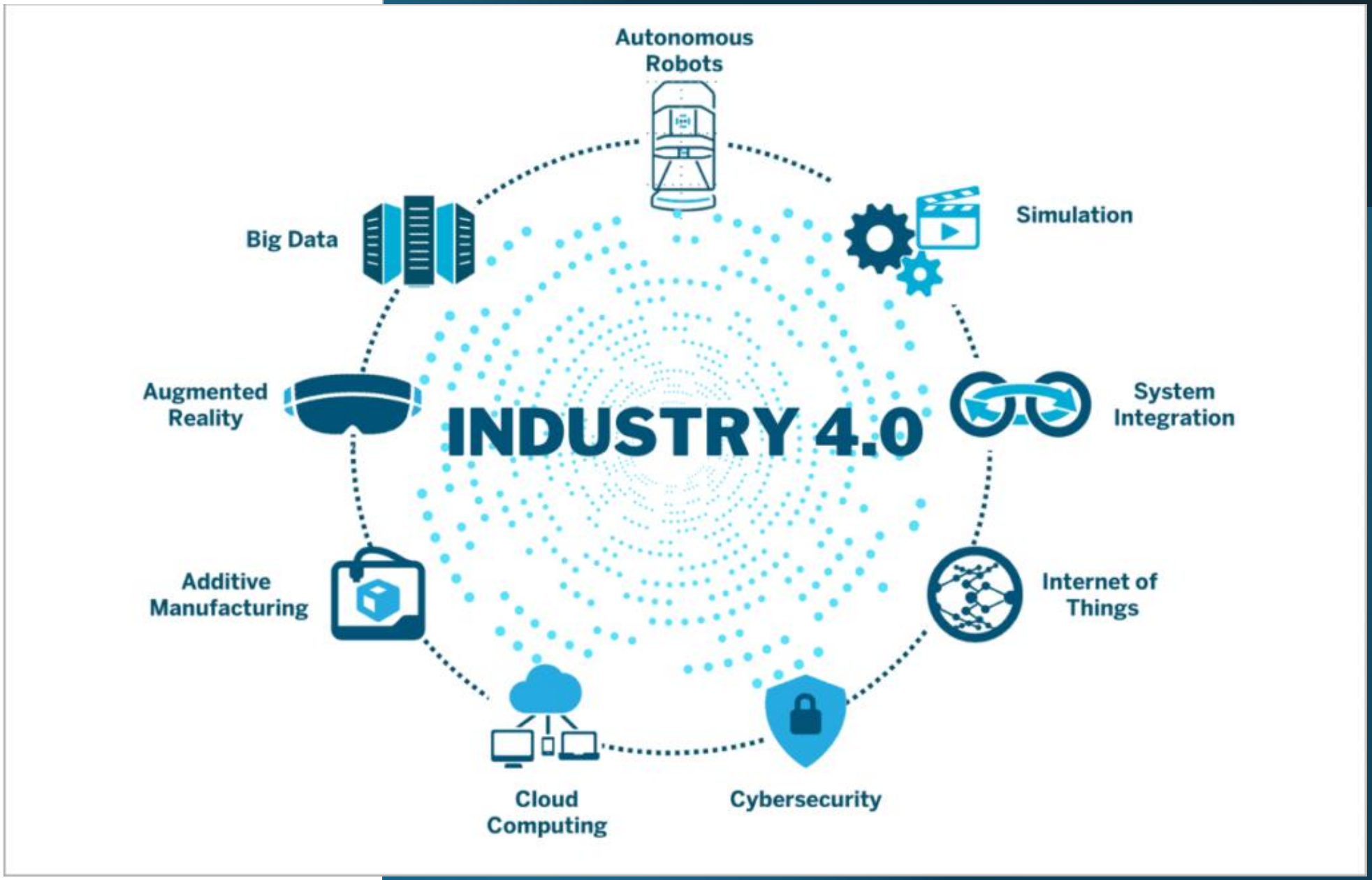
Rank			DBMS	Database Model	Score		
Sep 2024	Aug 2024	Sep 2023			Sep 2024	Aug 2024	Sep 2023
1.	1.	1.	Oracle	Relational, Multi-model	1286.59	+28.11	+45.72
2.	2.	2.	MySQL	Relational, Multi-model	1029.49	+2.63	-82.00
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	807.76	-7.41	-94.45
4.	4.	4.	PostgreSQL	Relational, Multi-model	644.36	+6.97	+23.61
5.	5.	5.	MongoDB	Document, Multi-model	410.24	-10.74	-29.18
6.	6.	6.	Redis	Key-value, Multi-model	149.43	-3.28	-14.26
7.	7.	↑ 11.	Snowflake	Relational	133.72	-2.25	+12.83
8.	8.	↓ 7.	Elasticsearch	Search engine, Multi-model	128.79	-1.04	-10.20
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model	123.05	+0.04	-13.67
10.	10.	↓ 9.	SQLite	Relational	103.35	-1.44	-25.85
11.	11.	↑ 12.	Apache Cassandra	Wide column, Multi-model	98.94	+1.94	-11.11
12.	12.	↓ 10.	Microsoft Access	Relational	93.76	-2.61	-34.81
13.	13.	↑ 14.	Splunk	Search engine	93.02	-3.08	+1.63
14.	↑ 15.	↑ 17.	Databricks	Multi-model	84.24	-0.22	+9.06
15.	↓ 14.	↓ 13.	MariaDB	Relational, Multi-model	83.44	-3.09	-17.01
16.	16.	↓ 15.	Microsoft Azure SQL Database	Relational, Multi-model	72.95	-2.08	-9.78
17.	17.	↓ 16.	Amazon DynamoDB	Multi-model	70.06	+1.15	-10.85
18.	↑ 19.	18.	Apache Hive	Relational	53.07	-2.17	-18.76
19.	↓ 18.	↑ 20.	Google BigQuery	Relational	52.67	-2.86	-3.80
20.	20.	↑ 21.	FileMaker	Relational	45.20	-1.47	-8.40
21.	21.	↑ 23.	Neo4j	Graph	42.68	-1.22	-7.71

https://db-engines.com/en/ranking

Big Data, Data Warehouse (DW) e Data Lake


MB to GB
GB to TB
Tb to PB
PB to EX
EX to YB
YB to BR
BR to GB
1970
1980
1990
2004
2005
2006
2008
2009
2010
2011
2012
2015
2020
2025
DataBase Machine
Parallel DataBase
Oracle/RDBMS & Data Warehouse
Facebook
Hadoop Created
Twitter
Yahoo/ Linkedin/ Digital Data
Apache Casandra Hive and Hbase
Spark(BSD)
Apache Storm, Apache Hadoop Yarn
Apache giraph, Cloudera Impala
Apache Spark
Digital Universe Data
Digital Universe Data


Complexity
Industry 1.0
(Mechanical loom)
– Water & steam
– Mechanical production
Industry 2.0
(Assembly line)
– Electrical energy
– Mass production
Industry 3.0
(Programmable Logic Control)
– Electronics
– IT system
– Automated production
Industry 4.0
(Cyber Physical System)
– Internet of Things
– Robotics & AI
– Big Data
– Cloud Computing
Industry 5.0
(Human-Robot co-working, Bio-economy)
– Robotics & AI
– Renewable Resou
– Bionics
– Sustainability
1784
1870
1969
2000
Future
Time in Years


Autonomous Robots
Big Data
Simulation
Augmented Reality
INDUSTRY 4.0
System Integration
Additive Manufacturing
Internet of Things
Cloud Computing
Cybersecurity

A Revolução Industrial 4

- A Revolução Industrial 4 está chegando e Baseia-se principalmente em **Big Data**

- **Big Data** será o componente mais importante de todas as inovações que acontecerão na RI4

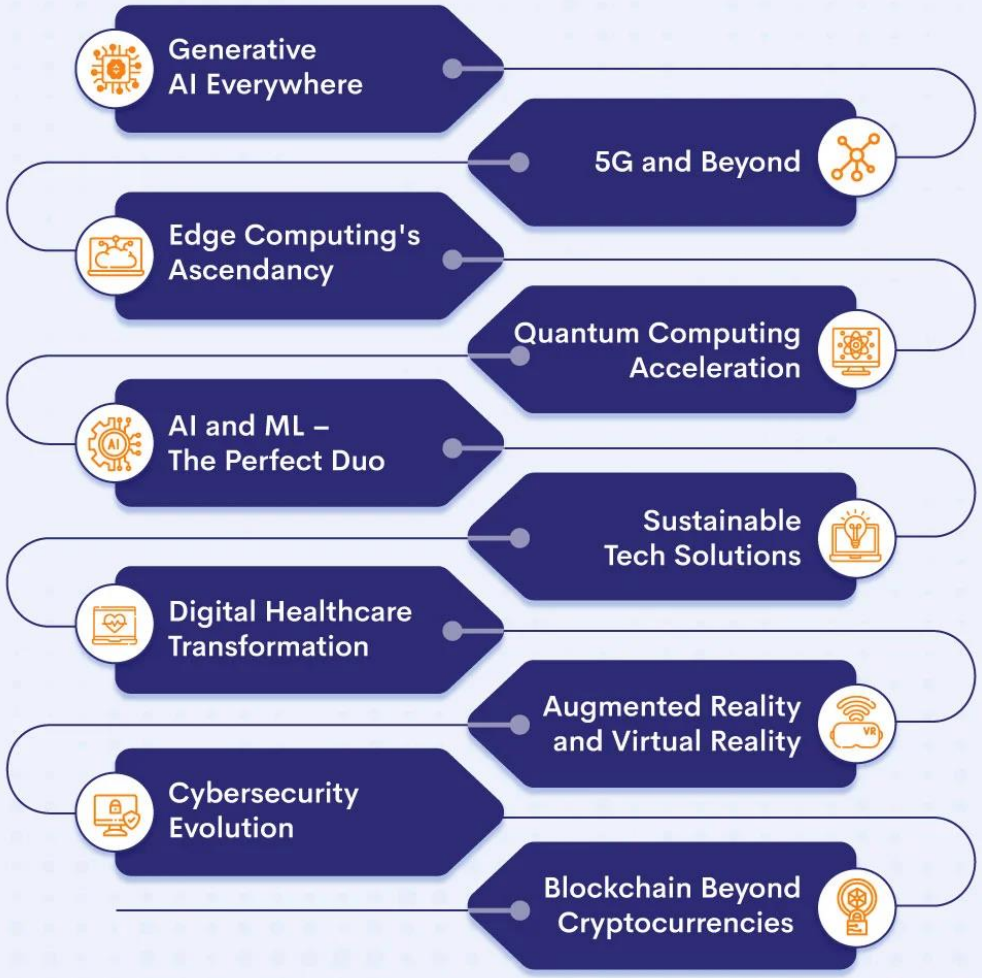
- Atualmente, estima-se que existam entre **6 a 7 milhões** de desenvolvedores, analistas, cientistas de dados, e profissionais relacionados a Big Data no mundo

- O mercado global de Big Data foi avaliado em cerca de **US$198,08** bilhões em 2023

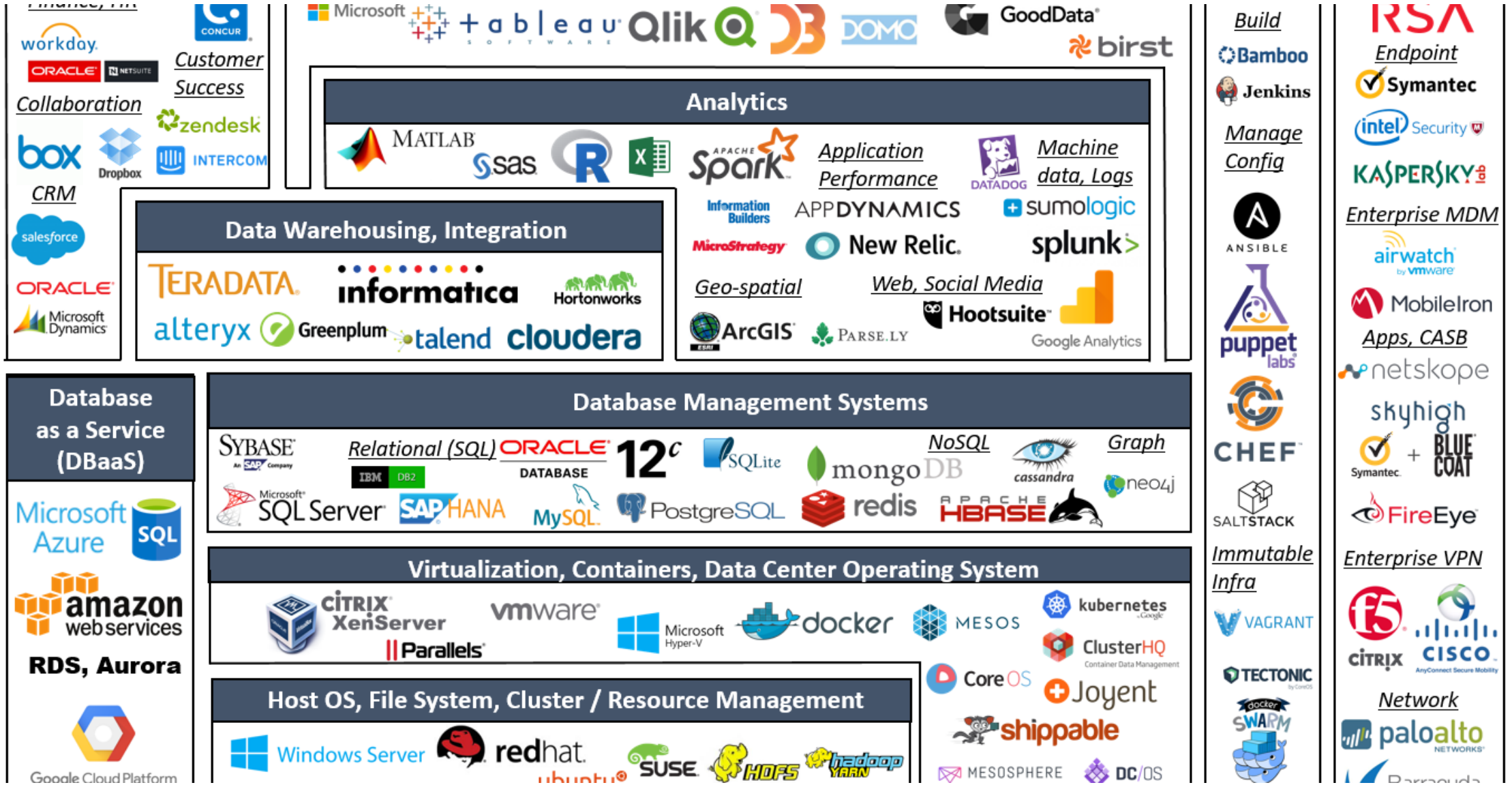
- Em 2023, a quantidade de dados gerados por todas as pessoas no mundo foi aproximadamente **120 trilhões de gigabytes (GB)** (120 Zetabytes)

- Por segundo - por pessoa na Terra gera **1,7 megabytes** de dados.

Top 10 Technology Trends to Watch in 2024

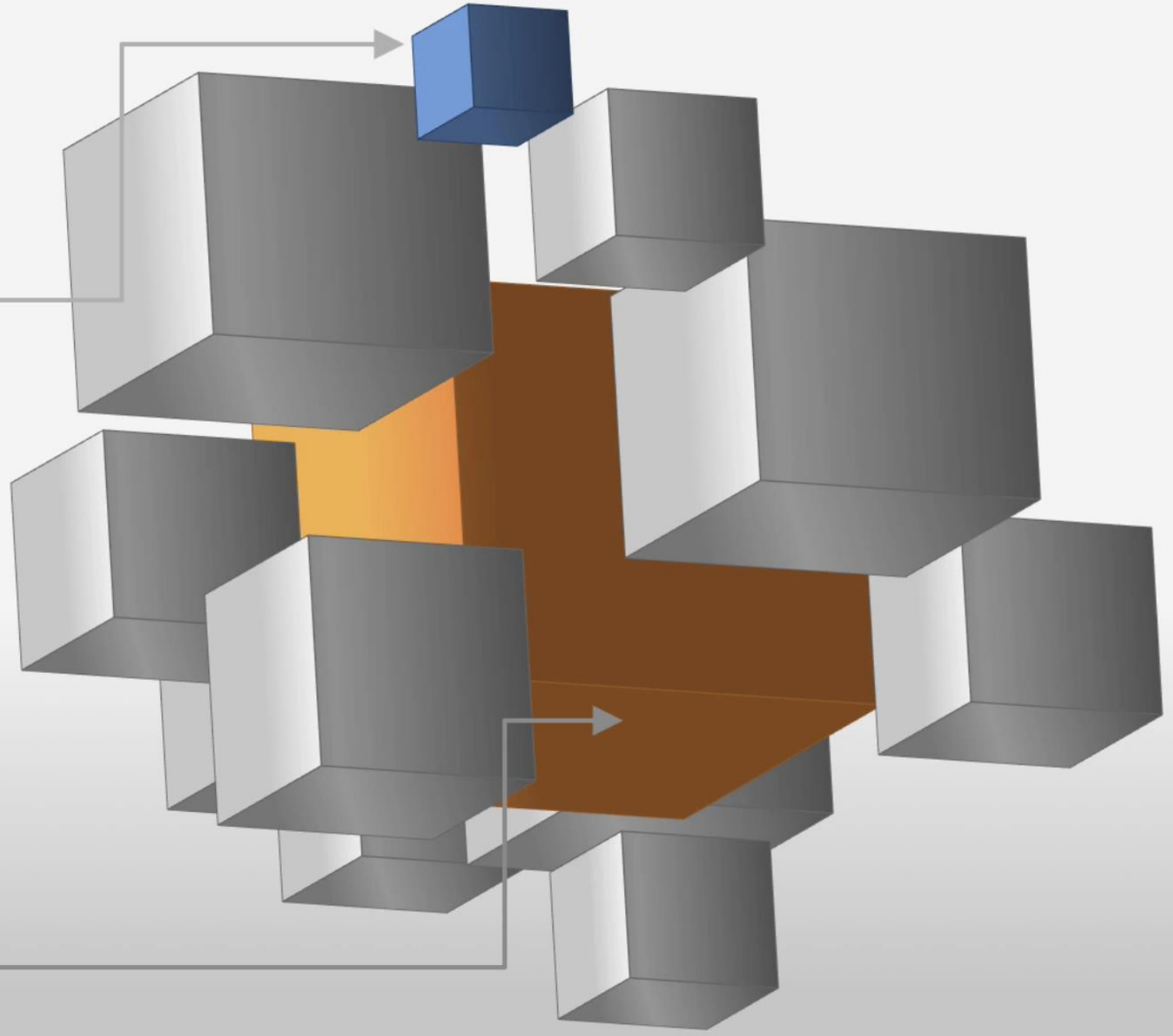


Big Data - Ecosistema



valencar@gmail.com

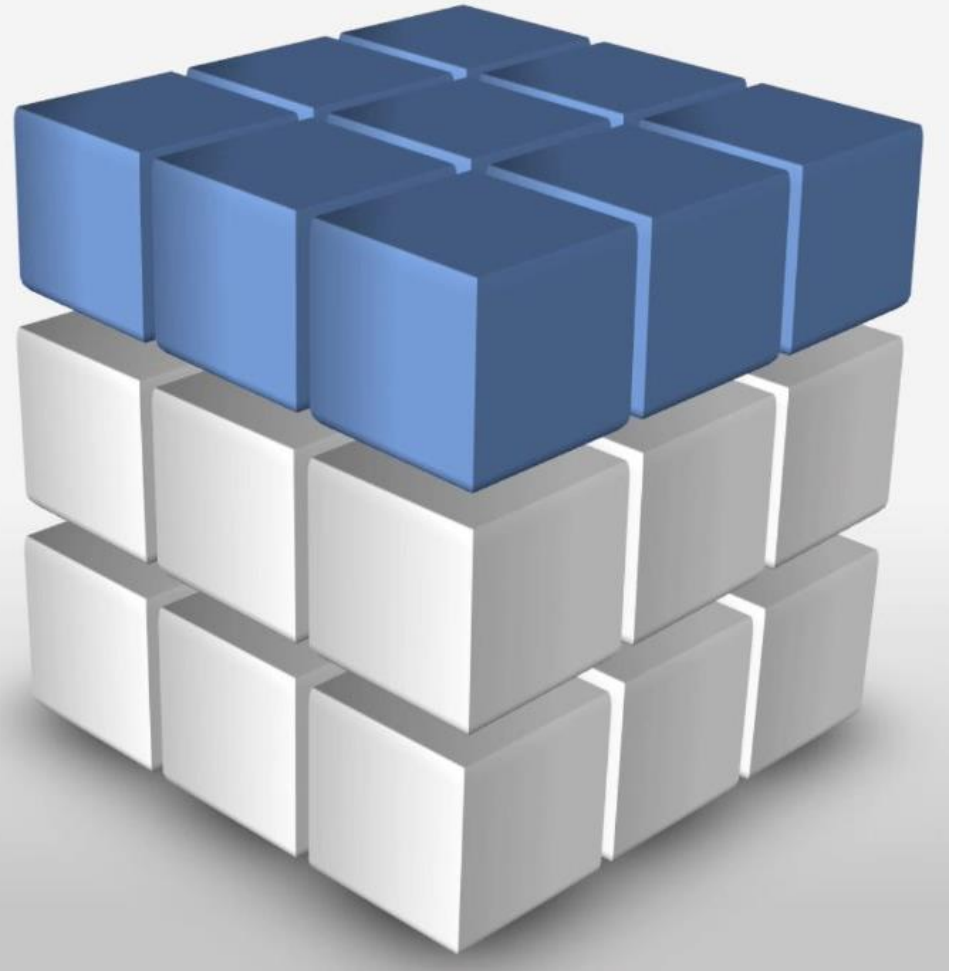
DATA SCIENCE



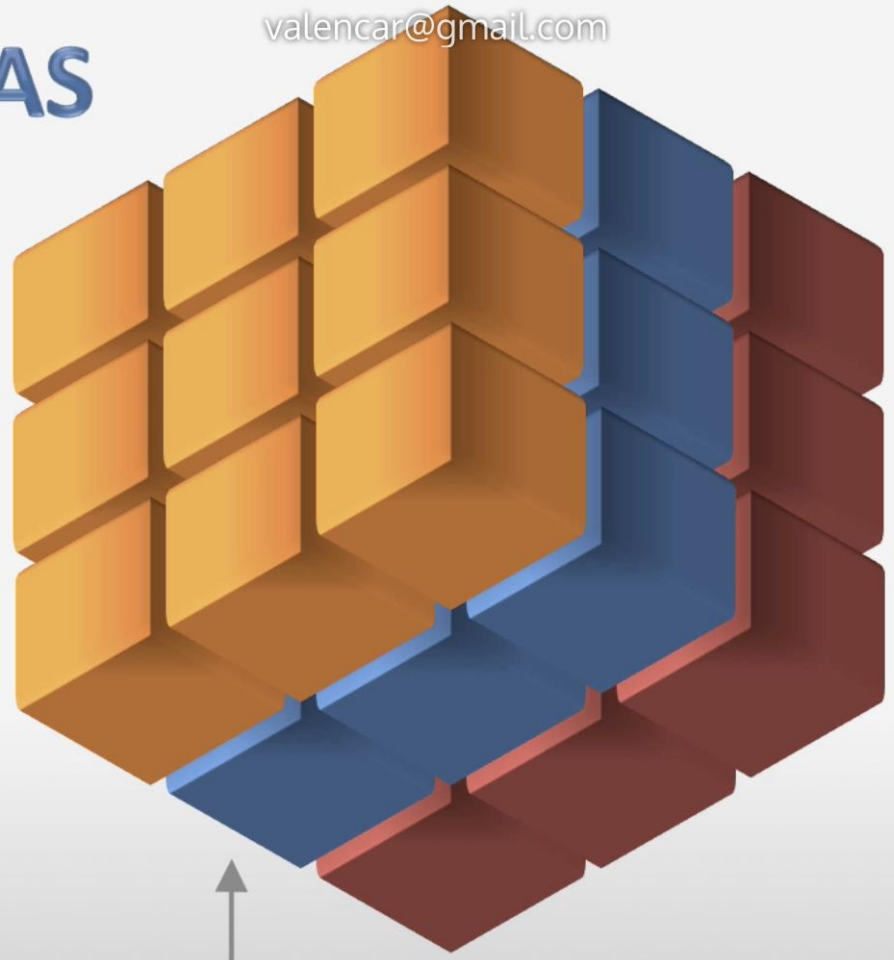
O QUE É DATA WAREHOUSE?

Um Data Warehouse é um banco de dados, usado para armazenar informações relativas às atividades de uma organização de forma consolidada.

Possibilita a análise de grandes volumes de dados, que são coletados a partir de sistemas transacionais (OLTP – Online Transaction Processing).



CARACTERÍSTICAS DO DW



Não Volátil

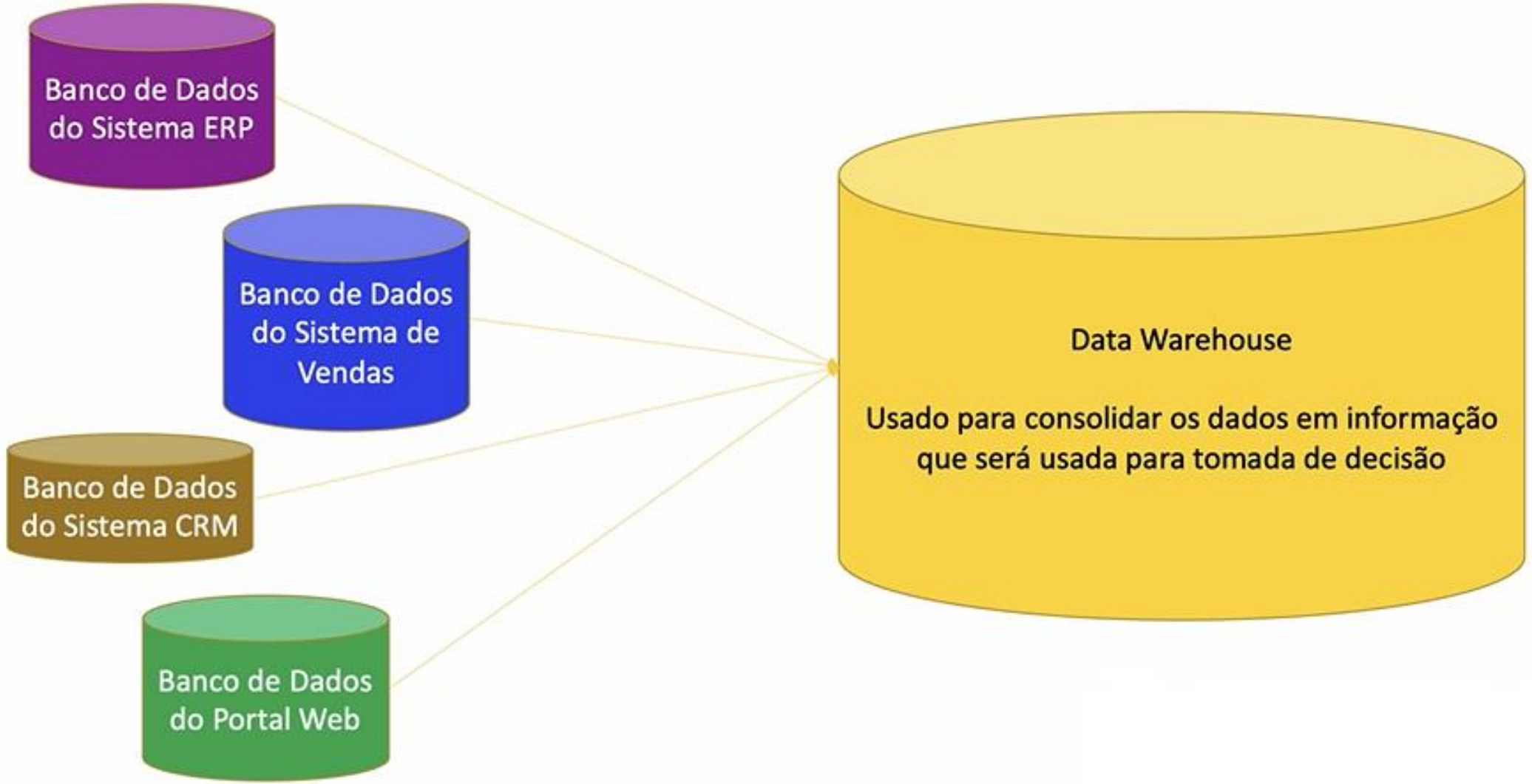
Orientado ao Negócio

Variante no Tempo


ERP
01
CONTROLE FINANCEIRO
02
CONTROLE DE ESTOQUE
03
EMISSÃO DE NOTAS FISCAIS E BOLETOS
04
PDV ONLINE
05
WEBDESK PORTAL DO CLIENTE
06
RELATÓRIOS DE PROCESSOS
07
CONCILIAÇÃO BANCÁRIA
08
CADASTRO DE CLIENTES E FUNCIONÁRIOS


Customer service
Sales
CRM
Promotion
Customers
Communication
Payment

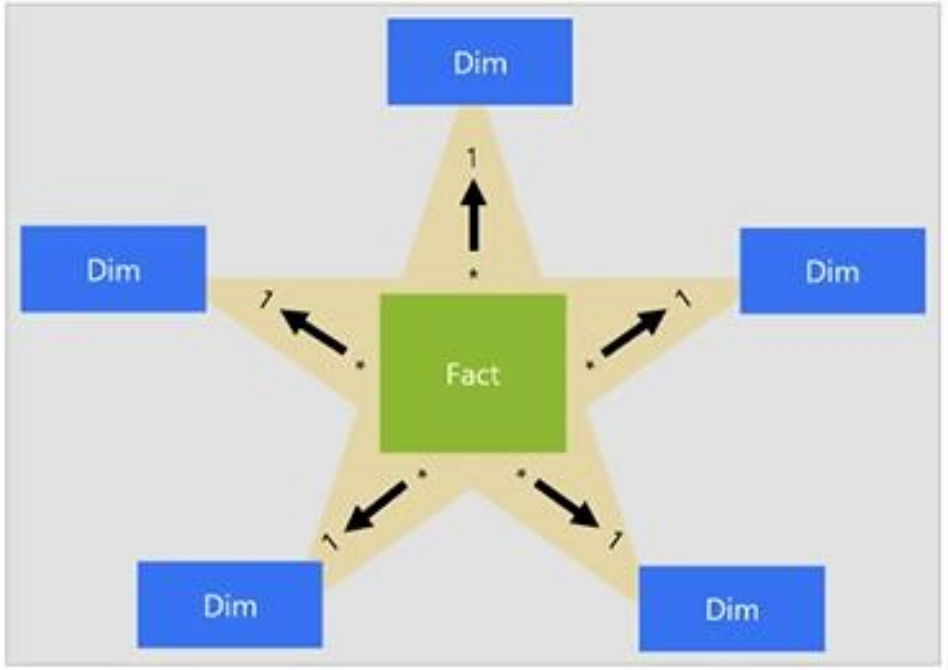
O Que é um Data Warehouse?

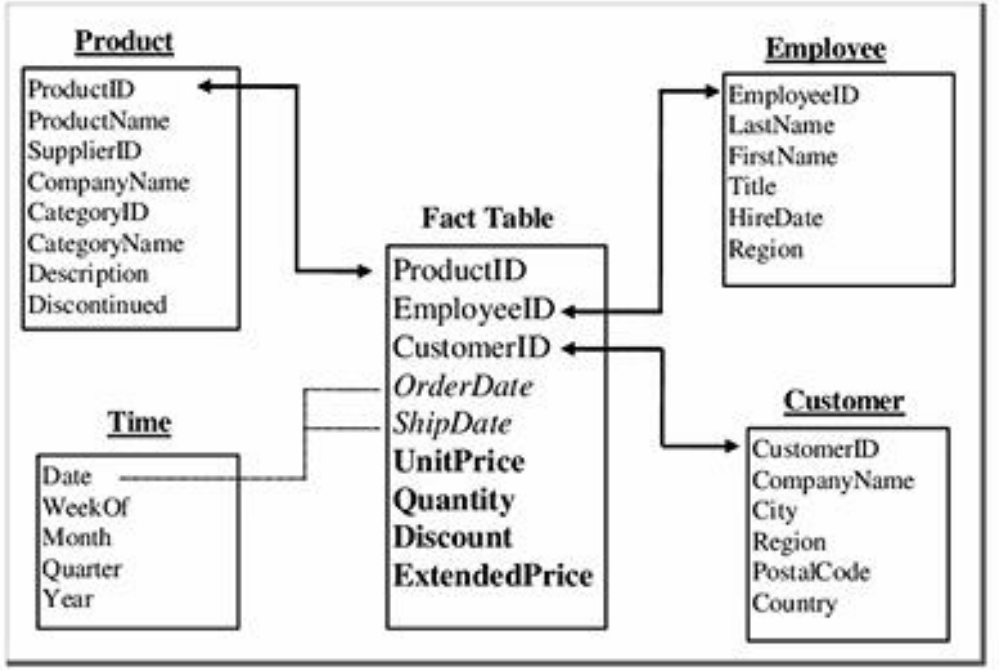


Exemplo de Modelo Físico

Compreender os tipos de dados
e relacionamentos entre as tabelas

Modelo Star Schema





Visualização de Dados, Relatórios e BI

Data Science / Machine Learning / Data Learning

Data Warehouse

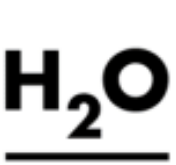
Data Lake

Data Warehouse

Visualização de Dados, Relatórios e BI

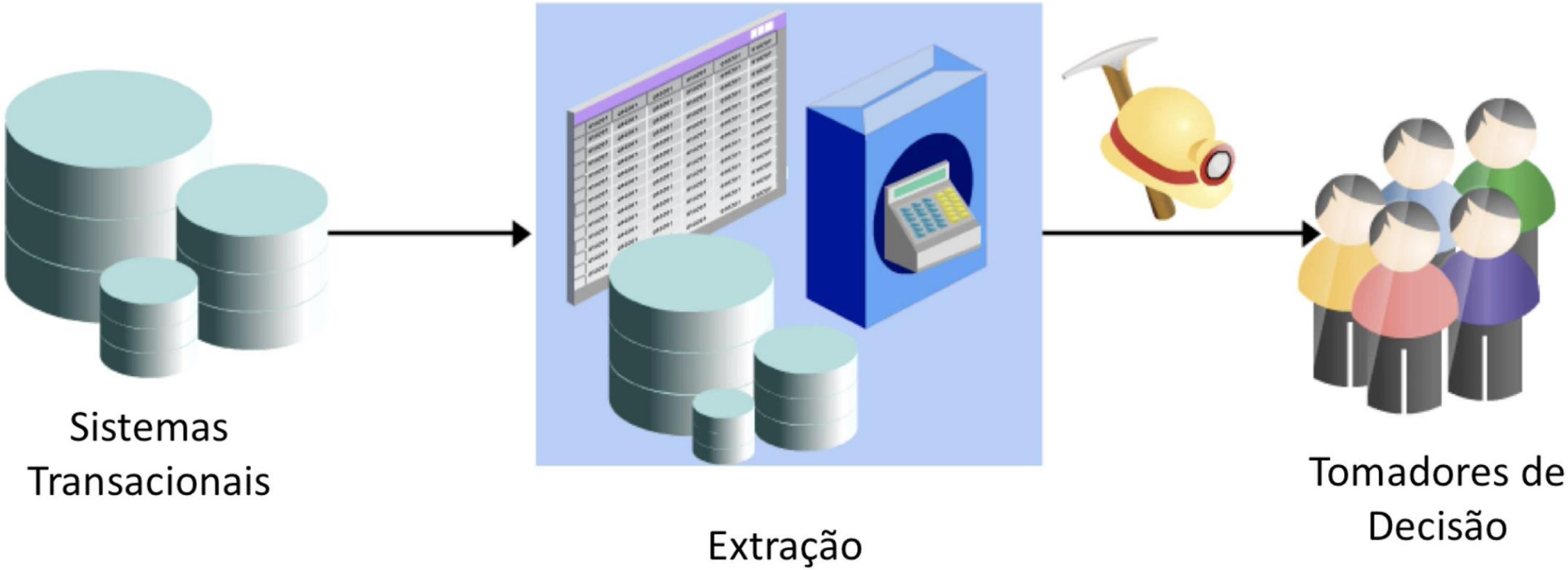
Data Lake

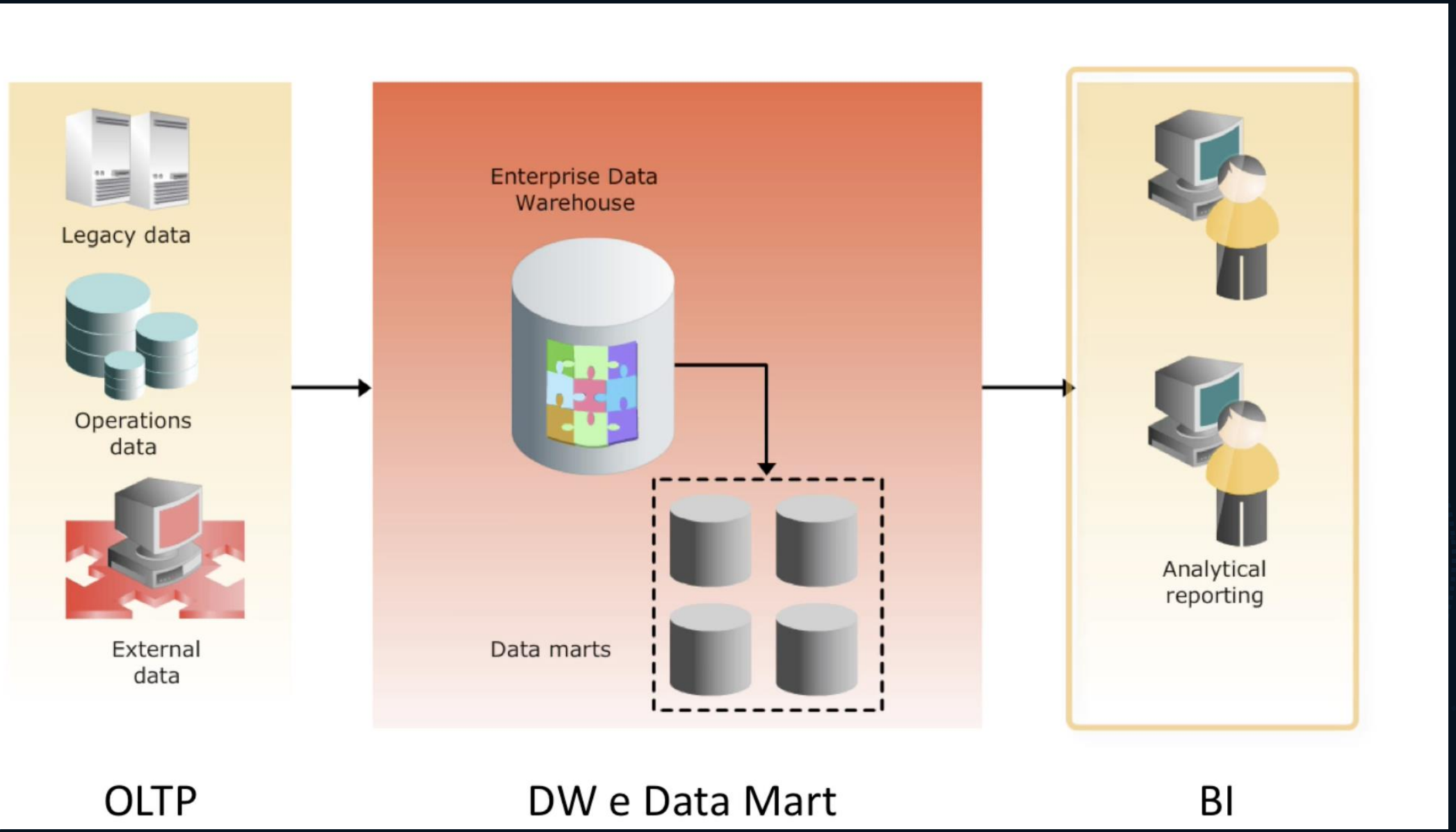
Data Science / Machine Learning / Deep Learning



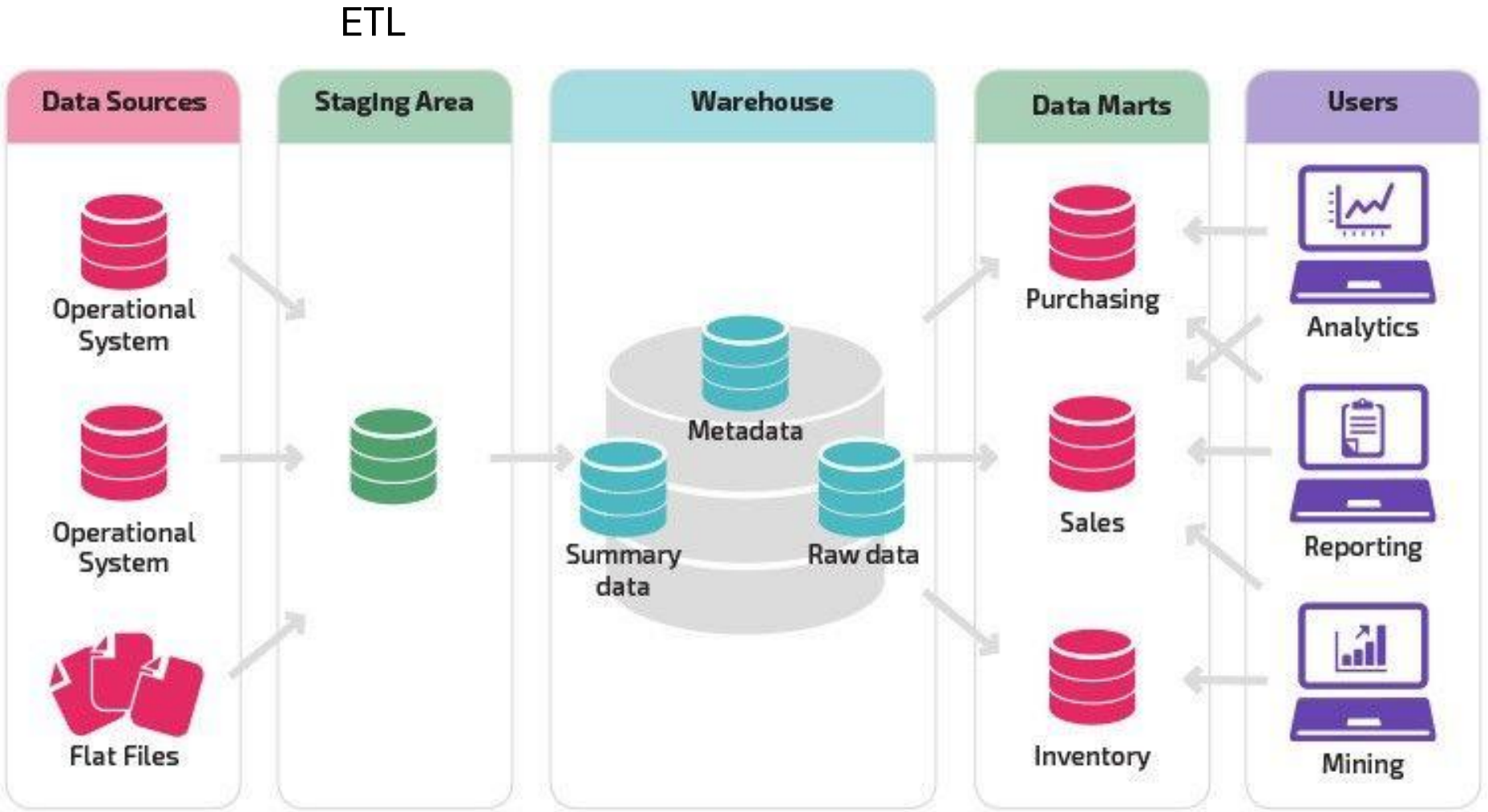
DATA WAREHOUSES, BUSINESS INTELLIGENCE, DATA MARTS E OLTP

O DW é o ponto central de uma infraestrutura de BI

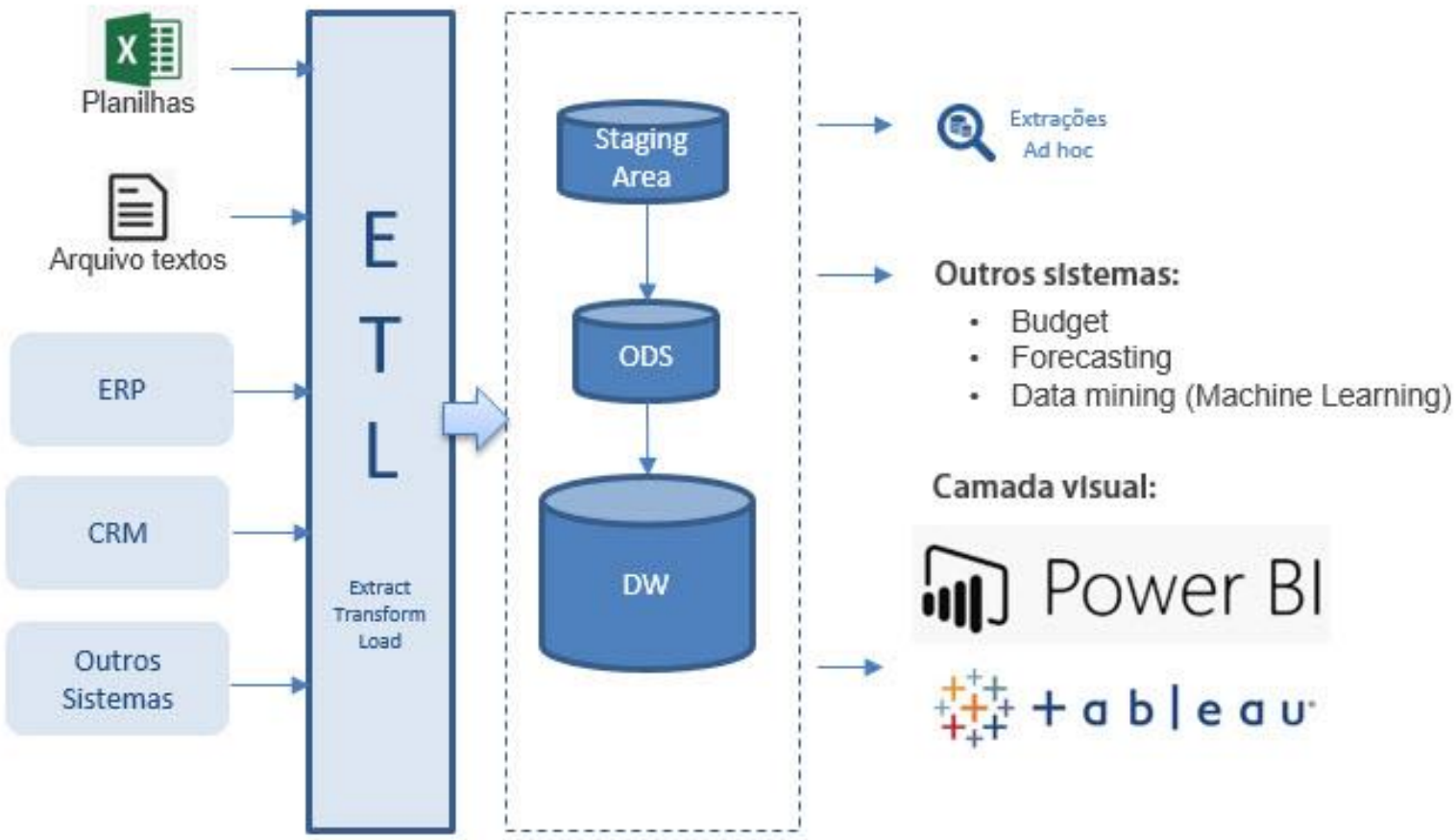

Sistemas
Transacionais
Extração
Tomadores de
Decisão


Legacy data
Operations data
External data
Enterprise Data Warehouse
Data marts
Analytical reporting
OLTP
DW e Data Mart
BI

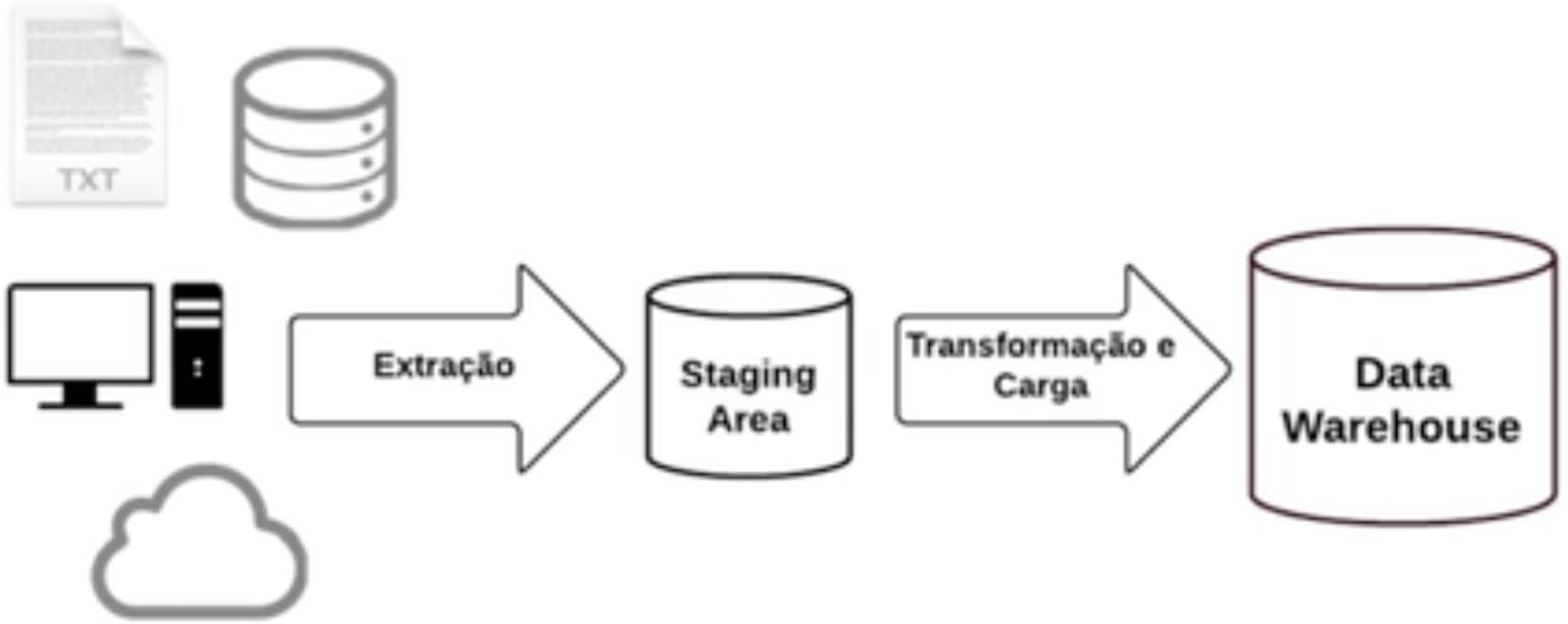
Data Warehouse (DW)



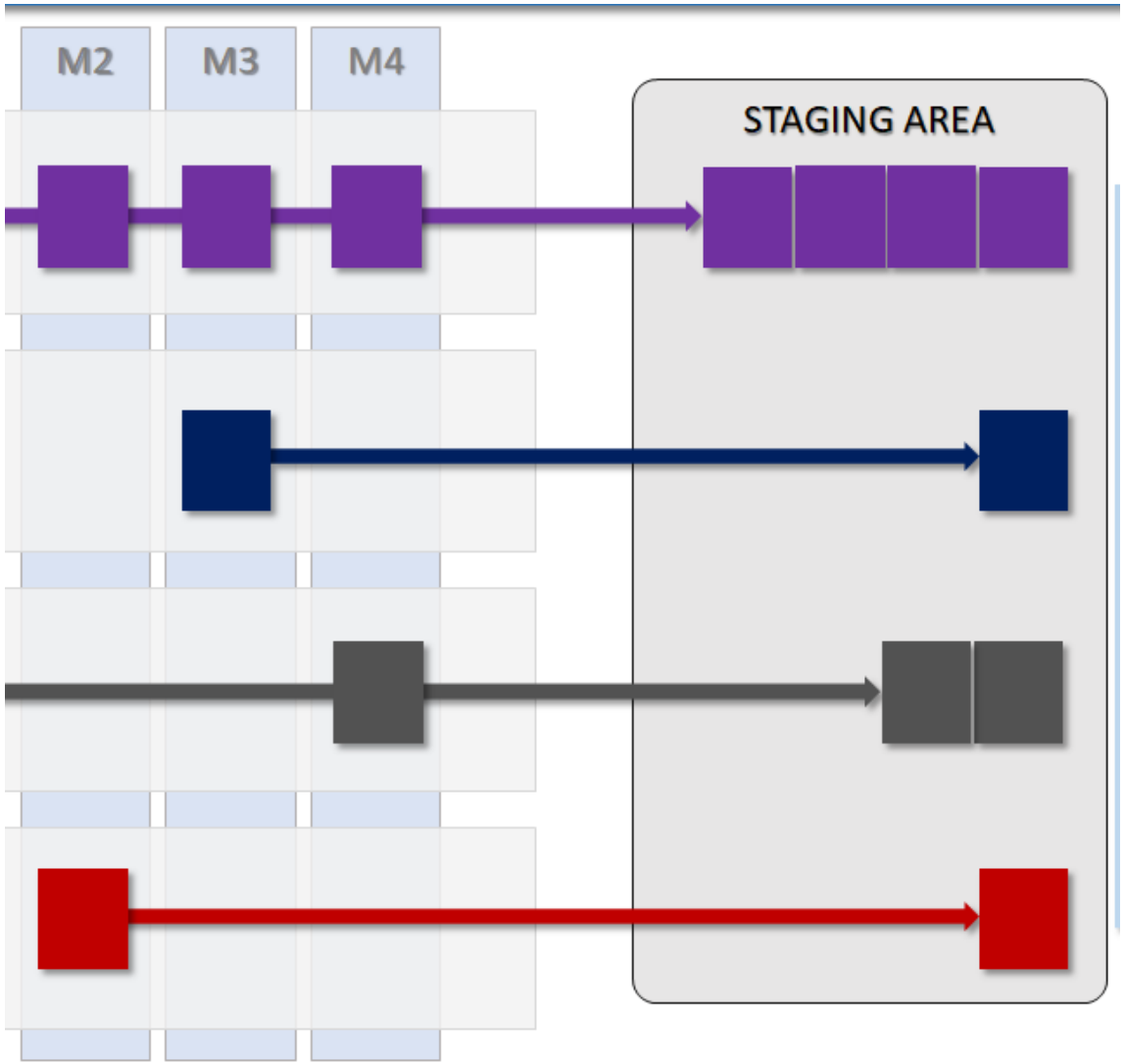
Data Warehouse (DW)



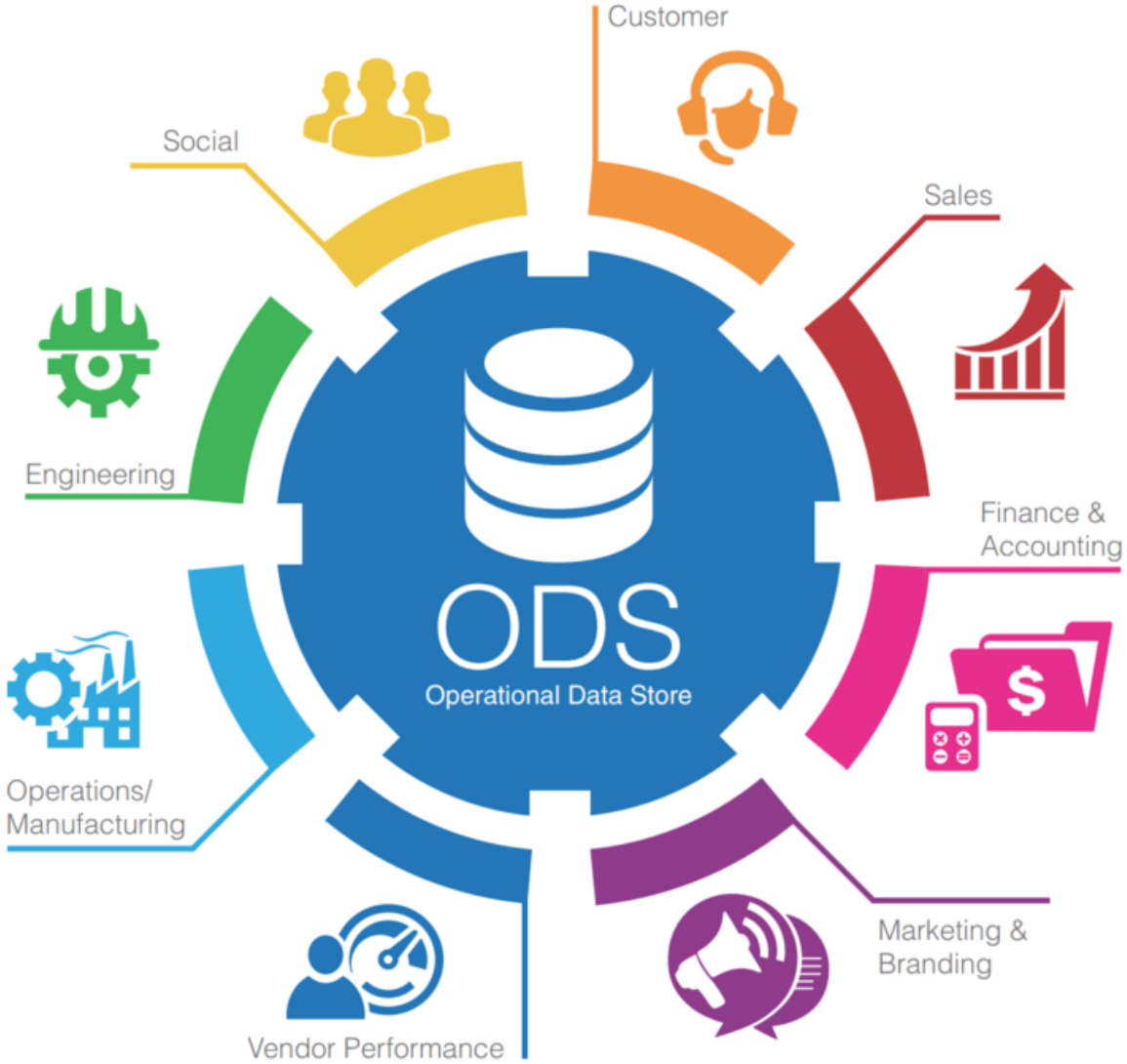
Extração, Transformação e Carga
(Extract, Transform and Load)



- ***Staging Area***, no contexto de *Data Warehouse*, corresponde a área de dados reservada para recepção de dados a partir de suas fontes.

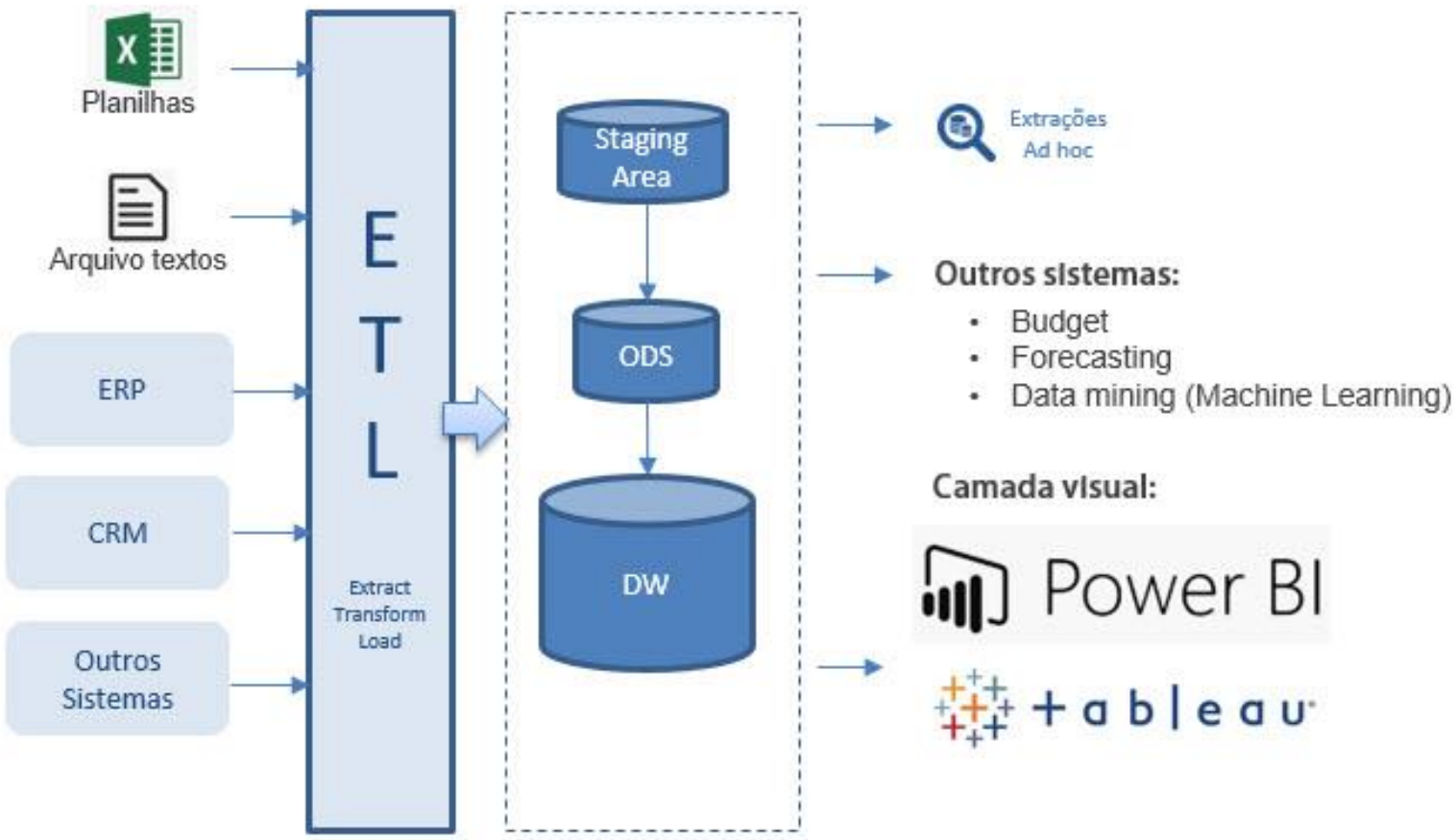


- ***Operational Data Store*** é um repositório de dados que visa atender consultas detalhadas e simplificar a criação de Data Warehouses.

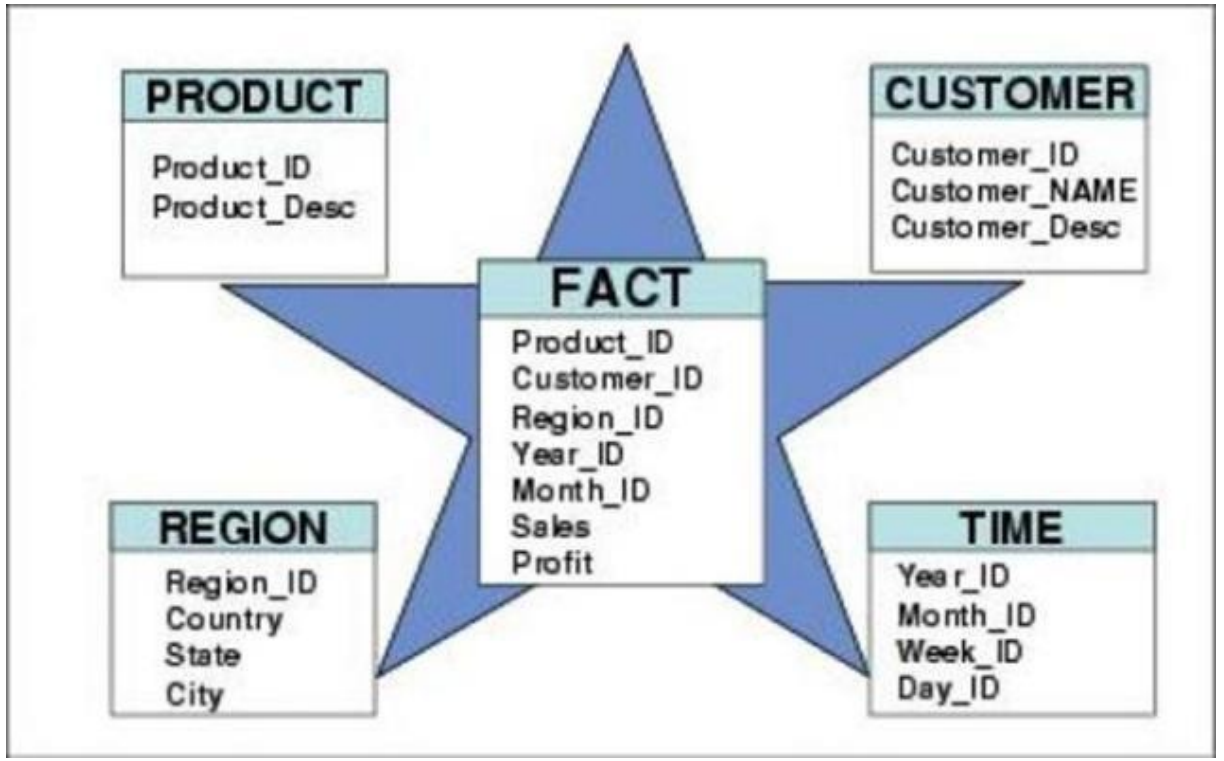


- Um **data mart** é um subconjunto de um **data warehouse** que normalmente é usado para acessar informações voltadas para o cliente

Data Warehouse (DW)



DW Modelo Estrela - Tabelas Fato e Dimensões



DW Modelo Estrela - Tabelas Fato e Dimensões

- A TABELA “FATO”

- A tabela **FATO** representa os dados históricos sobre o assunto que se deseja analisar.
- Exemplos de Tabela **Fato**:

– Itens vendidos ao longo do tempo;

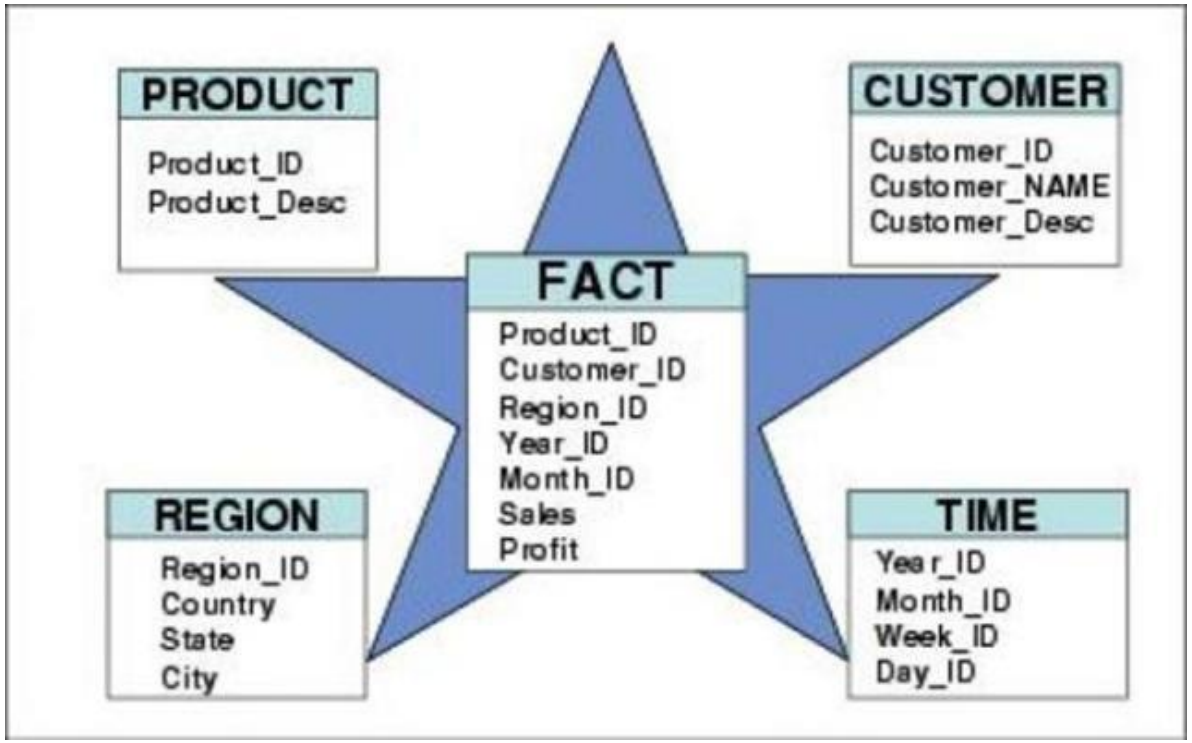
– Pagamento de Salários;

– Rentabilidade de investimentos;

– Variações em valores de ações;

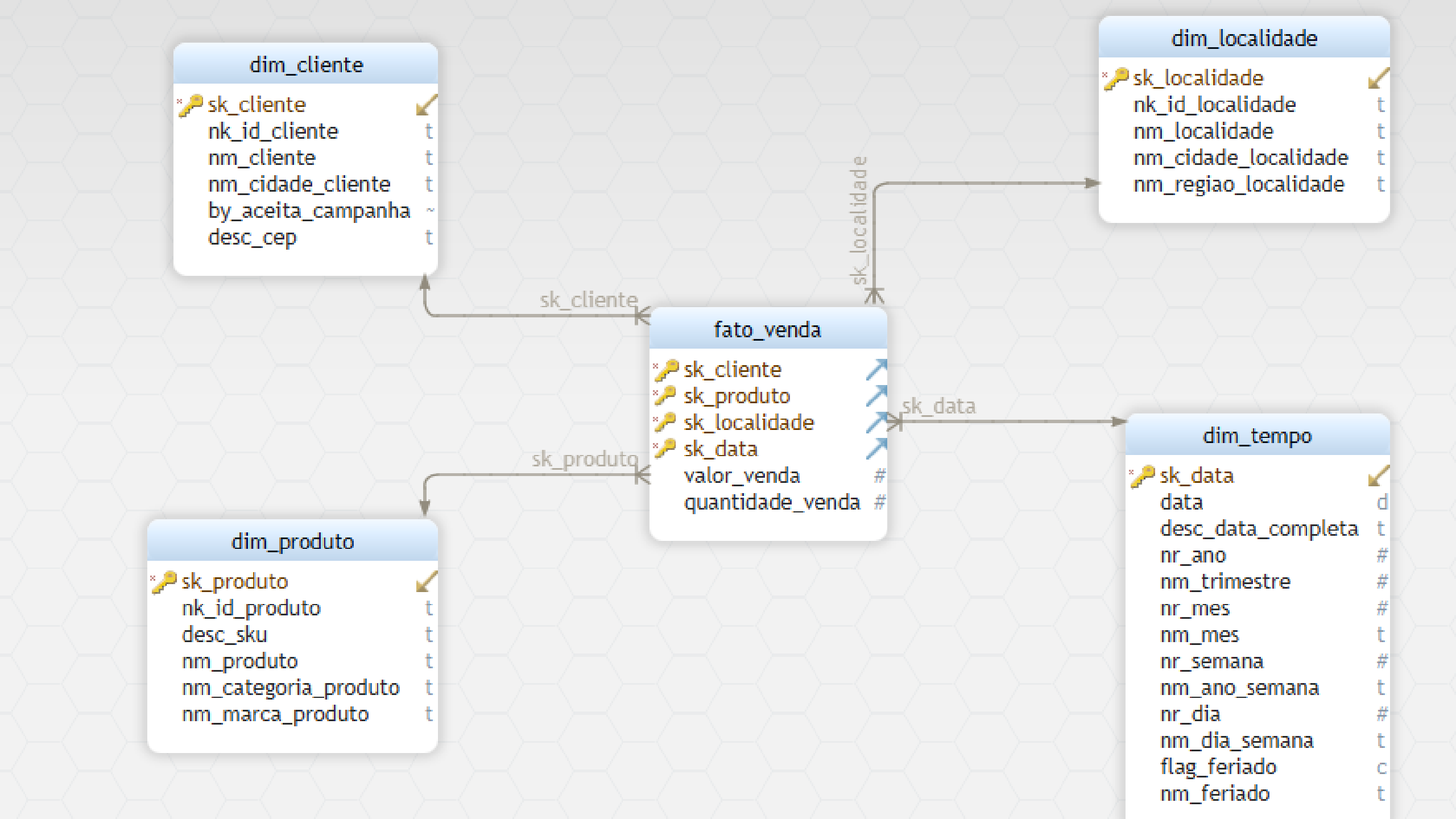
– Dados Cientificos;

– Informações Estatísticas.



Exemplo de Modelo Físico

Compreender os tipos de dados
e relacionamentos entre as tabelas


dim_cliente
sk_cliente
nk_id_cliente
nm_cliente
nm_cidade_cliente
by_aceita_campanha
desc_cep
dim_localidade
sk_localidade
nk_id_localidade
nm_localidade
nm_cidade_localidade
nm_regiao_localidade
sk_localidade
sk_cliente
fato_venda
sk_cliente
sk_produto
sk_localidade
sk_data
valor_venda
quantidade_venda
sk_data
sk_produto
dim_produto
sk_produto
nk_id_produto
desc_sku
nm_produto
nm_categoria_produto
nm_marca_produto
dim_tempo
sk_data
data
desc_data_completa
nr_ano
nm_trimestre
nr_mes
nm_mes
nr_semana
nm_ano_semana
nr_dia
nm_dia_semana
flag_feriado
nm_feriado

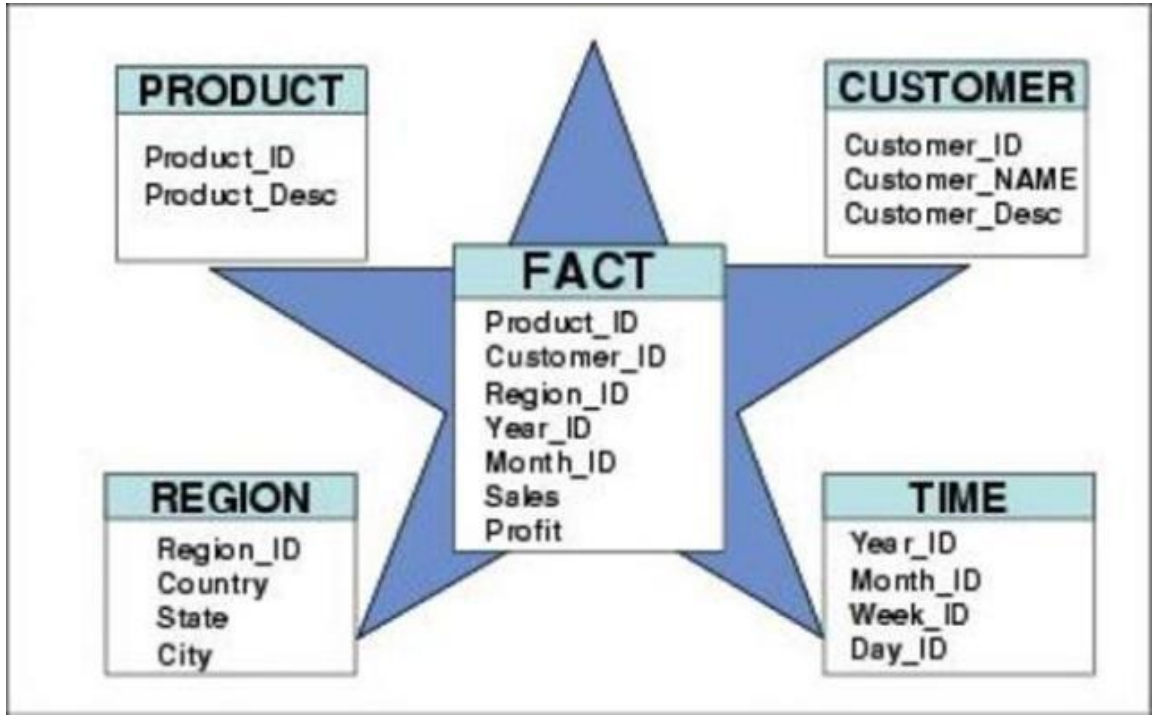
DW Modelo Estrela - Tabelas Fato e Dimensões

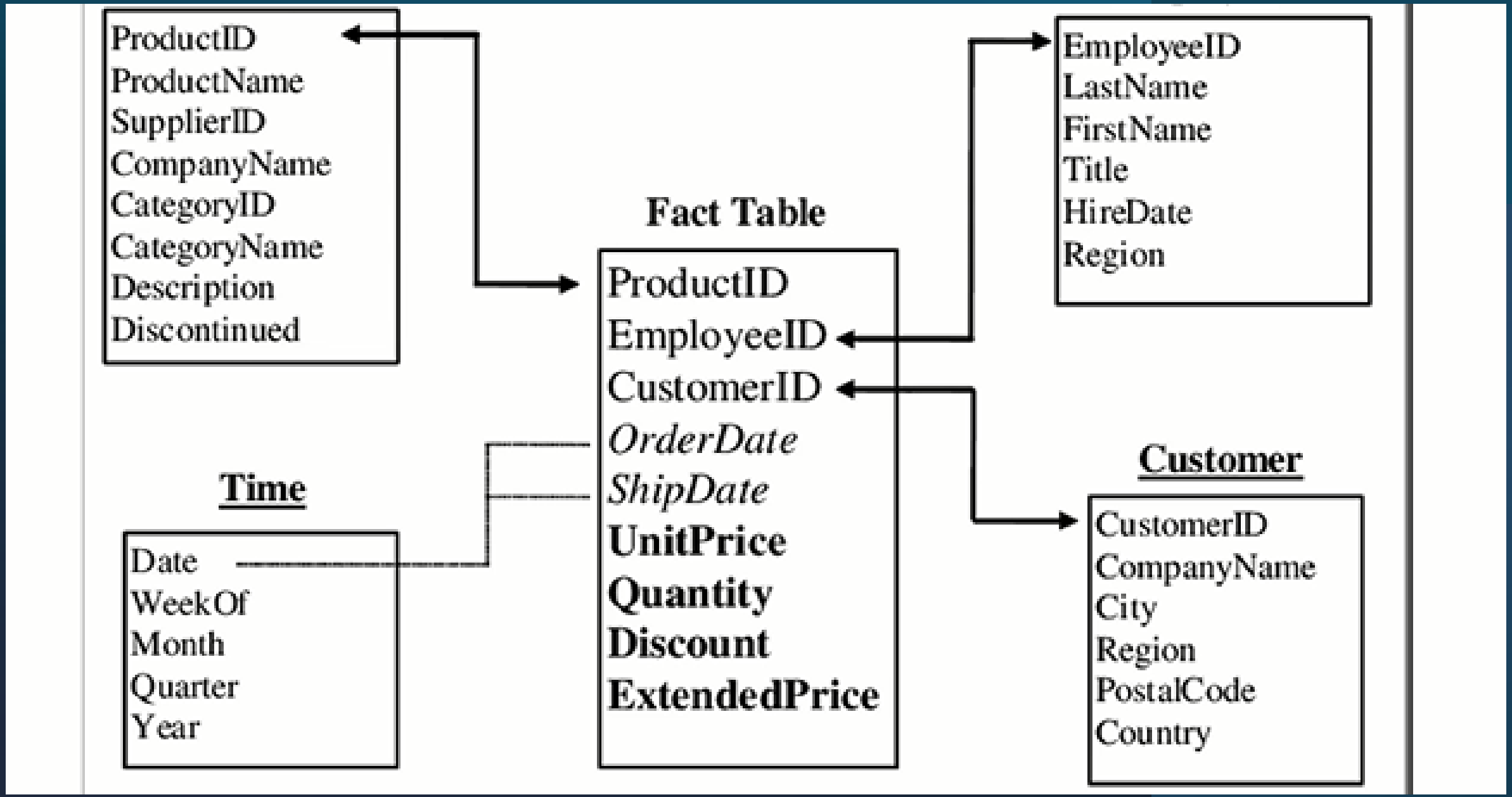
- CRIANDO UMA DIMENSÃO

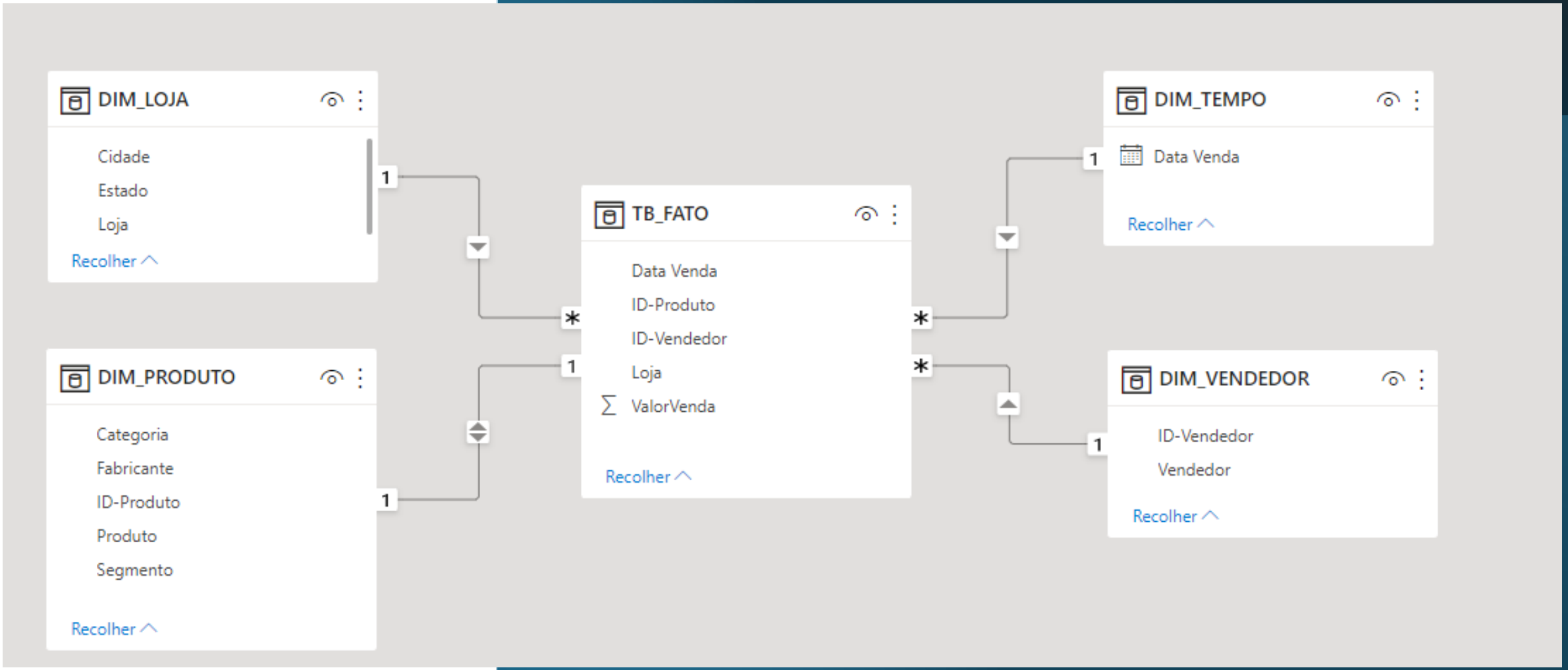
- A dimensão tem por objetivo descrever o conteúdo da tabela fato.

- Enquanto a tabela fato é composta basicamente de chaves estrangeiras e valores, a dimensão conterá dados detalhados sobre quem é o produto, cliente, onde ele está localizado, etc.

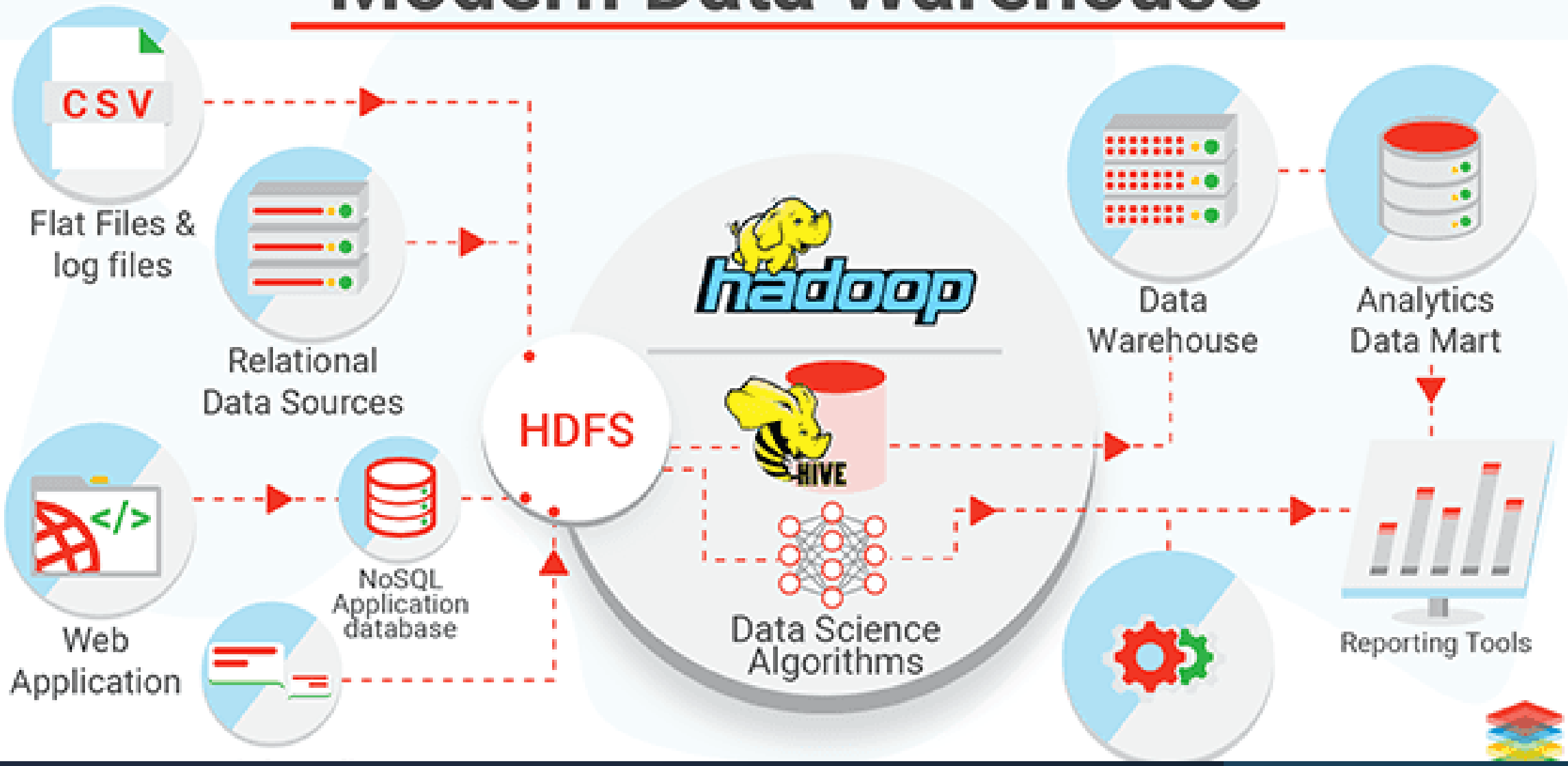
- Uma Dimensão de produtos, precisará conter campos como NOME DO PRODUTO, NOME DO FABRICANTE, DESCRIÇÃO DO MODELO, etc.




ProductID
ProductName
SupplierID
CompanyName
CategoryID
CategoryName
Description
Discontinued
EmployeeID
LastName
FirstName
Title
HireDate
Region
Fact Table
ProductID
EmployeeID
CustomerID
OrderDate
ShipDate
UnitPrice
Quantity
Discount
ExtendedPrice
Time
Date
WeekOf
Month
Quarter
Year
Customer
CustomerID
CompanyName
City
Region
PostalCode
Country


DIM_LOJA
Cidade
Estado
Loja
Recolher
DIM_PRODUTO
Categoria
Fabricante
ID-Produto
Produto
Segmento
Recolher
TB_FATO
Data Venda
ID-Produto
ID-Vendedor
Loja
ValorVenda
Recolher
DIM_TEMPO
Data Venda
Recolher
DIM_VENDEDOR
ID-Vendedor
Vendedor
Recolher
1
*

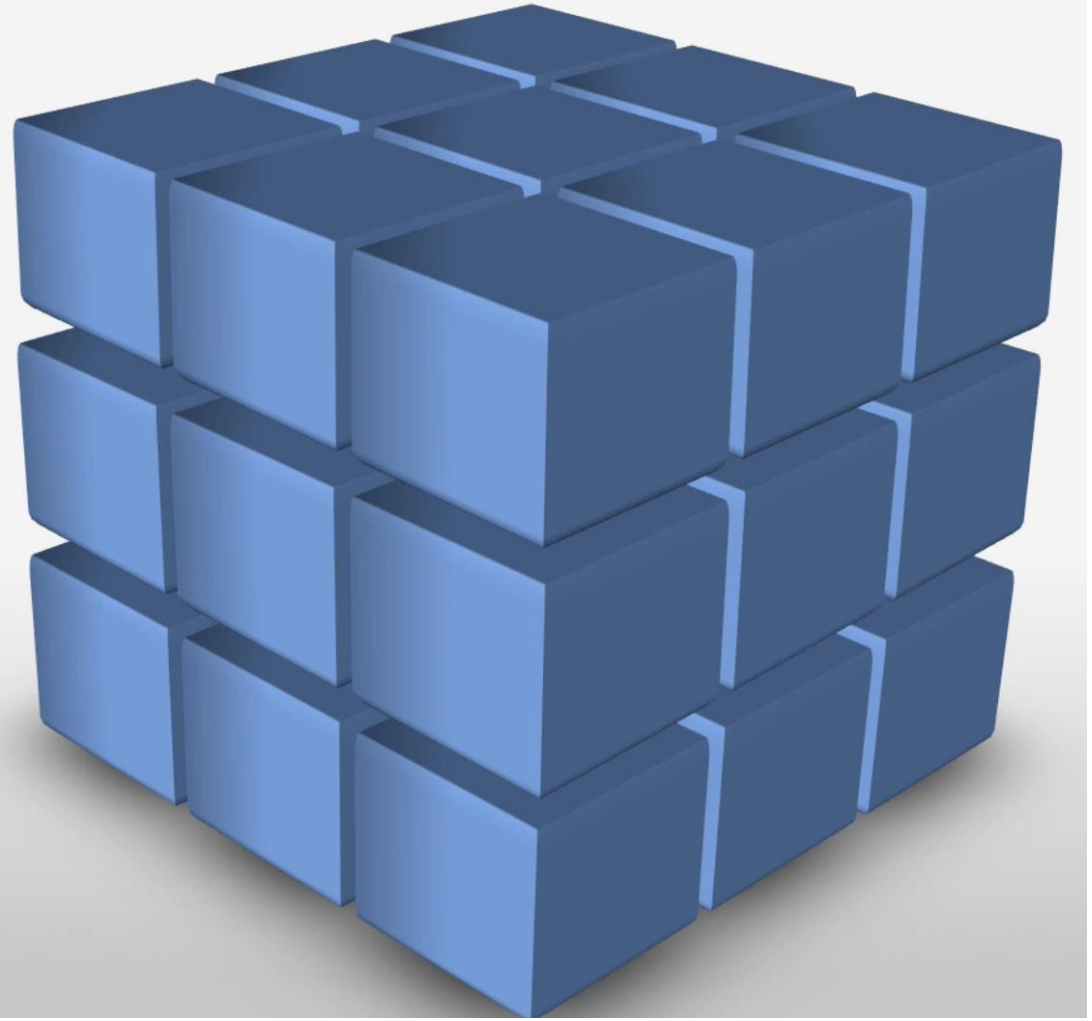
Modern Data Warehouse



QUANDO USAR UM DATA WAREHOUSE?

Data Warehouse

- Hardware
- Sistema Operacional
- Banco de Dados
- Ferramentas de ETL e Consulta
- Técnicas de Indexação
- Ferramentas de Análise
- Ferramentas de Relatórios, Gráficos e Dashboards

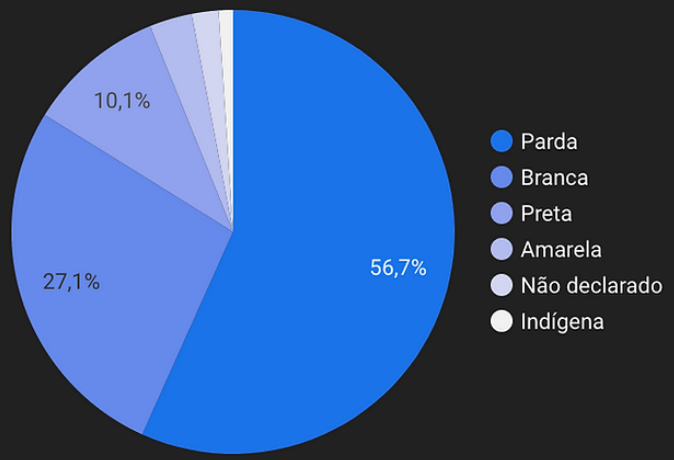


Dashboard dados ENEM Paraíba - 2020

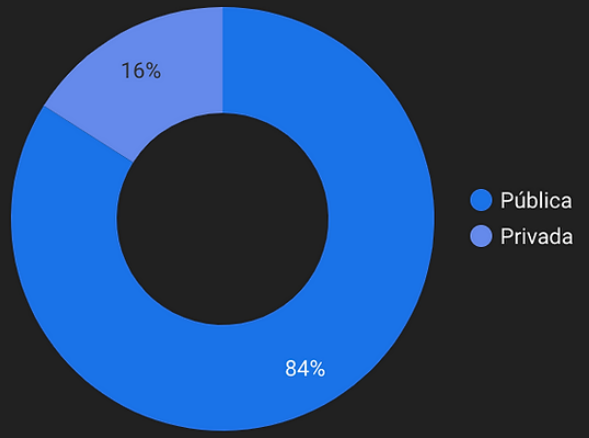
Numero de Inscritos
149.092

IDHM média
0,61

PIB per Capita
R$ 11.987



Tipo de Escola



Dashboard PowerBI e Google Analytics

Análise de dados e-commerce

country

Afghanistan | Albania | Algeria | Angola | Argentina | Armenia | Aruba | Australia | Austria | Azerbaijan | Bahamas

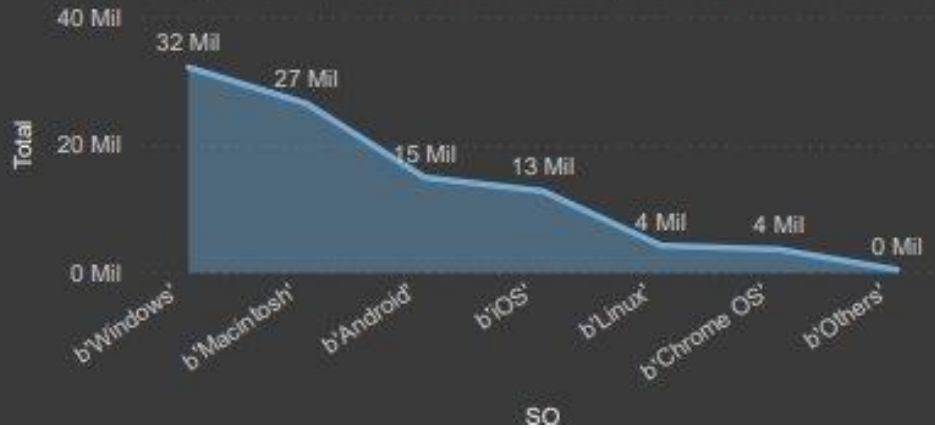
Tempo Médio no Site por Dia

KPI pageviews Dia

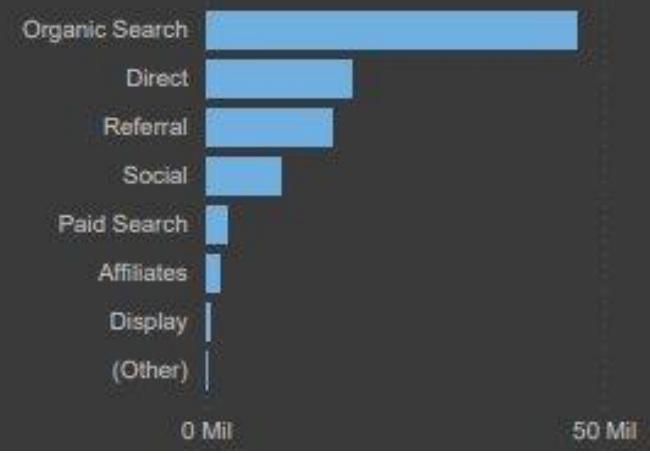
Total de Faturamento por Dia

Volume de Acesso por Dispositivo

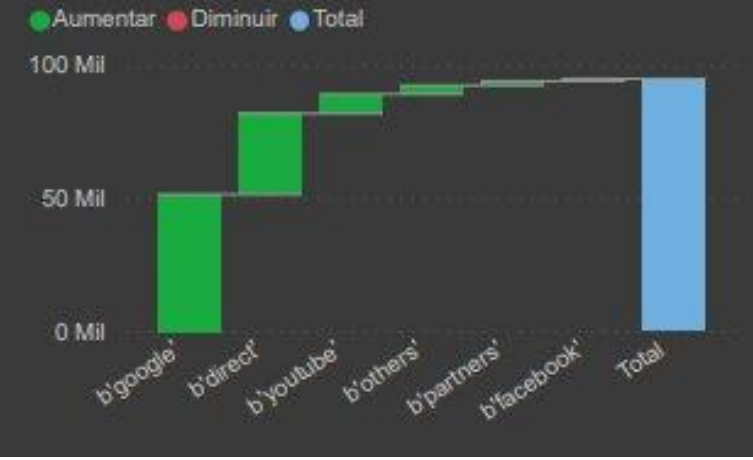
Sistemas Operacionais mais usados pelos visitantes

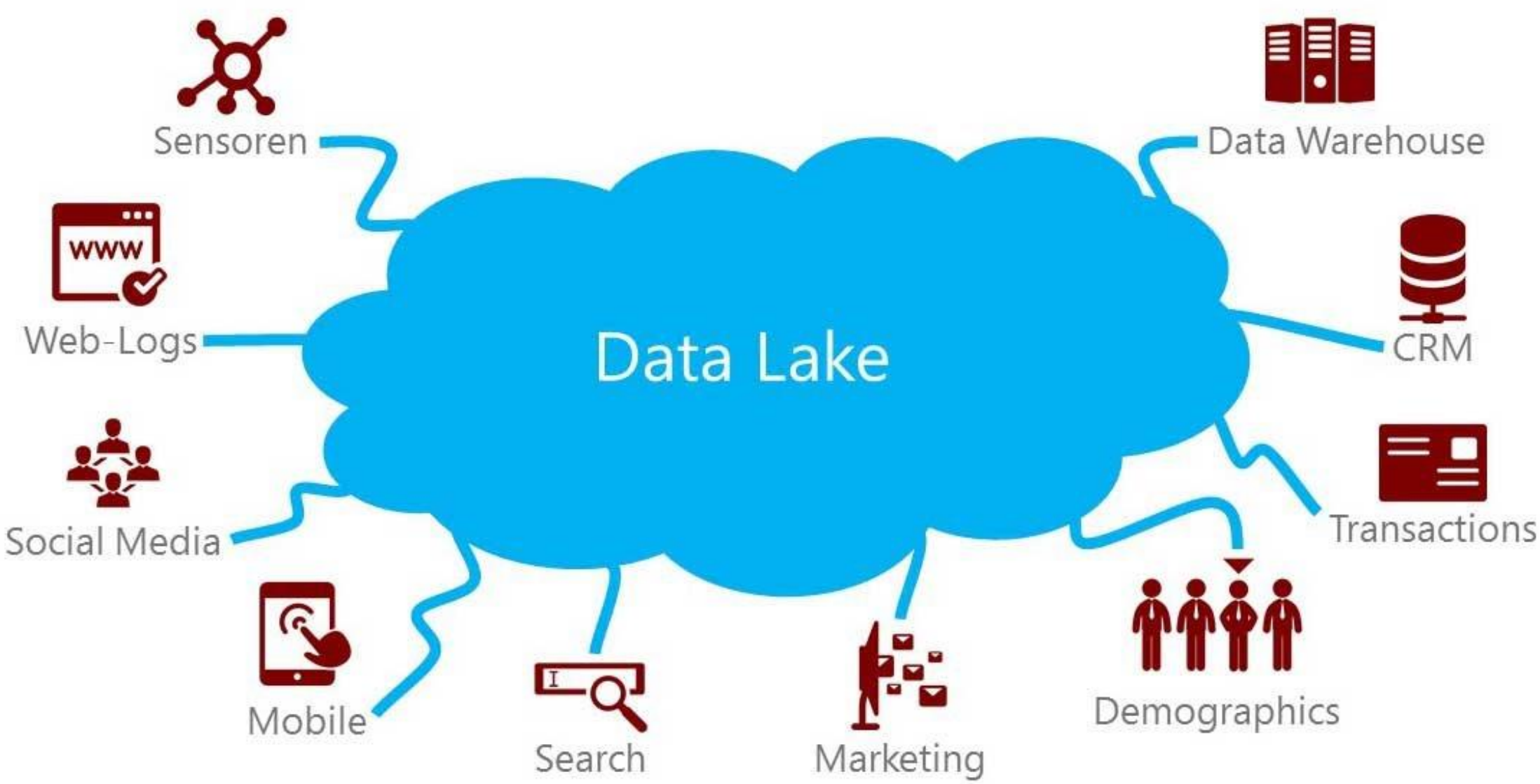


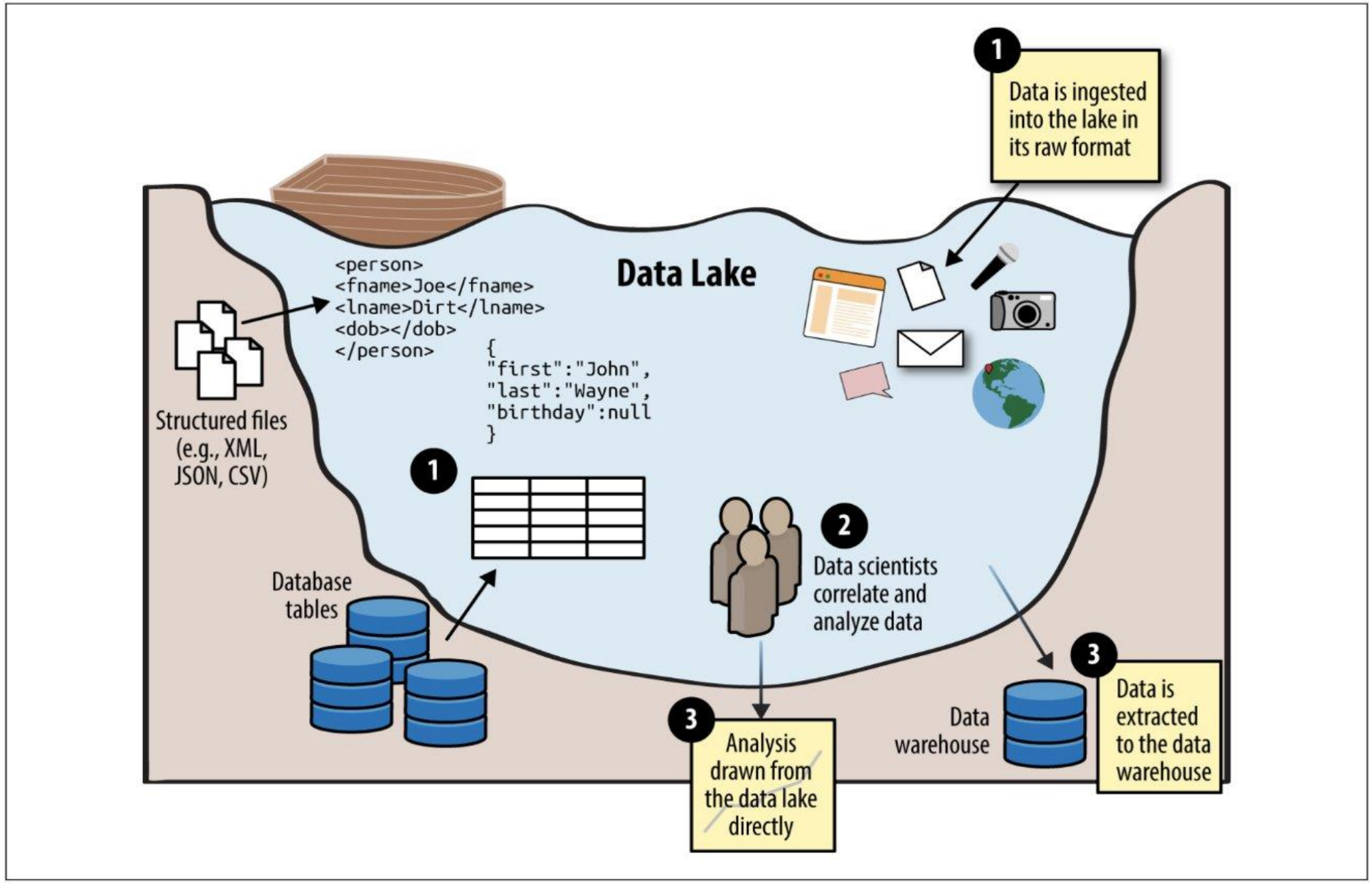
Principais Canais de Vendas



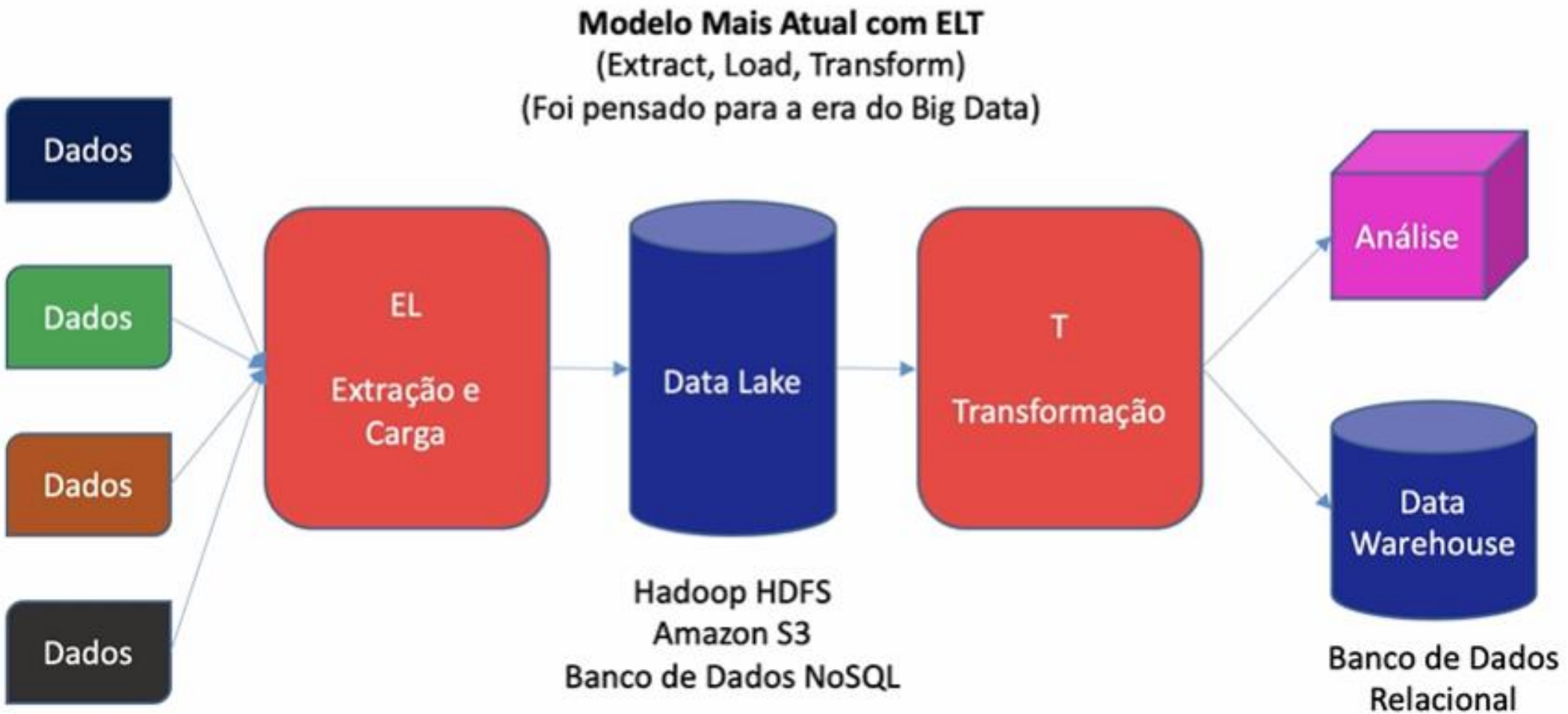
Principais Fontes de Acesso

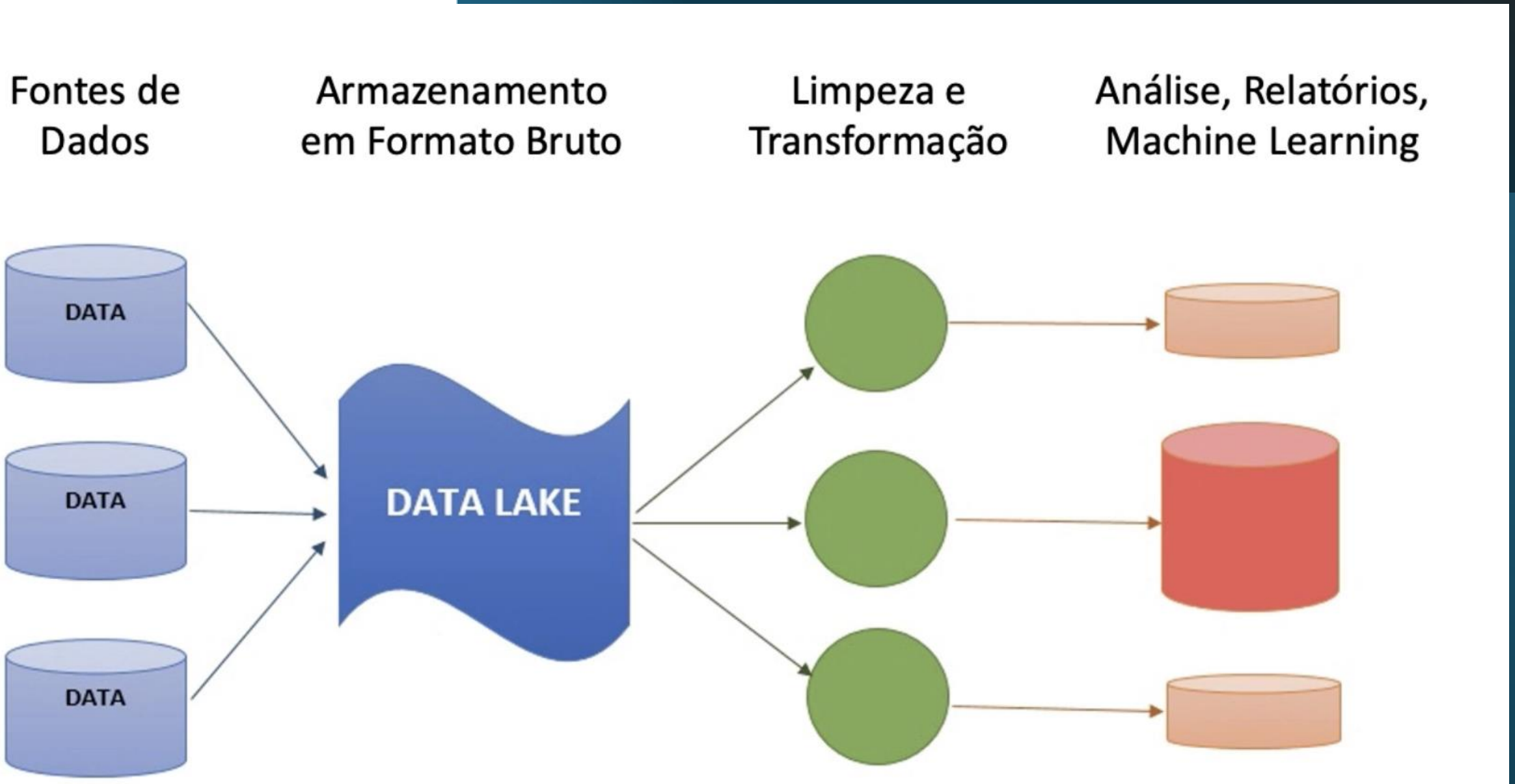



Sensoren
Data Warehouse
Web-Logs
Data Lake
CRM
Social Media
Transactions
Mobile
Search
Marketing
Demographics

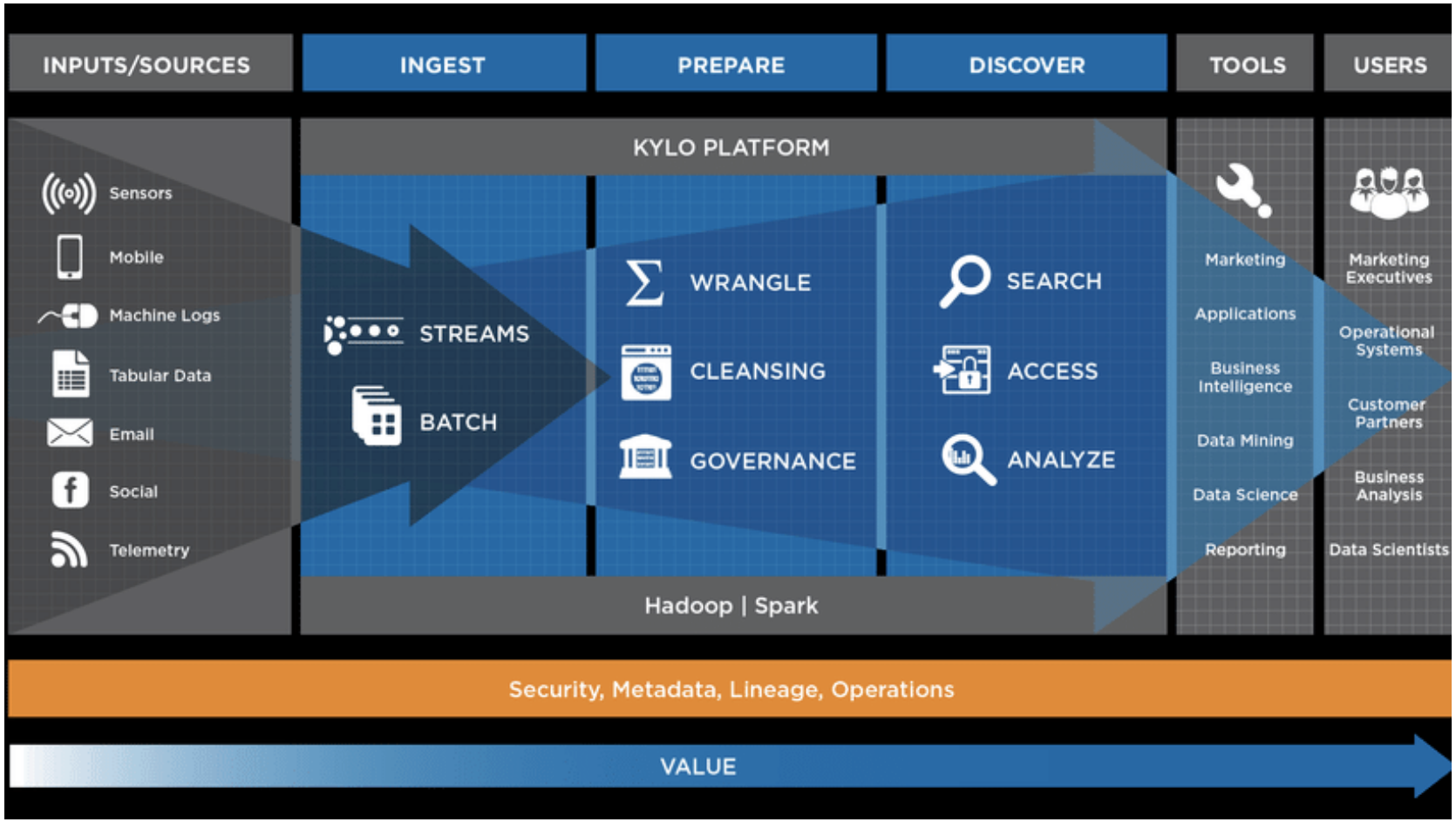

1
Data is ingested into the lake in its raw format
Data Lake
<person>
<fname>Joe</fname>
<lname>Dirt</lname>
<dob></dob>
</person>
{
"first":"John",
"last":"Wayne",
"birthday":null
}
Structured files (e.g., XML, JSON, CSV)
1
Database tables
2
Data scientists correlate and analyze data
3
Analysis drawn from the data lake directly
Data warehouse
3
Data is extracted to the data warehouse

O Que é um Data Lake?




Fontes de Dados
Armazenamento em Formato Bruto
Limpeza e Transformação
Análise, Relatórios, Machine Learning
DATA
DATA
DATA
DATA LAKE

Plataformas de Data Lake - kylo.io

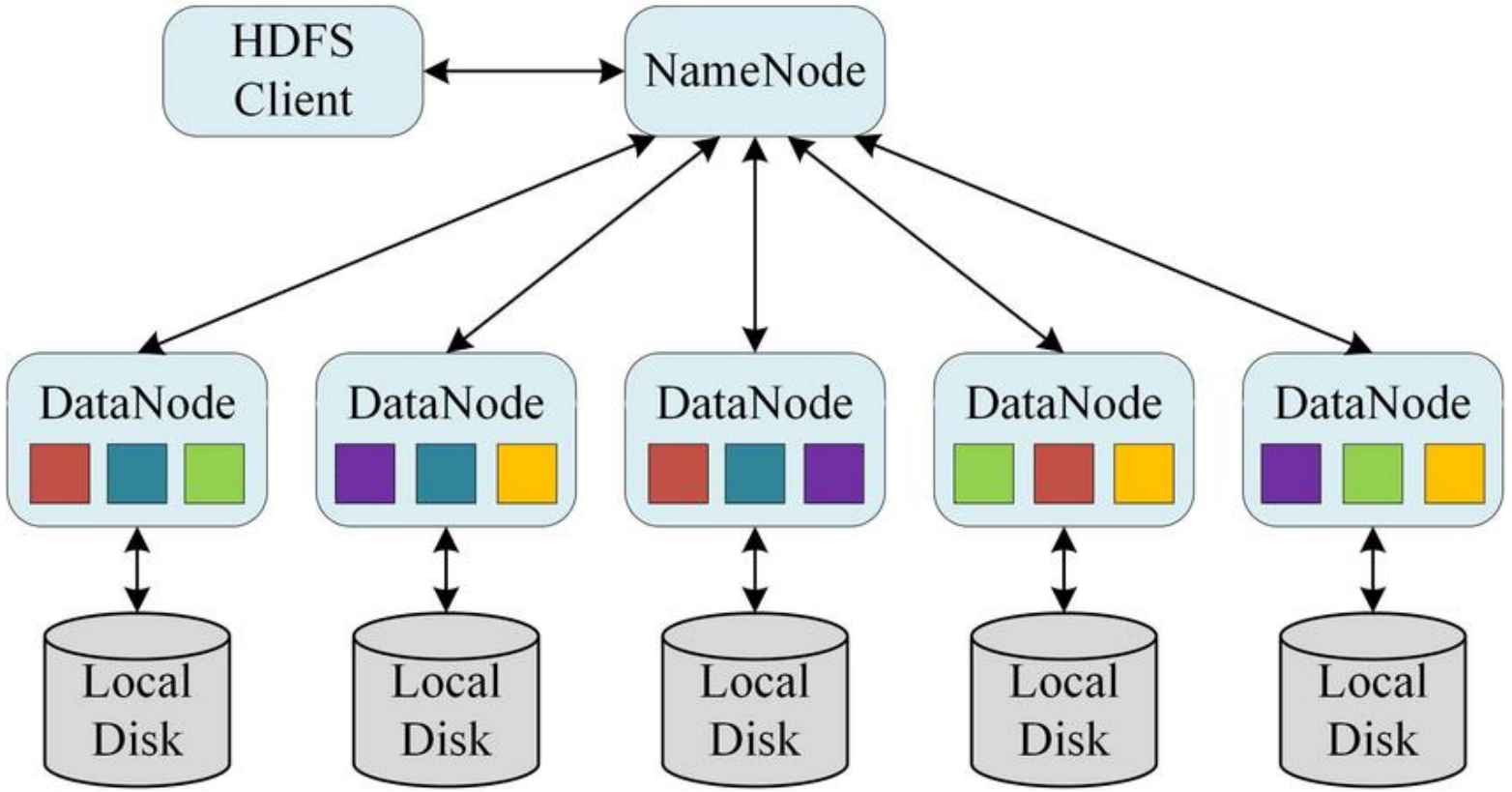


O projeto Apache® Hadoop® desenvolve software de código aberto para computação distribuída confiável, escalável.

A biblioteca de software Apache Hadoop é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores usando modelos de programação simples.

Ele foi projetado para escalar desde servidores únicos até milhares de máquinas, cada uma oferecendo computação e armazenamento local.

- Em vez de depender de hardware para fornecer alta disponibilidade, a própria biblioteca foi projetada para detectar e tratar falhas na camada de aplicação, fornecendo assim um serviço altamente disponível sobre um cluster de computadores, cada um dos quais pode estar sujeito a falhas.



HADOOP ECOSYSTEM TIMELIENE

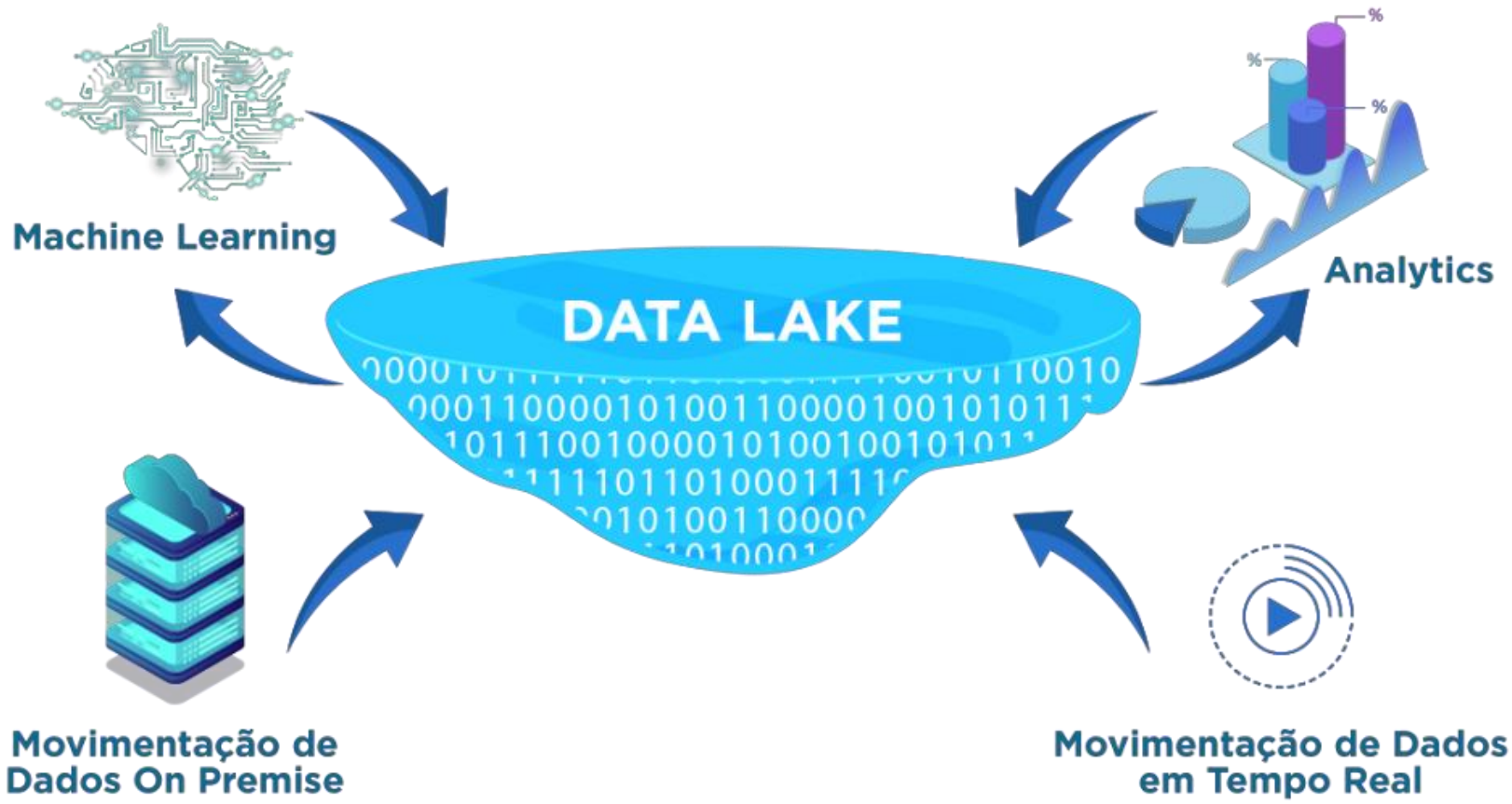
altexsoft

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	Flink Stream and batch processing	Apache Storm Stream processing

dremio

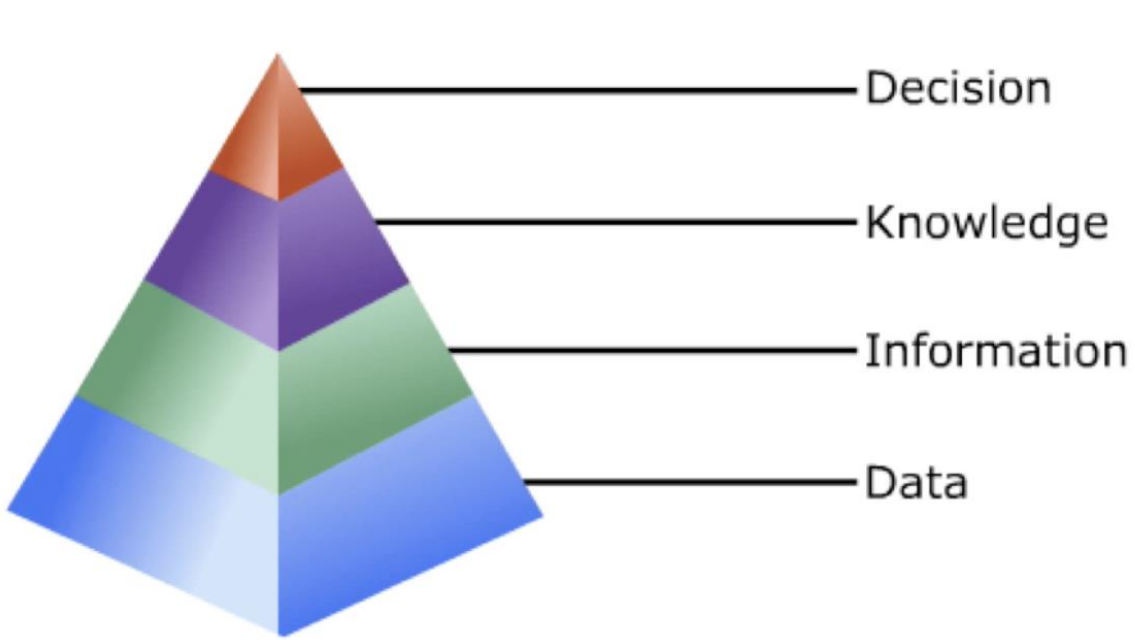
É uma plataforma que unifica as camadas de storage/bancos de dados com interfaces de consulta — como ferramentas de BI, códigos-fonte e etc

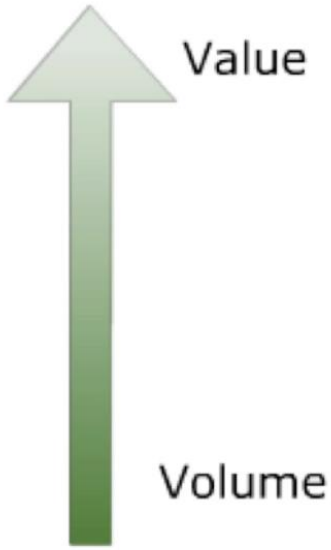
Big Data e Data Lake

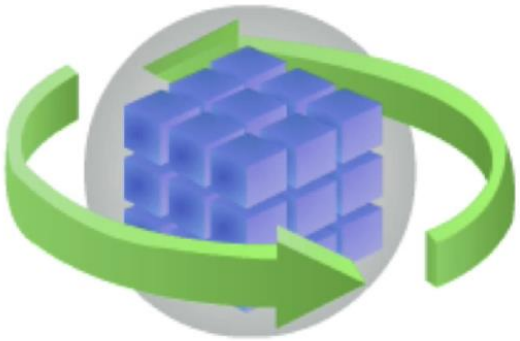


BUSINESS INTELLIGENCE X DATA SCIENCE

Business Intelligence é o processo de transformar dados em informação e, através de descobertas, transformar informação em conhecimento que suporte a tomada de decisões. (Definição do Gartner Group)







Visualização de Dados, Relatórios e BI

Data Science / Machine Learning / Data Learning

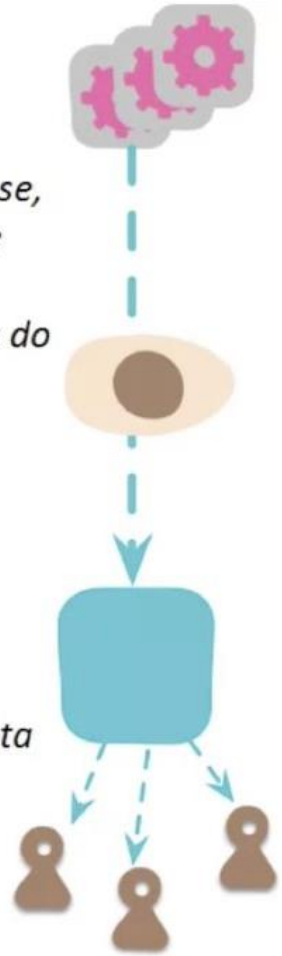
Data Warehouse

Data Lake

Data Lake x Data Warehouse

Com o Data Warehouse, os dados são limpos e organizados em um único esquema, antes do armazenamento

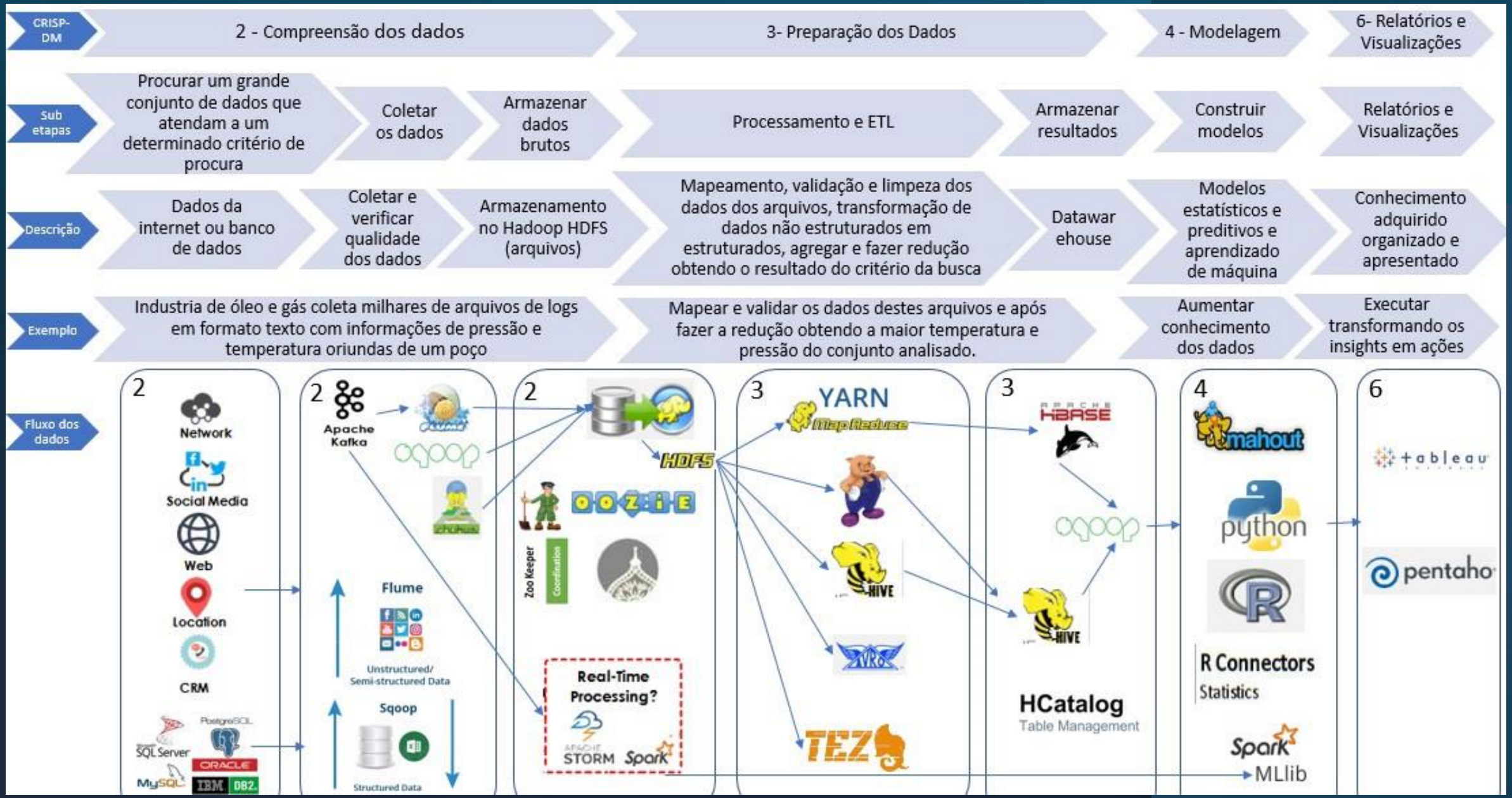
A análise é feita consultando diretamente no Data Warehouse



Com o Data Lake, os dados são armazenados em seu formato bruto

Os dados são selecionados e organizados de acordo com a necessidade

Big Data Stack - Ecossistema


CRISP-DM
2 - Compreensão dos dados
3- Preparação dos Dados
4 - Modelagem
6- Relatórios e Visualizações
Sub etapas
Procurar um grande conjunto de dados que atendam a um determinado critério de procura
Coletar os dados
Armazenar dados brutos
Processamento e ETL
Armazenar resultados
Construir modelos
Relatórios e Visualizações
Descrição
Dados da internet ou banco de dados
Coletar e verificar qualidade dos dados
Armazenamento no Hadoop HDFS (arquivos)
Mapeamento, validação e limpeza dos dados dos arquivos, transformação de dados não estruturados em estruturados, agregar e fazer redução obtendo o resultado do critério da busca
Datawarehouse
Modelos estatísticos e preditivos e aprendizado de máquina
Conhecimento adquirido organizado e apresentado
Exemplo
Industria de óleo e gás coleta milhares de arquivos de logs em formato texto com informações de pressão e temperatura oriundas de um poço
Mapear e validar os dados destes arquivos e após fazer a redução obtendo a maior temperatura e pressão do conjunto analisado.
Aumentar conhecimento dos dados
Executar transformando os insights em ações
Fluxo dos dados
2
Network
Social Media
Web
Location
CRM
SQL Server
PostgreSQL
ORACLE
MySQL
IBM DB2
2
Apache Kafka
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data
2
HDFS
OOZIE
Zoo Keeper
Coordination
Real-Time Processing?
STORM
Spark
3
YARN
Map Reduce
HIVE
AVRO
TEZ
3
HBASE
HIVE
HCatalog
Table Management
4
mahout
python
R
R Connectors
Statistics
Spark
MLlib
6
+ableau
pentaho

Clip slide

Hadoop Ecosystem

cloudera
Hortonworks
MAPR
ster
nagement

Apache
Zookeeper
ordination

OOZIE
workflow

Visualization
HUE
+ableau

Analysis


HIVE

Spark SQL
Spark MLlib
GraphX

Solr
Search
elasticsearch

Processing
hadoop MapReduce

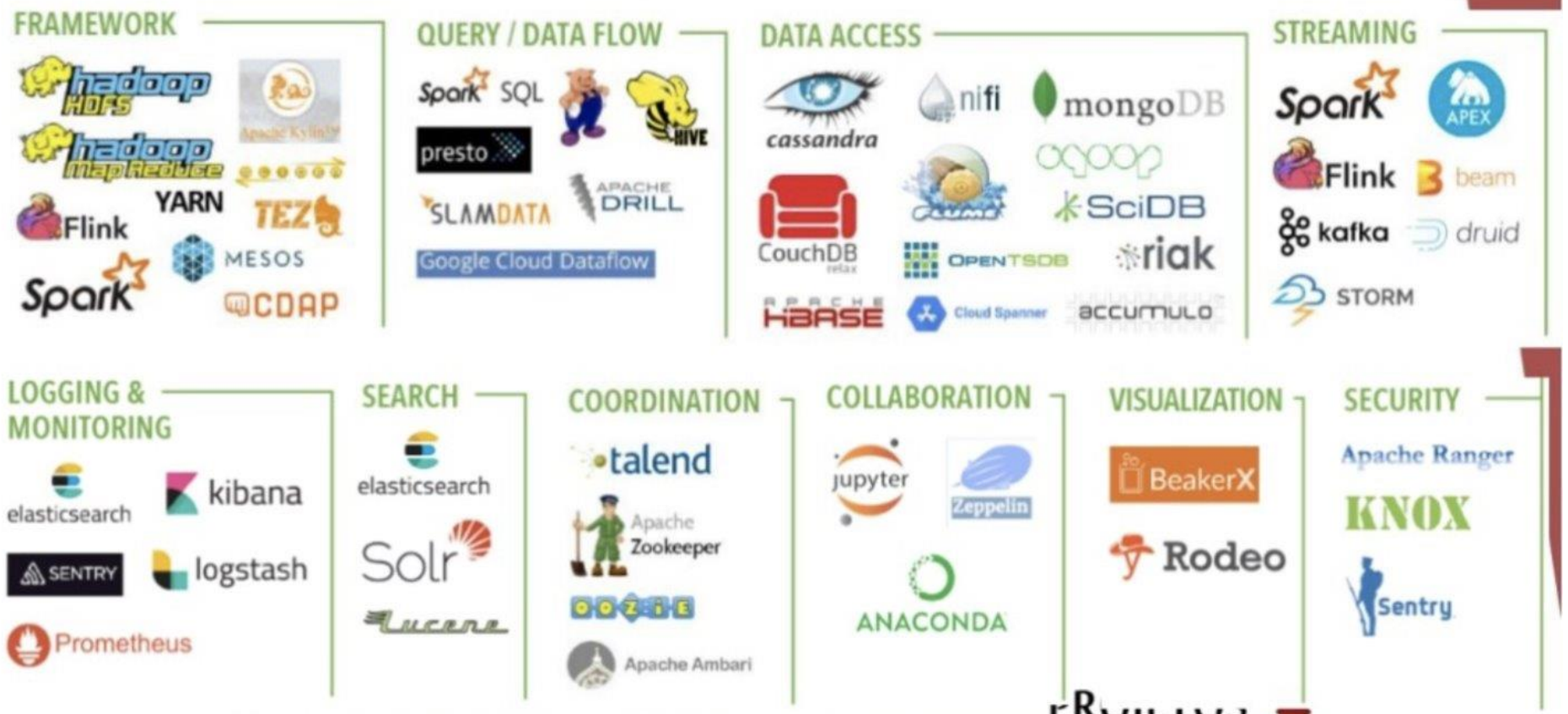
Resource Management
YARN

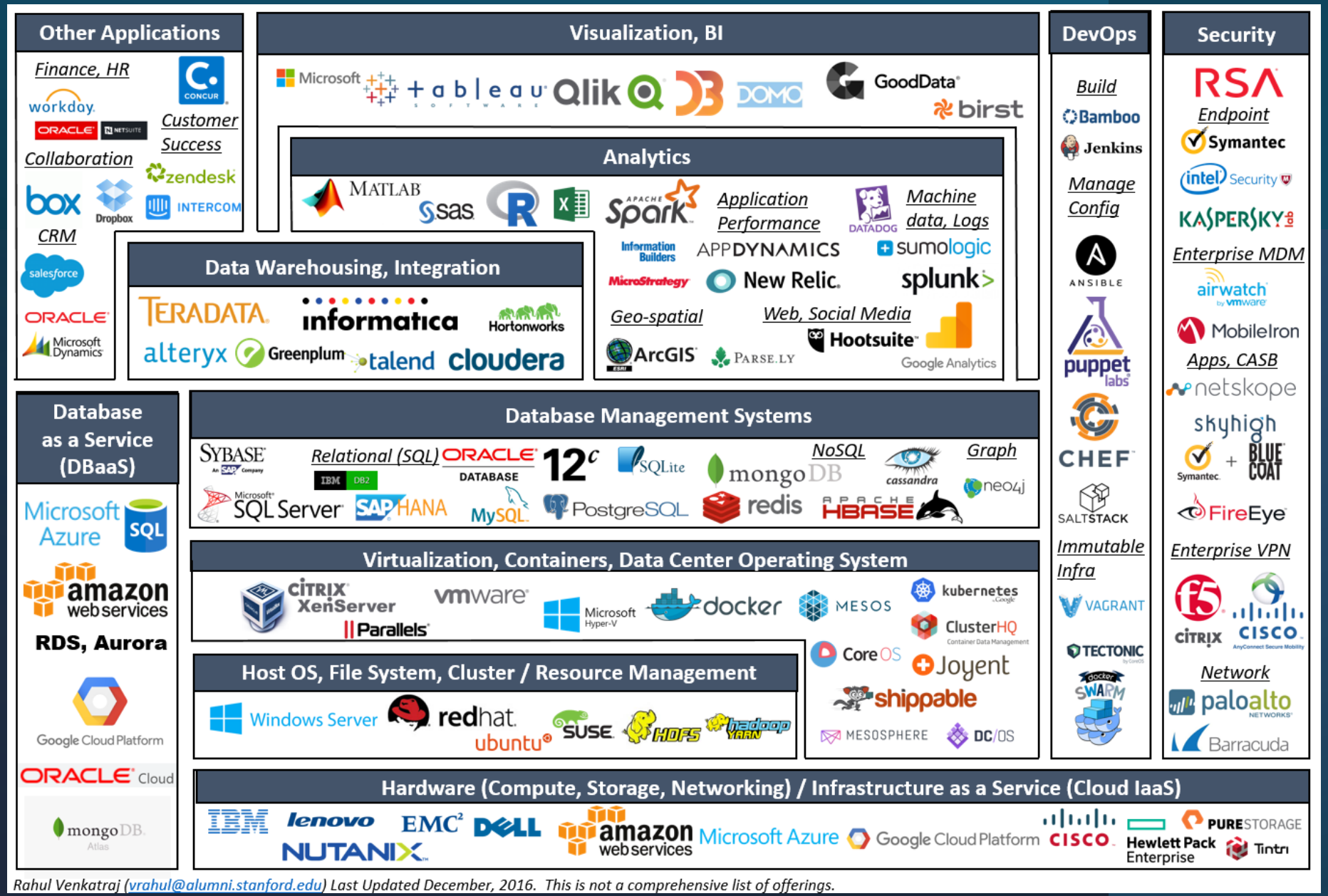
Storage
HDFS
APACHE HBASE

Data Formats
AVRO
Parquet

Data

Ecossistema Open Source

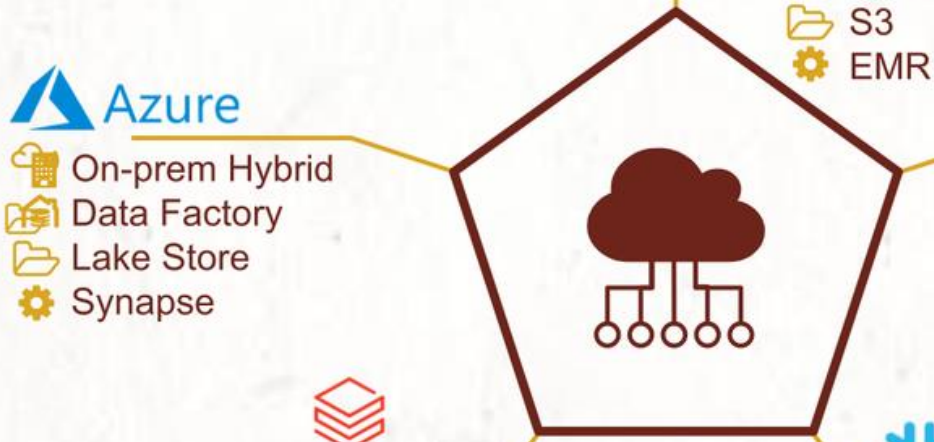




Rahul Venkatraj (vrahul@alumni.stanford.edu) Last Updated December, 2016. This is not a comprehensive list of offerings.

Big Data Infra: Cloud Alternatives

scgupta.link/big-data



aws
- Single Cloud
- RedShift
- S3
- EMR

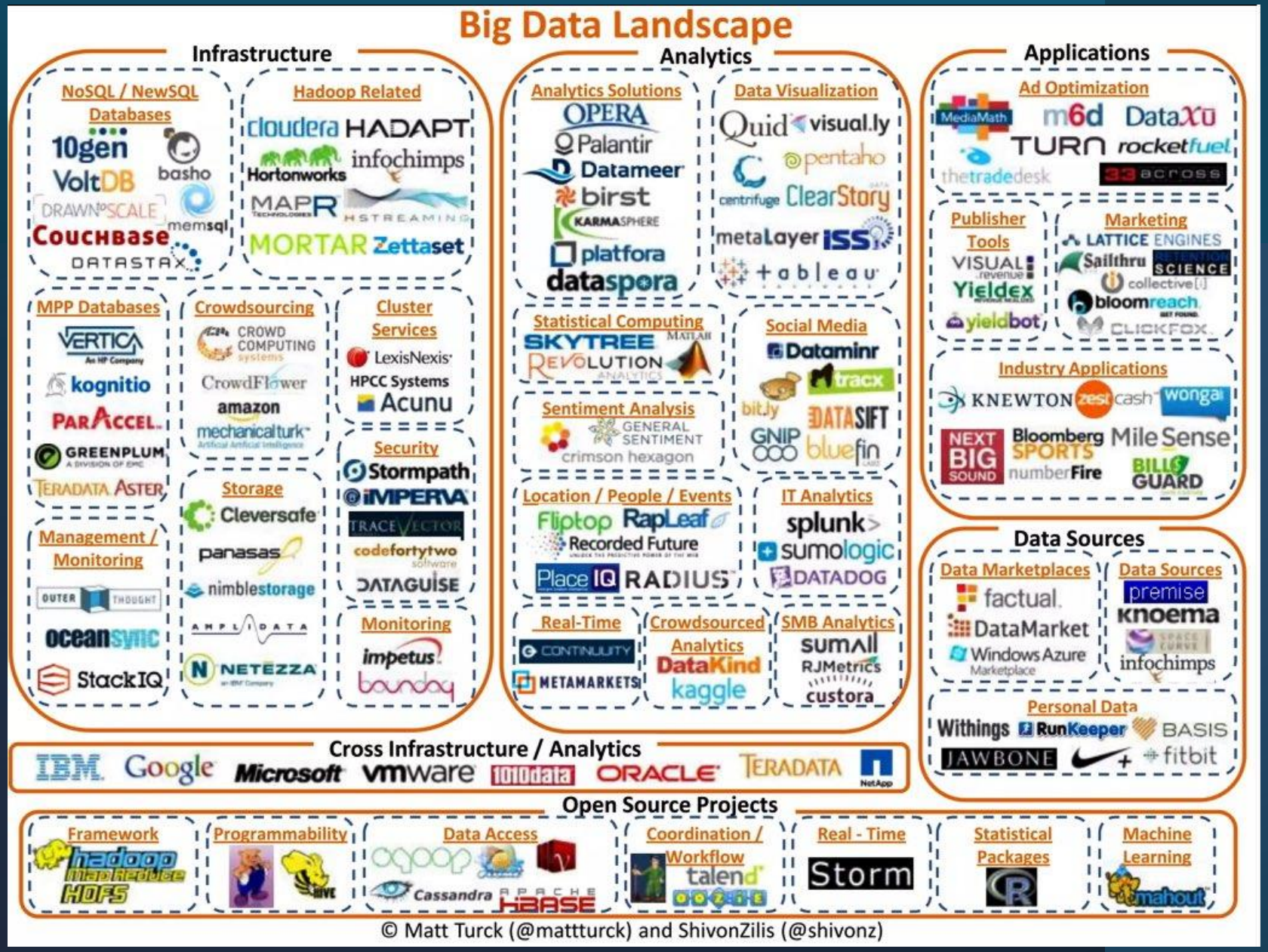
Azure
- On-prem Hybrid
- Data Factory
- Lake Store
- Synapse

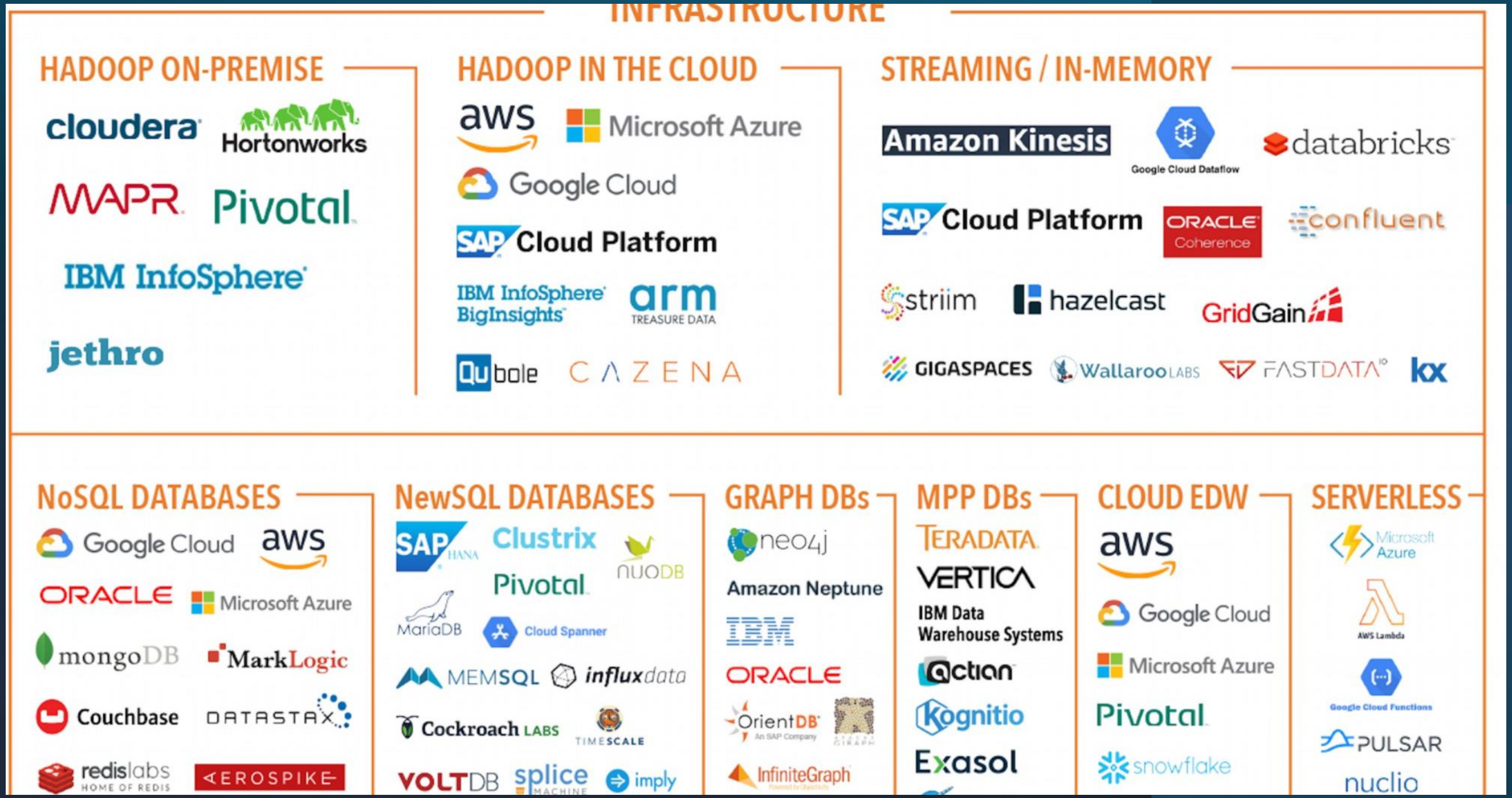
Google Cloud
- Single Cloud
- BigQuery
- Cloud Storage
- DataProc
- BigQuery Omni

databricks
- Multi Cloud
- Delta Lake
- Schema-on-Read
- ML Workloads

snowflake
- Multi Cloud
- Columnar Store
- Relational Schema
- BI Workloads

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta




INFRASTRUCTURE
HADOOP ON-PREMISE
cloudera
Hortonworks
MAPR
Pivotal
IBM InfoSphere
jethro
HADOOP IN THE CLOUD
aws
Microsoft Azure
Google Cloud
SAP Cloud Platform
IBM InfoSphere BigInsights
arm TREASURE DATA
Qubole
CAZENA
STREAMING / IN-MEMORY
Amazon Kinesis
Google Cloud Dataflow
databricks
SAP Cloud Platform
ORACLE Coherence
confluent
striim
hazelcast
GridGain
GIGASPACES
WallarooLABS
FASTDATA.IO
kx
NoSQL DATABASES
Google Cloud
aws
ORACLE
Microsoft Azure
mongoDB
MarkLogic
Couchbase
DATASTAX
redislabs HOME OF REDIS
AEROSPIKE
NewSQL DATABASES
SAP HANA
Clustrix
NUODB
Pivotal
MariaDB
Cloud Spanner
MEMSQL
influxdata
Cockroach LABS
TIMESCALE
VOLTDB
splice MACHINE
imply
GRAPH DBs
neo4j
Amazon Neptune
IBM
ORACLE
OrientDB An SAP Company
InfiniteGraph
MPP DBs
TERADATA
VERTICA
IBM Data Warehouse Systems
Actian
Kognitio
Exasol
CLOUD EDW
aws
Google Cloud
Microsoft Azure
Pivotal
snowflake
SERVERLESS
Microsoft Azure
AWS Lambda
Google Cloud Functions
PULSAR
nuclio