Escola Politécnica de Pernambuco

Especialização em Ciência de Dados e Analytics

Introdução à Ciência de Dados
Aula 3

Prof. Dr. Alexandre Maciel

alexandre.maciel@upe.br

QUAIS DADOS ANALISAR?

PRODUÇÃO POR INPUT

- Teclado.
- Mouse.
- *Touch Screen.*
- Scanners.
- Código de Barras.
- RFID.
- Câmeras.
- Filmadoras.

DADOS POR PROCESSAMENTO

- Análise ou Execução de Procedimentos.

- Criação de Modelos:

- Estatísticos;
- Machine Learning.

- Criação de Data Warehouses.

- Data Marts.

DADOS POR TRANSFORMAÇÃO

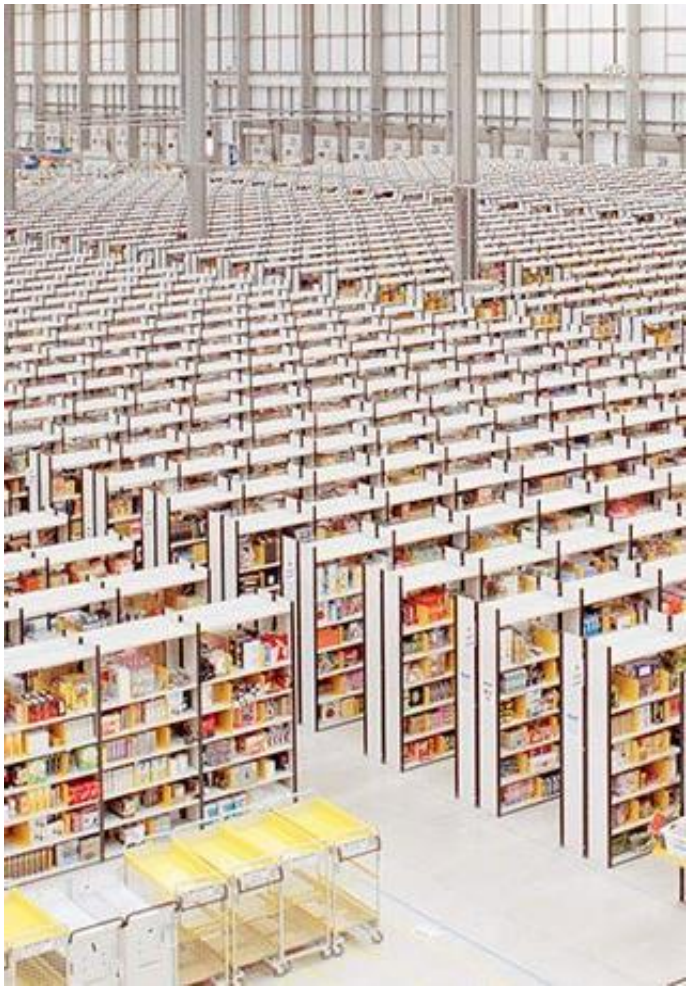
- **Data Warehouse**

- **Objetivo:**
  - Geração de informação e conhecimento.

- **OLAP versus OLTP.**

- **Data Marts:**
  - Unidades temáticas.

- **Base para os BI's.**

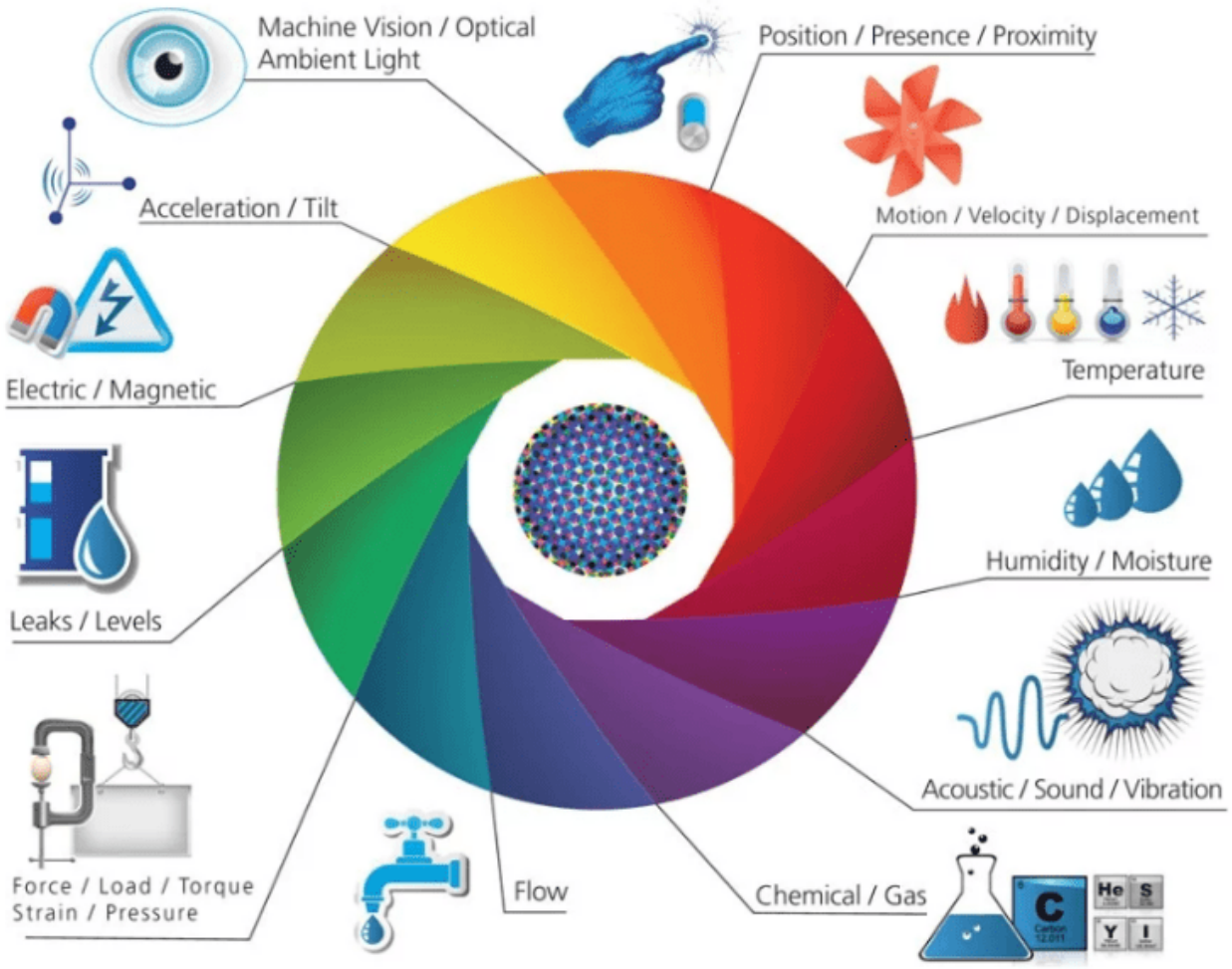


DEFINIÇÕES

Business Intelligence

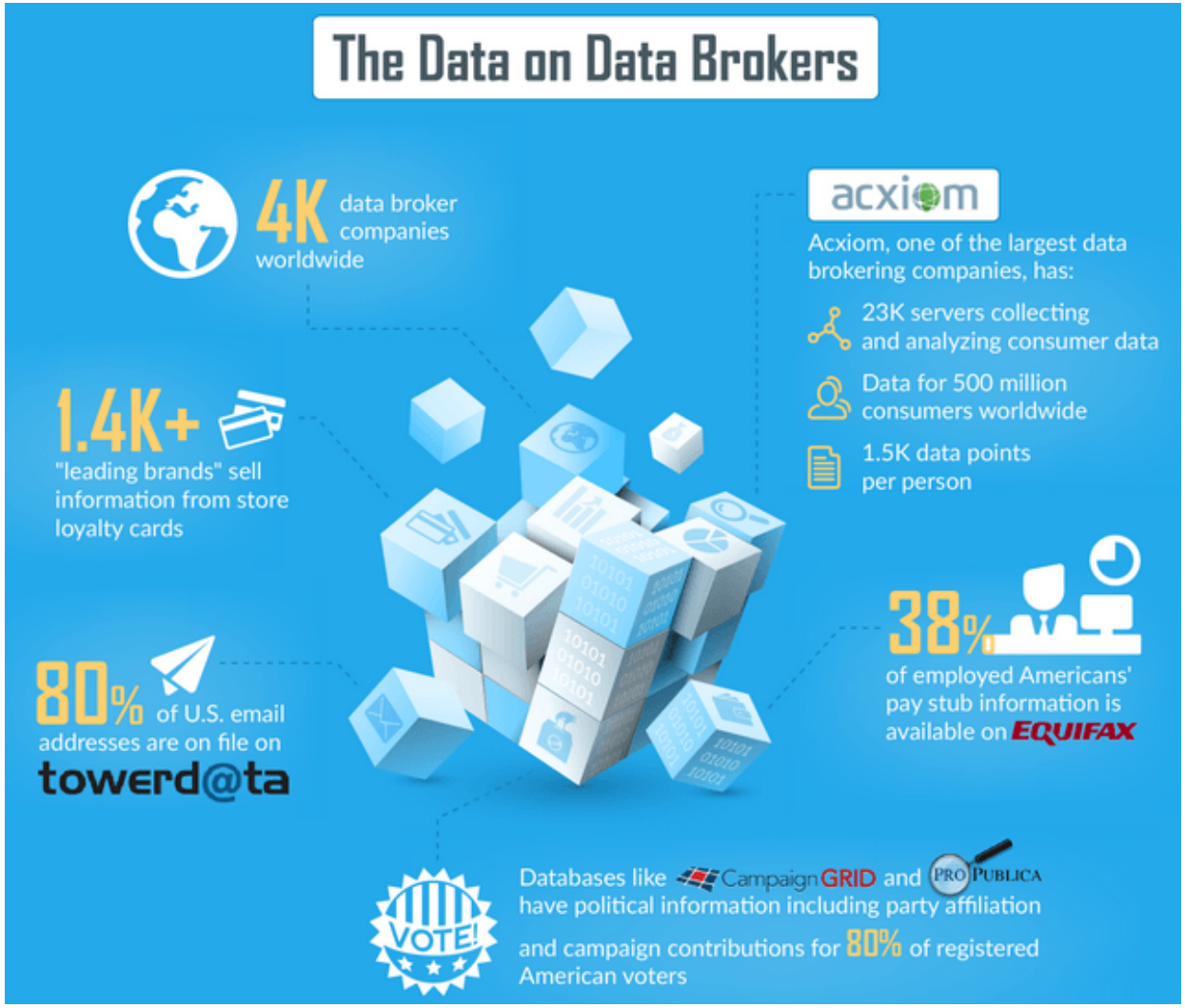
se refere à **metodologia**, **ferramentas**, **técnicas** de produzir dados para apoio à **decisões**.
(Amaral).

PRODUÇÃO POR SENSORES



https://v2com.com/2021/10/04/iot-sensores-inteligentes/

DADOS COMPRADOS

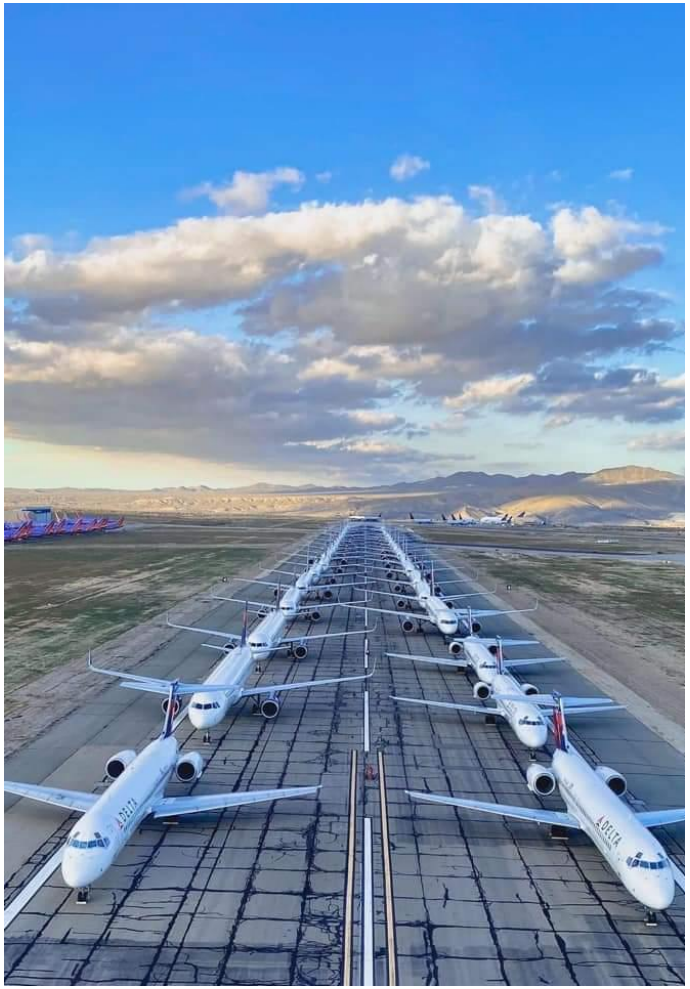


https://www.webfx.com/blog/general/what-are-data-brokers-and-what-is-your-data-worth-infographic/

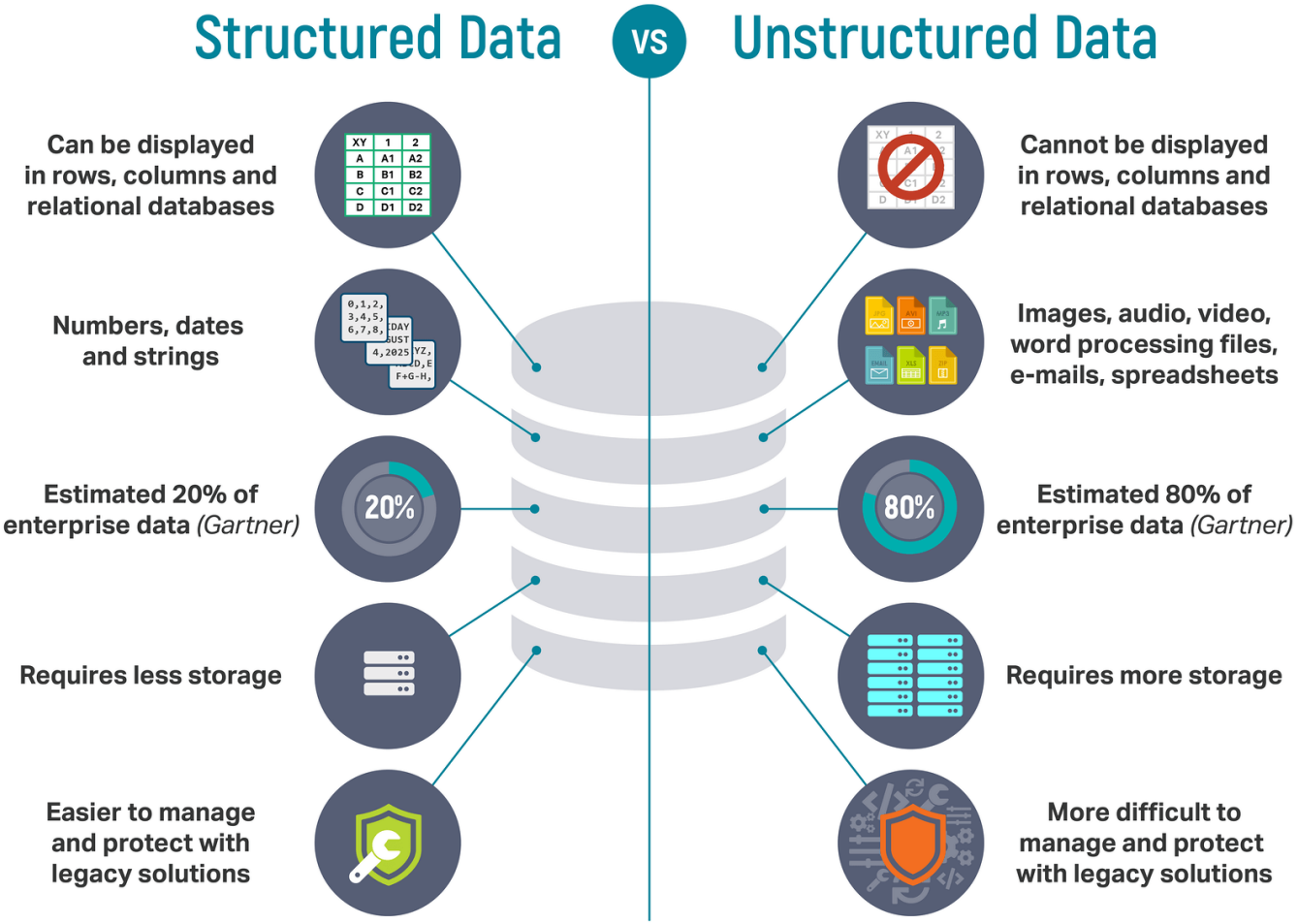
ARMAZENAMENTO

- Dado produzido, armazenado, recuperado, utilizado.

- Segurança, integridade, minimização de redundância, concorrência, otimização de espaço...

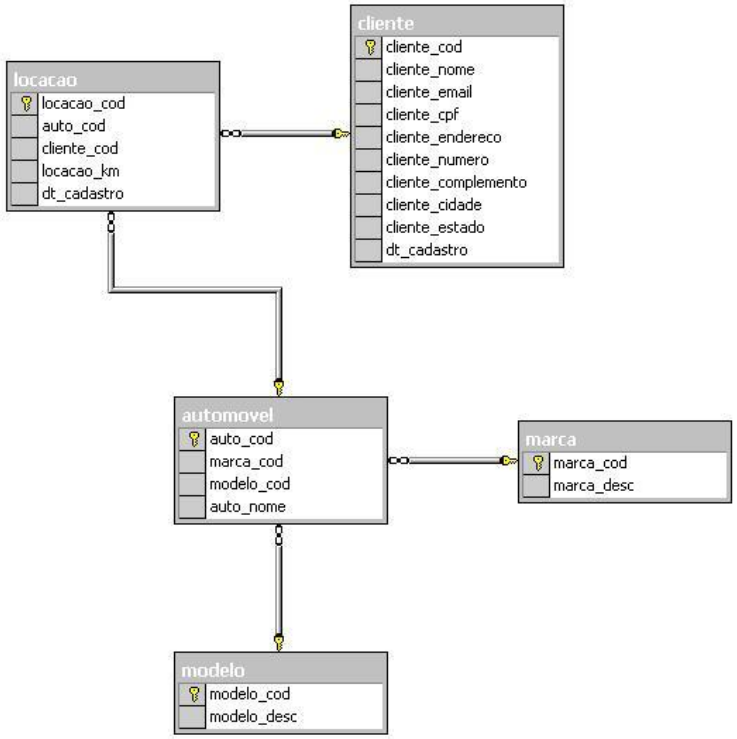


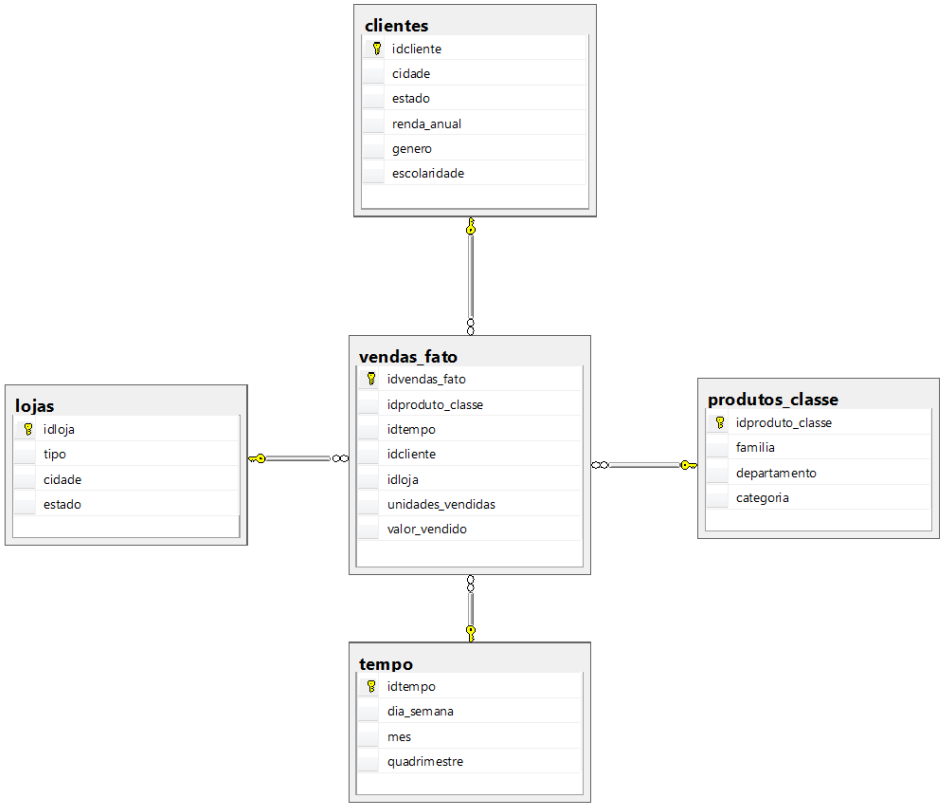
ARMAZENAMENTO



SEMI-ESTRUTURADO

MODELO RELACIONAL X DIMENSIONAL

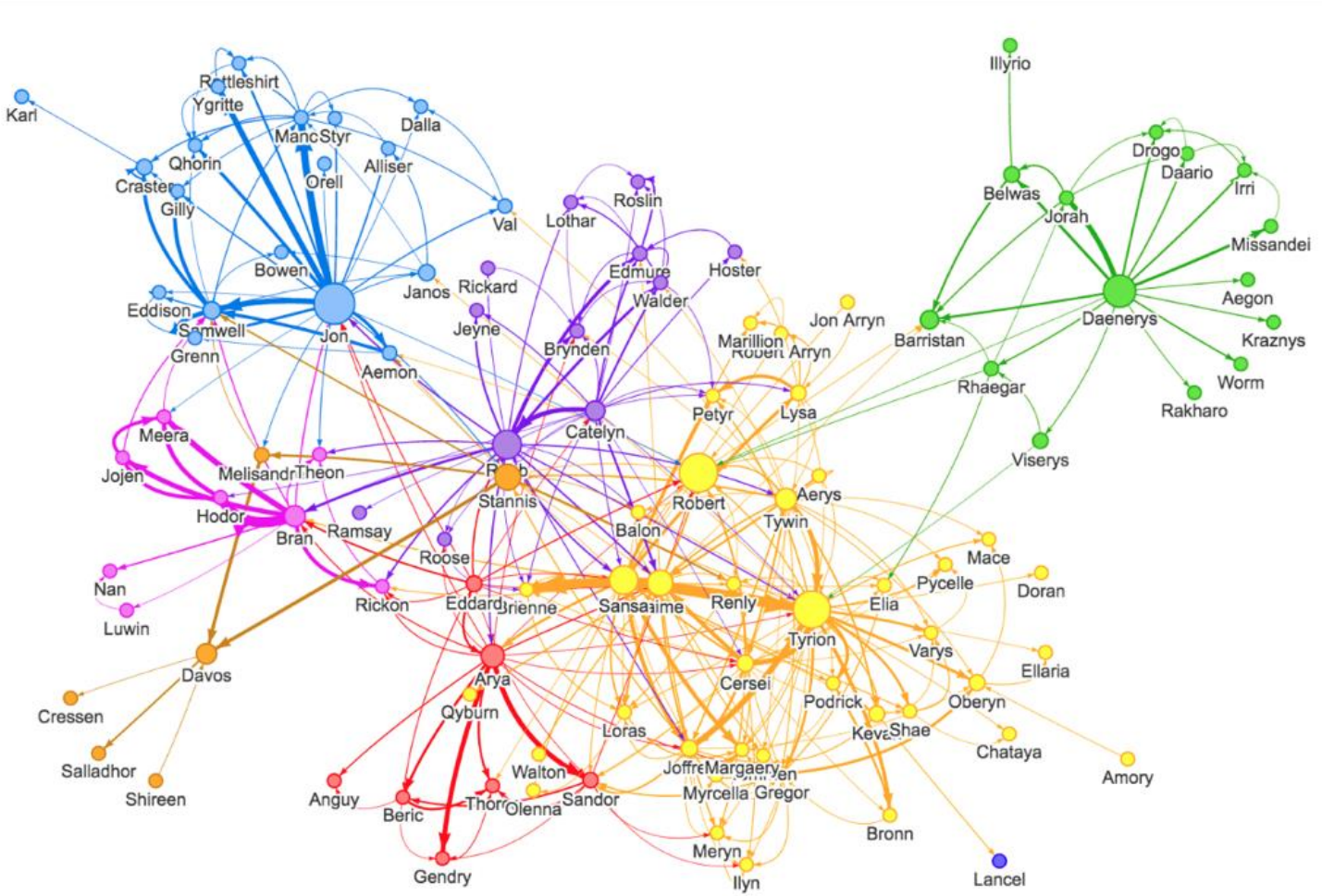




RELACIONAL

DIMENSIONAL

DADOS EM GRAFOS



DADOS ORDENADOS

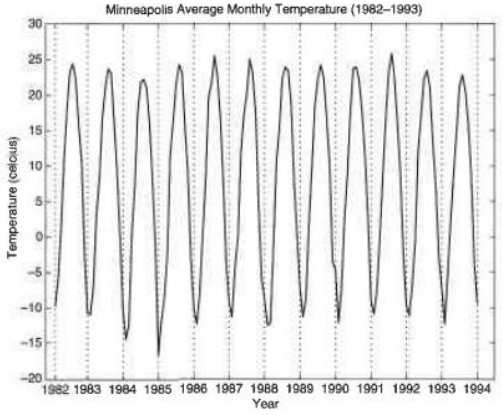
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

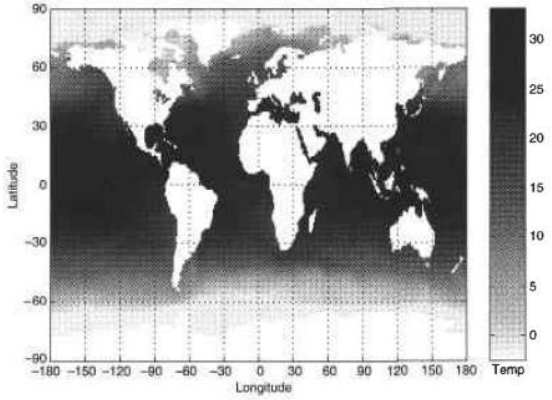
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



(c) Temperature time series.



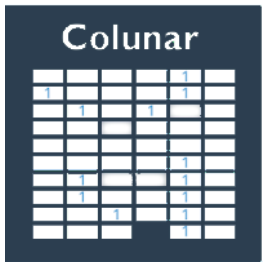
(d) Spatial temperature data.

NoSQL

- Not Only SQL: Bancos de dados não relacionais.

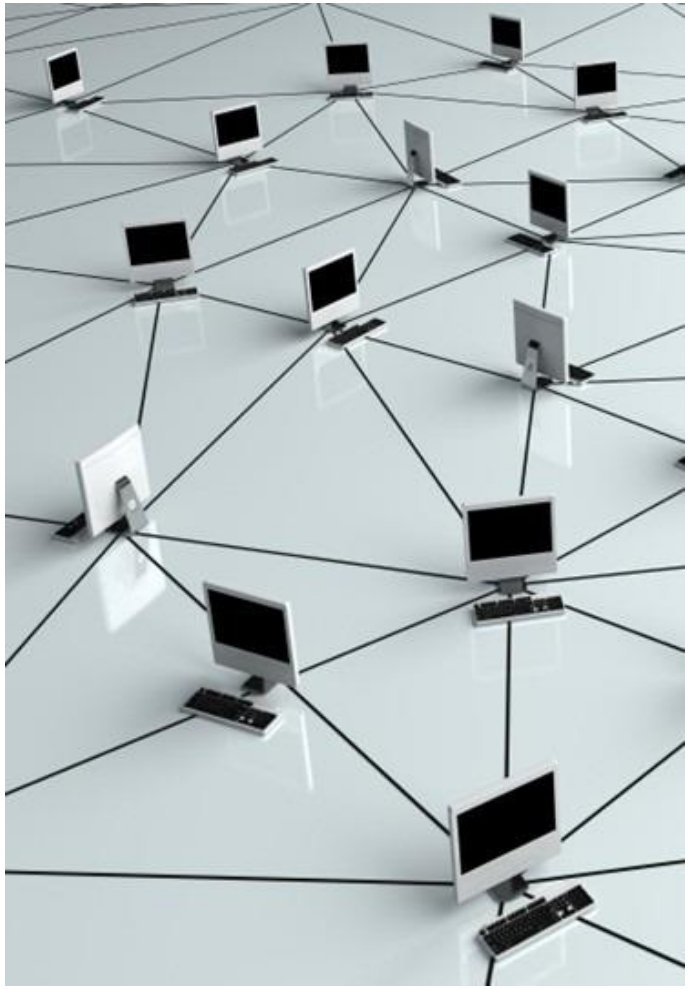
- Agrega grandes volumes de dados.

- Famílias:



ARMAZANAMENTO DISTRIBUÍDO

- **HDFS:**
  - *Hadoop Distributed File System*.
- **Armazena grande quantidade de dados.**
- **Acesso rápido e fácil:**
  - MapReduce.
- **Tolerância a falhas.**

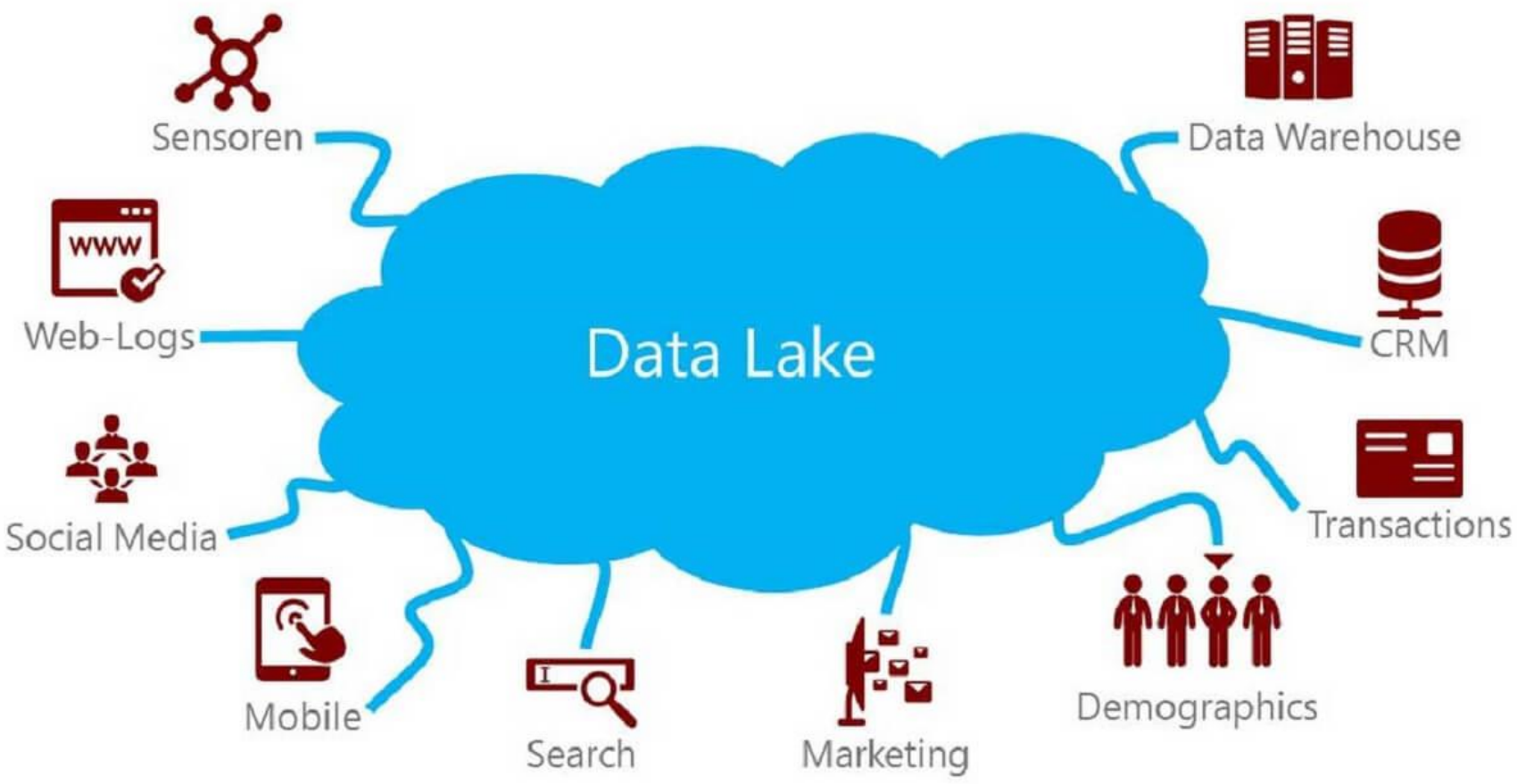


DEFINIÇÕES

Data Lake

repositório centralizado que permite armazenar todos os seus dados **estruturados** e **não estruturados** em **qualquer escala**.
(AWS).

DATA LAKE

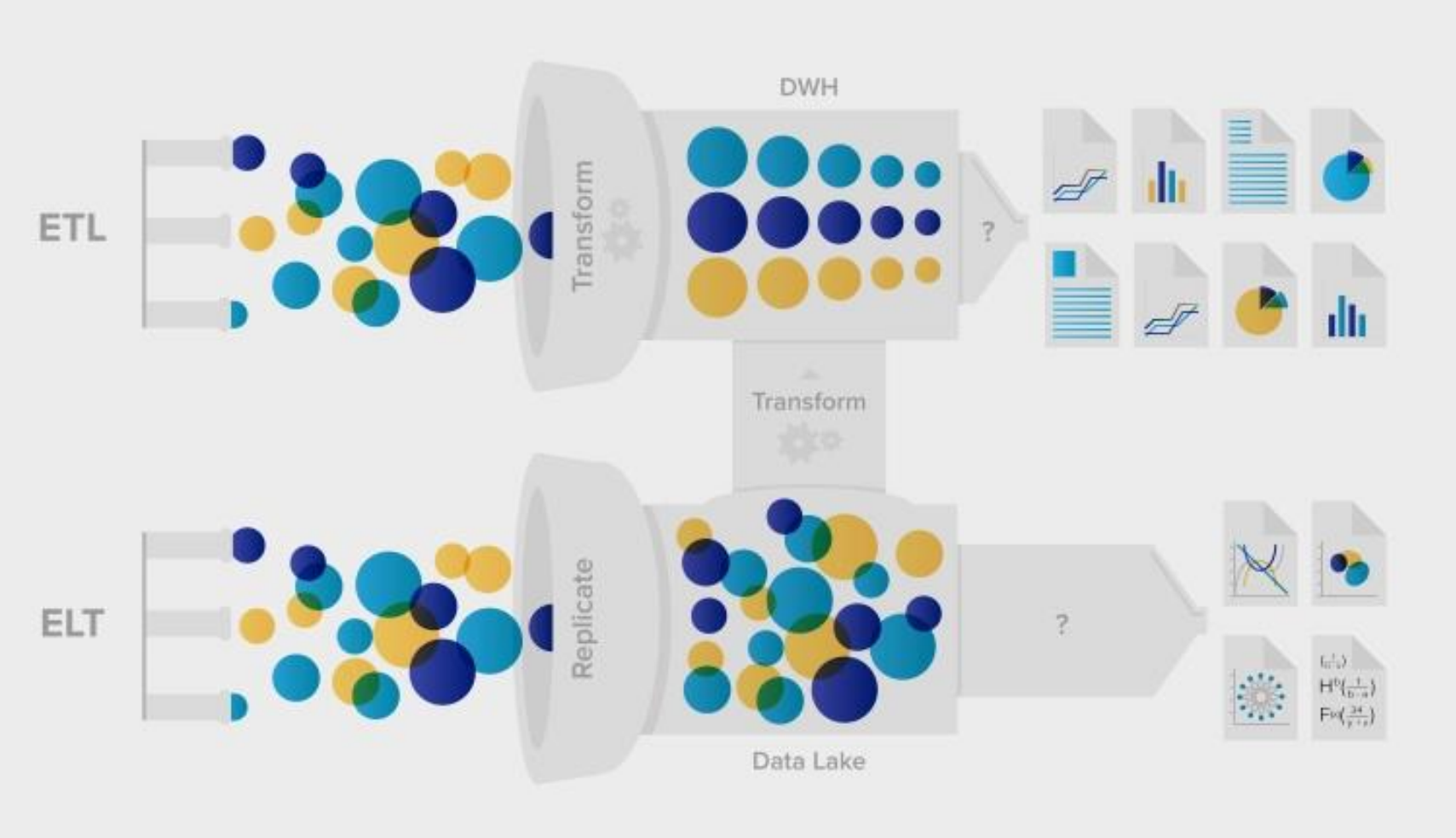

Sensoren
Data Warehouse
Web-Logs
Data Lake
CRM
Social Media
Transactions
Mobile
Search
Marketing
Demographics

Key Differences Between the Data Lake and Data Warehouse

DATA WAREHOUSE	vs.	DATA LAKE
Structured, processed	DATA	Structured/semi-structured/unstructured/raw
Schema-on-write	PROCESSING	Schema-on-read
Expensive for large data volumes	STORAGE	Designed for low-cost storage
Less agile, fixed configuration	AGILITY	Highly agile, configure and reconfigure as needed
Mature	SECURITY	Maturing
Business pros	USERS	Data scientists et al.

Analysis Source: "A Big Data Cheat Sheet: What Marketers Want to Know" by Tamara Dull

TRANSFORMAÇÃO DOS DADOS



ATIVIDADE 3

1. Apresente o dicionário de dados do seu projeto.

2. Descreva em mais detalhes os principais atributos que serão utilizados no seu projeto.

3. Entregue os slides (+2) de sua apresentação.

4. Faça o upload no Google Classroom.