TÉCNICO LISBOA

Bases de Dados

T29 - Data Management

Prof. Daniel Faria

Summary

- Motivation
- Data Management
- Data Governance & Architecture
- Data Quality & Cleansing
- Data Interoperability
- Metadata Management
- Master Data, Reference Data & Documents

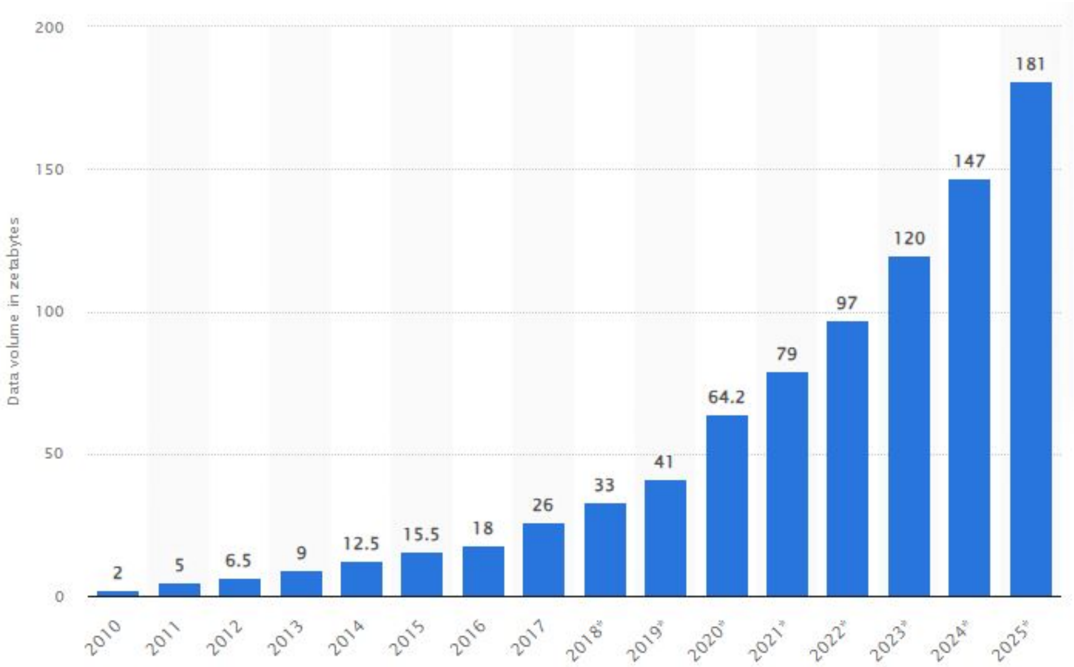
TÉCNICO LISBOA

Motivation

Data Growth

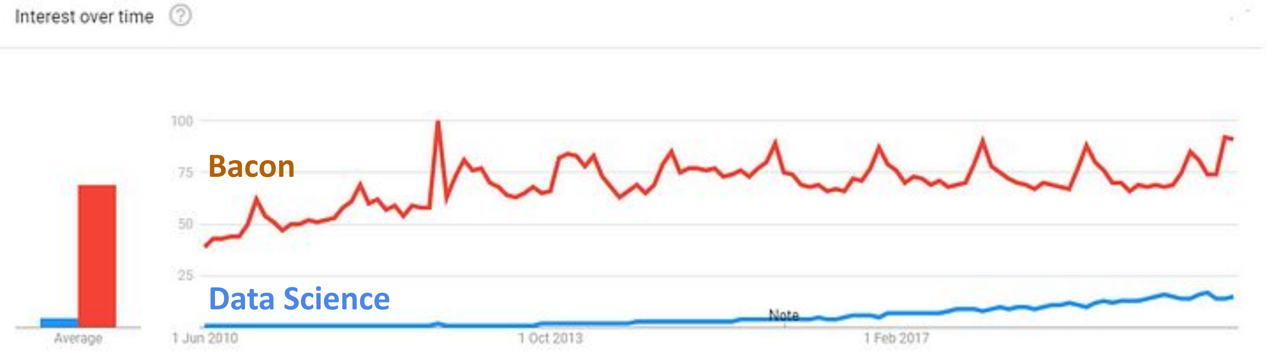
- Human data production is increasing exponentially (literally)
- This is true of virtually every domain of human endeavor
- Even small companies may have to handle Big Data

Worldwide Data Production/Consumption



Data Importance

- Data has also been gaining an increasingly central role in human activities, from business to science, from finance to politics
- This has led to a growing demand for professionals who can explore and exploit large quantities of data: data scientists



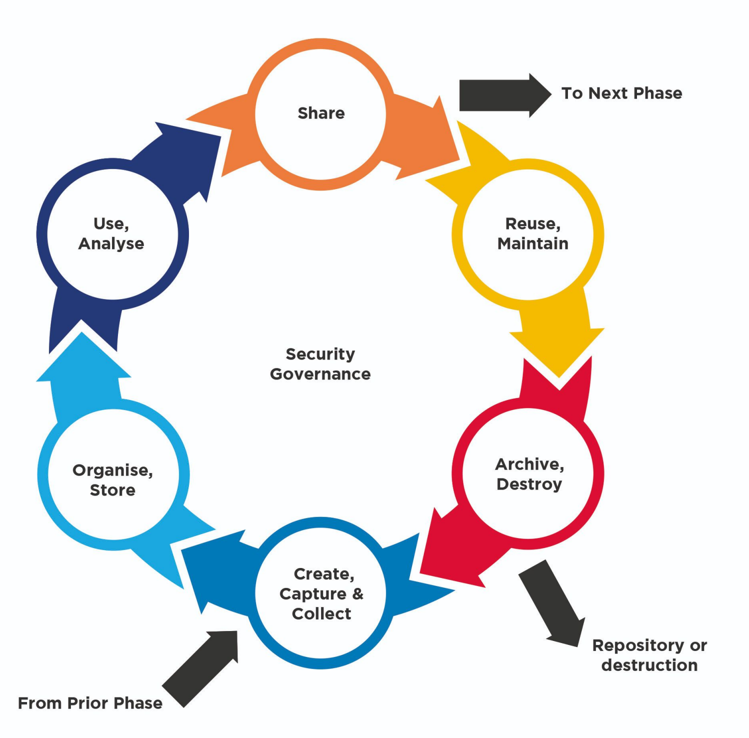
Not quite… but still notable increase in interest

Data Ubiquity

- Virtually every activity produces and/or consumes data
- Every decision should be backed by data
- Data is dispersed and heterogeneous
- There are heterogeneous needs for it but also a need for integration

Data Lifecycle

- The creation, maintenance and use of data involves many processes
- These can be grouped into stages which in turn form a lifecycle
- Effective and efficient use of data hinges on concerted management of all processes across the data lifecycle

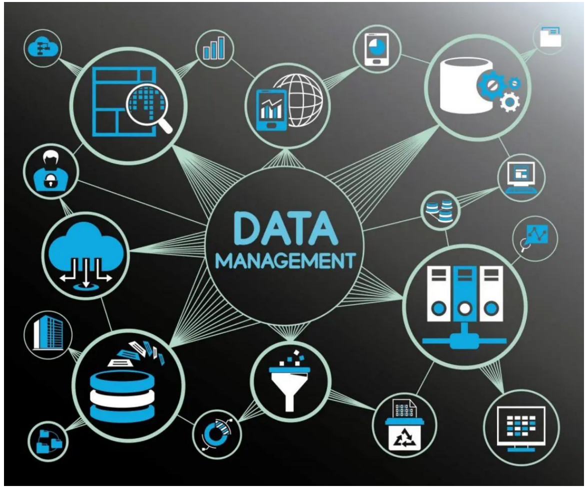


Data Management

Data Management

- Data management encompasses all processes pertaining to the creation, maintenance, and use of data in an organization, across the data lifecycle

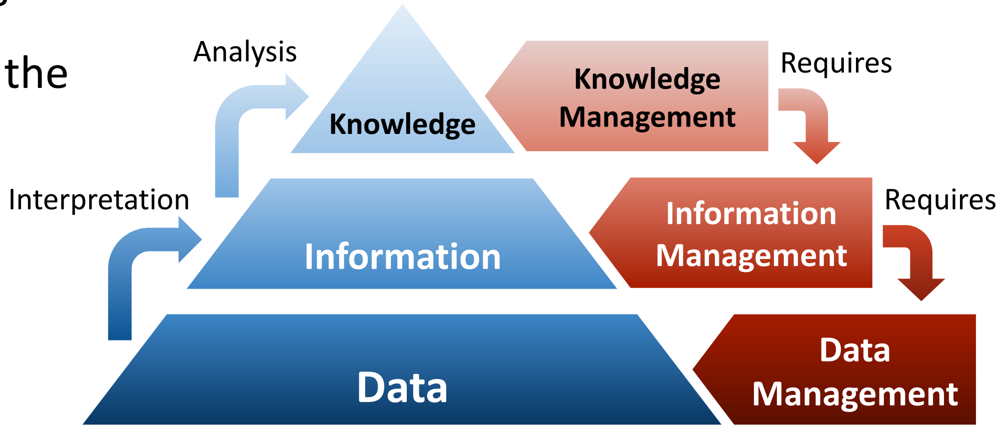
- It aims to increase effectiveness and efficiency of processes requiring or involving data, by ensuring that relevant data is available, accessible, and usable on-demand (by authorized users)



- It also aims to ensure security and reliability by protecting data assets from unauthorized access as well as from software and hardware failure

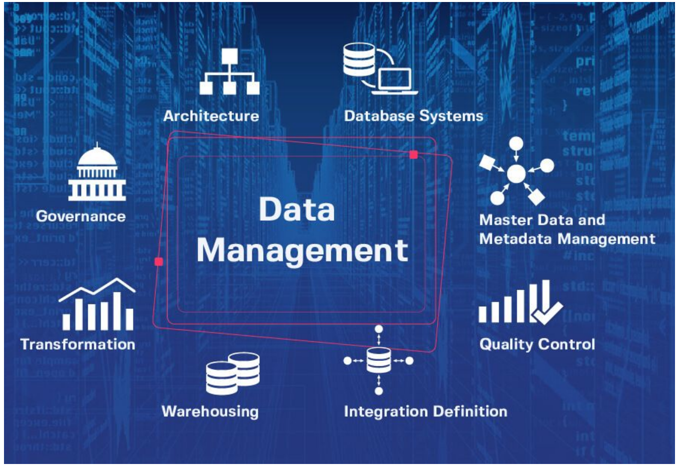
Data *vs.* Information *vs.* Knowledge

- The line is blurry between **data management** and **information management** and between the latter and **knowledge management**
    - Information management is concerned with the creation, maintenance, and use of information in an organization
    - Knowledge management is the equivalent for knowledge
- Some topics are common between data and information management, but many are exclusive to data management



Data Management Topics

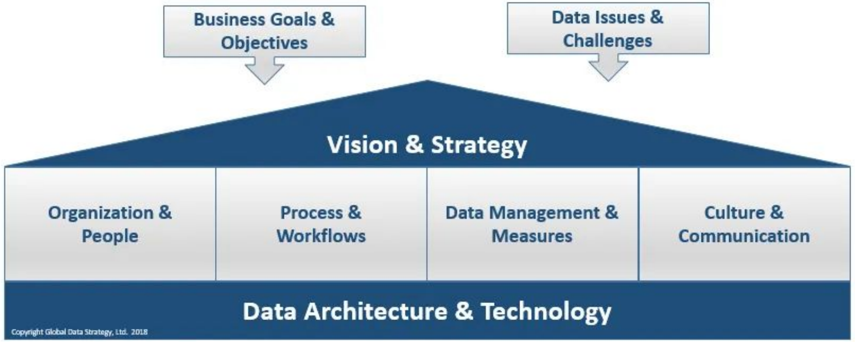
- Data governance & architecture
- ✓ Data modeling
- ✓ Database & storage management
- Data security & privacy
- Data quality & cleansing
- Data interoperability
- ✓ Data integration, warehousing & analytics
- Metadata management
- Master data & reference data management
- Document management



Data Governance & Architecture

Data Governance

- The processes, policies, guidelines and responsibilities for administering an organization's data
  - In compliance with policy and/or regulatory obligations

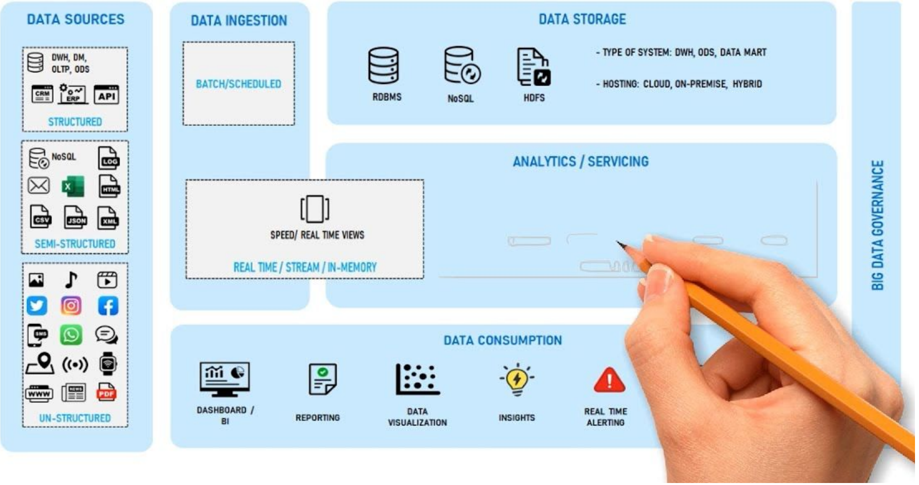


Copyright Global Data Strategy, Ltd. 2018

  - Providing the vision & strategy needed to ensure that data is managed as an asset and transformed into meaningful information
  - Ensuring that high quality data exists throughout the complete lifecycle of the data: availability, usability, consistency, integrity, security, standard compliance
  - Encompassing the people, processes, and IT

Data Architecture

- The models, policies, rules, and standards that govern which data is collected and how it is stored, arranged, integrated, and put to use in data systems and in organizations
  - Focusing on technology and infrastructure design



Questions To Answer

- **What** data will be collected/produced by the organization?
- **What** data types and formats?
- **Why** will it be collected/produced?
- **How** will it be collected/produced, processed, integrated, analyzed?
- **When** will it be collected/produced, needed, deleted?
- **Where** will it be stored, accessed, backed-up?
- **Who** will be involved in collection/production, processing, integration, analysis?
- **Who** owns the data, who can access it, who is responsible for each data process?

Data Stewards & Custodians

- **Data stewards** administer an organization’s data assets, ensuring the enactment of its data governance policies concerning the content and structure of the data
    - They are responsible for ensuring the quality and fitness for purpose of the data assets, including the metadata for those assets
    - They can participate in the development and implementation of data assets, as well as in the development of data governance policies
- **Data custodians** are responsible for the safe custody, transport, and storage of an organization’s data assets, ensuring the enactment of the technical aspects of its data governance policies (i.e. mostly the data architecture)

Data Quality & Cleansing

Data Quality

- Data is considered high quality if it:
  - Is fit for purpose, i.e. it can be adequately used in operations, decision making and planning, or by customers (meeting or exceeding expectations)
  - Is accurate, i.e. it correctly represents the "real-world" construct to which it refers
  - Complies with applicable standards with respect to structure, syntax, units, etc

Data Quality

Dimensions contemplated often include:

- **Completeness:** all necessary attributes are present
- **Uniqueness:** minimal or no data duplication
- **Timeliness:** available (up-to-date) when required
- **Validity:** conforms with rules and standards, within expected intervals
- **Accuracy:** agreement with verifiable source of “truth”
- **Consistency:** agreement between data records (in different datasets) for the same object

COMPLETENESS
CONSISTENCY
UNIQUENESS
Data Quality Dimensions
ACCURACY
TIMELINESS
VALIDITY

Data Cleansing

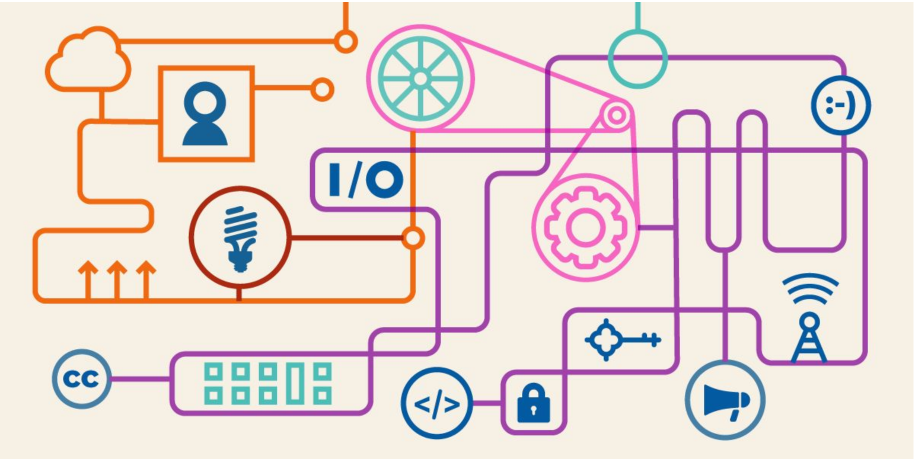
- The process of identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting them, involving:
    - Quality checks (e.g. identify missing values)
    - Deduplication to improve uniqueness
    - Data analysis to identify anomalies (e.g. outliers)
    - Standardization (e.g. conversion of units, formats)
    - Data normalization to ensure data are statistically comparable and easily interpretable

Data Deduplication
Quality Check
Data Analysis
Data Cleansing
Data Normalization
Data Standardization

TÉCNICO
LISBOA

Data Interoperability

Interoperability

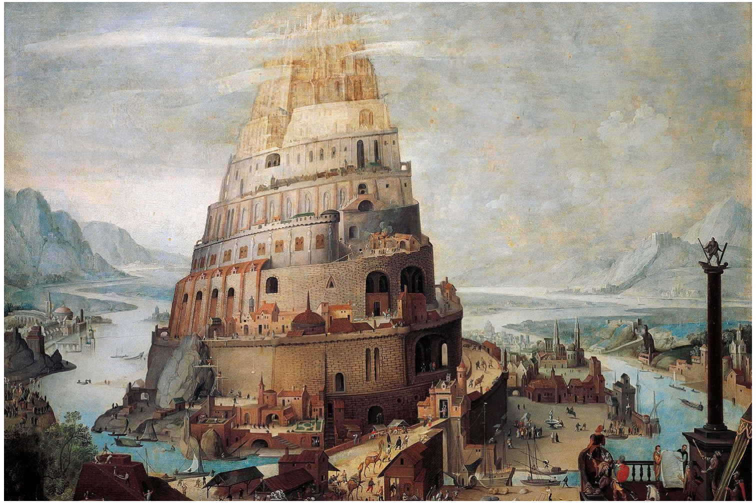


- The characteristic of a product or system to work with other products or systems
- Interoperability always relies on the adoption of **standards**
- **Standard:** a set of rules and definitions that specify how to carry out a process, how to produce a product, or what characteristics a product or object must conform to

Data Interoperability

- The characteristic of data to be readily exchangeable and usable across information systems, data workflows, analysis tools, etc

- **Syntactic Data Interoperability:** ability of data to be read by and processed by different systems (e.g. standard file formats, standard data organization)



- **Semantic Data Interoperability:** ability of data to be interpreted by different systems (e.g. controlled vocabularies, standard **metadata**)

TÉCNICO LISBOA

Metadata Management

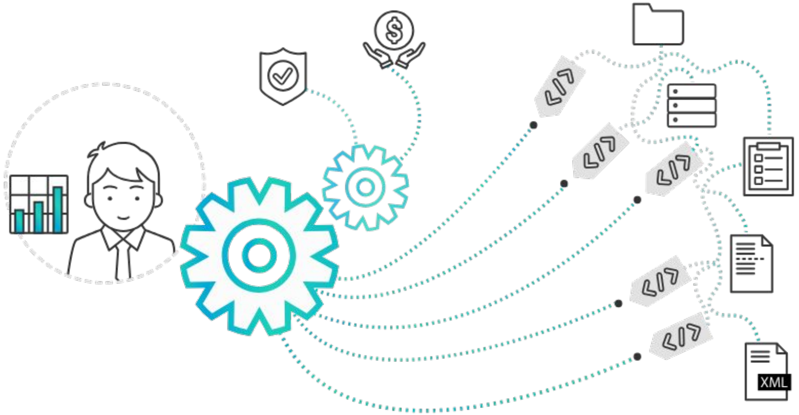
Metadata

- Metadata is data about data
  - Describing its provenance and all processes it underwent until its present form
  - Summarizing its content and scope
  - Describing its structure and organization
  - Stipulating access and usage restrictions
- It can provide context and interpretability, facilitate retrieval, and enable interoperability

Not all metadata is digital!!!

Metadata Management

- The process of managing the metadata associated with the data assets of an organization across the lifecycle of those assets



- End-to-end process that starts with governance policies and processes delineating what metadata should be collected at each stage of the data lifecycle, how it should be structured, and what vocabularies should be used, as well as designing the technical and human infrastructure to support metadata management

Master Data, Reference Data & Documents

Master Data & Reference Data

- **Master Data:** data about the business entities that provide context for business transactions
  - E.g. individuals and organisations and their roles (customers, suppliers, employees, etc.), products, locations
- **Reference Data:** data used to classify or categorize other data
  - E.g. units of measurement, country codes, calendar structure and constraints, product classification
- Both types of data provide context to business transactions, and the **quality** of both is critical to all aspects of an organisation, operational and analytical

Document Management

- Documents or files represent a significant part of the information collected/produced by an organization (e.g. receipts, procedures, policy statements)
- A centralized document management system is essential to minimize storage costs (by avoiding duplication of documents across PCs), enable versioning and metadata management, and ensure security, findability and accessibility
  - Cloud-based solutions have become the *de-facto* standard

TÉCNICO LISBOA