Mestrado em Engenharia Informática
KE

2022/2023

BigData

Programa

Large Scale Data Handling

1. Characteristics
2. Modeling and Management
3. MapReduce
4. Hadoop

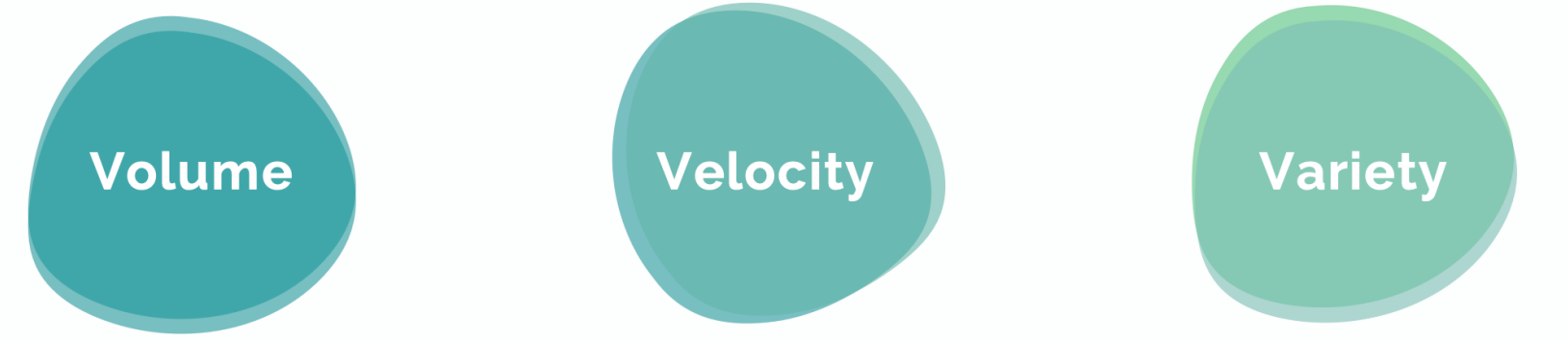
BIG DATA CONCEPT

The concept of Big Data remains so far a relative term with regard to the boundary between what is and is not considered Big Data.

For a company such as Google, the concept and size of Big Data is much different from that assumed for a medium-sized company.

BIG DATA CONCEPT

The most accepted definition was given by Douglas Laney. Laney observed that Big Data grew in three different dimensions:



BIG DATA CONCEPT

However, other authors have crossed these characteristics by adding several other V's to this definition, such as:

- Value,
- Veracity,
- Visualization,
- Viscosity,
- Virality,
- among others.

The 4th most consensual V's is undoubtedly the **veracity**.

BIG DATA CHARACTERISTICS

VOLUME

The volume of data gives the large amount of data, mostly described in several petabytes or even more. However, not even this definition is consensual among the authors, since the definition depends on the type of data being analyzed.

BIG DATA CHARACTERISTICS

VELOCITY

The velocity concerns both the rate of data generation and the speed of analysis they require. Big Data Velocity deals with the speed at which data flows in from sources.

BIG DATA CHARACTERISTICS

VARIETY

The variety of data has increased exponentially due to the diversity of collection sources. Data can have several organizations and reach the collection point in a structured, semi-structured or even unstructured way. In addition, data formats must be taken into account.

BIG DATA CHARACTERISTICS

VERACITY

Veracity encompasses the reliability inherent in some sources of data collection. For example, information taken from a social network cannot be given the same relevance as information taken from hospital software.

Why Is Data Modeling Necessary?

BIG DATA MODELING

Large amounts of data imply a system or method to keep everything in order.

The process of sorting and storing data is called "data modeling".

A data model is a method by which we can organize and store data.

BIG DATA MODELING

Proper models and storage environments offer the following benefits to large data:

- **Performance,**
- **Cost**,
- **Efficiency**, and
- **Quality**.

BIG DATA MODELING

Performance: Ensures fast query and reduces I/O output.

BIG DATA MODELING

Cost: Significantly reduces data redundancy, reducing storage and computing costs for the
large data system.

BIG DATA MODELING

Efficiency: They greatly improve the user experience as well as the efficiency of data use.

BIG DATA MODELING

Quality: They make data statistics more consistent and reduce the possibility of
computing errors.

6 TIPS FOR MODELING BIG DATA

01 DON'T IMPOSE TRADITIONAL MODELING

02 DESIGN A SYSTEM, NOT A SCHEMA

03 LOOK FOR BIG DATA MODELING TOOLS

04 FOCUS ON DATA THAT IS CORE TO YOUR BUSINESS

05 DELIVER QUALITY DATA

06 LOOK FOR KEY INROADS INTO THE DATA

BIG DATA MANAGEMENT

Big Data Management is a set of practices that promotes the

- **collection,**
- **organization,**
- **administration and**
- **interpretation**

of large volumes of data.

BIG DATA MANAGEMENT

- **Adequacy**

Ability to analyze a large amount of information, structured or not, allows the detection and correction of errors in stored information

BIG DATA MANAGEMENT

- **Integration**

Ability to filter and classify data so that it can later be handled assuming a standardized structure.

BIG DATA MANAGEMENT

- **Migration**

Ability to move data from one environment to another quickly and conveniently.

BIG DATA MANAGEMENT

- **Management**

Ensure the availability and security of data, ensuring that it follows all the organization's policies and standards.

BIG DATA MANAGEMENT: ADVANTAGES

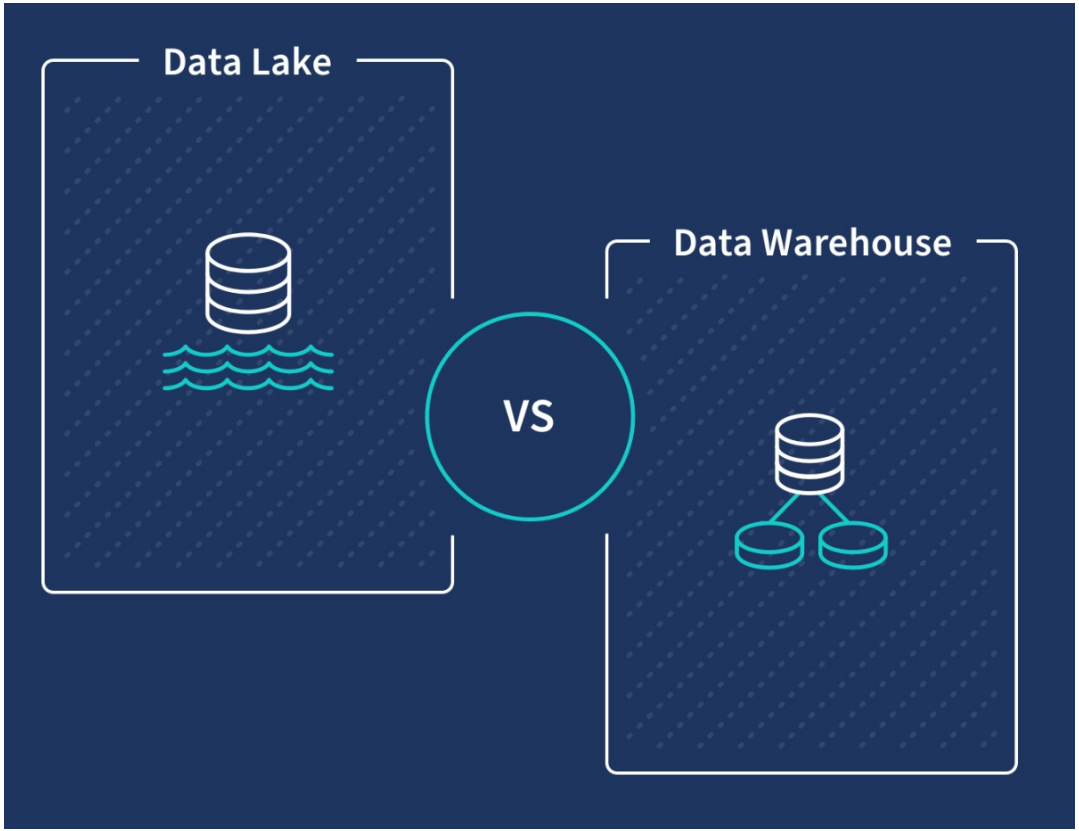
Increase in company revenue

More accurate decision making

Strategy improvement

Team productivity and efficiency

DATA WAREHOUSE VS DATA LAKE



DATA LAKE

What is a Data Lake?

A data lake is a repository that stores all of organization's data — both structured and unstructured. Think of it as a massive storage pool for data in its natural, raw state (like a lake).

Mestrado em Engenharia Informática
KE

2022/2023