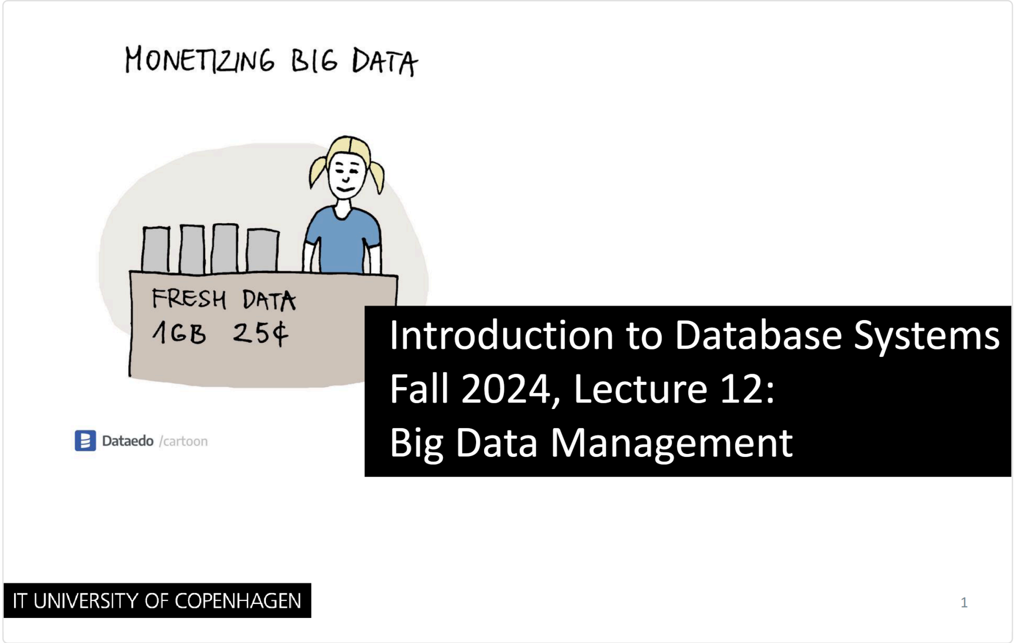
# Big Data Management



Introduction to Database Systems
Fall 2024, Lecture 12:
Big Data Management

IT UNIVERSITY OF COPENHAGEN

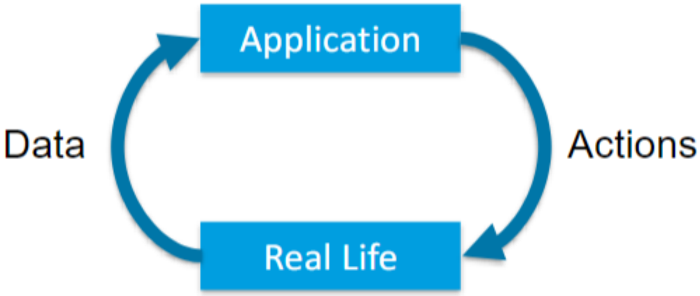
# The Five V's of Big Data

## Volume

**The amount of data generated, stored and processed** - Data management systems must scale to a given amount of data volume.

## Velocity

**The speed at which data is generated, collected and processed** - Requirement on the system to process data at a given speed.



## Variety

**The diversity of data types and sources** - Challenge of managing different types of data. Choose data management system for your use case.

Data can come in 3 structed ways:

- Structed
  - Relational Database Management Systems (Table structure)
- semi-structed
  - JSON, Logs, Graphs
- Unstructured formats
  - Photos, videos, music etc.

## Veracity

**The accuracy and trustworthiness of data** - Data quality varies, and not at all collected data can be trusted. It's essential to filter, clean and validate data to ensure accurate insights.

## Value

**The ability to turn data into insights for decision making and creating business value** - Bid data is only beneficial if it delivers insights.

This could be for new science discoveries, to serve peoples needs and optimize business operations and earn money.

# Big Data Processing

## ETL

Extraction, Transformation and Loading (ETL) of data (from OLTP, Docs, Feeds, IoT etc.)

## MapReduce

2003 by google

# Apache Spark

Apache Spark a new programming paradigm centered on a data structure called the resilient distributed dataset, or RDD, which can be distributed across a cluster of machines and is maintained in a fault-tolerant way.

- Spark can process unstructured data (See also NoSQL)
- SQL is usable with spark
- spark can handle large amount of data
- RAM
  Improvement to MapReduce

# Data Warehouses vs. Data Lakes

Principles of database management the practical guide to storing, managing and analyzing big and small data - PDF Room, page 1.256

| | Data Warehouse | Data Lake |
|---|---|---|
| Data | Structured | Often unstructured |
| Processing | Schema-on-write | Schema-on-read |
| Storage | Expensive | Low cost |
| Transformation | Before entering the data warehouse | Before analysis |
| Agility | Low | High |
| Security | Mature | Maturing |
| Users | Decision-makers | Data scientists |

## Data Warehouses

A **Data warehouse** is a centralized repository that stores structured data (database tables, Excel sheets) and semi-structured data (XML files, webpages) for the purposes of analysis.

## Data lakes

A **data lake** a large data repository that **holds data in their raw format**, which can be structured, unstructured, or semi-structured.