Introduction to Big Data

Definition

> **Big data** is a field that treats ways to
> - analyze,
> - systematically extract information from,
> - or otherwise deal with
>
> data sets that are **too large or complex** to be dealt with by traditional data-processing application software.

≈ data does not fit into the RAM

source: https://en.wikipedia.org/wiki/Big_data

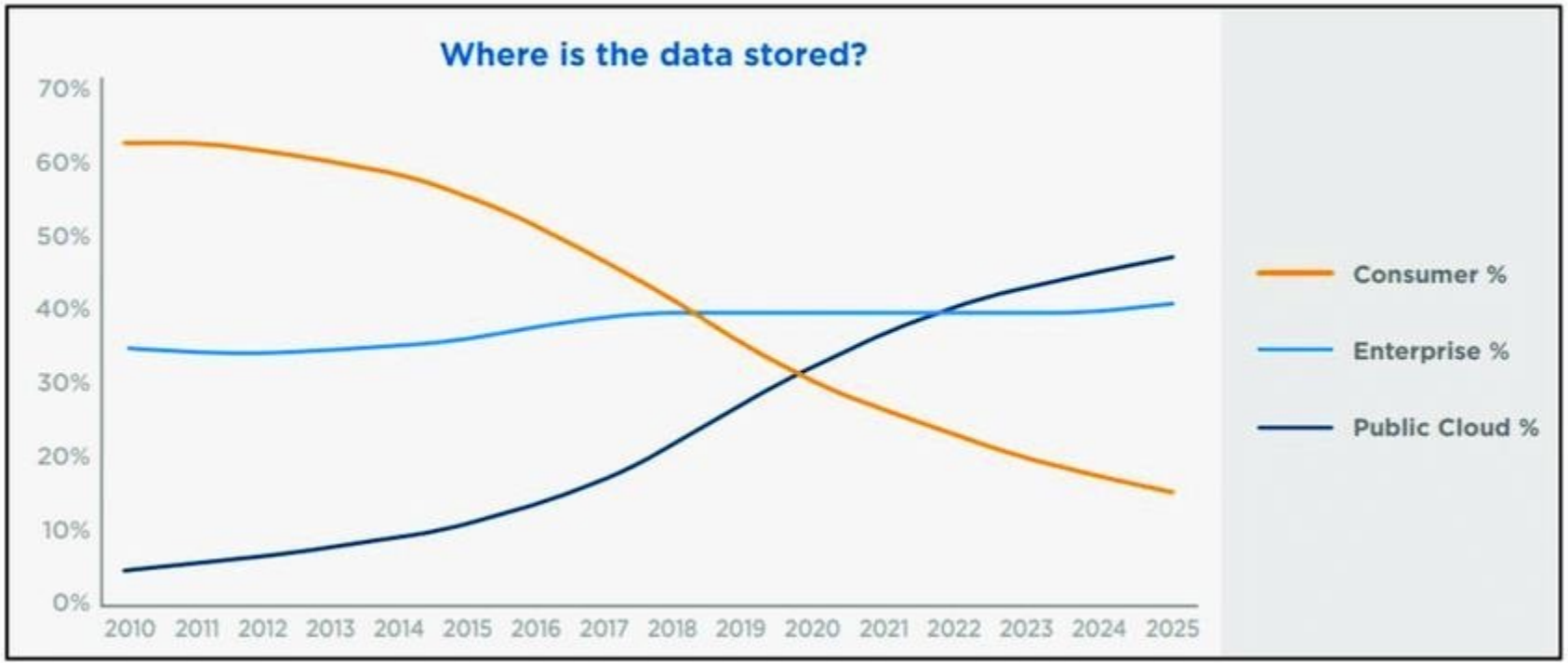
Big data statistics and economic impact

- In 2020, every person will generate 1.7 MB in just a second
- Internet users generate about 2.5 EB (10^{18}) of data each day
- Big data and business analytics market is set to reach $274 billion by 2022 (IDC source)
- In 2019, Big Data market grew by 20% from previous year
- 91% of organizations are investing in Big Data and AI
- Using Big Data, Netflix saves $1 billion per year on customer retention

Big data driving factors

- Digital
- Smartphones
- Social networks
- Internet of Things (IoT)

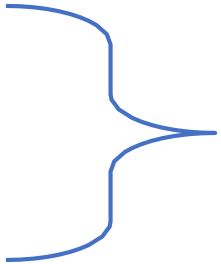
Where is data stored?



source: Reinsel, D.; Gantz, J.; Rydning, J. Data Age 2025: The Digitization of the World from Edge to Core; IDC Analyze the Future: Framingham, MA, USA, 2018; pp. 1–28

Big data 5Vs

- Volume
- Variety
- Velocity
- Veracity
- Value



original 3Vs by Gartner, 2012

"Big data is high **volume**, high **velocity**, and/or high **variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Increasing Volumes of data

- The quantity of generated and stored data.
- The size of the data determines the value and potential insight, and whether it can be considered big data or not.
- The size of big data is usually larger than terabytes and petabytes.

Increasing Variety of data types

- The type and nature of the data.
- The earlier technologies like RDBMSs were capable to handle **structured data** efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies.
- The Big Data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed(velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, and sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

Increasing Velocity at which data changes

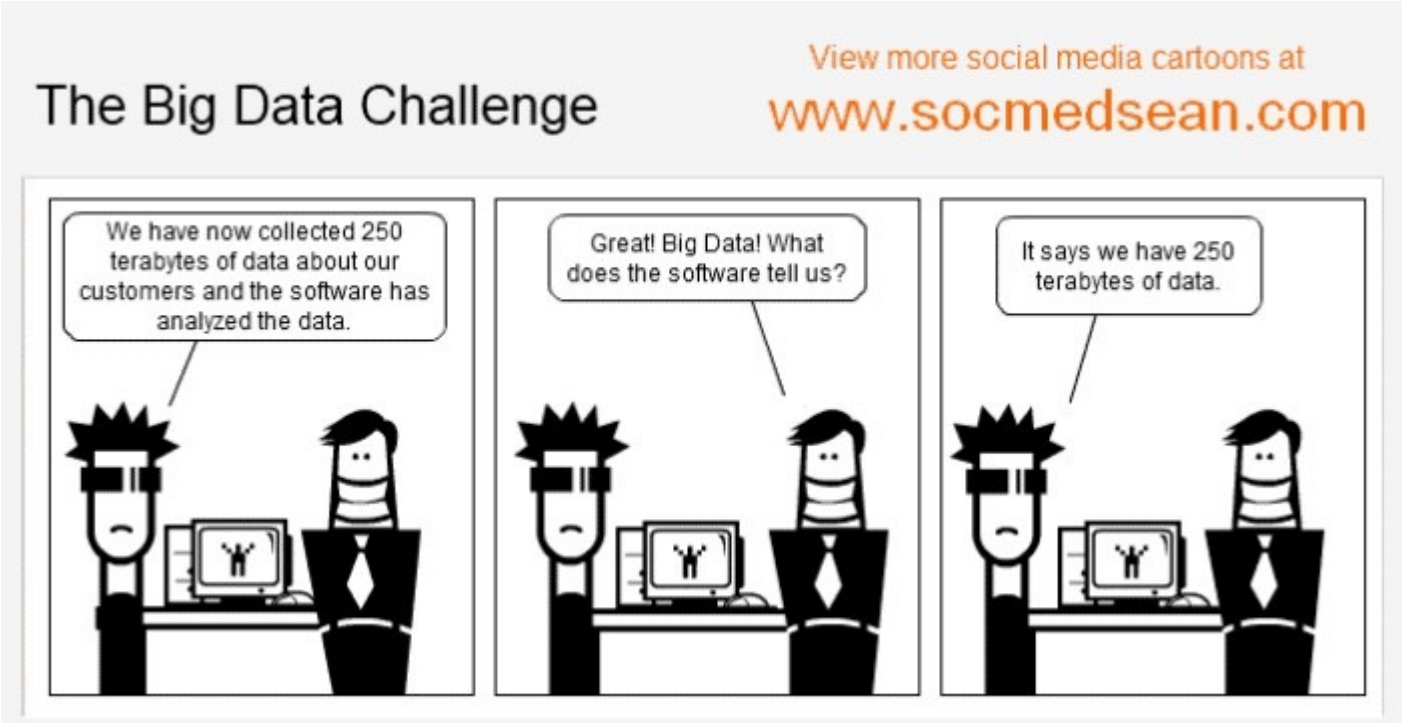
- The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- Big data is often **available in real-time**.
- Compared to small data, big data is produced more continually.
- Two kinds of velocity related to big data are the frequency of **generation** and the frequency of **handling**, recording, and publishing.

Veracity = data quality

- The data quality of captured data can vary greatly, affecting the accurate analysis.
- With many forms of big data quality and accuracy are less controllable

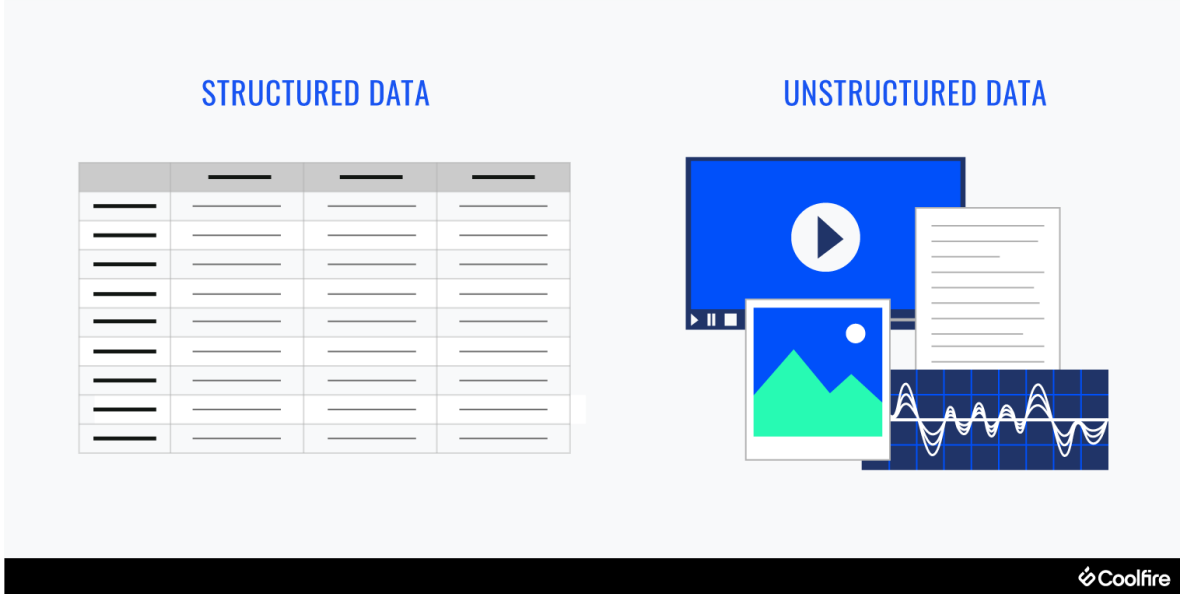
Value

- The utility that can be extracted from the data.

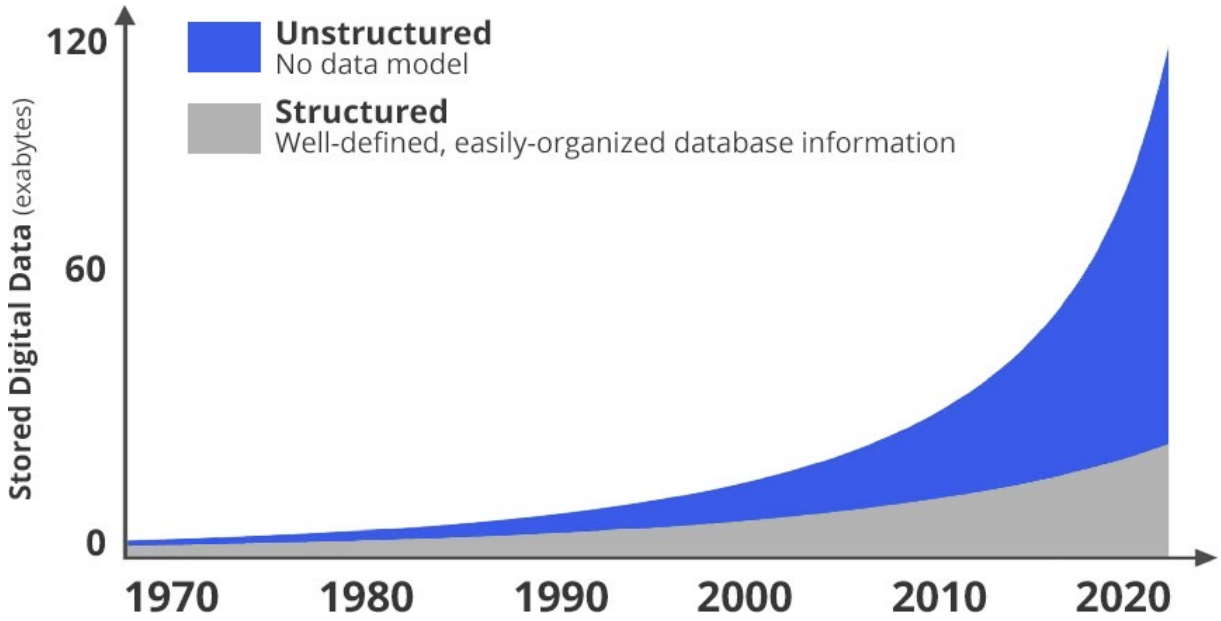


Structured vs unstructured data

- Structured data ≈ data organized in tables
- Unstructured data ≈ everything else (e.g. text, images, video, …)

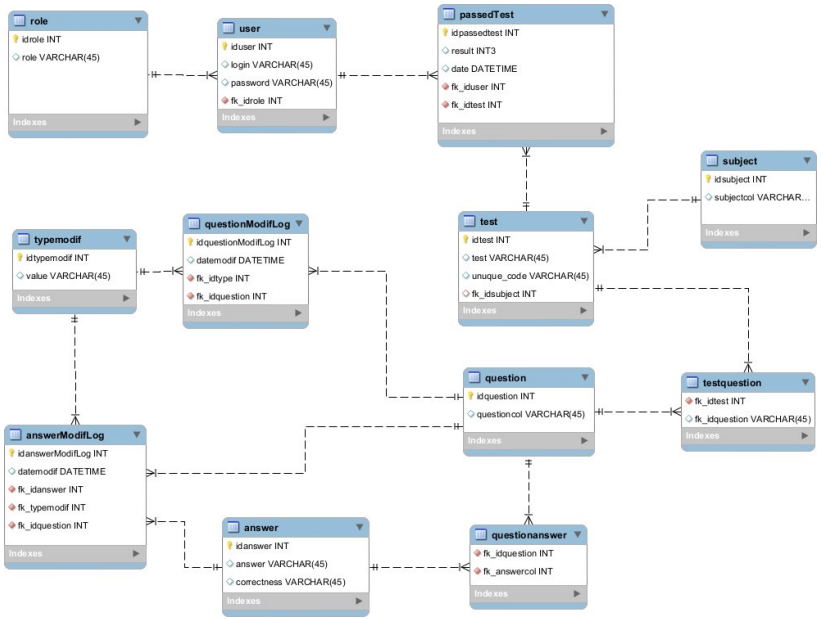


Structured vs unstructured data (2)

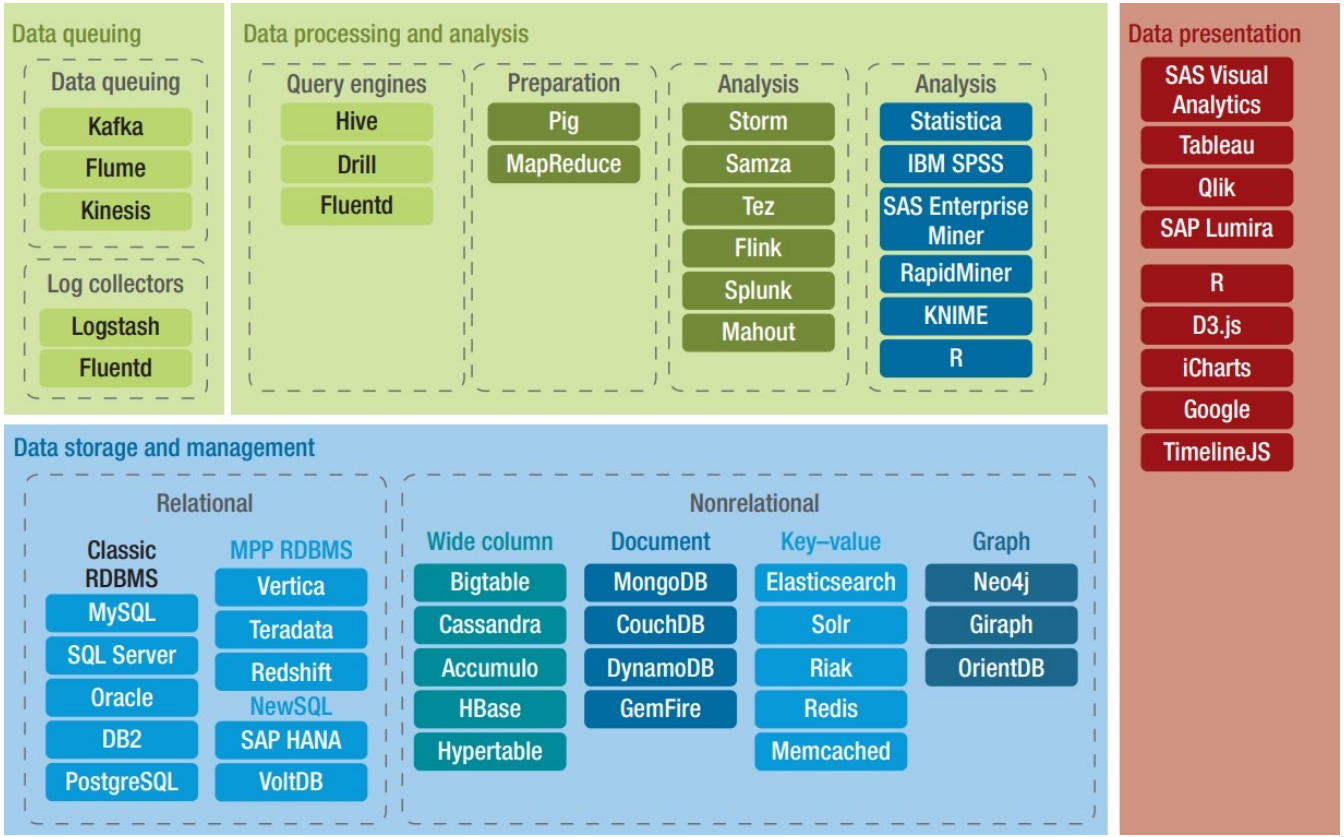


Traditional RDBMS

- RDMBS = Relational Data Base Management Systems
- Gold standard to organize and store structured data

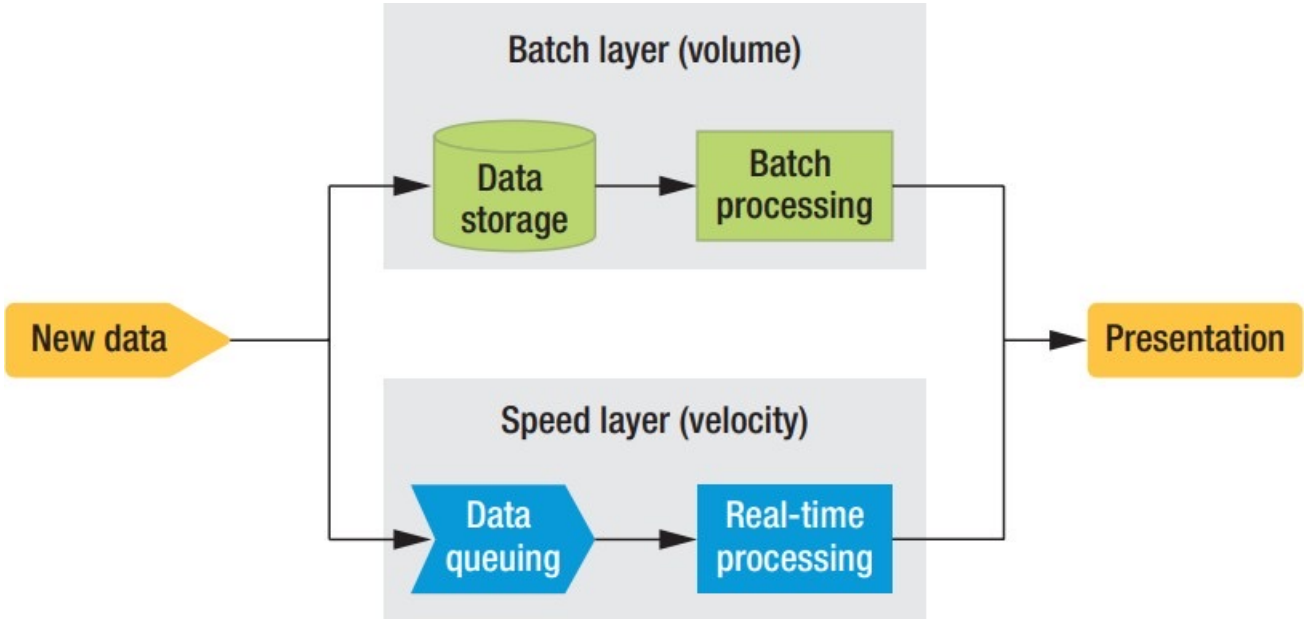


Big data technologies landscape



source: J. Heidrich, A. et al., Exploiting Big Data's Benefits, in IEEE Software, 2016

Big data analytics



source: J. Heidrich, A. et al., Exploiting Big Data's Benefits, in IEEE Software, 2016

Reading list

J. Heidrich, A. Trendowicz and C. Ebert, Exploiting Big Data's Benefits, in *IEEE Software*, 2016

(documents available on Blackboard)

Questions, comments?