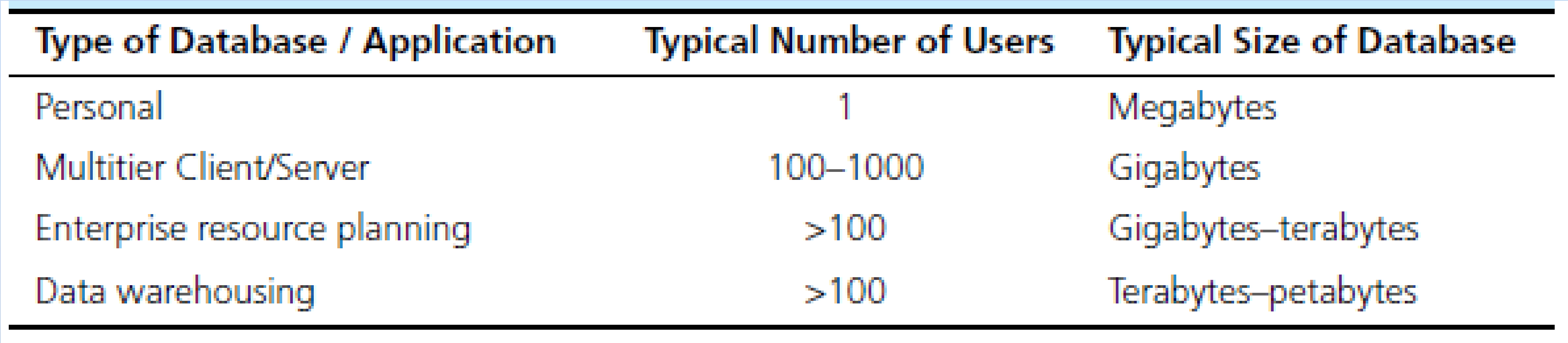
Big Data Introduction

Dr. Deepak Saxena, SME IIT Jodhpur

The size of small/traditional databases



Type of Database / Application	Typical Number of Users	Typical Size of Database
Personal	1	Megabytes
Multitier Client/Server	100–1000	Gigabytes
Enterprise resource planning	>100	Gigabytes–terabytes
Data warehousing	>100	Terabytes–petabytes

Enter Big Data

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

Why Big Data?

- Proliferation of devices that generate digital data
- Content generation and self-publishing
- Consumer Activity
- Machine data and Internet of Things
- Advances in natural science
- Plummeting cost of storage and processing power
- Open-source platforms
- Cloud computing

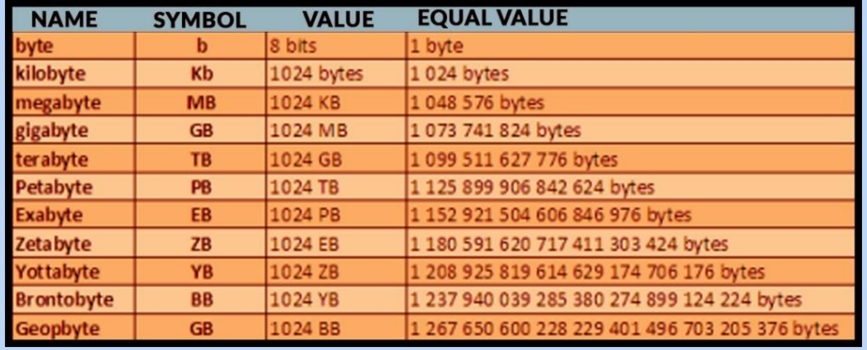
I'm gonna ask you this one time. Where is Big Data?
AVN
Yeah, I'll do you one better, What is Big Data?
I'll do you one better, WHY is Big Data?

3 (or 5) (or 7) Vs of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value
- Variability
- Visualization

Original Vs

Volume

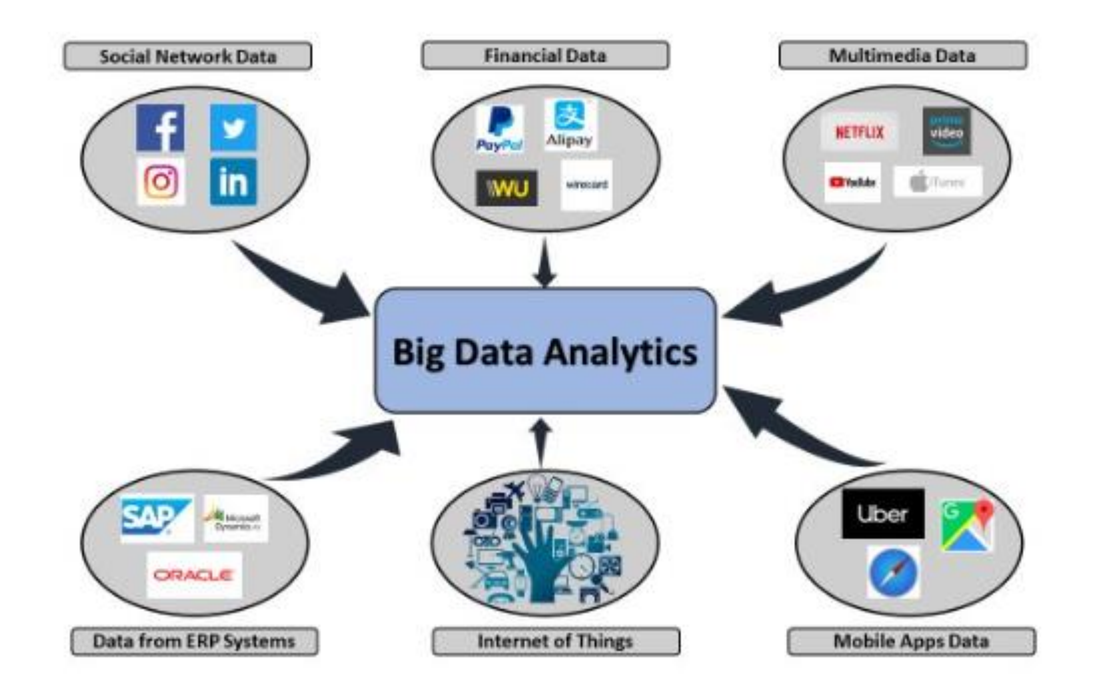


NAME	SYMBOL	VALUE	EQUAL VALUE
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

- Refers to the generation of large chunk of data.
- It is estimated that the volume of digital data would be around 123 Zettabytes (1 Zettabyte equals 10^{21} bytes) in 2023, reaching a staggering 181 Zettabytes by 2025.
- Data Centres
  - The average data center occupies approximately 100,000 square feet of space.
  - The International Energy Agency estimates that 1% of all global electricity is used by data centers and that by 2025, data centers will consume 1/5 of the world’s power supply.

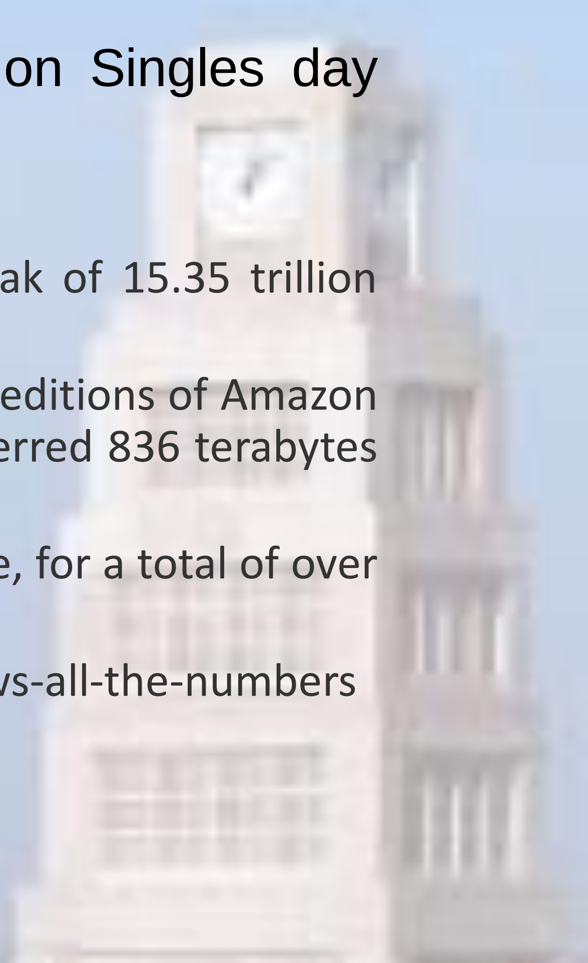
Variety

- Not just relational structured data.
- Various forms of data in disparate formats from disparate sources are combined to generate a holistic picture.
- For instance, Google collects and combines data from various sources (Android operating system, Chrome browser, Gmail, Maps, Search history, Voice, and YouTube activity to name a few) to personalize its offerings to the users
- Social media is a big contributor of Big Data.



Velocity

- Refers to the speed with which data is generated.
- Correct and rapid processing of data in the real time has become extremely crucial for companies.
- Chinese retailer Alibaba processed around 544,000 orders per second on Singles day 2019.
- Amazon on Prime Day 2023
  - An incremental 163 petabytes of EBS storage capacity allocated – generating a peak of 15.35 trillion requests and 764 petabytes of data transfer per day.
  - 5,835 database instances running the PostgreSQL-compatible and MySQL-compatible editions of Amazon Aurora processed 318 billion transactions, stored 2,140 terabytes of data, and transferred 836 terabytes of data.
  - Amazon CloudFront handled a peak load of over 500 million HTTP requests per minute, for a total of over 1 trillion HTTP requests during Prime Day.
  - For more details: https://aws.amazon.com/blogs/aws/prime-day-2023-powered-by-aws-all-the-numbers



Veracity and Value

Veracity

- Refers to the authenticity and truthfulness of the data.
- This is more relevant in case of unstructured data, for instance making sure that service reviews are coming from authentic users and not from automated bots.

Value

- Refers to the outcomes, for instance efficiency or reputational gains, made possible by Big data analytics.
- Beyond operational gains, however, Big data capabilities may act as an enabler of digital enterprise transformation.

Variability and Visualization

Variability

- refers to the variability in the meaning when processing natural language data.
- For example, the term ‘great service’ would differ in meaning depending on being preceded by ‘got the reply within two hours’ or ‘still waiting for reply since last 15 days’.
- It may also refer to inconsistent speed of the data

Visualization

- Primarily relates to the appropriation of data as opposed to its inherent nature.
- Since Big data is unstructured and comes from a variety of sources, visualization of the trends helps in making sense of the vast tapestry of data.

Data Lake

@timoelliott

*"What's a data lake for?
So you can drown in more data even faster!"*