# Concept of Big Data

Dr. V. K. Patle

Associate Professor

S. o. S. Computer Science & IT

Pt. Ravishankar Shukla University, Raipur (C.G.)

# What is “big data”?

- "Big Data are 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”  (Gartner 2012)
- Complicated (intelligent) analysis of data may make a small data “appear” to be “big”
- Bottom line: Any data that exceeds our current capability of processing can be regarded as “big”

# Big Data Everywhere!

# BIG DATA

Data that is TOO LARGE & TOO COMPLEX for conventional data tools to capture, store and analyze.

Shares traded on US Stock Markets each day:

7 Billion

Data generated in one flight from NY to London:

10 Terabytes

Number of tweets per day on Twitter:

400 Million

Number of 'Likes' each day on Facebook:

3 Billion

## The 3V's of Big Data

VOLUME

VARIETY

VELOCITY

90%

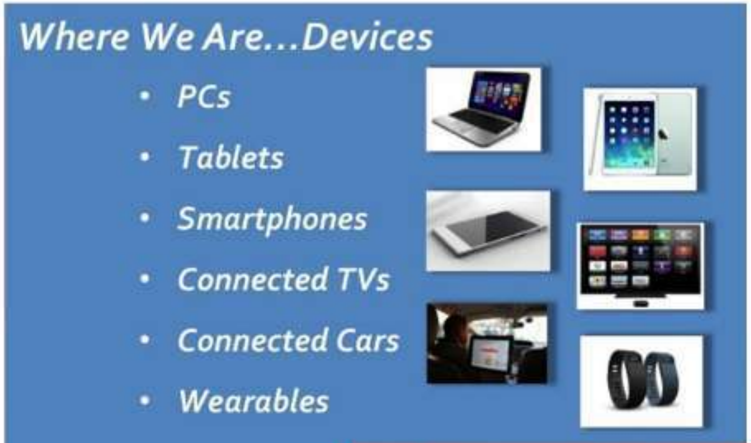
OF THE WORLD'S DATA WAS GENERATED IN THE **LAST TWO YEARS**

- Big Data are “data sets that are so big they cannot be handled efficiently by common database management systems” (Dasgupta, 2013).

- Big Data have volume of 100 terabytes to petabytes, have structured and unstructured formats, and have a constant flow of data (Davenport, 2014).

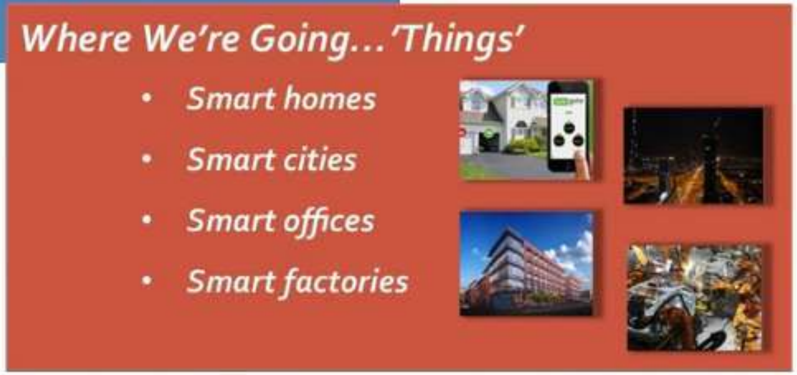
- A 2016 definition states that "Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value".
- Similarly, Kaplan and Haenlein define big data as "data sets characterized by huge amounts (volume) of frequently updated data (velocity) in various formats, such as numeric, textual, or images/videos (variety)”.
- A 2018 definition states "Big data is where parallel computing tools are needed to handle data”.

# Big Data coming from "Smart Things"…



*Where We Are…Devices*

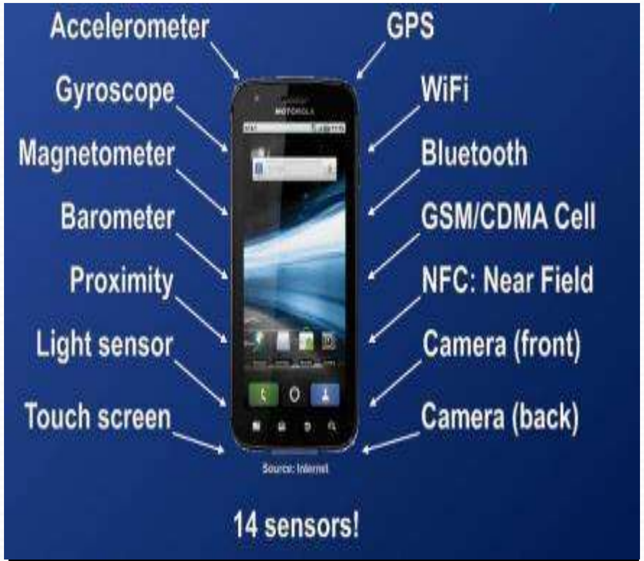
- *PCs*
- *Tablets*
- *Smartphones*
- *Connected TVs*
- *Connected Cars*
- *Wearables*

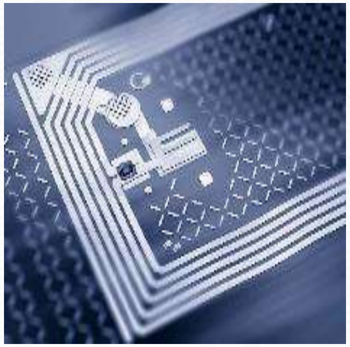


*Where We're Going…'Things'*

- *Smart homes*
- *Smart cities*
- *Smart offices*
- *Smart factories*

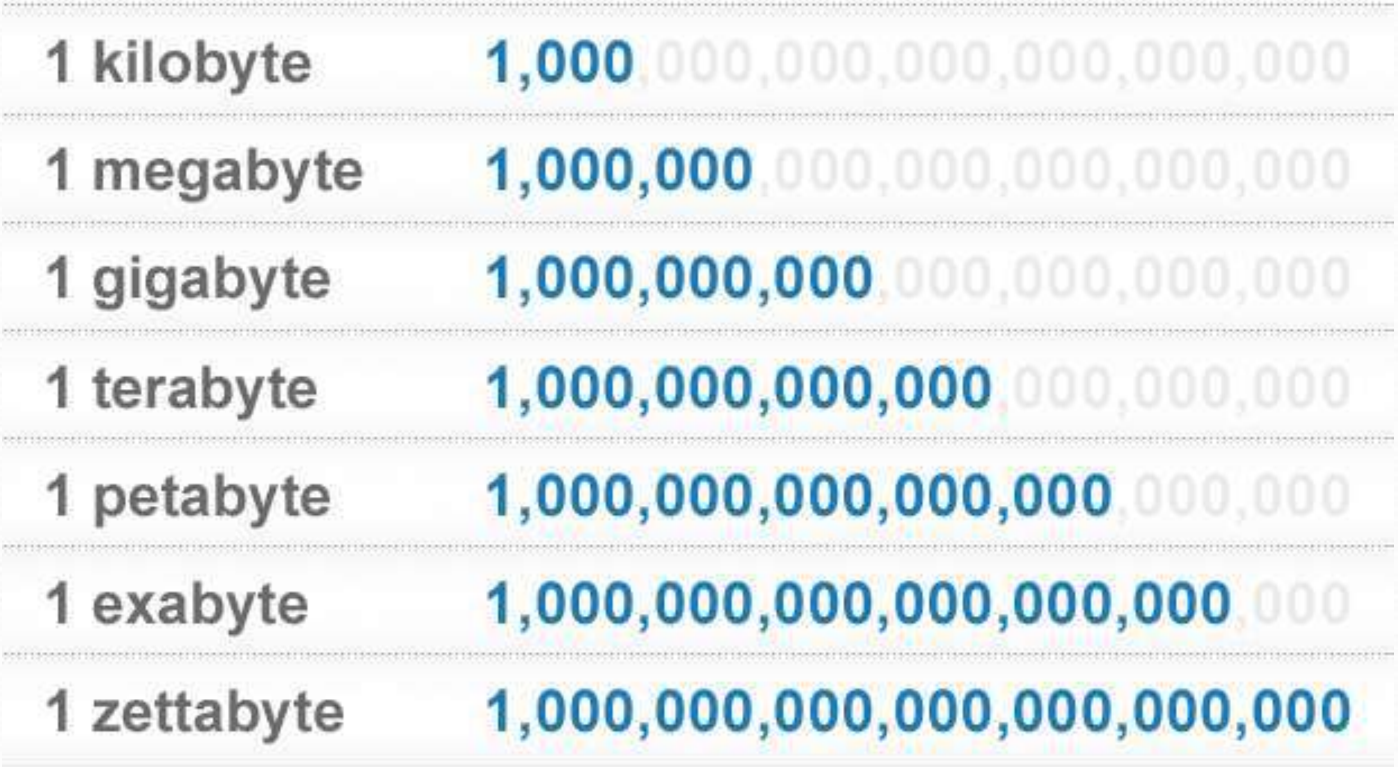
# Big Data coming form Sensor Data…





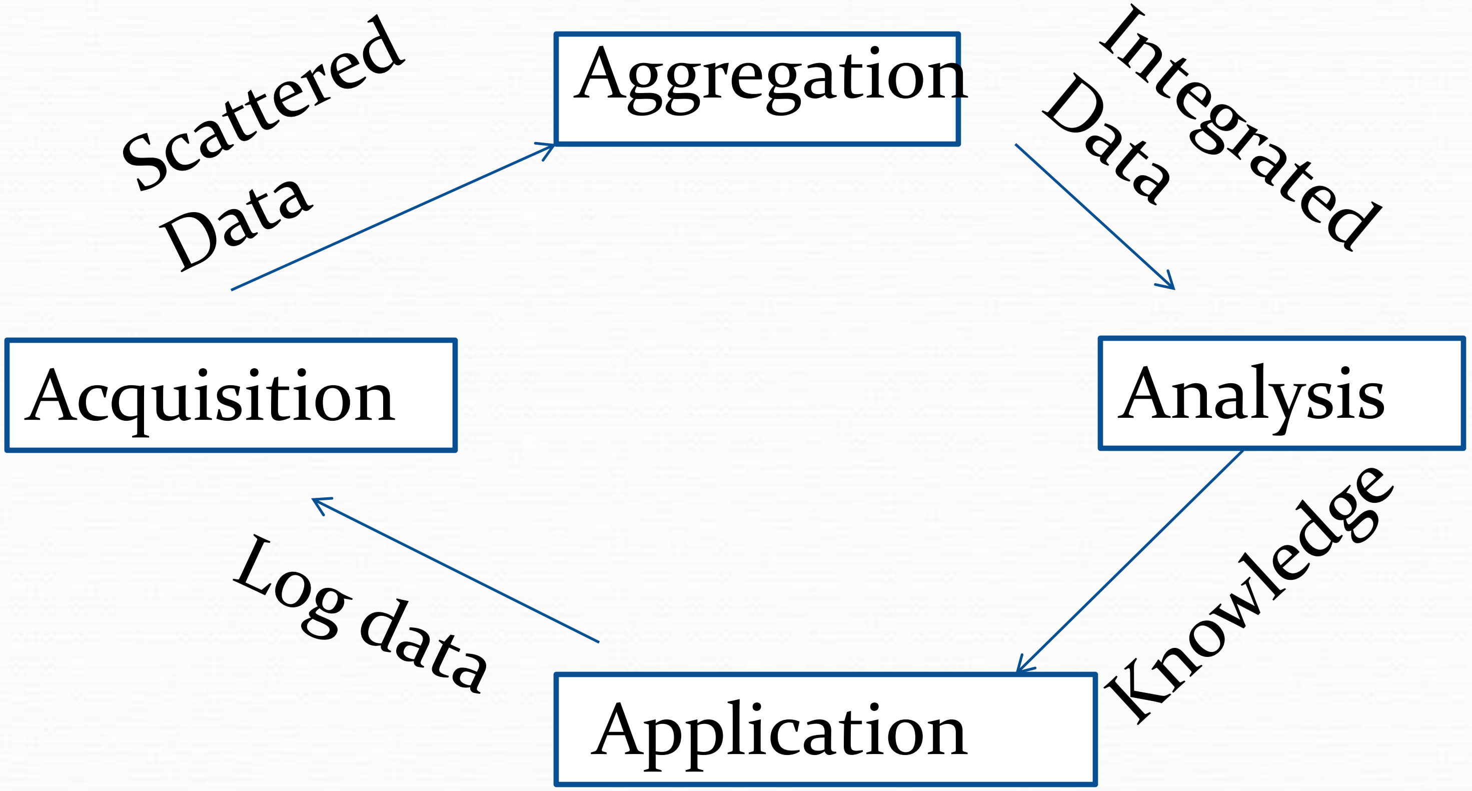
# Big Data – A Brief Review

So, we know that "big data" is BIG



| | |
|---|---|
| 1 kilobyte | 1,000,000,000,000,000,000,000 |
| 1 megabyte | 1,000,000,000,000,000,000,000 |
| 1 gigabyte | 1,000,000,000,000,000,000,000 |
| 1 terabyte | 1,000,000,000,000,000,000,000 |
| 1 petabyte | 1,000,000,000,000,000,000,000 |
| 1 exabyte | 1,000,000,000,000,000,000,000 |
| 1 zettabyte | 1,000,000,000,000,000,000,000 |

# Lifecycle of Data: 4 “A”s


Aggregation

Integrated Data

Analysis

Knowledge

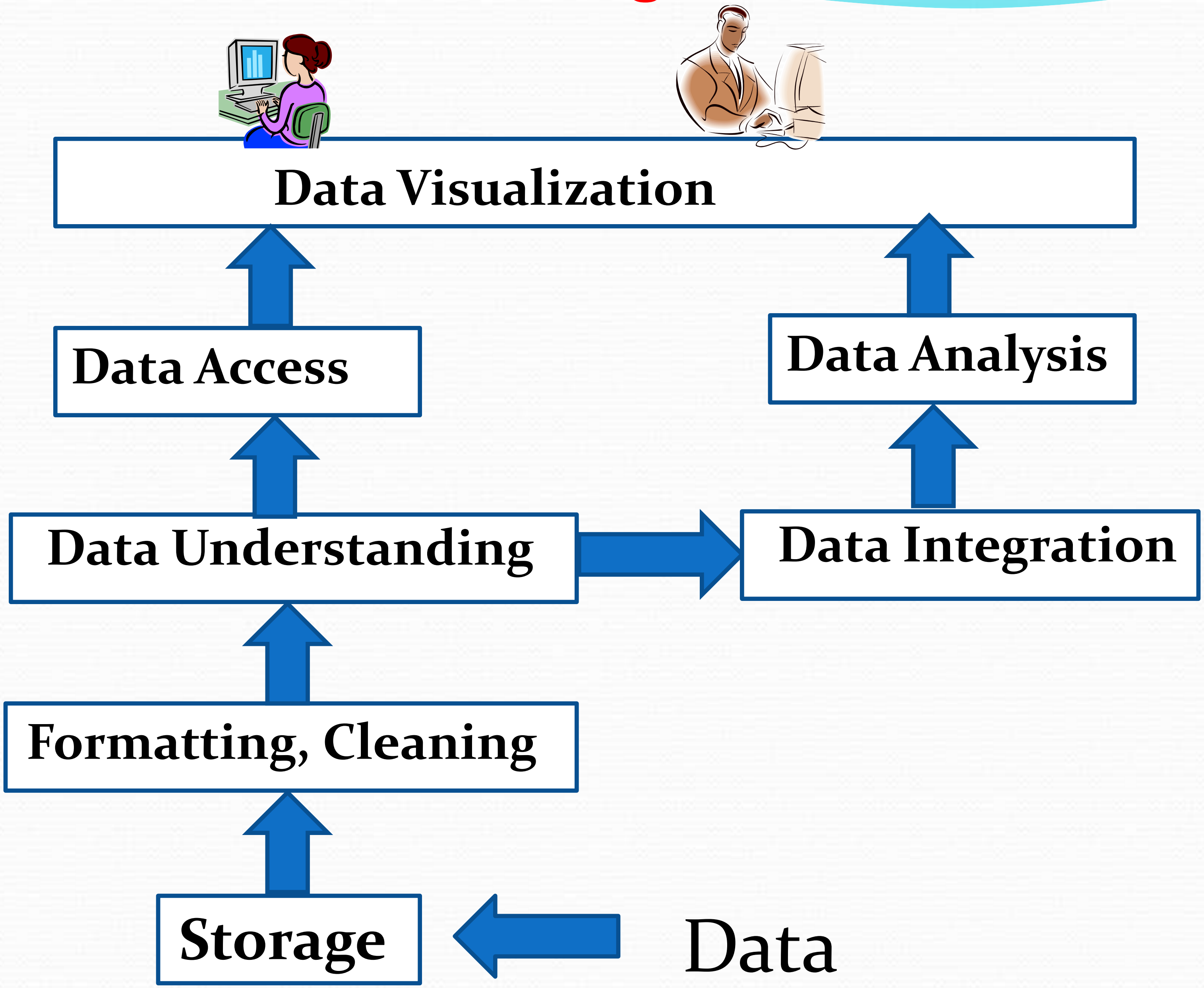
Application

Log data

Acquisition

Scattered Data

# Computational View of Big Data



Data Visualization

Data Access

Data Analysis

Data Understanding

Data Integration

Formatting, Cleaning

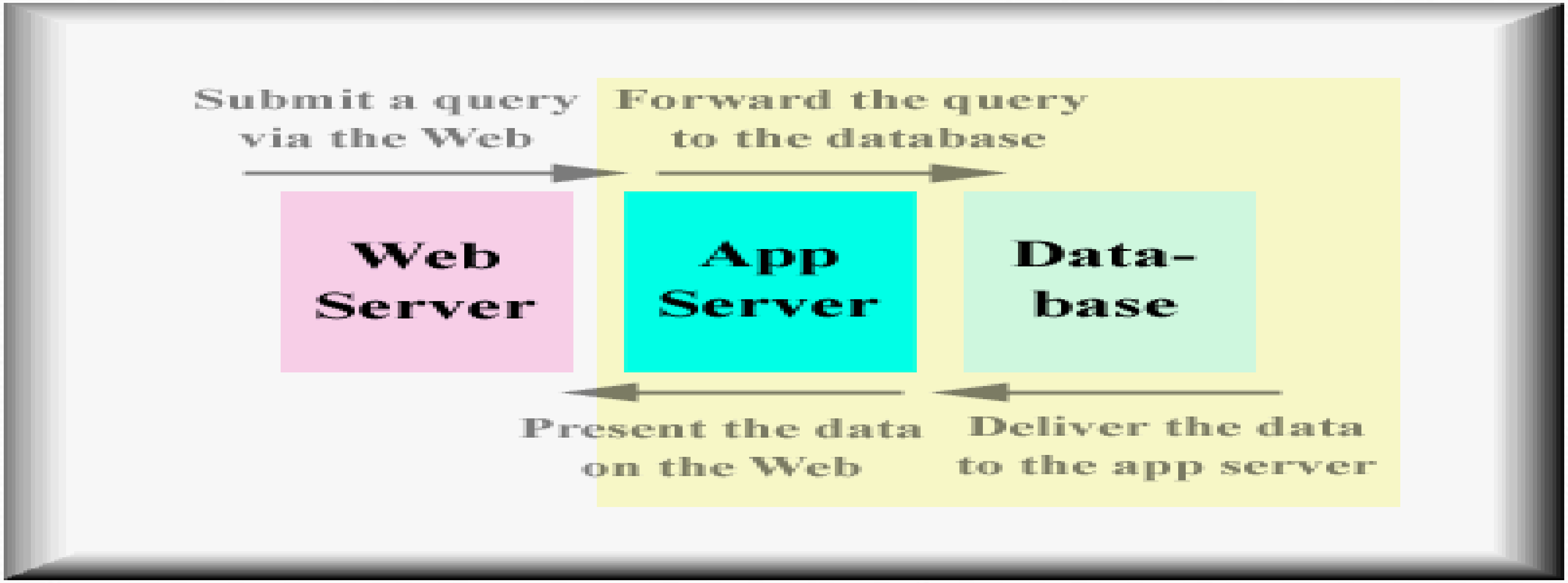
Storage

Data

# Web Data

"Web data is a collective term which refers to any type of data you might pull from the internet, whether to study for research purposes or otherwise." That might be data on what your competitors are selling, published government data, football scores, etc. It's a catchall for anything you can find on the web that is public facing (ie not stored in some internal database). Studying this data can be very informative, especially when communicated well to management.

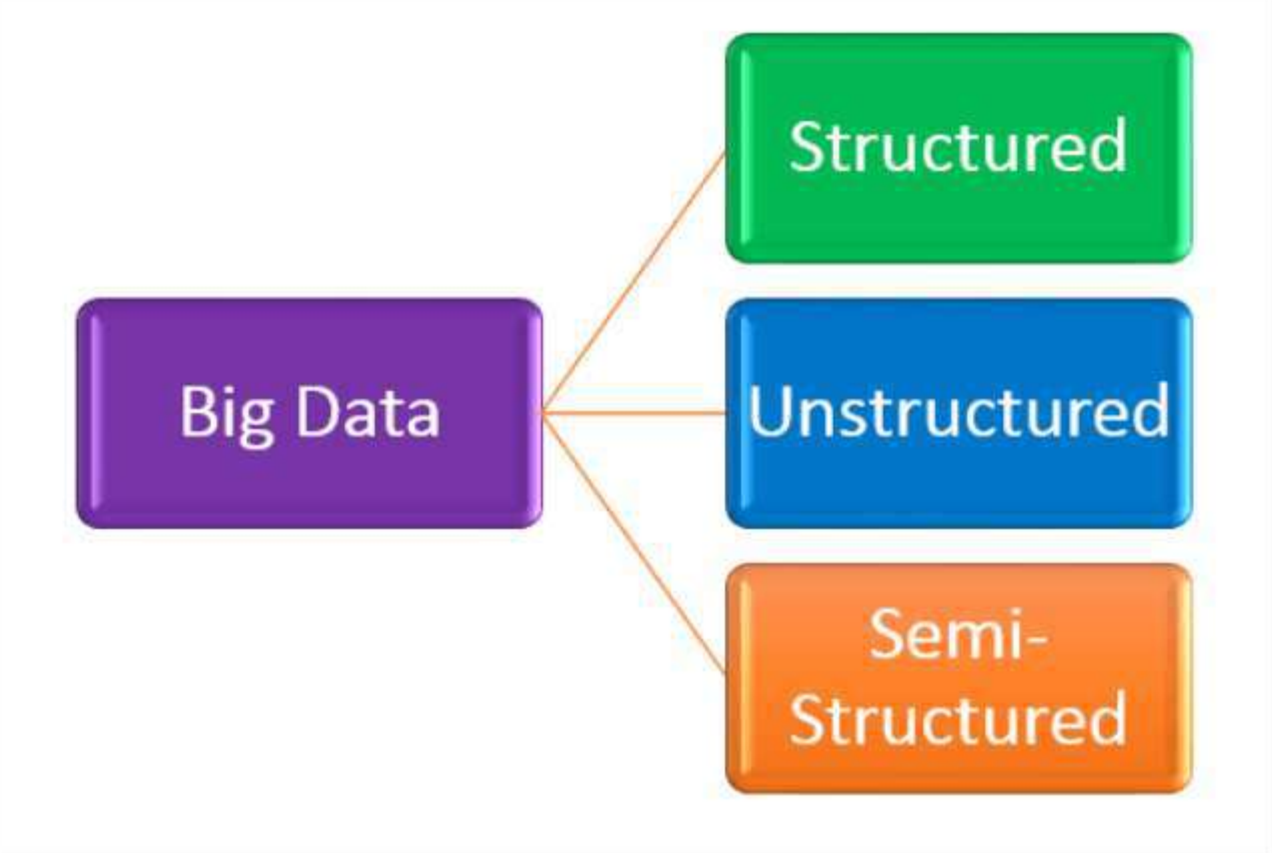
Web data is important because it's one of the major ways businesses can access information that isn't generated by themselves. When creating quality business models and making important BI decisions, businesses need information on what is happening internally and externally within their organization and what is happening in the wider market.


Submit a query via the Web
Forward the query to the database
Web Server
App Server
Data-base
Present the data on the Web
Deliver the data to the app server

Web data can be used to monitor competitors, track potential customers, keep track of channel partners, generate leads, build apps, and much more. It’s uses are still being discovered as the technology for turning unstructured data into structured data improves.

Web data can be collected by writing web scrapers to collect it, using a scraping tool, or by paying a third party to do the scraping for you. A web scraper is a computer program that takes a URL as an input and pulls the data out in a structured format – usually a JSON feed or CSV.

# Classifications of Big Data



Big Data

Structured

Unstructured

Semi-Structured

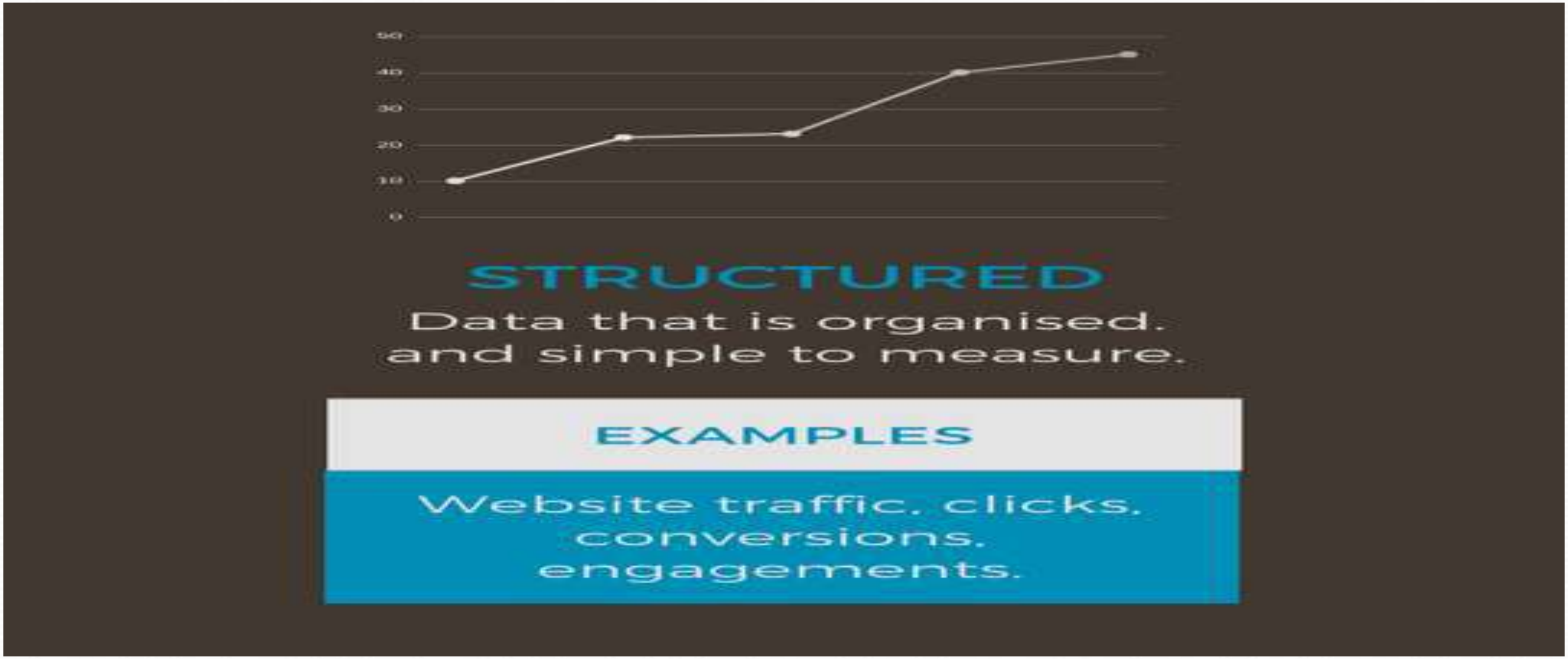
# 1. Structured data

Structured Data is used to refer to the data which is already stored in databases, in an ordered manner. It accounts for about 20% of the total existing data and is used the most in programming and computer-related activities.

There are two sources of structured data- machines and humans.

➢All the data received from sensors, weblogs, and financial systems are classified under machine-generated data. These include medical devices, GPS data, data of usage statistics captured by servers and applications and the huge amount of data that usually move through trading platforms, to name a few.

- ➢Human-generated structured data mainly includes all the data a human input into a computer, such as his name and other personal details. When a person clicks a link on the internet, or even makes a move in a game, data is created- this can be used by companies to figure out their customer behavior and make the appropriate decisions and modifications.



STRUCTURED

Data that is organised and simple to measure.

EXAMPLES

Website traffic, clicks, conversions, engagements.

# 2. Unstructured data

While structured data resides in the traditional row-column databases, unstructured data is the opposite- they have no clear format in storage. The rest of the data created, about 80% of the total account for unstructured big data. Most of the data a person encounters belong to this category- and until recently, there was not much to do to it except storing it or analyzing it manually.

Unstructured data is also classified based on its source, into machine-generated or human-generated.

➢Machine-generated data accounts for all the satellite images, the scientific data from various experiments and radar data captured by various facets of technology.

- Human-generated unstructured data is found in abundance across the internet since it includes social media data, mobile data, and website content. This means that the pictures we upload to Facebook or Instagram handle, the videos we watch on YouTube and even the text messages we send all contribute to the gigantic heap that is unstructured data.

UNSTRUCTURED

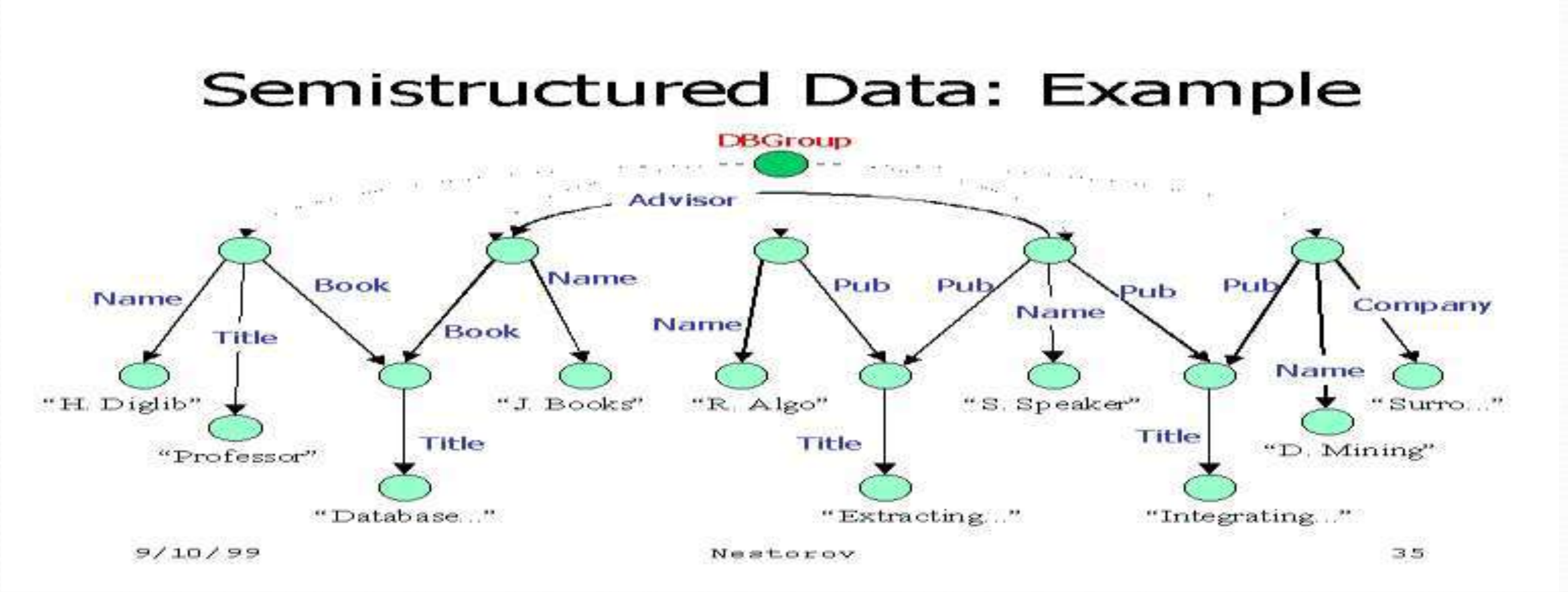
Data that is unorganised. and hard to measure.

EXAMPLE

Sentiment - what people are saying (and meaning) in Tweets, emails, chats.

# 3. Semi-structured data

The line between unstructured data and structured data has always been unclear since most of the semi-structured data appear to be unstructured at a glance. Information that is not in the traditional database format as structured data, but contains some organizational properties which make it easier to process, are included in semi-structured data. For example, NoSQL documents are considered to be semi-structured, since they contain keywords that can be used to process the document easily.



Semistructured Data: Example

9/10/99 Nestorov 35

# Difference between Structured, Semi-structured and Unstructured data.

| Factors | Structured data | Semi-structured data | Unstructured data |
|---|---|---|---|
| Flexibility | It is dependent and less flexible | It is more flexible than structured data but less than flexible than unstructured data | It is flexible in nature and there is an absence of a schema |
| Transaction Management | Matured transaction and various concurrency technique | The transaction is adapted from DBMS not matured | No transaction management and no concurrency |
| Query performance | Structured query allow complex joining | Queries over anonymous nodes are possible | An only textual query is possible |
| Technology | It is based on the relational database table | It is based on RDF and XML | This is based on character and library data |

# Challenges of Conventional system

- **Three Challenges That big data face.**

➢ Data or Volume

➢ Process

➢ Management

## 1. Data or Volume

✓ The volume of data, especially machine-generated data, is exploding,

✓ How fast that data is growing every year, with new sources of data that are emerging.

✓ For example, in the year 2000, 800,000petabytes (PB) of data were stored in the world, and it is expected to reach 35zetta bytes(ZB) by2020 (according to IBM).

## 2. Processing

✓ More than 80% of today’s information is unstructured and it is typically too big to manage effectively.

✓Today, companies are looking to leverage a lot more .

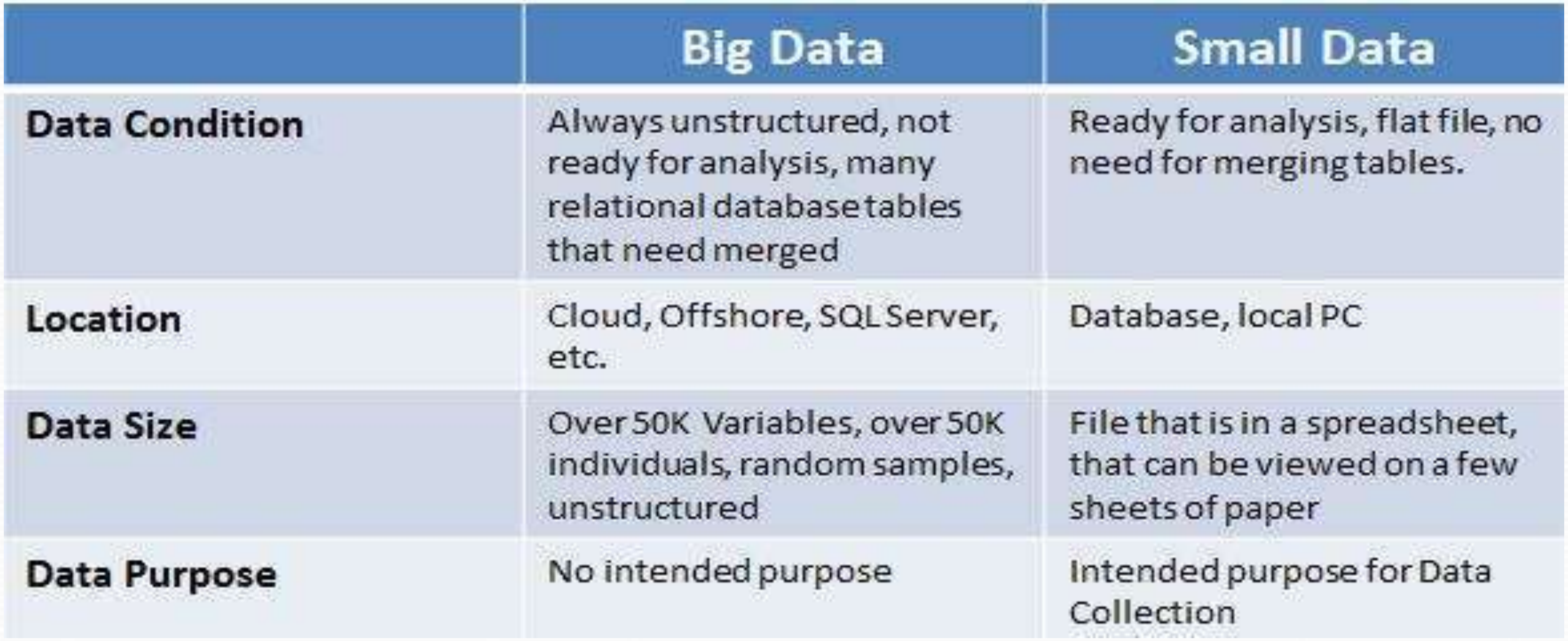
✓ data from a wider variety of sources both inside and outside the organization.

✓ Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

## 3. Management

✓ A lot of this data is unstructured, or has a complex structure that’s hard to represent in rows and columns.

# Difference between Big Data and Small Data



| | Big Data | Small Data |
|---|---|---|
| **Data Condition** | Always unstructured, not ready for analysis, many relational database tables that need merged | Ready for analysis, flat file, no need for merging tables. |
| **Location** | Cloud, Offshore, SQL Server, etc. | Database, local PC |
| **Data Size** | Over 50K Variables, over 50K individuals, random samples, unstructured | File that is in a spreadsheet, that can be viewed on a few sheets of paper |
| **Data Purpose** | No intended purpose | Intended purpose for Data Collection |

- **Small data** is collected with an intended purpose for analysis. It is a sample size that is determined by the data scientist that is collected to answers the problem at hand. With Small Data, there is control of the data. It is ready and conditioned for analysis once the data is collected.

- **Big Data**, does not have an intended purpose other than data mining. For this reason, the data takes a long time to clean and processed by the machine learning algorithms. The data scientist lets the machine do all the work to come up with relationships in the data structures. Then uses different algorithms to verify the findings.

## Conclusion

Time, data complexity, and cleaning processes are the main differences in Big Data vs. Small Data. Here is an example of a decision tree machine learning data model built with small data. I will be writing about ways to process big data machine learning on this blog in the near future.

# Thank You