# Introduction to Big Data and and Data Analytics

Chapter 5

**Key Concepts:**

1. Introduction to Big Data
2. Types of Big Data
3. Advantages and Disadvantages of Big Data
4. Characteristics of Big Data
5. Big Data Analytics
6. Working on Big Data Analytics
7. Mining Data Streams
8. Future of Big Data Analytics

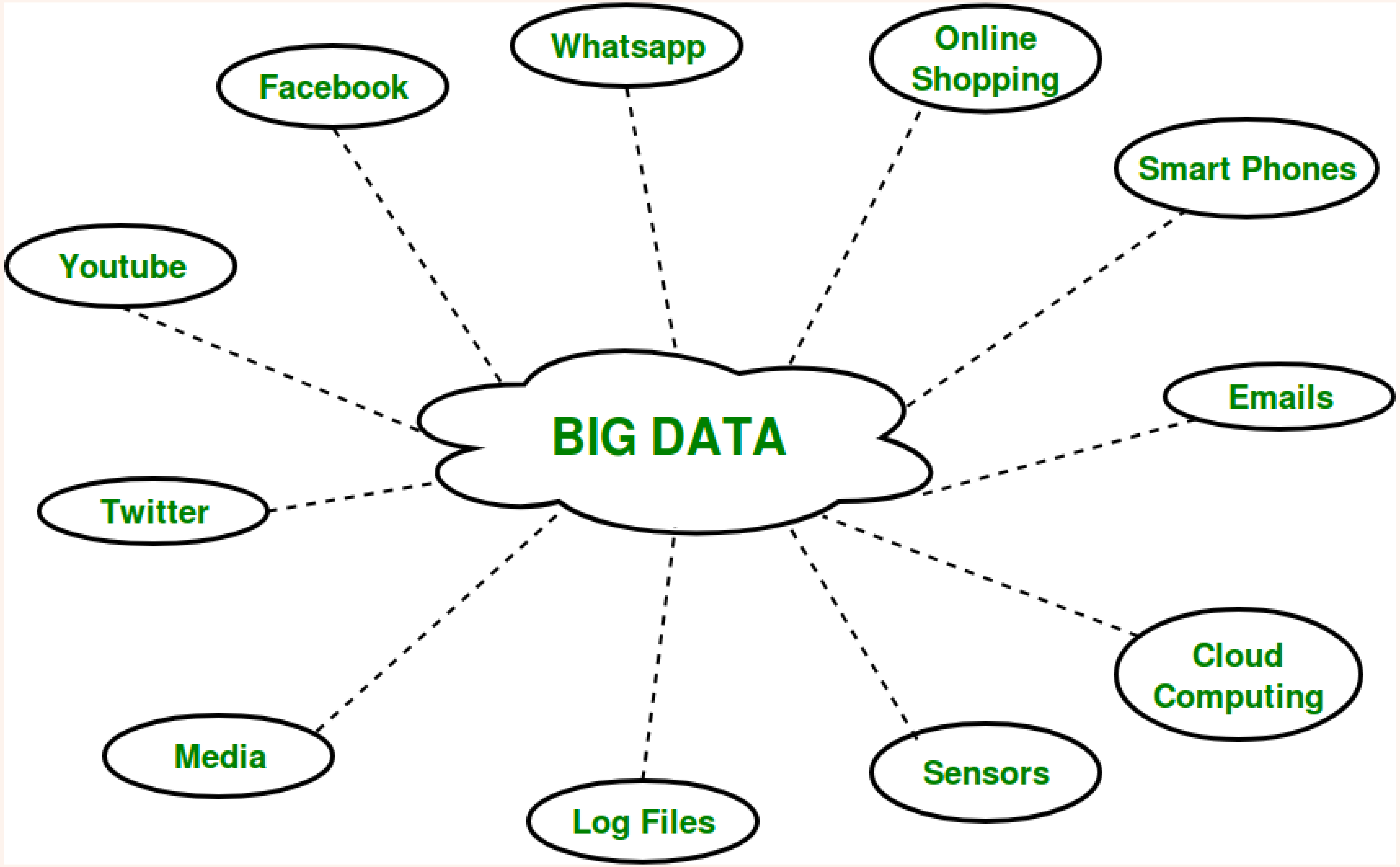
# 5.1. What is Small Data?

- Datasets that are easily comprehendible by people
- Easily accessible, informative, and actionable.
- Ideal for individuals and businesses to find useful information and make better choices in everyday tasks.
- For example, a small store might track daily sales to decide what products to restock.

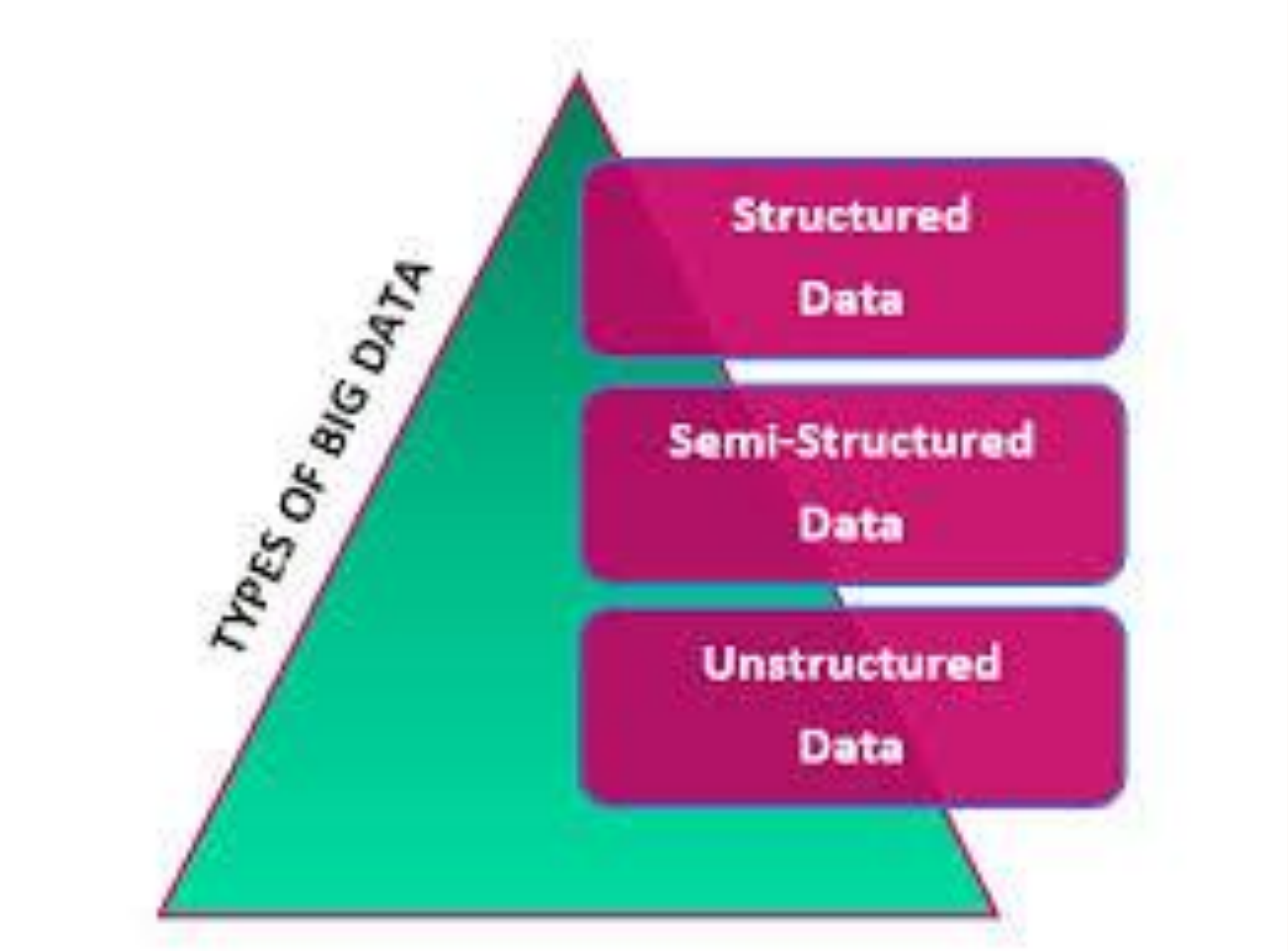
# 5.1. What is Big Data?

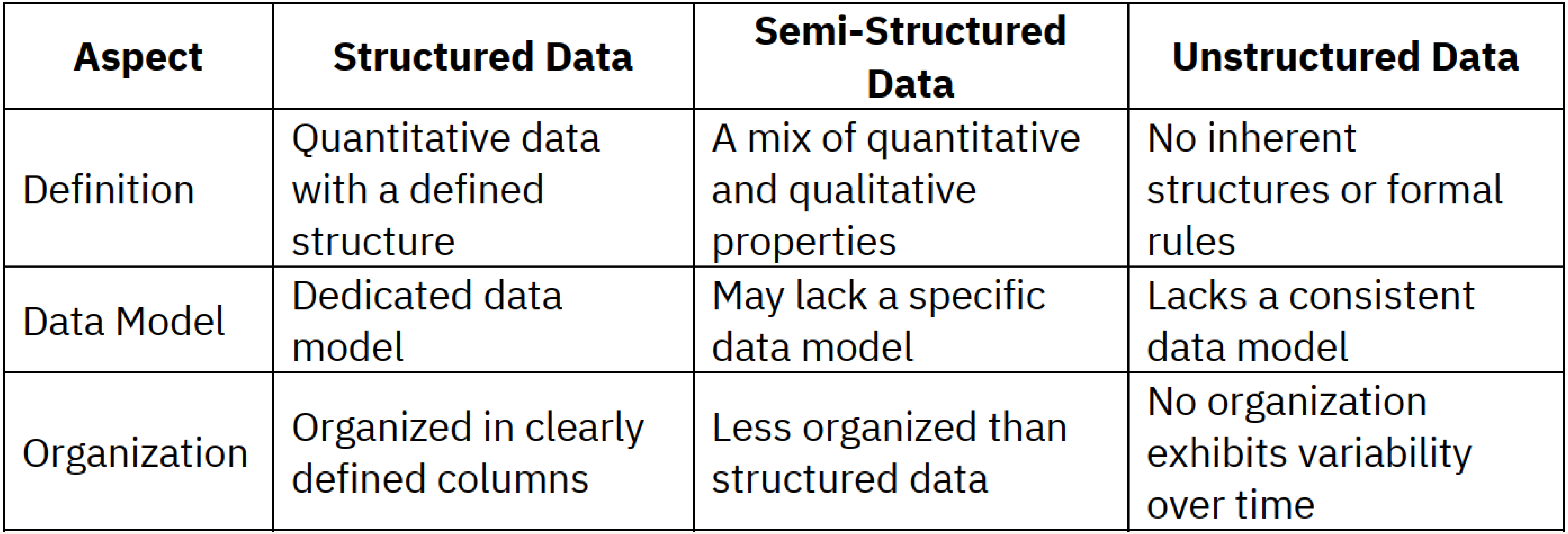
- Extremely large and complex datasets that regular computer programs and databases cannot handle.
- It comes from three main sources:
    - transactional data (e.g., online purchases)
    - machine data (e.g., sensor readings)
    - social data (e.g., social media posts)
- Special tools and techniques are required.
- Help organizations find valuable insights hidden in the data, which lead to innovations and better decision-making.
- For example, companies like Amazon and Netflix use Big Data to recommend products or shows based on users’ past activities.

# Sources of Big Data

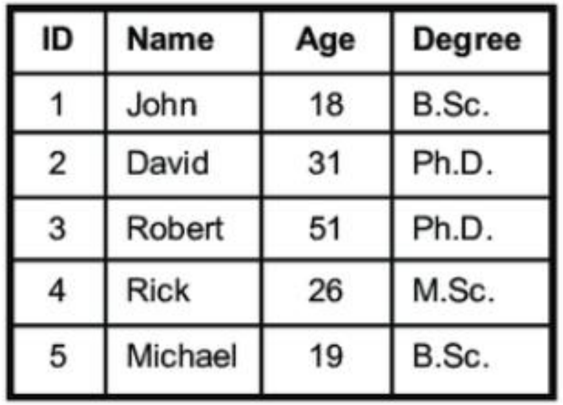
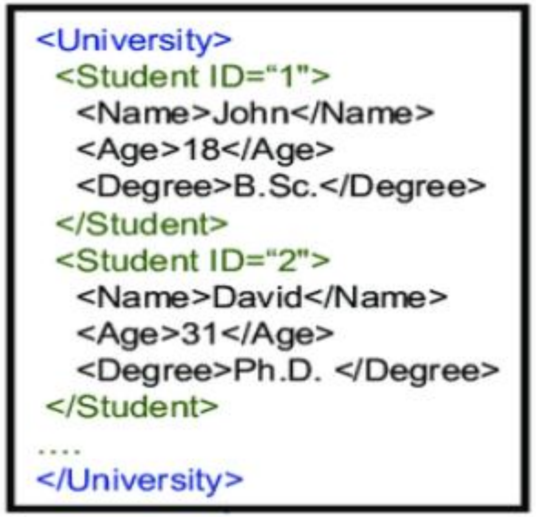
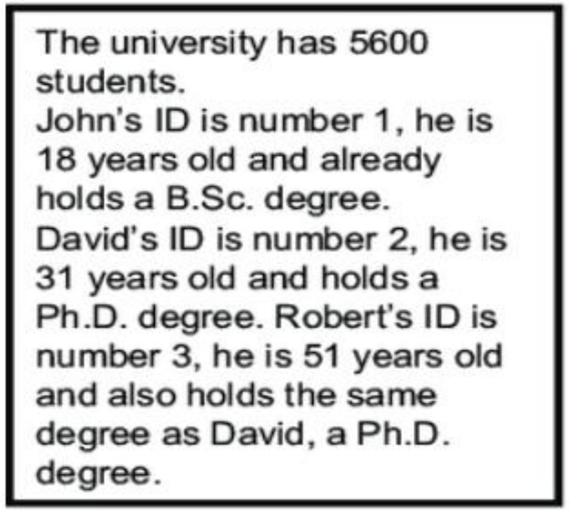


# 5.2. Types of Big Data





| Aspect | Structured Data | Semi-Structured Data | Unstructured Data |
|---|---|---|---|
| Definition | Quantitative data with a defined structure | A mix of quantitative and qualitative properties | No inherent structures or formal rules |
| Data Model | Dedicated data model | May lack a specific data model | Lacks a consistent data model |
| Organization | Organized in clearly defined columns | Less organized than structured data | No organization exhibits variability over time |

<table>
<tr>
<td>Accessibility</td>
<td>Easily accessible and searchable</td>
<td>Accessible but may be harder to analyze</td>
<td>Accessibility depends on the specific data format</td>
</tr>
<tr>
<td>Examples</td>
<td>Customer information, transaction records, product directories<br>Structured data<br>
<table>
<tr><th>ID</th><th>Name</th><th>Age</th><th>Degree</th></tr>
<tr><td>1</td><td>John</td><td>18</td><td>B.Sc.</td></tr>
<tr><td>2</td><td>David</td><td>31</td><td>Ph.D.</td></tr>
<tr><td>3</td><td>Robert</td><td>51</td><td>Ph.D.</td></tr>
<tr><td>4</td><td>Rick</td><td>26</td><td>M.Sc.</td></tr>
<tr><td>5</td><td>Michael</td><td>19</td><td>B.Sc.</td></tr>
</table>
</td>
<td>XML files, CSV files, JSON files, HTML files, semi-structured documents<br>Semi-structured data<br>
<pre>
&lt;University&gt;
 &lt;Student ID="1"&gt;
  &lt;Name&gt;John&lt;/Name&gt;
  &lt;Age&gt;18&lt;/Age&gt;
  &lt;Degree&gt;B.Sc.&lt;/Degree&gt;
 &lt;/Student&gt;
 &lt;Student ID="2"&gt;
  &lt;Name&gt;David&lt;/Name&gt;
  &lt;Age&gt;31&lt;/Age&gt;
  &lt;Degree&gt;Ph.D. &lt;/Degree&gt;
 &lt;/Student&gt;
….
&lt;/University&gt;
</pre>
</td>
<td>Audio files, images, video files, emails, PDFs, social media posts<br>Unstructured data<br><br>The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.</td>
</tr>
</table>

# Advantages of Big Data

### Enhanced Decision Making

Data-driven decisions from large, diverse datasets.

### Improved Efficiency & & Productivity

Identify inefficiencies, streamline processes, optimize resources.

### Better Customer Insights

Deeper understanding of behavior, preferences, and needs.

### Competitive Advantage

Uncover market trends, identify opportunities, stay ahead.

### Innovation & Growth

Develop new products, services, and business models.

# Disadvantages of Big Data

### Privacy & Security Concerns

Risks of unauthorized access, breaches, and misuse of personal info.

### Data Quality Issues

Challenges in ensuring accuracy, reliability, and completeness.

### Technical Complexity

Requires specialized skills and expertise for infrastructure and tools.

### Regulatory Compliance

Challenges in meeting data protection laws like GDPR.

### Cost & Resource Intensiveness

High costs for acquisition, storage, processing, and skilled staff.

# Characteristics of Big Data: The 6Vs

### Velocity

Speed of data generation, delivery, and analysis (e.g., 40,000 Google searches/sec).

### Volume

Huge amount of data generated daily, typically exceeding gigabytes (e.g., 328.77 million TB/day).

### Variety

Data in various formats: structured, unstructured, semi-structured (text, images, audio, video).

### Veracity

Consistency, accuracy, quality, and trustworthiness of data; requires cleaning.

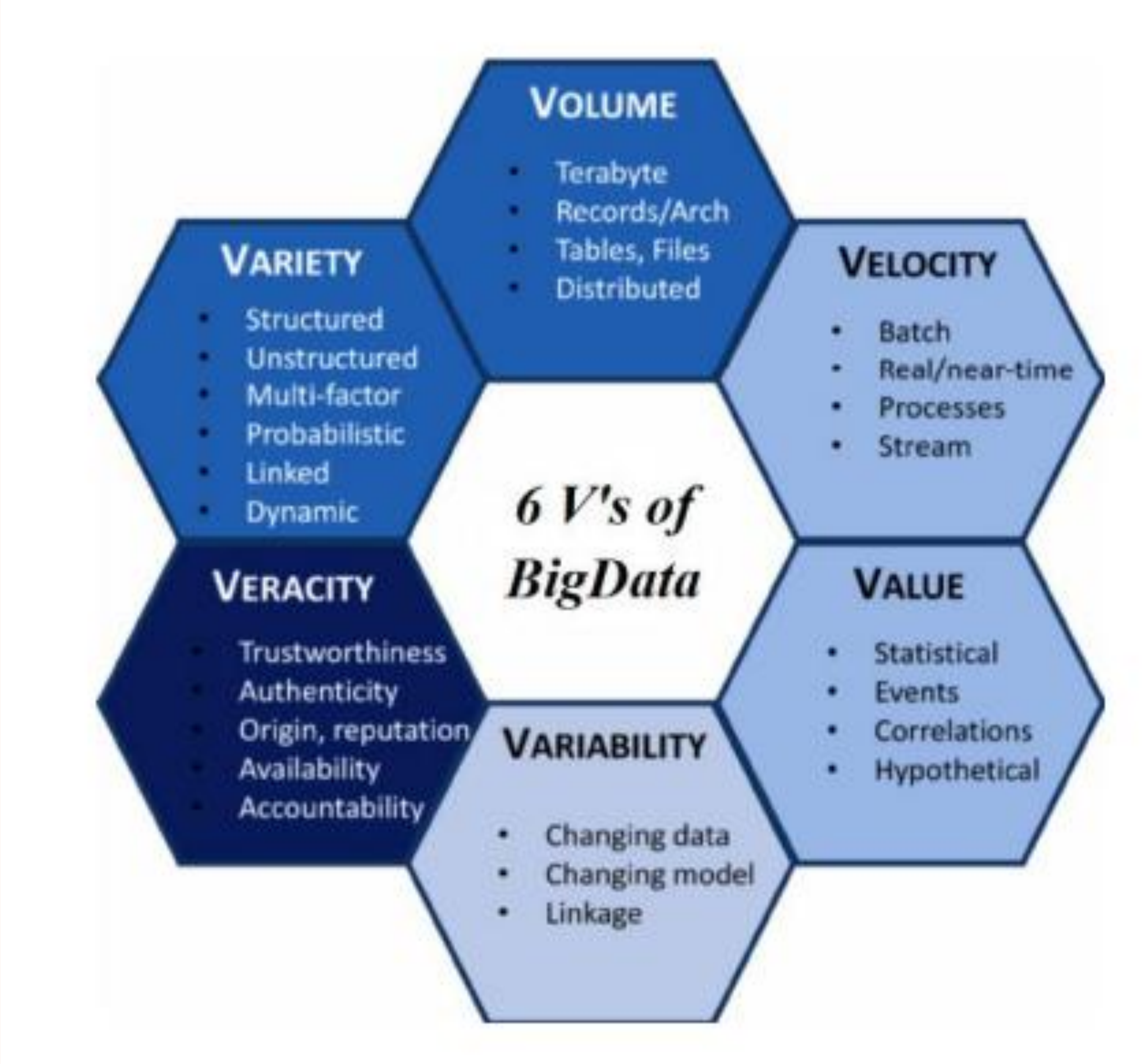
### Value

Business value derived from data analysis; critical for insights.

### Variability

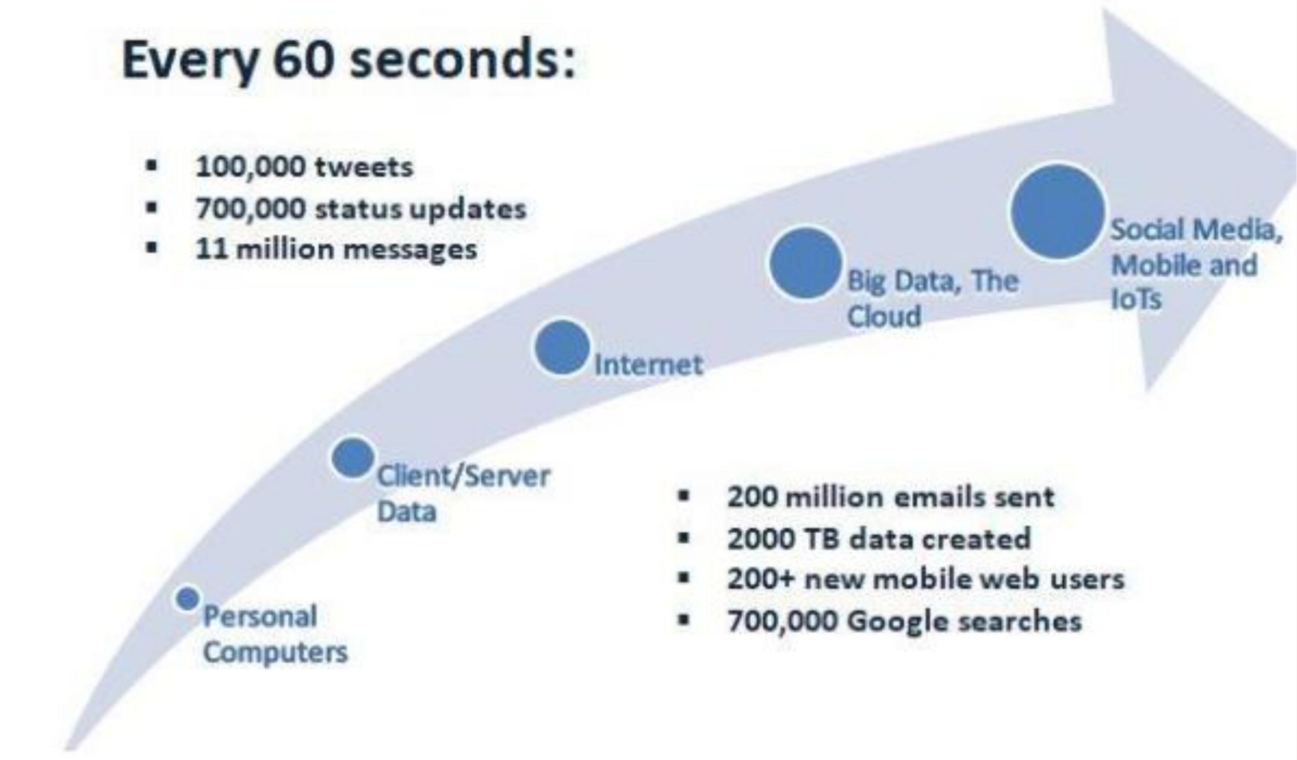
Regularity and dependability of data stream structure despite unpredictability.

# 5.3. Advantages and Disadvantages of Big Data:



# Velocity

Velocity refers to the speed at which data is generated, delivered, and analyzed.



Every 60 seconds:

- 100,000 tweets
- 700,000 status updates
- 11 million messages

- 200 million emails sent
- 2000 TB data created
- 200+ new mobile web users
- 700,000 Google searches

ig. 5.4 Speed of data generation from various sources

# Volume

Every day a huge volume of data is generated as the number of people using online platforms has increased exponentially. Such a huge volume of data is considered Big Data.

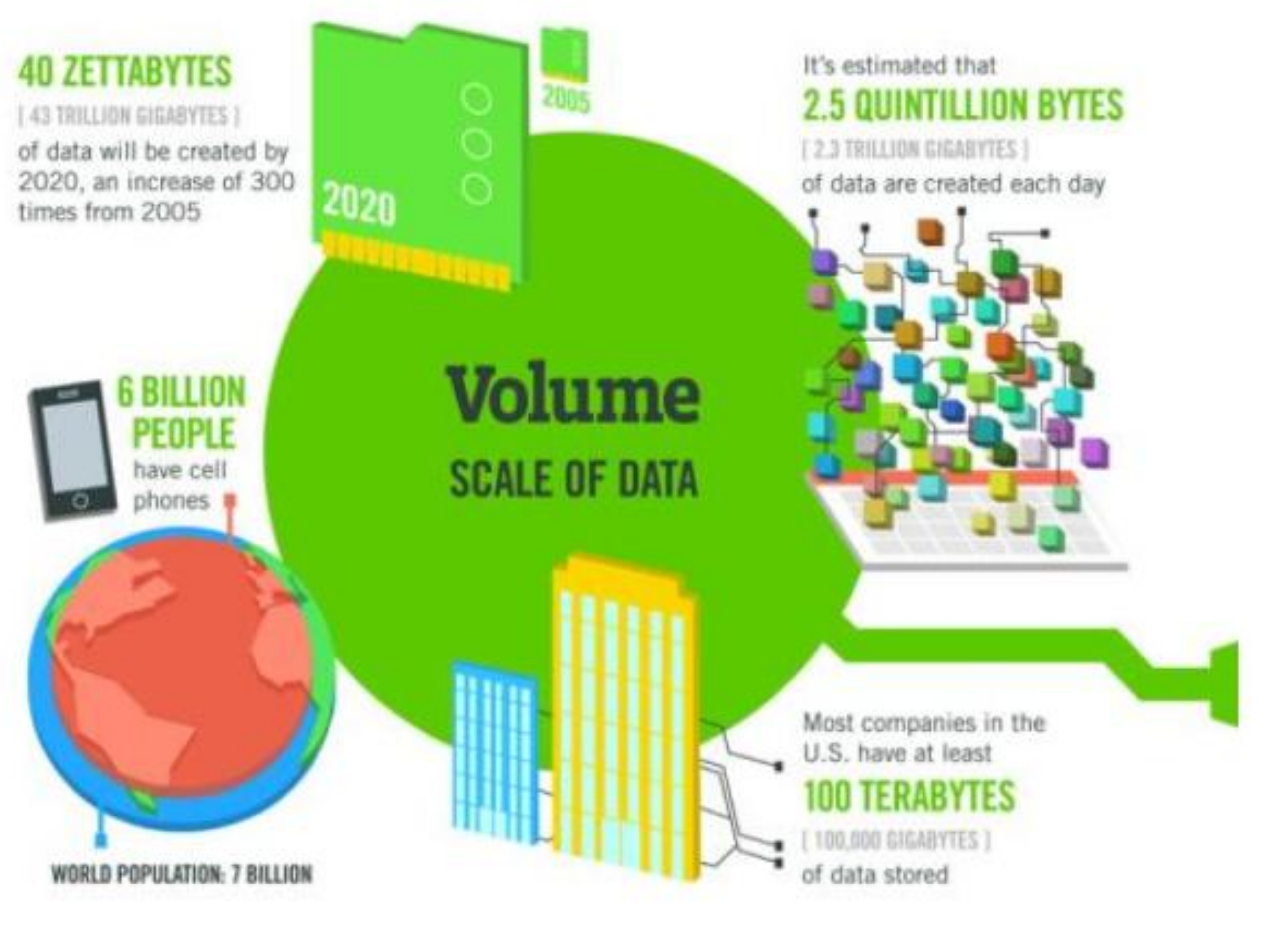

40 ZETTABYTES
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[ 2.3 TRILLION GIGABYTES ]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[ 100,000 GIGABYTES ]
of data stored

Fig.5.5 Volume of data

# Variety

- Big data encompasses data in various formats, including structured, unstructured, semi-structured, or highly complex structured data.
- These can range from simple numerical data to complex and diverse forms such as text, images, audio, videos, and so on.

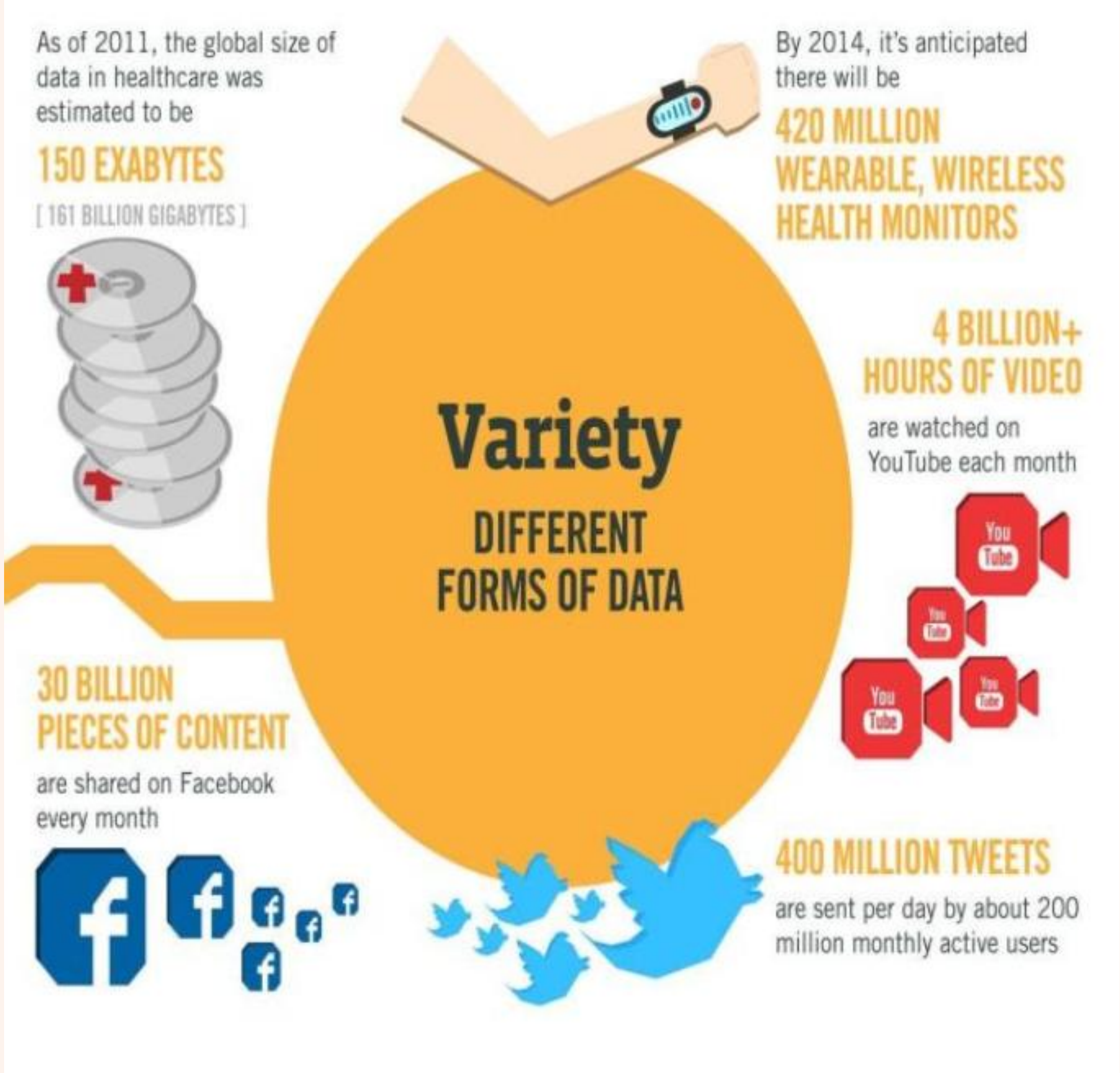

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[ 161 BILLION GIGABYTES ]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Fig.5.6 Varieties in Big data

# Veracity

- Veracity is a characteristic in Big Data related to consistency, accuracy, quality, and trustworthiness.
- Not all data that undergoes processing holds value.

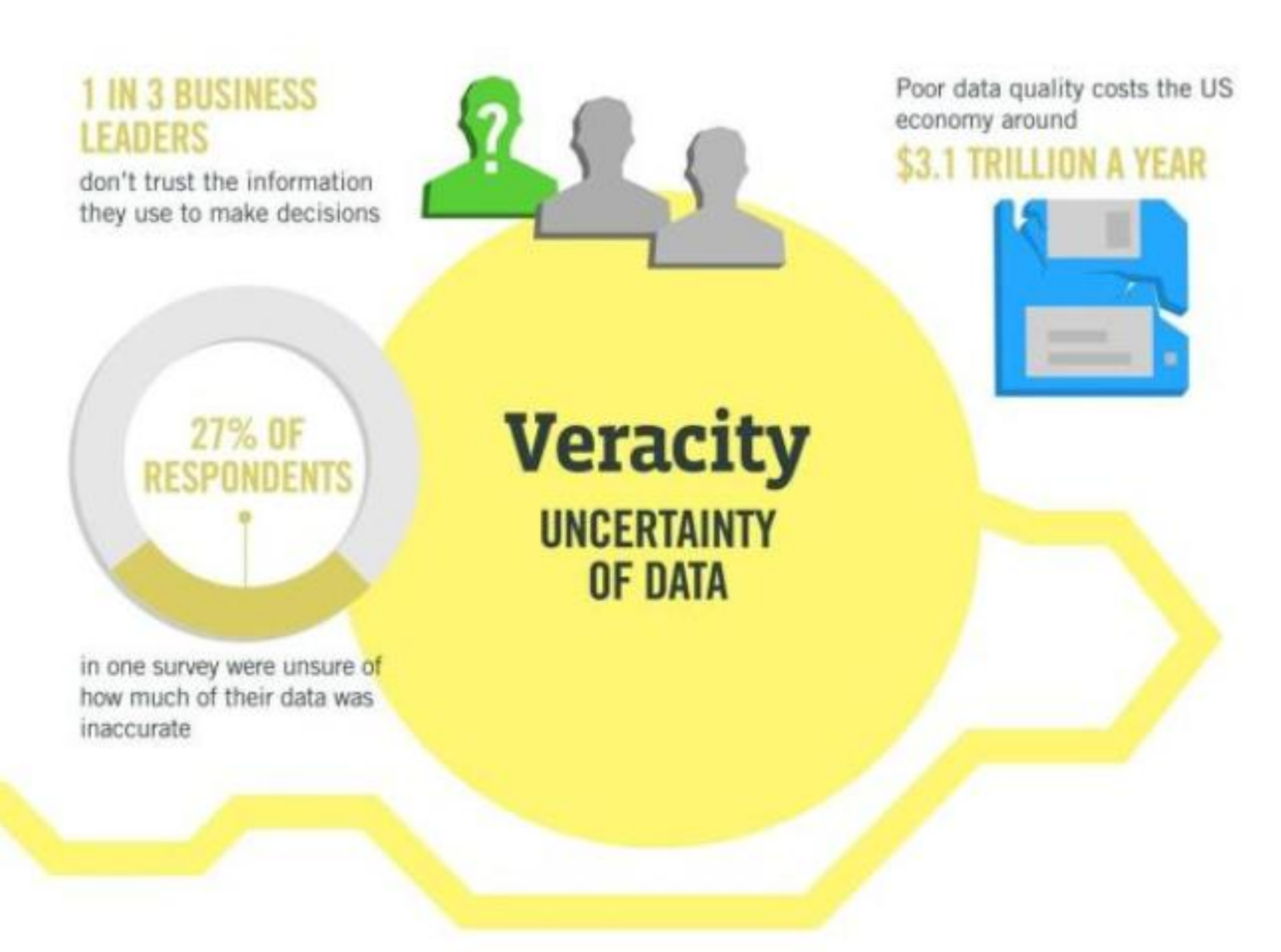

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Fig. 5.7

# Value

- The goal of big data analysis lies in extracting business value from the data.
- Hence, the business value derived from big data is perhaps its most critical characteristic.

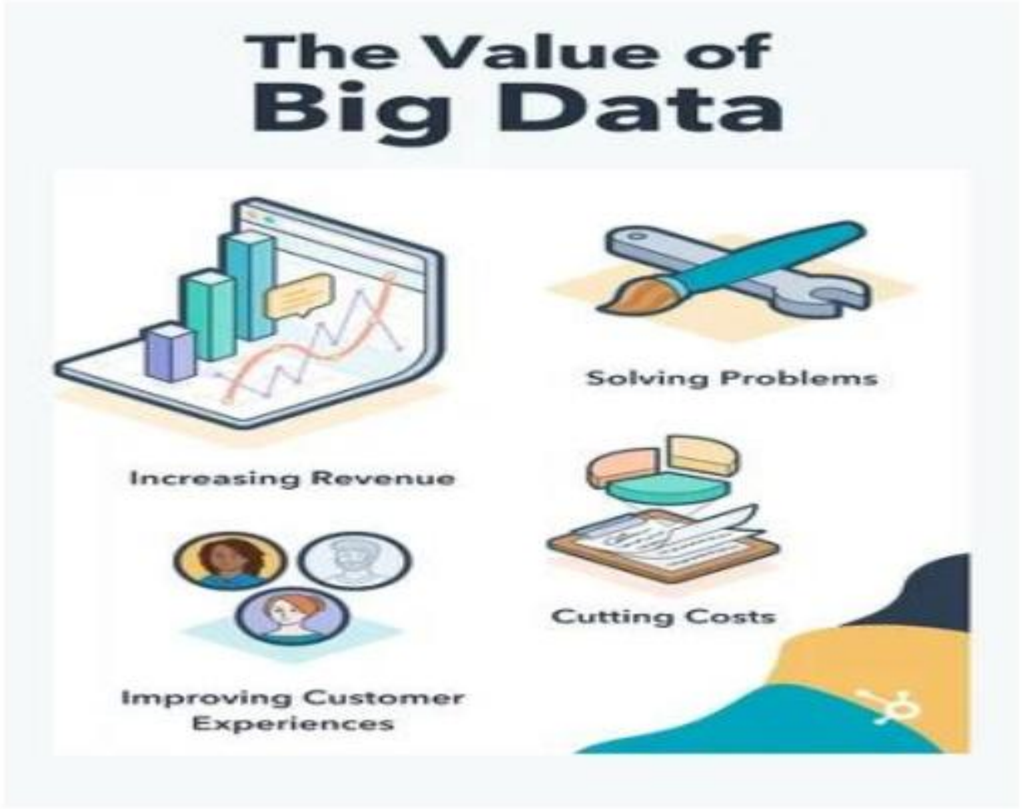
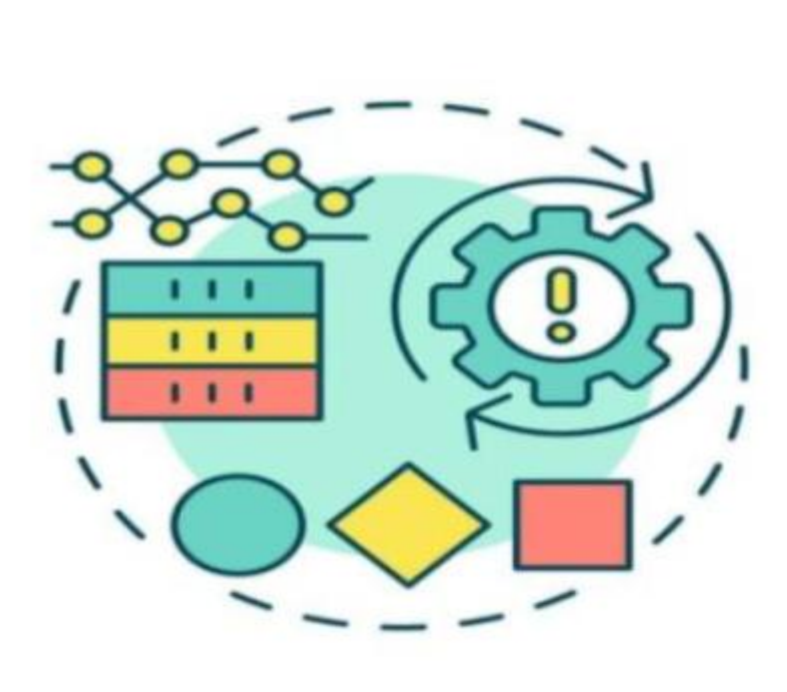


Fig. 5.8 The value of Big Data

# Variability

- This refers to establishing if the contextualizing structure of the data stream is regular and dependable even in conditions of extreme unpredictability.



# Big Data Analytics

- Involves analyzing datasets to uncover insights, trends, and patterns.
- Technologies commonly used in data analytics include statistical analysis software, data  visualization tools, and relational database management systems (RDBMS).

SAP Analytics Cloud | Features and Capabilities

# Big Data Analytics

- Big data analytics uses advanced analytic techniques against huge, diverse datasets that include
- structured,
- semi-structured, and
- unstructured data,
- from different sources, and
- in various sizes from terabytes to zettabytes.

# Big Data Analytics

- Big-Data Analytics encompasses the methodologies, tools, and practices involved in analyzing and managing data, covering tasks such as data collection, organization, and storage.
- The primary objective of data analytics is to utilize statistical analysis and technological methods to uncover patterns and address challenges.
- It provides valuable insights and forecasts that help businesses make informed decisions to improve their operations and outcomes.

# Big Data Analytics

- Some of the common types are:
- Descriptive analytics,
- Diagnostic analytics,
- Predictive analytics, and
- Prescriptive analytics

# Global Trends of Big Data

### **Moore’s Law**:

The exponential growth of computing power as per Moore's Law has enabled the handling and analysis of massive datasets, driving the evolution of Big Data Analytics.

### **Mobile Computing**:

With the widespread adoption of smartphones and mobile devices, access to vast amounts of data is now at our fingertips, enabling real-time connectivity and data collection from anywhere.

### **Social Networking**:

Platforms such as Facebook, Foursquare, and Pinterest facilitate extensive networks of user-generated content, interactions, and data sharing, leading to the generation of massive datasets ripe for analysis.

### **Cloud Computing**:

This paradigm shift in technology infrastructure allows organizations to access hardware and software resources remotely via the Internet on a pay-as-you-go basis, eliminating the need for extensive on-premises hardware and software investments.

# Working on Big Data Analytics

### 1. Gather Data

Collect structured and unstructured data from diverse sources like cloud, mobile apps, IoT sensors.

### 2. Process Data

Prepare data for accurate results using batch or stream processing.

### 3. Clean Data

Improve data quality by correcting formatting and eliminating duplicates/irrelevant entries.

### 4. Analyze Data

Turn cleaned data into insights using advanced analytics processes.

Tools: Tableau, Apache Hadoop, Cassandra, MongoDB, SAS.

# Using Orange Data Mining for Big Data Analytics

# Mining Data Streams

### What is a Data Stream?

A continuous, real-time flow of data from sources like sensors, satellite images, and web traffic.

### What is Mining Data Streams?

Extracting patterns and knowledge from real-time data flows without complete storage. Example: Analyzing website traffic spikes for election results.

# Future of Big Data Analytics

### Real-Time Analytics

Instantaneous data processing for immediate insights and actions.

### Advanced Predictive Models

Integration of sophisticated ML/AI for precise trend forecasting.

### Quantum Computing

Revolutionary processing power to solve complex problems faster.