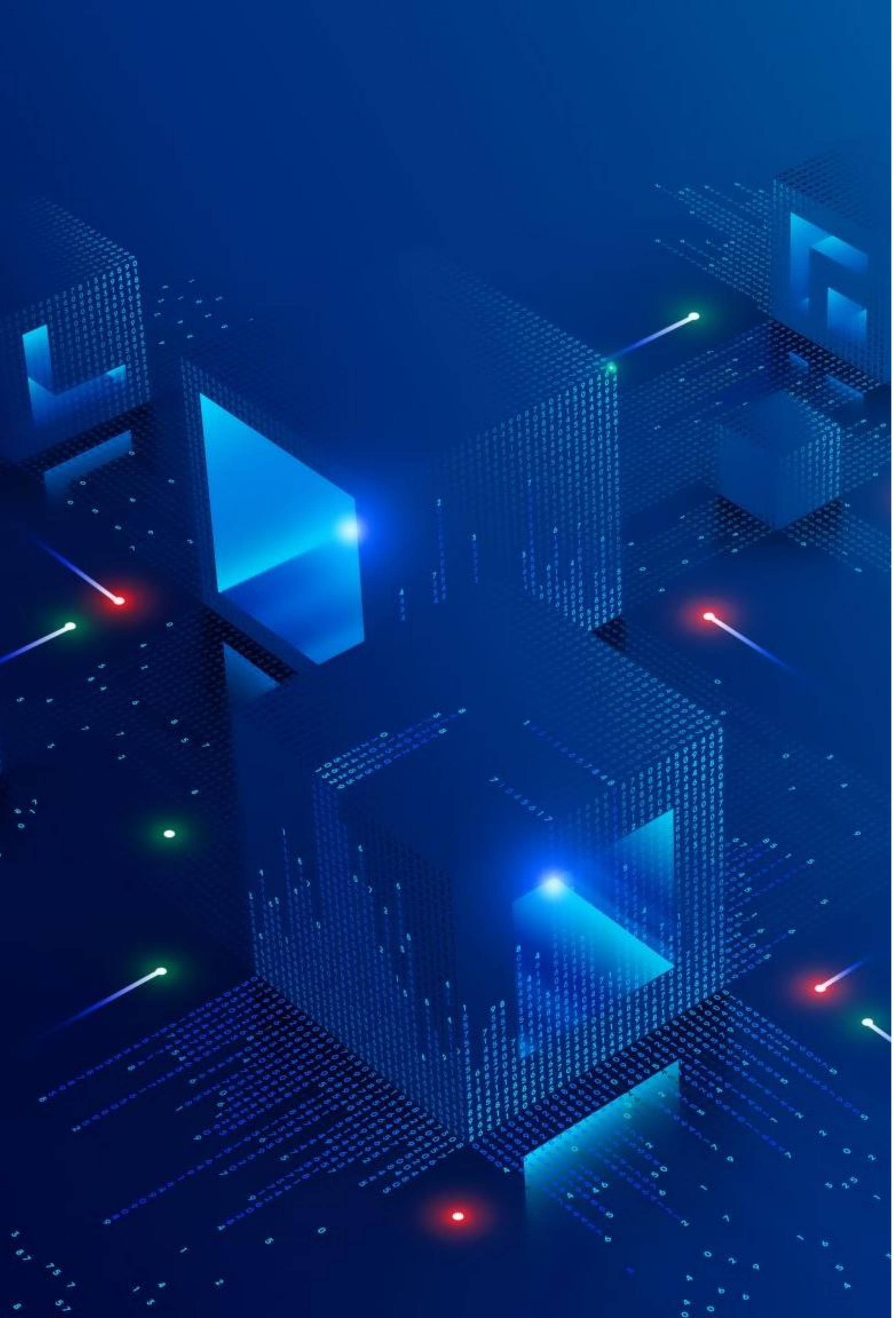
# UNIT 1
# Introduction to Big Data & Hadoop

Dr. Nilesh M. Patil

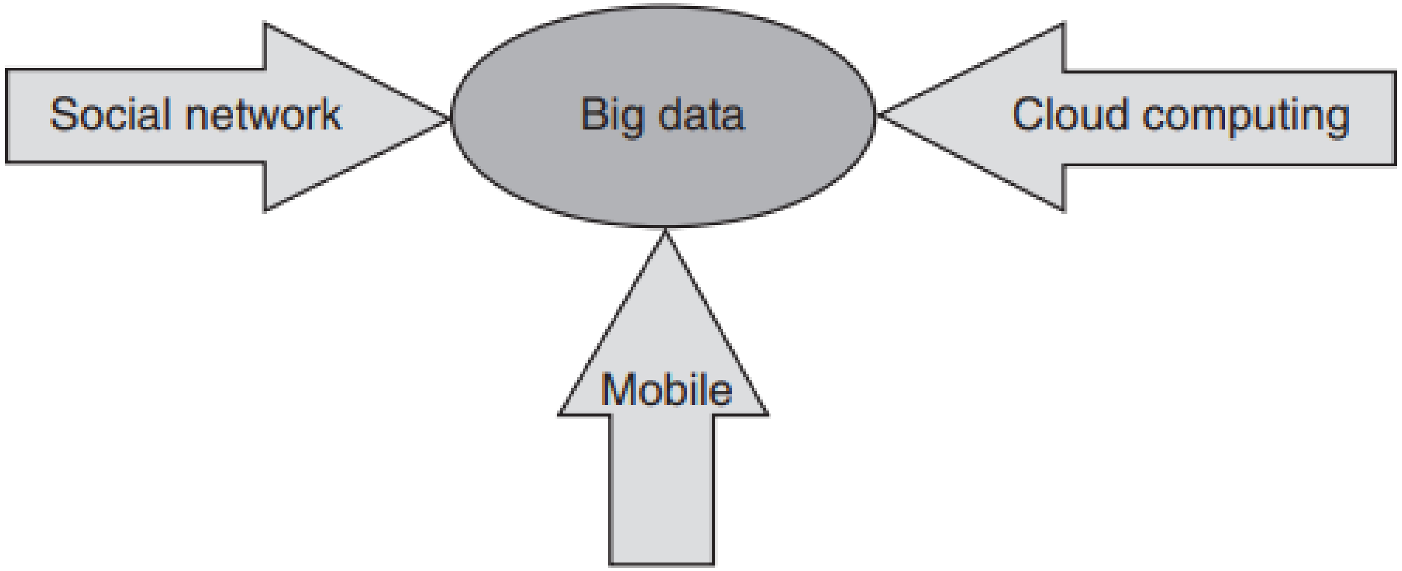


# Introduction

- Big Data is a massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools.
- "Big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

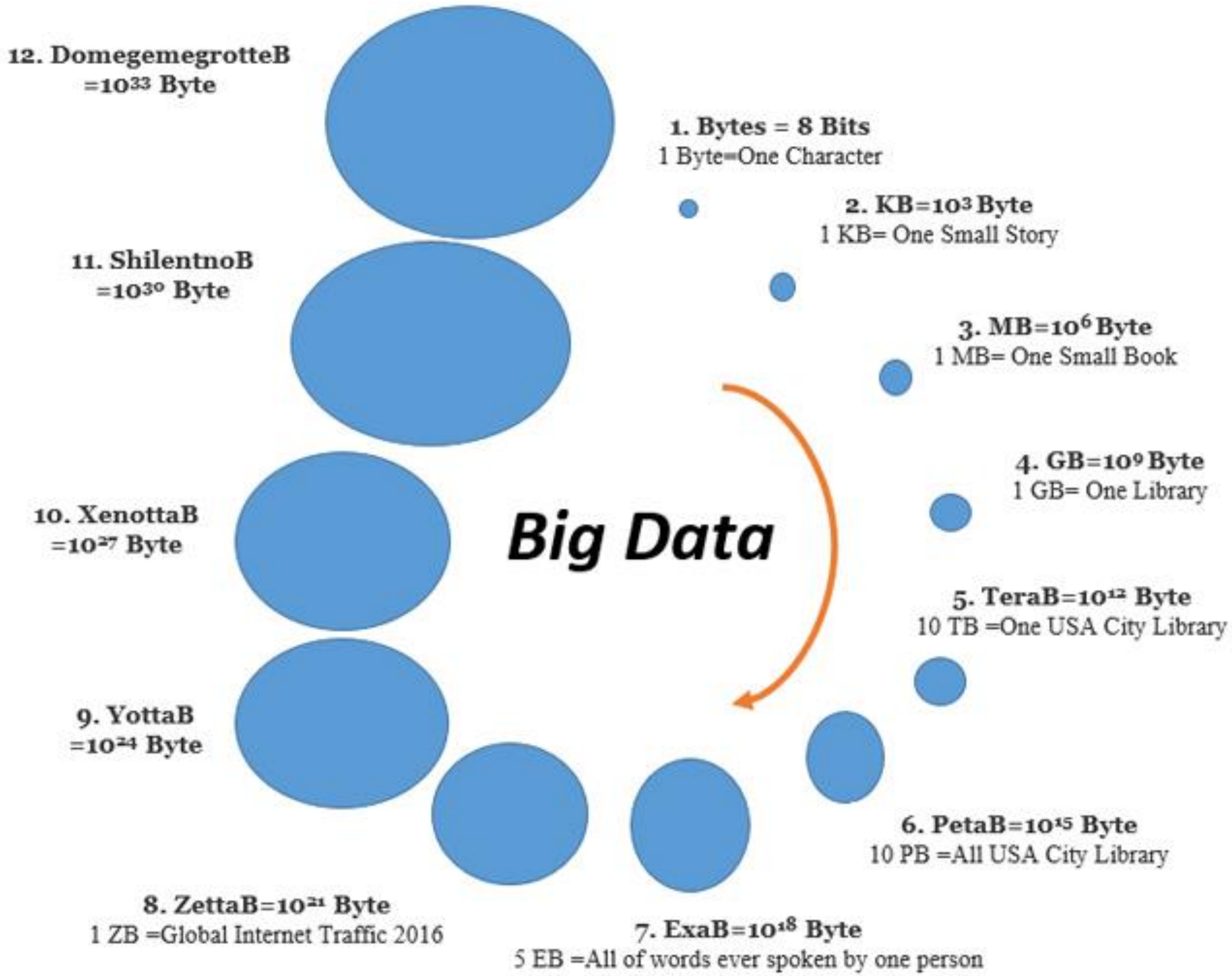
  - Gartner, Research and Advisory Company
- Data also exists in different formats, like structured data, semi-structured data, and unstructured data.
- Structured Data: Excel Sheet
- Semi-structured Data: Email
- Unstructured Data: Pictures and Videos

# Big Data – Result of Three Computing Trends



Mobile Computing uses hand-held devices, such as smartphones and tablets; Social Networking, such as Facebook and Pinterest; and Cloud Computing by which one can rent or lease the hardware setup for storing and computing.

# Big Data Sizes



1. **Bytes = 8 Bits**
1 Byte=One Character

2. **KB=$10^3$ Byte**
1 KB= One Small Story

3. **MB=$10^6$ Byte**
1 MB= One Small Book

4. **GB=$10^9$ Byte**
1 GB= One Library

5. **TeraB=$10^{12}$ Byte**
10 TB =One USA City Library

6. **PetaB=$10^{15}$ Byte**
10 PB =All USA City Library

7. **ExaB=$10^{18}$ Byte**
5 EB =All of words ever spoken by one person

8. **ZettaB=$10^{21}$ Byte**
1 ZB =Global Internet Traffic 2016

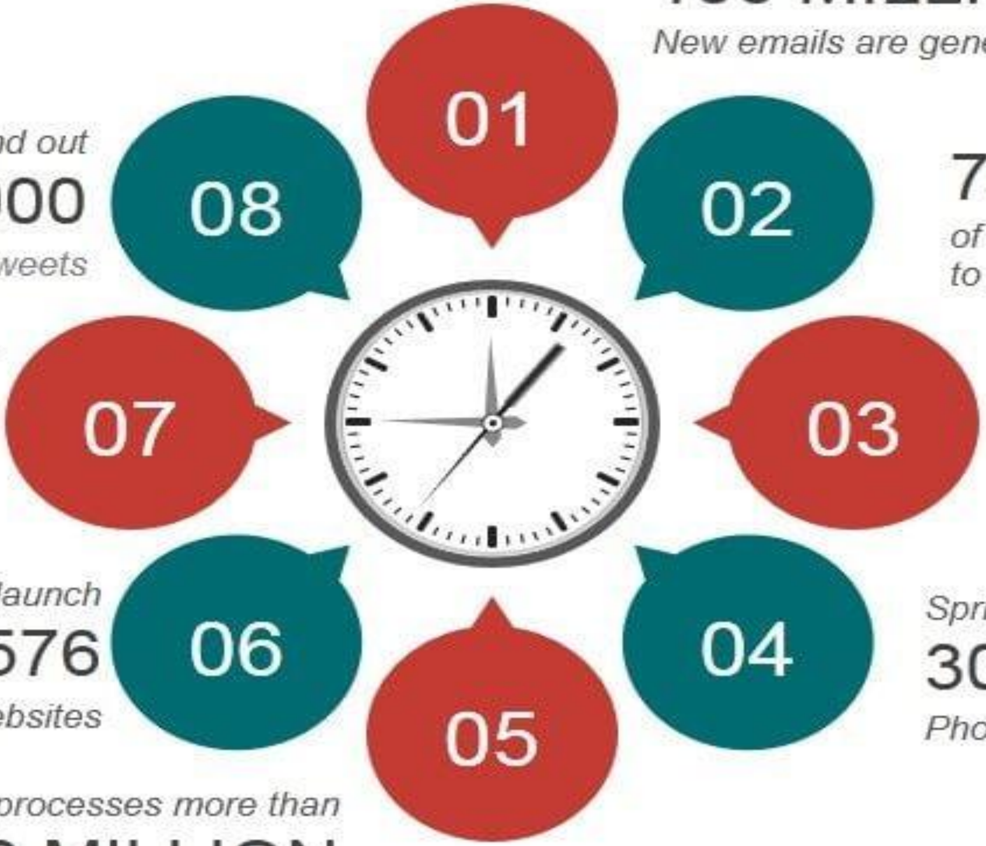
9. **YottaB =$10^{24}$ Byte**

10. **XenottaB =$10^{27}$ Byte**

11. **ShilentnoB =$10^{30}$ Byte**

12. **DomegemegrotteB =$10^{33}$ Byte**

# Big Data Facts-how Big Is Big Data



*More than*
100 MILLION
*New emails are generated*

72 HOURS
*of new video are uploaded to you tube*

*Walmart processes almost*
15,000
*Transactions*

*Sprint processes more than*
300,000
*Phone calls*

*Google processes more than*
2 MILLION
*Search queries*

*Individuals and organizations launch*
576
*New websites*

*Facebook processes almost*
350GB
*of data*

*Twitter users send out*
285,000
*Tweets*

# How Big is Big Data

50 Million
*Tweets per day*

700 Billion
*Total minutes spent on Facebook each month*

1.3 Exabytes
*Data sent and received by mobile internet users*

72.9 Items
*Products ordered on amazon per second*

24 Petabytes
*Data per day processed by Google*

20 Hours
*Video upload to YouTube every minute*

375 Megabytes
*Data consumed by households each day*

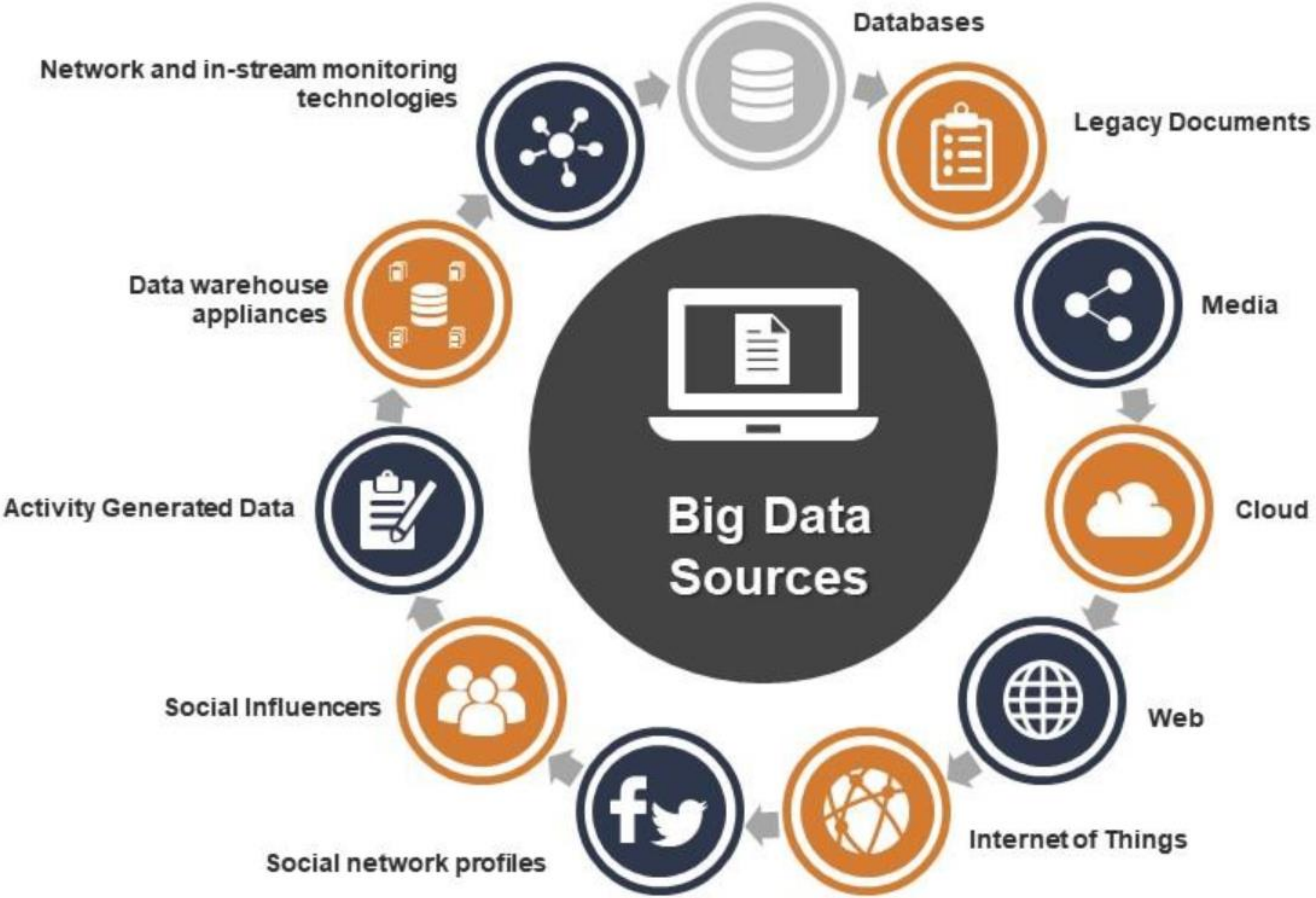
2.9 Million
*Number of emails sent every second*

# Example 1

- If we try to attach a document that is of 100 megabytes in size to an email we would not be able to do so.
- As the email system would not support an attachment of this size.
- Therefore this 100 megabytes of attachment with respect to email can be referred to as Big Data.

# Example 2

- Let’s say we have around 10 terabytes of image files, and we want to resize and enhance these images within a given time frame.
- Therefore this 10 terabytes of image files can be referred to as Big Data with respect to processing them on a desktop computer.



Big Data Sources

Databases

Legacy Documents

Media

Cloud

Web

Internet of Things

Social network profiles

Social Influencers

Activity Generated Data

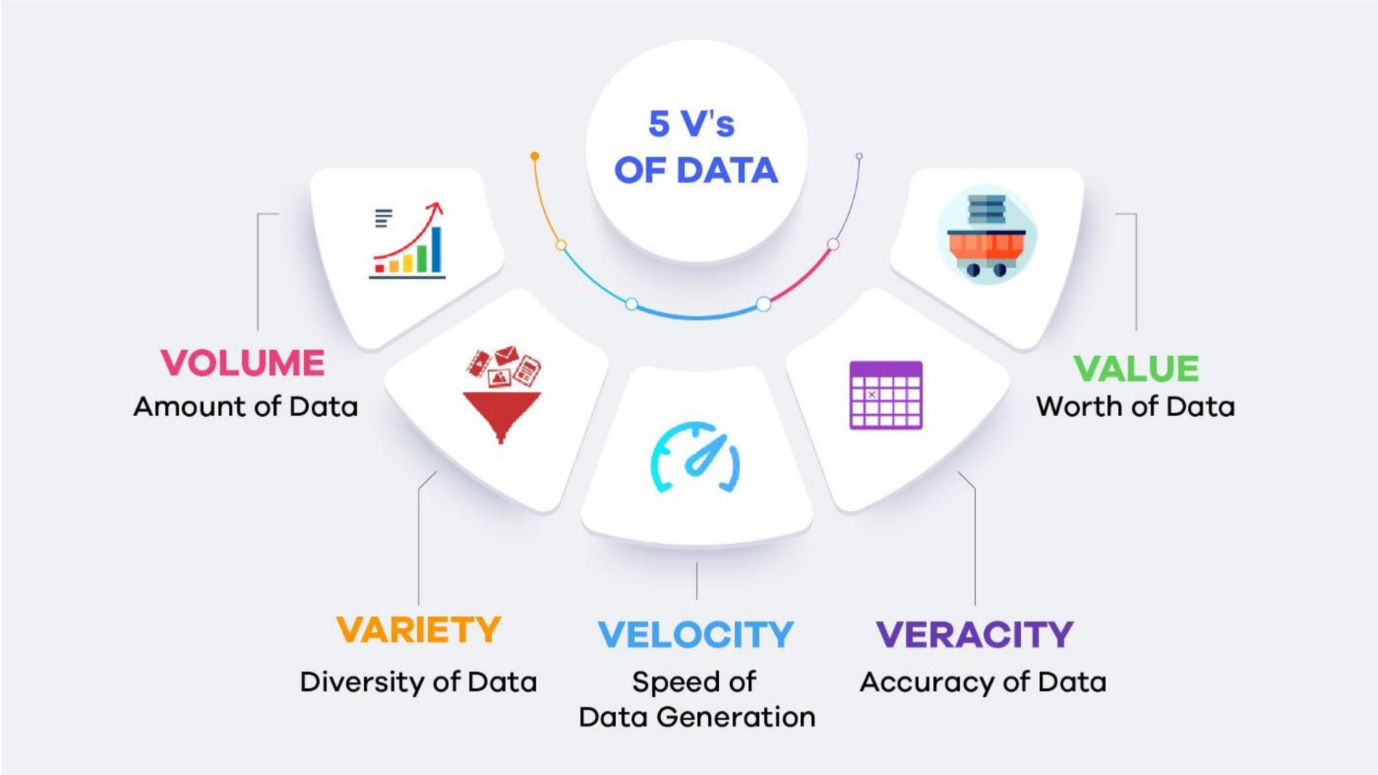
Data warehouse appliances

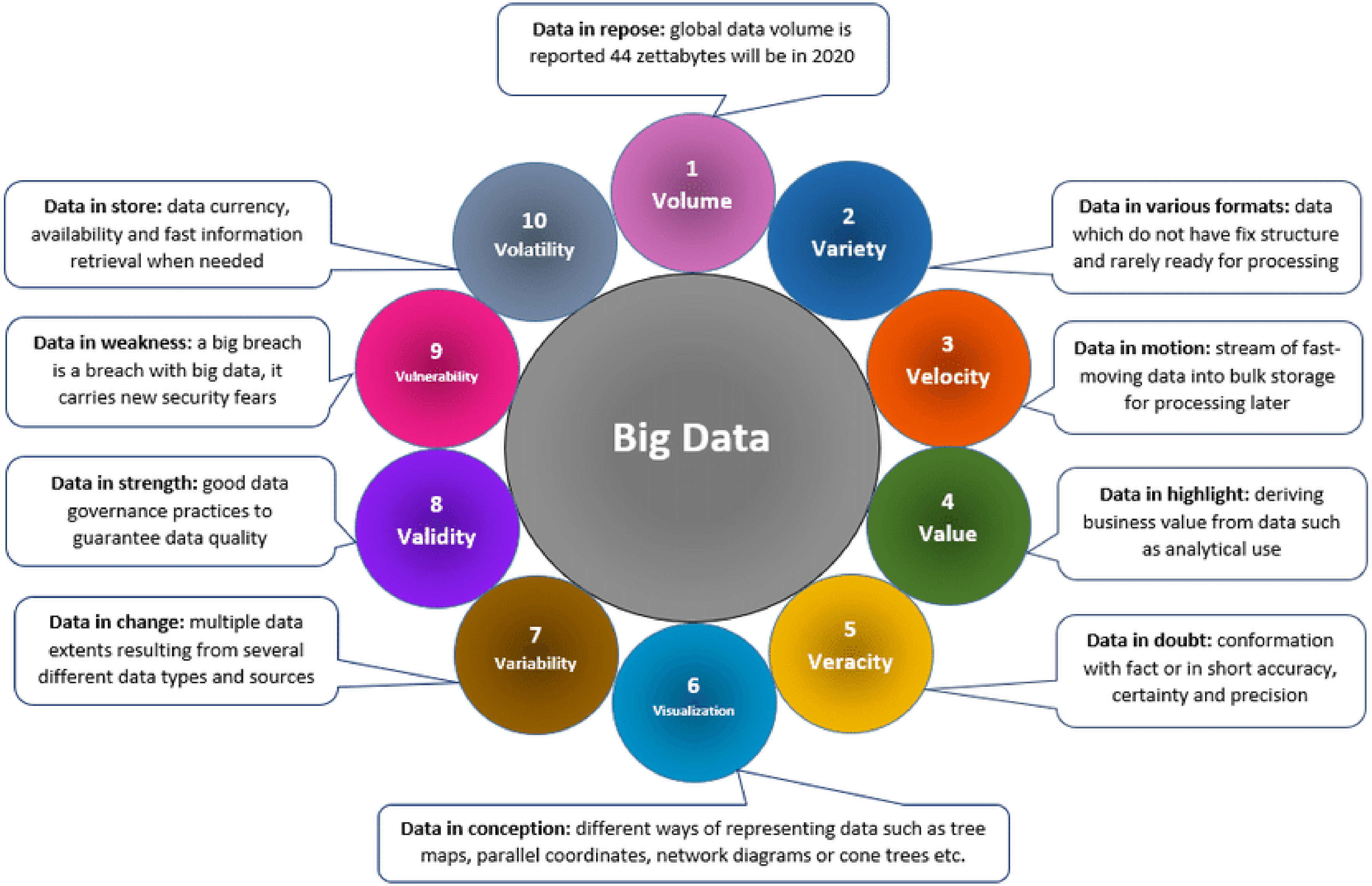
Network and in-stream monitoring technologies

# The Sources of Big Data

1. Black Box Data: This is the data generated by airplanes, including jets and helicopters. Black box data includes flight crew voices, microphone recordings, and aircraft performance information.
2. Social Media Data: This is data developed by such social media sites as Twitter, Facebook, Instagram, Pinterest, and Google+.
3. Stock Exchange Data: This is data from stock exchanges about the share selling and buying decisions made by customers.
4. Power Grid Data: This is data from power grids. It holds information on particular nodes, such as usage information.
5. Transport Data: This includes possible capacity, vehicle model, availability, and distance covered by a vehicle.
6. Search Engine Data: This is one of the most significant sources of big data. Search engines have vast databases where they get their data.

# Characteristics of Big Data/ 5V's of Big Data





Big Data

1 Volume

**Data in repose:** global data volume is reported 44 zettabytes will be in 2020

2 Variety

**Data in various formats:** data which do not have fix structure and rarely ready for processing

3 Velocity

**Data in motion:** stream of fast-moving data into bulk storage for processing later

4 Value

**Data in highlight:** deriving business value from data such as analytical use

5 Veracity

**Data in doubt:** conformation with fact or in short accuracy, certainty and precision

6 Visualization

**Data in conception:** different ways of representing data such as tree maps, parallel coordinates, network diagrams or cone trees etc.

7 Variability

**Data in change:** multiple data extents resulting from several different data types and sources

8 Validity

**Data in strength:** good data governance practices to guarantee data quality

9 Vulnerability

**Data in weakness:** a big breach is a breach with big data, it carries new security fears

10 Volatility

**Data in store:** data currency, availability and fast information retrieval when needed

# Types of Big Data

## Structured

Pre-defined data models like databases

Usually text only

Easy to search and filter

Examples: Dates, phone numbers, transaction information

## Semi-Structured

Both structured & unstructured qualities

Considerably easier to analyze than unstructured data

Examples: Emails, CSV files, JSON files

## Unstructured

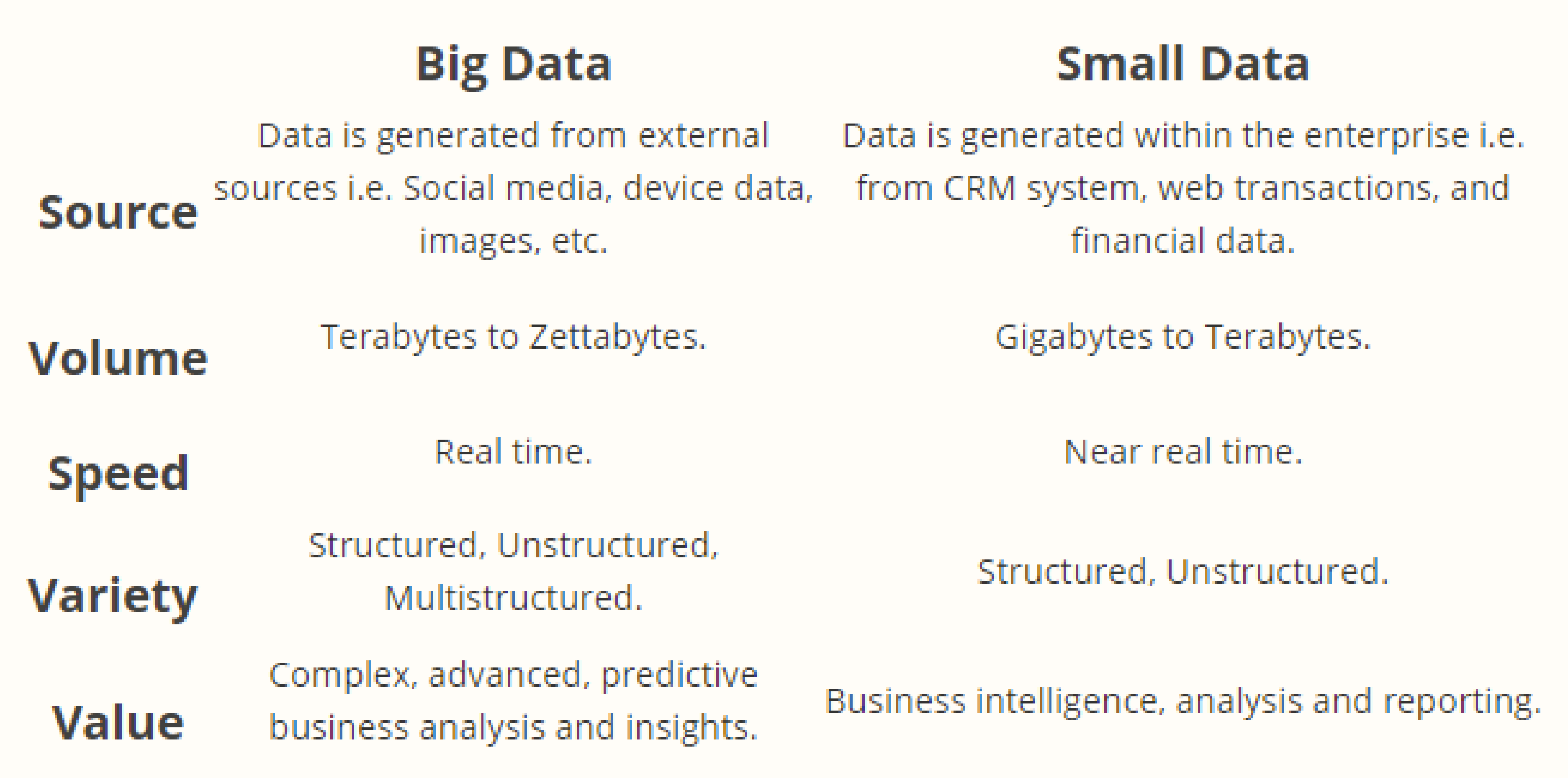
No Pre-defined data models

Difficult to search through

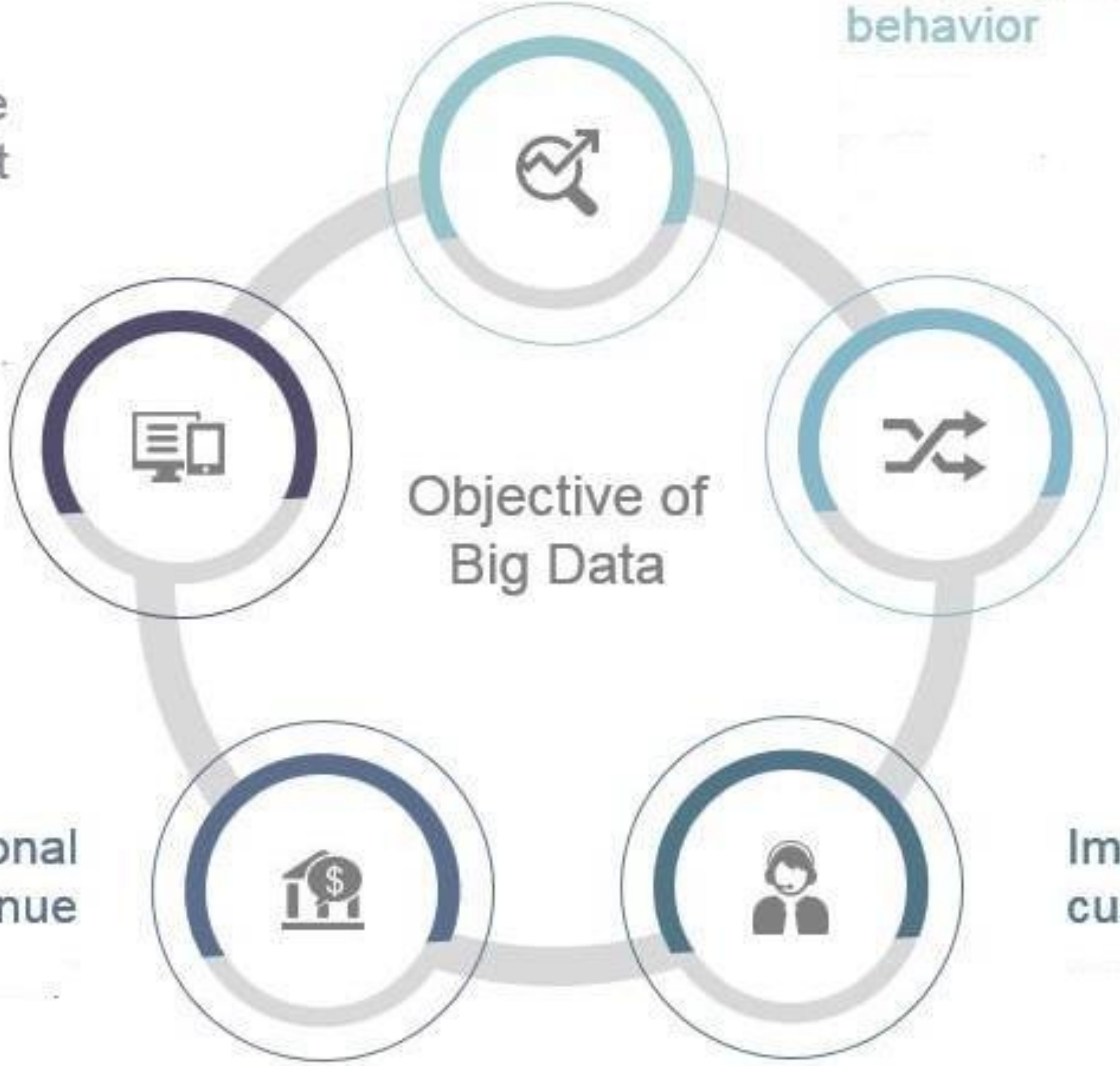
Usually stored as different types of files

Examples: Social media data, audio files, images

# Big Data Vs Small Data



| | Big Data | Small Data |
|---|---|---|
| **Source** | Data is generated from external sources i.e. Social media, device data, images, etc. | Data is generated within the enterprise i.e. from CRM system, web transactions, and financial data. |
| **Volume** | Terabytes to Zettabytes. | Gigabytes to Terabytes. |
| **Speed** | Real time. | Near real time. |
| **Variety** | Structured, Unstructured, Multistructured. | Structured, Unstructured. |
| **Value** | Complex, advanced, predictive business analysis and insights. | Business intelligence, analysis and reporting. |



Objective of Big Data

Analyzing customer behavior

Be more responsive to the market

Combining multiple data sources

Generate additional revenue

Improving customer service

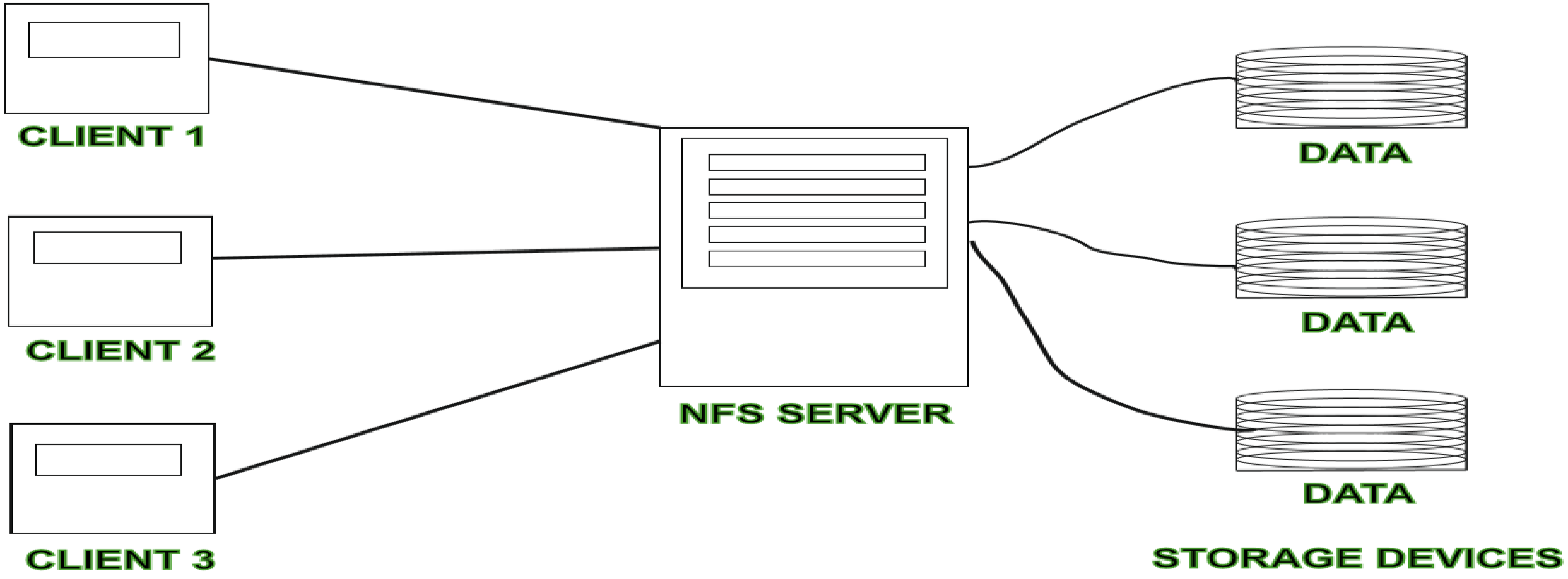
# Traditional data management Approach

**Traditional data systems like relational databases and data warehouses**

- Clearly defined fields organized in records
- Schema-on-write
- A design to get data from the disk and load the data into memory to be processed by applications
- Structured Query Language (SQL)
- Relational and warehouse database systems often read data in 8k or 16k block sizes
- In a number of traditional siloed environments data scientists can spend 80% of their time looking for the right data and 20% of the time doing analytics

# Distributed file system

- A **Distributed File System (DFS)** as the name suggests, is a file system that is distributed on multiple file servers or multiple locations.
- It allows programs to access or store isolated files as they do with the local ones, allowing programmers to access files from any network or computer.

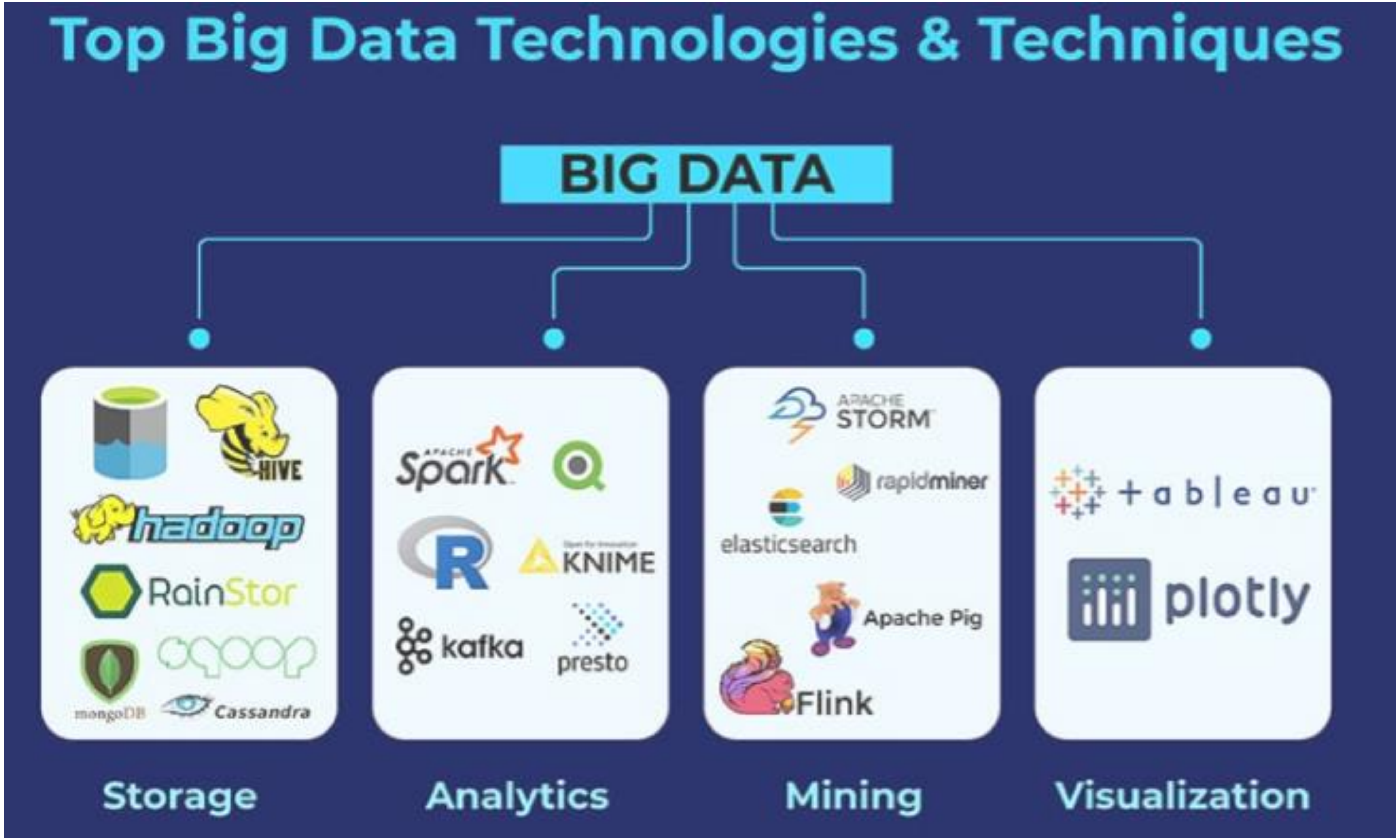


# Big Data Approach

- **An approach that involves managing Big Data from different sources or databases.**
- **Many IT tools** are available for Big Data projects.

–**Hadoop-** Storage requirement

–**Apache Spark-** Stream Processing



# Top Big Data Technologies & Techniques

BIG DATA

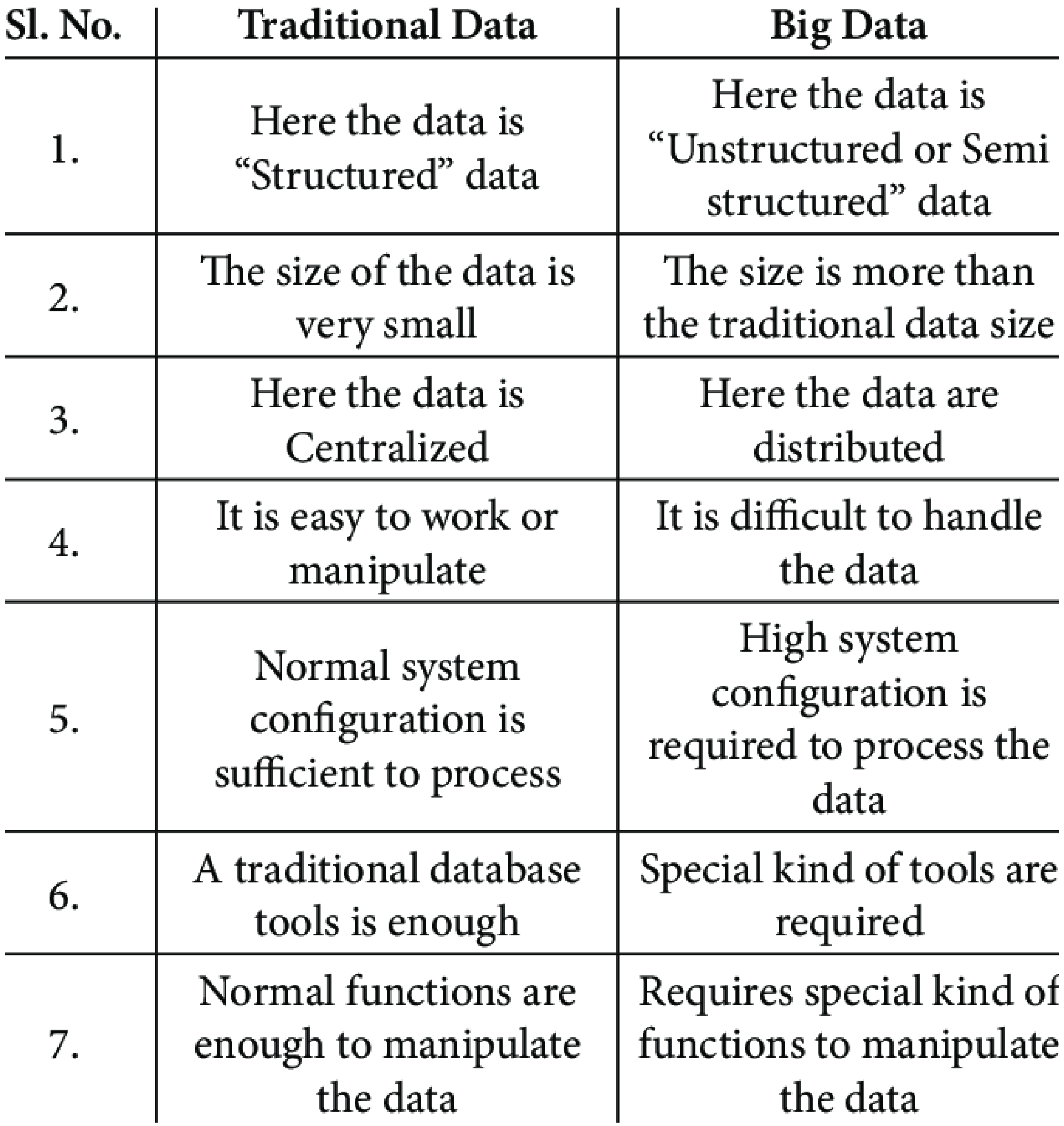
Storage

Analytics

Mining

Visualization

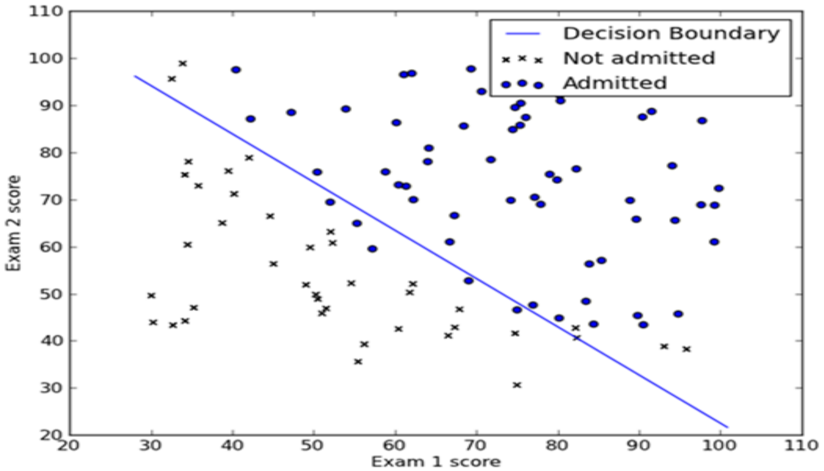
# Traditional Versus Big Data Approach

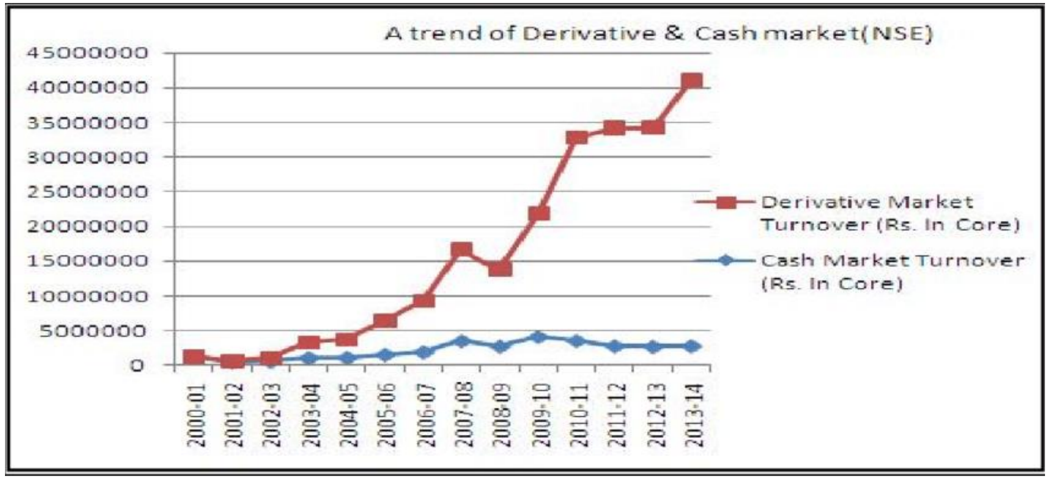


| Sl. No. | Traditional Data | Big Data |
|---|---|---|
| 1. | Here the data is “Structured” data | Here the data is “Unstructured or Semi structured” data |
| 2. | The size of the data is very small | The size is more than the traditional data size |
| 3. | Here the data is Centralized | Here the data are distributed |
| 4. | It is easy to work or manipulate | It is difficult to handle the data |
| 5. | Normal system configuration is sufficient to process | High system configuration is required to process the data |
| 6. | A traditional database tools is enough | Special kind of tools are required |
| 7. | Normal functions are enough to manipulate the data | Requires special kind of functions to manipulate the data |

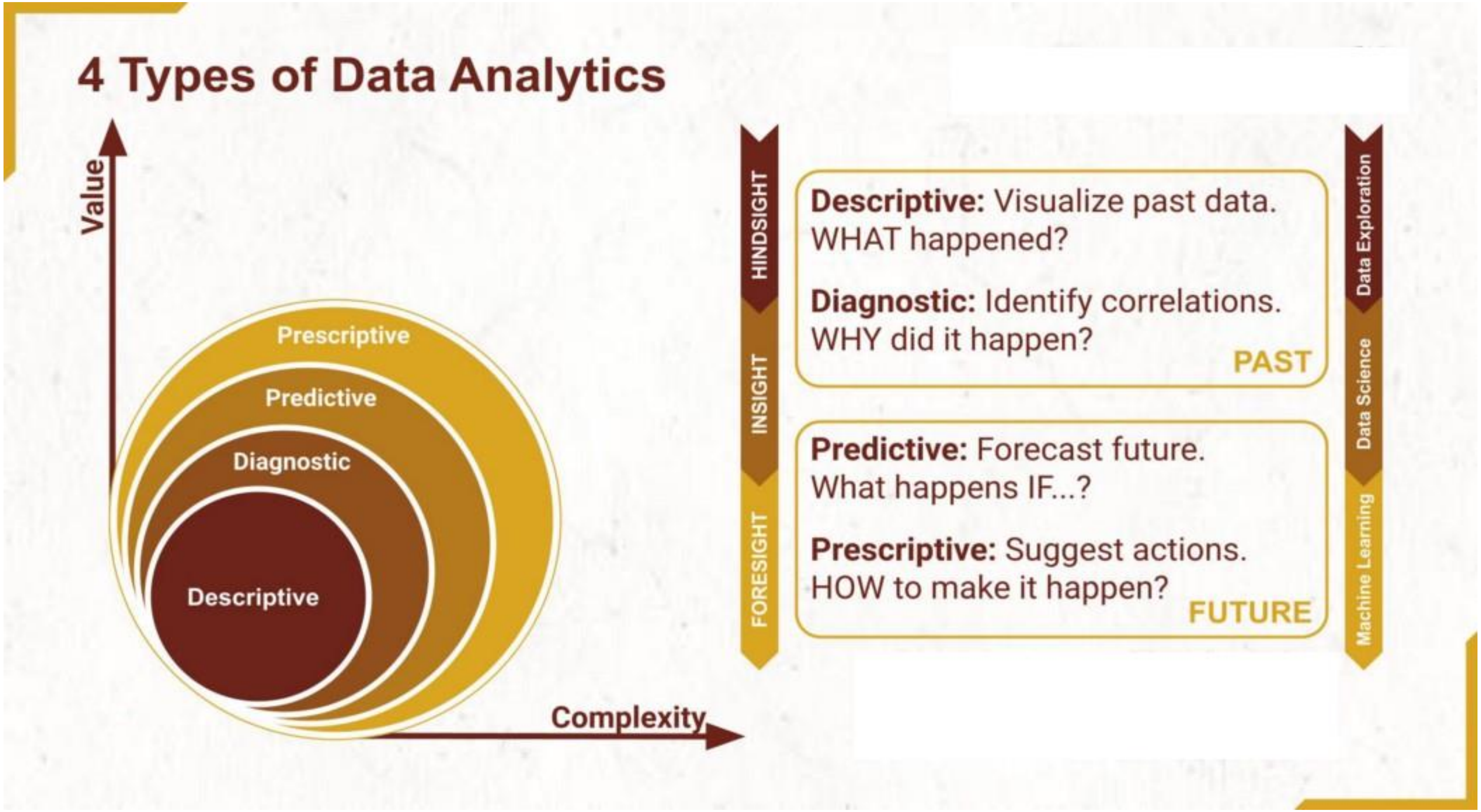
# Big Data Analytics

- Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences.
- Big data analytics enables analysts, researchers, and business users to leverage big data, which was previously inaccessible and unusable, for faster and better decision-making.





# Types of Big Data Analytics



## 4 Types of Data Analytics

**Descriptive:** Visualize past data.
WHAT happened?

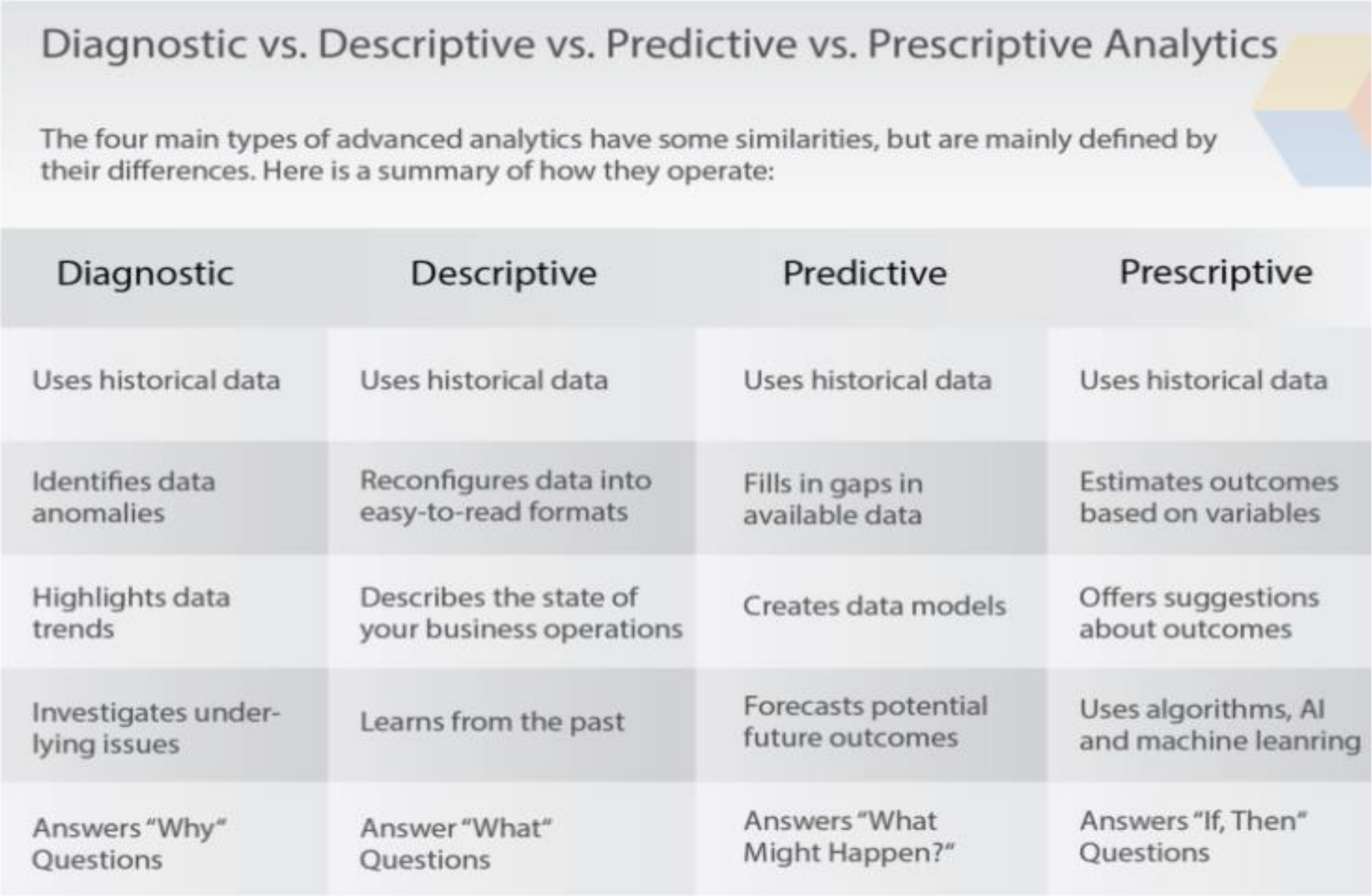
**Diagnostic:** Identify correlations.
WHY did it happen?

PAST

**Predictive:** Forecast future.
What happens IF...?

**Prescriptive:** Suggest actions.
HOW to make it happen?

FUTURE

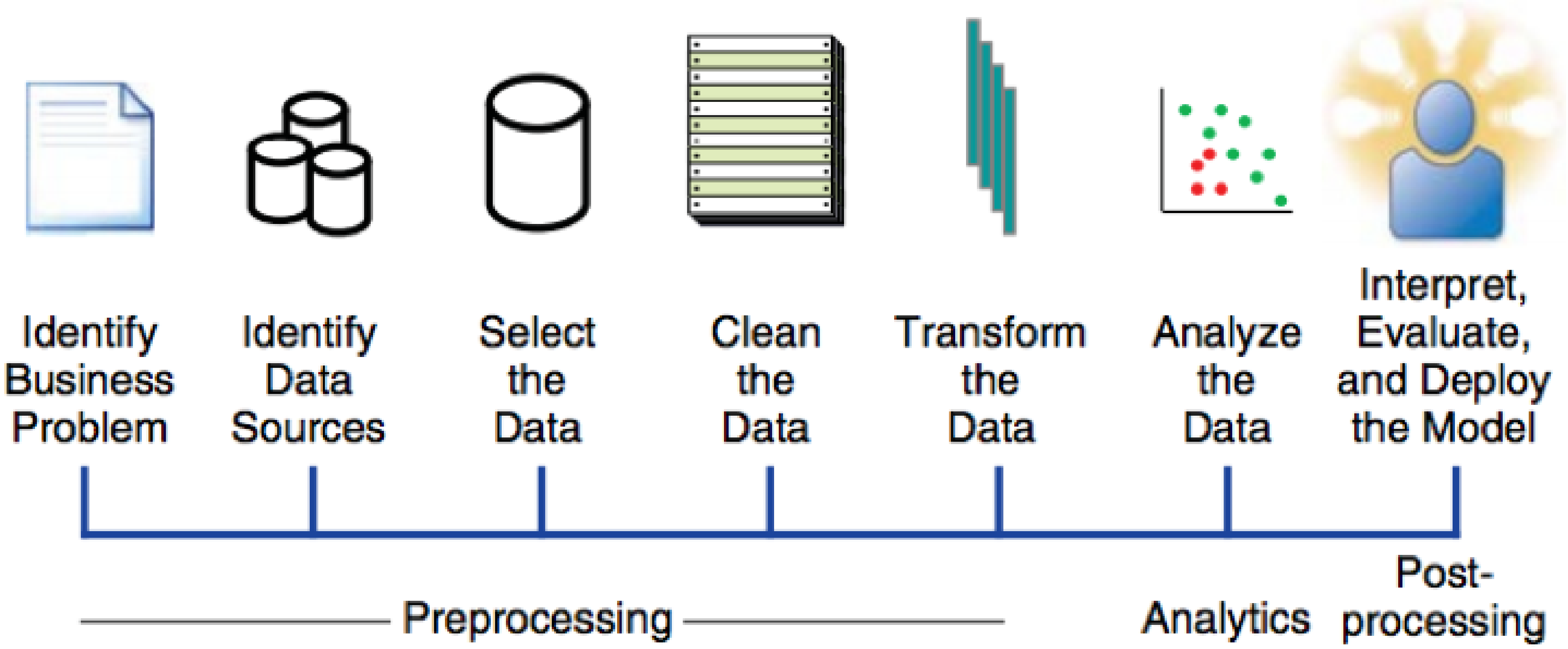


## Diagnostic vs. Descriptive vs. Predictive vs. Prescriptive Analytics

The four main types of advanced analytics have some similarities, but are mainly defined by their differences. Here is a summary of how they operate:

| Diagnostic | Descriptive | Predictive | Prescriptive |
|---|---|---|---|
| Uses historical data | Uses historical data | Uses historical data | Uses historical data |
| Identifies data anomalies | Reconfigures data into easy-to-read formats | Fills in gaps in available data | Estimates outcomes based on variables |
| Highlights data trends | Describes the state of your business operations | Creates data models | Offers suggestions about outcomes |
| Investigates underlying issues | Learns from the past | Forecasts potential future outcomes | Uses algorithms, AI and machine leanring |
| Answers “Why“ Questions | Answer “What“ Questions | Answers “What Might Happen?“ | Answers “If, Then“ Questions |

# Overview of the Analytics Process Model



Identify Business Problem

Identify Data Sources

Select the Data

Clean the Data

Transform the Data

Analyze the Data

Interpret, Evaluate, and Deploy the Model

Preprocessing

Analytics

Post-processing

# Benefits of Big Data Analytics

- Data accumulation from multiple sources, including the Internet, social media platforms, online shopping sites, company databases, external third-party sources, etc.
- Real-time forecasting and monitoring of business as well as the market.
- Identify crucial points hidden within large datasets to influence business decisions.
- Promptly mitigate risks by optimizing complex decisions for unforeseen events and potential threats.
- Identify issues in systems and business processes in real-time.
- Unlock the true potential of data-driven marketing.
- Dig in customer data to create tailor-made products, services, offers, discounts, etc.
- Facilitate speedy delivery of products/services that meet and exceed client expectations.
- Diversify revenue streams to boost company profits and ROI.
- Respond to customer requests, grievances, and queries in real-time.
- Foster innovation of new business strategies, products, and services.

# Challenges in Big Data Analytics

- **Data Sources**: Integration challenges when it comes to combining data from sources such as social media pages, financial reports, documents by employees, customer logs, presentations, emails, etc., to create insightful reports.
- **Data Growth**: Data is growing exponentially with time, and with that, enterprises are struggling to store large amounts of data.
- **Real-time Insights**: Data sets are a treasure trove of insights. However, data sets are of no value if no real-time insights are drawn from them.
- **Data Validation**: Data validation on a Big Data scale can be rather difficult. An organization can get similar sets of data from different sources but the data from these sources may not always be similar.
- **Data Security**: Data security is usually put on the back burner, which is not a wise move at all as unprotected data can fast become a serious problem. Stolen records can cost an organization millions.
- **Big Data Skills**: Running Big Data tools requires expertise that is possessed by data scientists, data engineers, and data analysts. Although organizations are spending on recruiting professionals with such skills, organizations are also investing in training their existing staff as well.

# Desired Properties of a Big Data System

- Robustness and fault tolerance
- Low latency reads and updates
- Scalability
- Generalization
- Extensibility
- Ad hoc queries
- Minimal maintenance
- Debuggability

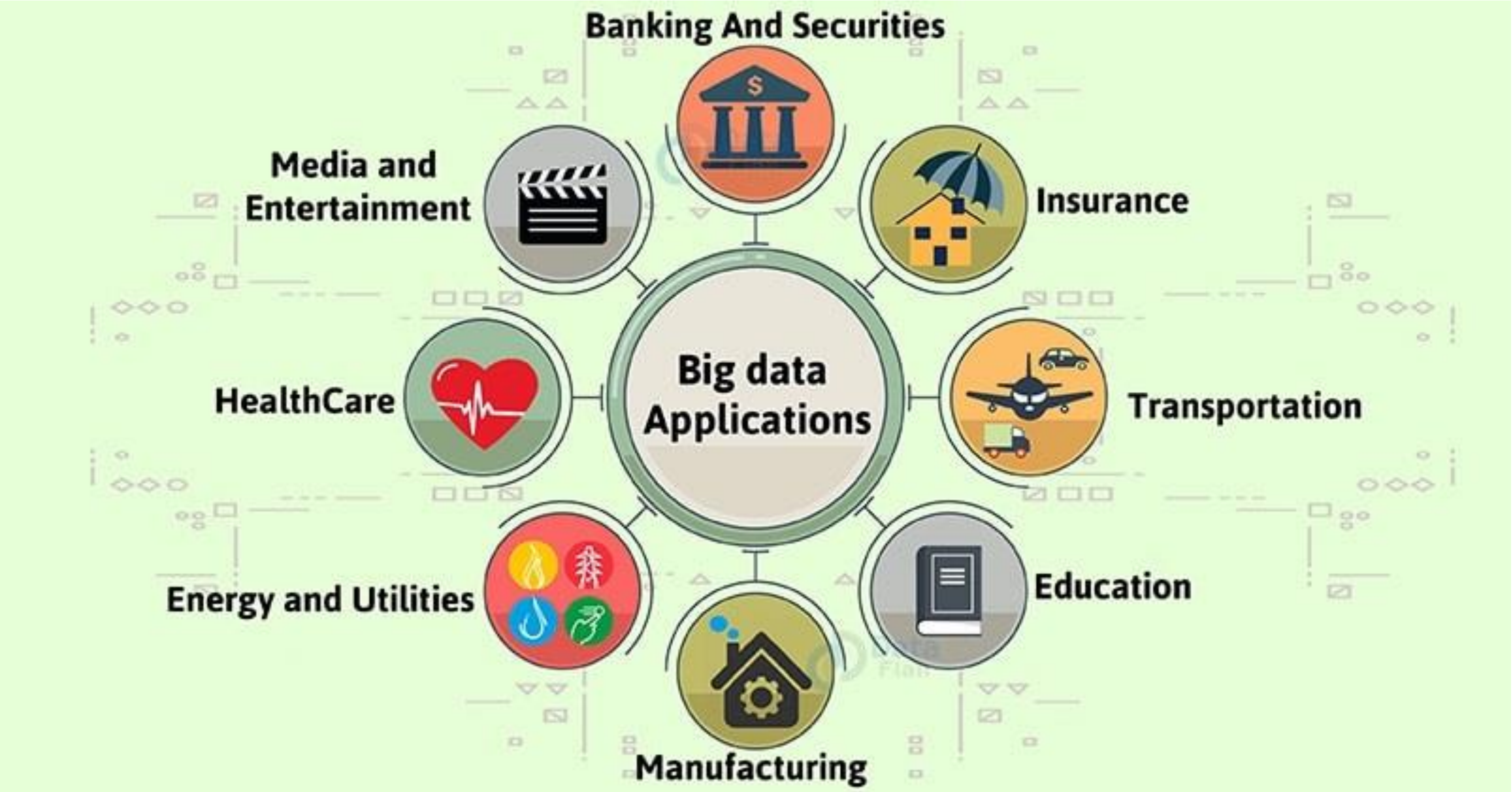
# Ethical Issues of Big Data

- **Private customer data and identity should remain private**
- **Shared private information should be treated confidentially**
- **Customers should have a transparent view**
- **Big Data should not interfere with human will**
- **Big data should not institutionalize unfair biases**

# Privacy Issues in Big Data

- Privacy breaches and embarrassments
- Anonymization could become impossible
- Data masking could be defeated to reveal personal information
- Unethical actions based on interpretations
- Big data analytics are not 100% accurate
- Discrimination
- Few (if any) legal protections exist for the involved individuals
- Big data will probably exist forever
- Concerns for e-discovery
- Making patents and copyrights irrelevant

# Application Areas of Big Data



# Case Study 1 : Feedback analysis using word count

- What problem does Hadoop solve?
- Businesses and governments have a large amount of data that needs to be analyzed and processed very quickly.
- If this data is fragmented into smaller chunks and spread over many machines, all those machines process their portion of the data in parallel and the results are obtained extremely fast.
- For example, a huge data file containing feedback mails is sent to the customer service department.
- The objective is to find the number of times goods were returned and refund requested. This will help the business to find the performance of the vendor or supplier.
- It is a simple word count exercise.
- The client will load the data into the cluster (Feedback.txt), submit a job describing how to analyze that data (word count), the cluster will store the results in a new file (Returned.txt), and the client will read the results file.

# Case Study 2: Clickstream Analysis

- Everyday millions of people visit organizations’ website and this forms the face of the organization to the public.
- The website informs the public about various products and services that are available with them.
- Some organizations allow people to transact through their website.
- Clickstream data is generated when the customers or visitors interact through the website.
- These data include the pages they load, the time spent by them on each page, the links they clicked, the frequency of their visit, from which page do they exit, etc.
- This gives good amount of information about their customers and helps in understanding them better.
- Eventually website content, its organization, navigation and transaction completion can be improved.
- This ensures that the customers can easily find what they want.
- The website owner can understand the customers’ buying behavior.
- Researchers are working on the mathematical model to predict the customer loyalty. This can help the sales and marketing team for preparing their campaigning strategies.
- Analysts use these kind of data for finding correlation.

# Big Data Analytics Techniques

- **Association Rule Learning**
- **Classification Tree Analysis**
- **Genetic Algorithms**
- **Machine Learning**
- **Clustering**
- **Regression Analysis**
- **Neural Networks**
- **Text Mining and Natural Language Processing (NLP)**
- **Dimensionality Reduction**
- **Time Series Analysis**

The End