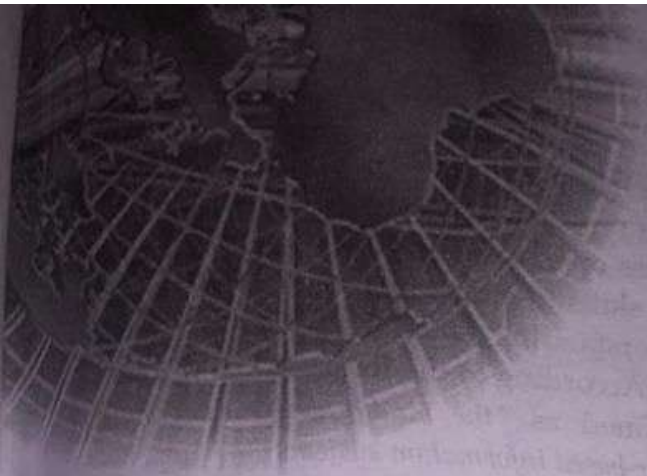


1

Getting an Overview of Big Data

"Information is the oil of the 21st century, and analytics is the combustion engine."–Peter **Sondergaard of the Gartner Group**.

The 21st century is characterized by the rapid advancement in the field of information technology. IT has become an integral part of daily life as well as various other industries, be it health, education, entertainment, science and technology, genetics, or business operations. In today's competitive and global economy, organizations must possess a number of skills to create their place and sustain in the market. One of the most crucial of these skills is an understanding of and the ability to utilize and harness the immense potential of information technology. According to the Information Technology Association of America, information technology is defined as *"the study, design, development, application, implementation, support or management of computer-based information systems."*

This is truly an information age where data is being generated at an alarming rate. This huge amount of data is often termed as Big Data. Organizations use data generated through various sources to run their businesses. They analyze the data to understand and interpret market trends, study customer behavior, and take financial decisions. The term Big Data is now widely used, particularly in the IT industry, where it has generated various job opportunities.

Big Data consists of large datasets that cannot be managed efficiently by the common database management systems. These datasets range from terabytes to exabytes. Mobile phones, credit cards, Radio Frequency Identification (RFID) devices, and social networking platforms, create huge amounts of data that may reside unutilized at unknown servers for many years. However, with the evolution of Big Data, this data can be accessed and analyzed to generate useful information.

The chapter introduces you to Big Data—the big buzzword of the IT industry–and its growing importance in almost every sector of human existence, be it education, health, science, technology, defense, lifestyle, etc.

What is Big Data?

Think of the following:

- Every second, there are around 8,22 tweets on Twitter.
- Every minute, nearly 510 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded on Facebook.
- Every hour, Walmart, a global discount departmental store chain, handles more than 1 million customer transactions.
- Every day, consumers make around 11.5 million payments by using PayPal.

According to IBM, "Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data."

Data is everywhere, in every industry in the form of numbers, images, videos, and text. As data continues to grow, so does the need to organize it. Collecting such huge amount of data would just be a waste of time, effort, and storage space if it cannot be put to any logical use. The need to sort, organize, analyze, and offer this critical data in a systematic manner leads to the rise of the much discussed term, Big Data.

The process of capturing or collecting Big Data is known as 'datafication'. Big Data is 'datafied' so that it can be used productively. Big Data cannot be made useful by simply organizing it, rather the data's usefulness lies in determining what we can do with it.

According to IBM, Big data is being generated by nearly everything around us at all times at an alarming velocity, volume, and variety. To extract meaningful value from Big Data, you need optimal processing power, analytical capabilities, and skills.

Figure 1.1 shows the features of Big Data:

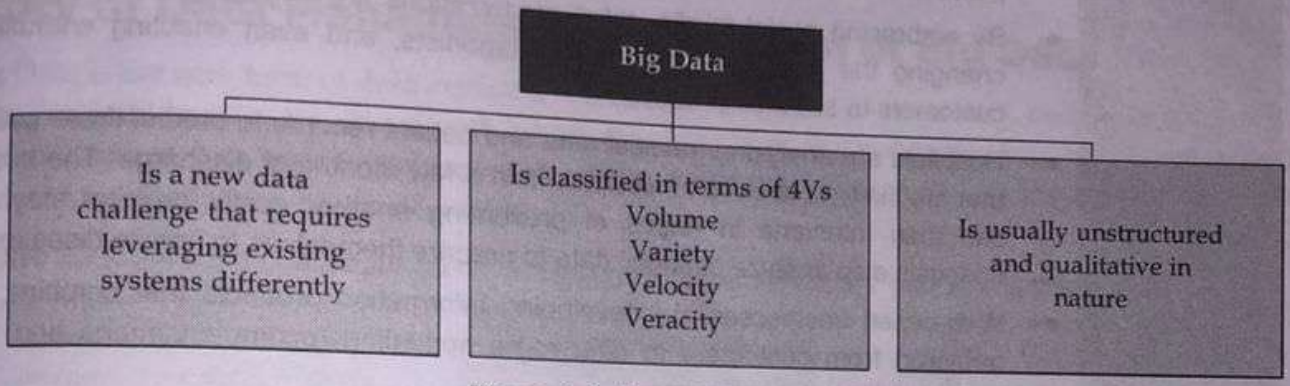


Figure 1.1: Features of Big Data

NOTE

By large or huge datasets or Big Data, we mean anything from a petabyte (1 PB = 1000 TB) to an exabyte (1 EB = 1000 PB) of data.

SCENARIO

Consider the scenario of an organization, Argon Technology, which provides Big Data analytical solutions to customers. Mr. Smith, the data analyst of the Argon Technology, is studying about Big Data and the ways in which it can be utilized in various sectors. He shares some examples with his team to enhance their knowledge.

EXHIBIT 1: Real-World Examples of Big Data

Some real-world examples of Big Data include:

- Consumer product companies and retail organizations are observing data on social media websites such as Facebook and Twitter. These sites help them to analyze customer behavior, preferences, and product perception. Accordingly, the companies can line up their upcoming products to gain profits. This phenomenon is also known as social media analytics.
- Manufacturers are monitoring minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money and replacing it too late triggers an expensive work stoppage.
- Manufacturers are also monitoring social networks, but with a different goal than marketers. They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- Financial service organizations are using the data mined from customer

interactions to slice and dice their users into finely-tuned segments. This enables these financial institutions to create increasingly relevant and sophisticated offers.

- Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions and other advertising mediums.
- Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed and which ones need a validating in-person visit from an agent.
- By embracing social media, retail organizations are engaging brand advocates, changing the perception of brand antagonists, and even enabling enthusiastic customers to sell their products.
- Hospitals are analyzing medical data and patient records to predict those patients that are likely to seek readmission within a few months of discharge. The hospital can then intervene in hopes of preventing another costly hospital stay. The hospitals also analyze patients' data to prepare themselves to handle diseases.
- Web-based businesses are developing information products that combine data gathered from customers to offer more appealing recommendations and more successful coupon programs.
- The government is making data public at the national, state, and city level for users to develop new applications that can generate public good. For example, weather data that is helpful for various industries.
- Sports teams are using data for tracking ticket sales and even for tracking team strategies. This is known as sports analytics.

Source: https://www.acquia.com/examples-big-data-projects

As already discussed, Big Data is a pool of huge amounts of data of all types, shapes, and formats collected from varied sources. Table 1.1 lists some common types of data and their sources:

Table 1.1: Types and Sources of Data

Type	Description	Source
Social Data	Refers to the information collected from various social networking sites and online portals	Facebook, Twitter, and LinkedIn
Machine Data	Refers to the information generated from (RFID) chips, bar code scanners, and sensors	RFID chip readings, Global Positioning System (GPS) results
Transactional Data	Refers to the information generated from online shopping sites, retailers, and Business to Business (B2B) transactions	Retail websites like eBay and Amazon

Google Inc. applied its expertise of data collecting, analyzing, and deriving conclusions to raise warnings for the flu plagues in the US approximately two weeks in advance of the existing public health services. To do this, Google monitored millions of users' health-tracking behaviors, and followed a cluster of queries on themes such as symptoms about flu, congestion in chest, and incidences of buying a thermometer. Google analyzed this collected data and generated consolidated results that revealed strong indications of flu levels across America. To determine the accuracy of this data, Google did further research and data comparison before sharing it with the general public.

History of Data Management – Evolution of Big Data

Big Data is the new term of data evolution directed by the enormous velocity, variety, and volume of data. Velocity implies the speed with which the data flows in an organization, variety refers to the varied forms of data such as structured or unstructured, and volume defines the amount or quantity of data an organization has to deal with.

Figure 1.2 shows the challenges faced while handling data over the past few decades:

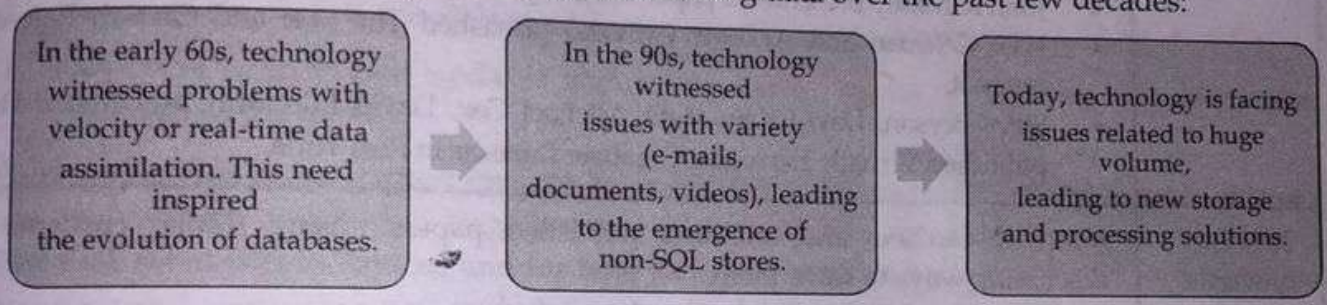


Figure 1.2: Evolution of Big Data

The advent of IT, the Internet, and globalization has facilitated increased volumes of data and information generation at an exponential rate, which has led to 'information explosion'. This, in turn, fueled the evolution of Big Data that started in 1940s and continues till date.

Information explosion is described as a continuous increase in the volume of the published information or data and the effects of this abundant information.

Table 1.2 lists some major milestones in the evolution of Big Data:

Table 1.2: Evolution of Big Data

Year	Milestone
1940s	An American librarian speculated the potential shortfall of shelves and cataloging staff, realizing the rapid increase in information and limited storage.
1960s	Automatic Data Compression was published in the Communications of the ACM. It states that the explosion of information in the past few years makes it necessary that requirements for storing information should be minimized. The paper described 'Automatic Data Compression' as a complete automatic and fast three-part compressor that can be used for any kind of information in order to reduce the slow external storage requirements and increase the rate of transmission from a computer system.

Table 1.2: Evolution of Big Data	
Year	**Milestone**
1970s	In Japan, the Ministry of Posts and Telecommunications initiated a project to study information flow in order to track the volume of information circulating in the country.
1980s	A research project was started by the Hungarian Central Statistics Office to account for the country's information industry. It measured the volume of information in bits.
1990s	Digital storage systems became more economical than paper storage. Challenges related to the amount of data and the presence of obsolete data became apparent. Some papers that discussed this concern are as follows: • Michael Lesk published How much information is there in the world? • John R. Masey presented a paper titled Big Data… and the Next Wave of InfraStress. • K.G. Coffman and Andrew Odlyzko published The Size and Growth Rate of the Internet. • Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes published Visually Exploring Gigabyte Datasets in Real Time.
2000 onwards	Many researchers and scientists published papers raising similar concerns and discussing ways to solve them. Various methods were introduced to streamline information. Techniques for controlling the Volume, Velocity, and Variety of data emerged, thus introducing 3D data management. A study was carried out in order to estimate the new and original information created and stored worldwide in four types of physical media: paper, film, optical media, and magnetic media.

Table 1.2 is only a synopsis of the evolution. The need for an adequate space and storage of data was always felt, and with time, Big Data grew into a technological phenomenon.

The next chapter will help you understand the business applicability of Big Data in various industries. 'Big Data' as a concept has been used for long. When researchers used computers to analyze huge volumes of data, they were actually analyzing the Big Data. The demand for faster access to data and the applications and programs to process this data led to the present concept of Big Data and Big Data analytics in the IT industry.

Suppose a bank plans to establish self-service kiosks in a major metro area. The marketing department wants to determine the busiest spots for establishing the self-service kiosks, on the basis of the traffic patterns of customers across the city. This information is not available in the existing data warehouse of the bank. In this situation, the bank can acquire the GPS location data of the customers through a third party, and thereby gather the information about the mobility patterns of its customers.

Thus, by using the right set of Big Data with the right technique of data extraction, preparation, and integration, today banks can identify the busiest spots in the city for establishing their self-service kiosks.

NOTE

World Wide Web (WWW) is the biggest source of generating data. More than 60% evolution happening in the field of data is just because of the use of the Internet.

Structuring Big Data

Structuring of data, in simple terms, is arranging the available data in a manner so that it becomes easy to study, analyze, and derive conclusion from it. But, why is structuring required?

In daily life, you may have come across questions like:

- How do I use to my advantage the vast amount of data and information I come across?
- Which news articles should I read of the thousands I come across?
- How do I choose a book of the millions available on my favorite sites or stores?
- How do I keep myself updated about new events, sports, inventions, and discoveries taking place across the globe?

Today, solutions to such questions can be found by information processing systems. These systems can analyze and structure a large amount of data specifically for you on the basis of what you searched, what you looked at, and for how long you remained at a particular page or website, thus scanning and presenting you with the customized information as per your behavior and habits. In other words, structuring data helps in understanding user behaviors, requirements, and preferences to make personalized recommendations for every individual.

When a user regularly visits or purchases from online shopping sites, say eBay, each time he/she logs in, the system can present a recommended list of products that may interest the user on the basis of his/her earlier purchases or searches, thus presenting a specially customized recommendation set for every user. This is the power of Big Data analytics.

Today, various sources generate a variety of data such as images, text, audios, etc. All such different types of data can be structured only if it is sorted and organized in some logical pattern. Thus, the process of structuring data requires one to first understand the various types of data available today.

Types of Data

Data that comes from multiple sources such as databases, Enterprise Resource Planning (ERP) systems, weblogs, chat history, and GPS maps, varies in its format. However, different formats of data need to be made consistent and clear to be used for analysis. Data is obtained primarily from the following types of sources:

- Internal sources, such as organizational or enterprise data
- External sources, such as social data

Table 1.3 compares the internal and external sources of data:

Table 1.3: Comparison between the Internal and External Sources of Data

Data Source	Definition	Examples of Sources	Application
Internal	Provides structured or organized data that originates from within the enterprise and helps run business	• Customer Relationship Management (CRM) • Enterprise Resource Planning (ERP) systems • Customers, details • Products and sales data • Generally OLTP and operational data	This data (current data in the operational system) is used to support daily business operations of an organization
External	Provides unstructured or unorganized data that originates from the external environment of an organization	• Business partners • Syndicate data suppliers • Internet • Government • Market research organizations	This data is often analyzed to understand the entities mostly external to the organization such as customers, competitors, market, and environment

On the basis of the data received from the aforementioned sources, Big Data comprises:

- Structured data
- Unstructured data
- Semi-structured data

In a real-world scenario, typically, the unstructured data is larger in volume than the structured and semi-structured data. Figure 1.3 illustrates the types of data that comprise Big Data:

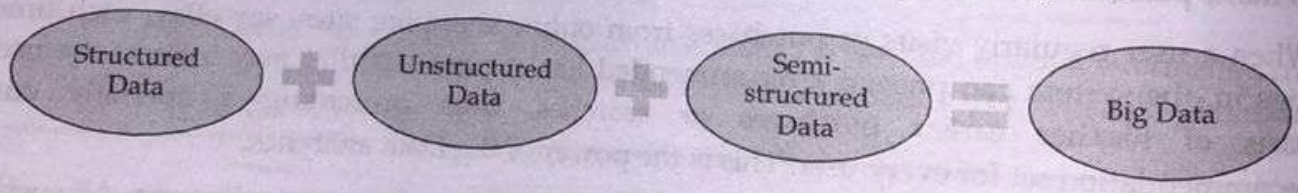


Figure 1.3: Types of Data

Structured Data

Structured data can be defined as the data that has a defined repeating pattern. This pattern makes it easier for any program to sort, read, and process the data. Processing structured data is much easier and faster than processing data without any specific repeating patterns.

SCENARIO

Reconsider the scenario of Mr. Smith, the Big Data analyst of Argon Technology, who is sharing his observations on the applications of Big Data with his team. In one such example, he tells that the data in most publishing houses is often captured by suitable software tools and maintained in a relational database, such as Oracle. The data stored in a relational database is in a structured format; therefore, it can directly be put to analysis, and the outcome con be used to take various organizational decisions.

Structured data:

- Is organized data in a predefined format
- Is the data that resides in fixed fields within a record or file
- Is formatted data that has entities and their attributes mapped
- Is used to query and report against predetermined data types

Some sources of structured data include:

- Relational databases (in the form of tables)
- Flat files in the form of records (like csv and tab-separated files)
- Multidimensional databases (majorly used in data warehouse technology)
- Legacy databases

Table 1.4 shows a sample of structured data in which the attribute data for every customer is stored in the defined fields:

Table 1.4: Sample of Structured Data

Customer ID	Name	Product ID	City	State
12365	Smith	241	Graz	Styria
23658	Jack	365	Wolfsberg	Carinthia
32456	Kady	421	Enns	Upper Austria

Unstructured Data

Unstructured data is a set of data that might or might not have any logical or repeating patterns.

SCENARIO

To better understand the concept of unstructured data, let us go back to the meeting of Mr. Smith. He explains that the publishing house also collects data from various blogs and websites. The data obtained from Web blogs or social media sites is considered as unstructured data because it does not follow any specific pattern and is inconsistent. The analysis of such data helps the organization to know more about customer preferences, feedback, and demands.

Unstructured data:

- Consists typically of metadata, i.e., the additional information related to data
- Comprises inconsistent data, such as data obtained from files, social media websites, satellites, etc.
- Consists of data in different formats such as e-mails, text, audio, video, or images

Some sources of unstructured data include:

- **Text both internal and external to an organization**—Documents, logs, survey results, feedbacks, and e-mails from both within and across the organization
- **Social media**—Data obtained from social networking platforms, including YouTube, Facebook, Twitter, LinkedIn, and Flickr
- **Mobile data**—Data such as text messages and location information

About 80 percent of enterprise data consists of unstructured content.

CASELET

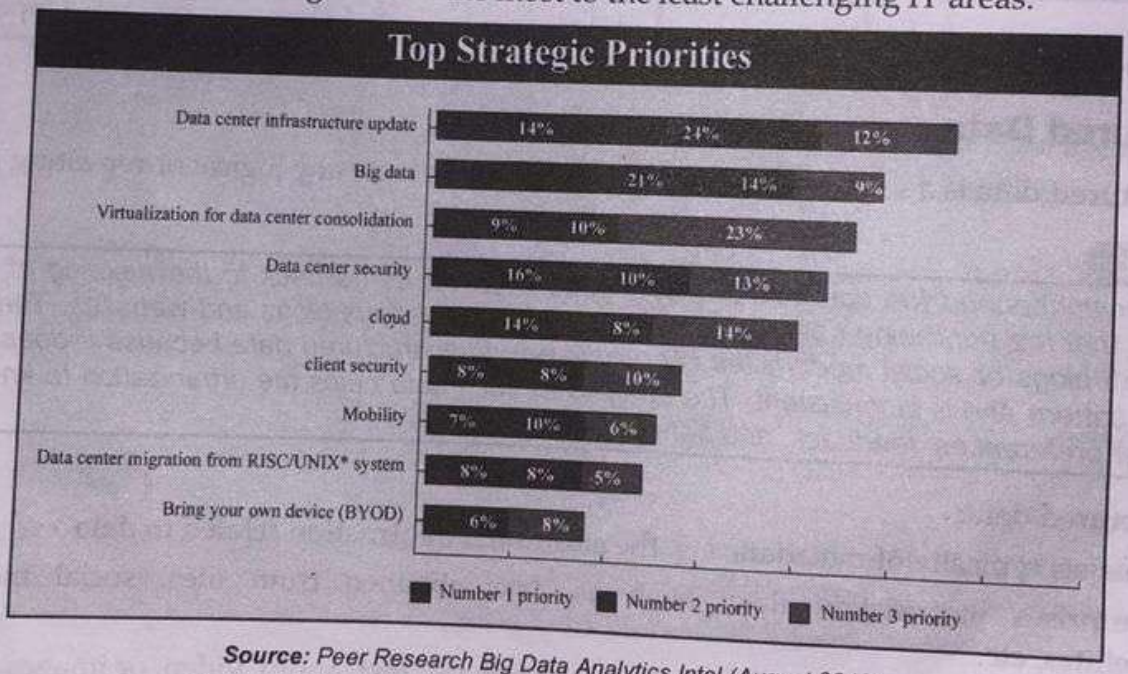
We all know that nowadays CCTV cameras are installed in almost every supermarket, and its footage is thoroughly analyzed by the management for various purposes. Some focus points of the analysis include the routes customers take to navigate through the store, customer behavior during a bottleneck, such as network traffic; and places where customers typically halt while shopping. This unstructured information from the CCTV footage is combined with structured data, comprising the details obtained from the bill counters, products sold, the amount and nature of payments, etc. to arrive at a complete data-driven picture of customer behavior. The analysis of the obtained information helps the management to provide a pleasant shopping experience to customers, as well as improve sales figures.

Challenges Associated with Unstructured Data

Working with unstructured data poses certain challenges which are as follows:

- Identifying the unstructured data that can be processed
- Sorting, organizing, and arranging unstructured data in different sets and formats
- Combining and linking unstructured data in a more structured format to derive any logical conclusions out of the available information
- Costing in terms of storage space and human resource (data analysts and scientists) needed to deal with the exponential growth of unstructured data

Figure 1.4 shows the result of a survey conducted to ascertain the challenges associated with unstructured data in percentage—from the most to the least challenging IT areas:



Source: *Peer Research Big Data Analytics Intel (August 2013)*

Figure 1.4: Challenges in Handling Unstructured Data

The survey reveals that the Big Data is the second biggest challenge followed by virtualization to manage the volume of data. Unstructured data is also generated from files that often have the same name and extension. For example, video files are generally stored with the extension .mp4 or .3gp; whereas, audio files have extension .wav or .mp3. As different files of the same category can have the same file name in different sources, merely a name and an extension do not help in data identification, classification, or even basic searches.

Semi-Structured Data

Semi-structured data, also known as having a schema-less or self-describing structure, refers to a form of structured data that contains tags or markup elements in order to separate elements and generate hierarchies of records and fields in the given data. Such type of data does not follow the proper structure of data models as in relational databases. In other words, data is stored inconsistently in rows and columns of a database.

Some sources for semi-structured data include:

- File systems such as Web data in the form of cookies
- Data exchange formats such as JavaScript Object Notation (JSON) data

SCENARIO

Mr. Smith also observes the presence of some semi-structured data saved in the database system of the publishing house. This data contained personal details of the authors working for the publishing house, as shown in Table 1.5:

Table 1.5: Semi-Structured Data

Sl. No	Name	E-Mail
1.	Sam Jacobs	smj@xyz.com
2.	First Name: David Last Name: Brown	davidb@xyz.com

As you can notice from Table 1.5, semi-structured data indicates that the entities belonging to the same class can have different attributes even if they are grouped together. In this case, different names and different e-mails are grouped under a common column name.

EXHIBIT 2: Types of Data in Big Data

There are three types of data we need to consider in Big Data: structured, unstructured, and semi-structured. Of these, the last two are new in Big Data.

- **Structured Data**—Your current data warehouse contains structured data and only structured data. It's structured because when you placed it in your relational database system, a structure was enforced on it, so we know where it is, what it means, and how it relates to other pieces of data in there. It may be text (a person's name) or numerical (their age) but we know that the age value goes with a specific person, hence structured.
- **Unstructured Data**—Essentially, everything else that has not been specifically structured is considered unstructured. The list of truly unstructured data includes free text such as documents produced in your company, images and videos, audio files, and some types of social media. If the object to be stored carries no tags (metadata about the data) and has no established schema, ontology, glossary, or consistent organization, it is unstructured. However, in the same category as unstructured data, there are many types of data that do have at least some organization.

- **Semi-Structured Data**—The line between unstructured data and semi-structured is a little fuzzy. If the data has any organizational structure (a known schema) or carries a tag (like XML used for documents on the Web), then it is somewhat easier to organize and analyze. Some types of data that appear to be unstructured but are actually semi-structured include:
  - **Emails or electronic data interchange messages (EDI)**—These lack formal structure but do contain tags or a known structure that separate semantic elements.
  - **Web Server Logs and Search Patterns**—An individual's journey through a website, whether searching, consuming content, or shopping is recorded in detail in electronic Web server logs.
  - **Sensor Data**—There is a huge explosion in the number of sensors producing streams of data all around us. This includes RFIDs, infrared and wireless technology, and GPS location signals among others. Your cell phone puts out a constant stream of signals that are being captured for location-based marketing. In-store sensors are monitoring consumer shopping behavior. While a great deal of attention is being paid to new types of analysis for social media, in the next two or three years at most, we will reach a crossover point where the volume of data available from sensors will exceed new social media postings, and sensor data volumes are likely to grow 10 or 20 times faster than social media sources.

Source: http://data-magnum.com/the-big-deal-about-big-data-whats-inside-structured-unstructured-and-semi-structured-data/