BIG DATA

A V KRISHNA MOHAN
DEPT. of CSE , SIT

INDEX

INDEX

Aim of studying Big Data Subject

This subject provides you an in-depth knowledge on

what the Big Data is all about,

its various sources and types,

why it has become so important to store, sort, and analyze Big Data

how it can be handled,

the various tools and techniques employed to handle Big Data,

and

much more.

What is Big Data ? **No single definition;**

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

What is Big Data?

1) Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

2) Big Data is ubiquitous , means being everywhere or appearing/ found everywhere and is most popular IT trending buzzword.

3) From Wikipedia: Big data is the term for a collection of data sets so large and complex that it becomes very difficult to process using on-hand database management tools or traditional data processing applications.

4) Big Data is a data that cannot be stored , processed and analyzed using today's conventional user computers. That is Big Data requires storage capacity and processing capability which are beyond to the existing computers.

5) According to IBM , **the Big Data is defined as 3 Vs**. That is the term Big Data is characterized by its three fundamental elements such as VOLUME , VELOCITY and VARIETY.

WHY Big Data is Needed for EVERYONE ?

Why Big Data Engineering?

The adoption of Big Data is growing across industries,

which has resulted in an increased demand for Big Data Engineers.

However, the supply is inadequate, leading to a large number of job opportunities. In this environment, professionals with the appropriate skills can command higher salaries.

The trend or need to Big Data or larger data sets is

due to the additional information [value or decision] derivable from analysis of a single large data set of related data,

as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to

"spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Demerit of RDBMS

What is the problem ?

Traditional RDBMS queries are not sufficient to get useful information out of big data [the huge volume of data]

That is traditional RDBMS queries are — in-efficient to Extract useful insight from Big data.

Traditional RDBMS is in-capable to process and Extract ~~large~~ data in Real-time.

Pre-requirements for studying Big Data Subject

RDBMS / SQL Essentials

Java Essentials for Big Data

Linux Fundamentals

Distributed Computing basics

Where it is coming from ? SOURCES OF BIG DATA

Big data is a popular term that describes the immense growth and availability of data in both structured and unstructured formats.

We all know that Big Data is being generated by nearly FROM everything around us at all times.

Various sources of Big Data include

Digital systems, sensors, satellites & Radio Frequency Identification (RFID) chips ,

Mobile phones, social media platforms & online transaction processing systems (OLTPs) etc.

Now all these sources generate Big Data at **an alarmingly high rate**, which need to be processed and analyzed[IN REAL-TIME] efficiently **almost at the same rate at which it is being generated**.

Facts and Figures related to Big Data

→ Every second, there are around 822 tweets on Twitter

→ Every minute, nearly
- 510 comments are posted
- 2,93,000 statuses are updated
- 1,36,000 photos are uploaded

} Facebook

500 TB of data is inserted into Facebook Every data

→ Every Hour, Walmart, Global discount departmental Store chain, handly more than 1 million Customer transactions.

→ Every data, Consumers make around 11.5 million payments using PayPal

→ ~~PayPal users~~ In Every minute of the data
- Youtube users upload 72 hours of new video
- Apple users download nearly 50,000 Apps
- Amazon generates around $80,000 online sale
- E-mail users send 200 million [nearly 20 coores] messages
- over 20,00,000 search

→ A Flight generates 240 TB of flight data in 6-8 hours of flight

→ Facebook Stores, accesses, and analyzes 30+ petabytes [PB] of user generated data.

An Insight [Volume of Data] or [Size of data]

Unit of Measuring Data.

Bit : 0 or 1

Nibble : Sequence of 4 bits.

Byte : —u— 8 bits [one grain of Rice]

KB [10^3 Bytes] : 10^3 bytes [one cup of rice]

kilo

MB : 10^6 bytes [8 bags of Rice] Desktop amt of data that flow thru desktop

Mega

GB : 10^9 : 3 Semi trucks of rice

Giga

TB : 10^{12} : 2 Container Ships of rice : Internet, the amt of data that flows thru Internet

Tera

PB : 10^{15} : Blankets in ½ of California

peta

EB : 10^{18} : Blankets west coast : Big Data

Exabyte

ZB : zettabyte : 10^{21} : Fully pacific ocean : Future

YB : yottabyte : 10^{24} : An Earth sized rice ball

BB : Brontobyte : 10^{27} : Astronomical size

Challenges in handling Big Data

The challenges include capture, curation, storage, search, sharing, transfer, analysis, visualization & **Taking final decision or Value.**

Needless to say, the manual handling of such a huge and ever expanding pool of varied data is neither possible nor feasible.

To derive any form of meaningful information [value] from Big Data generated from various sources, we need a set analytical tools and computational techniques.

However, first we need to have an extensive knowledge about each of these tools and techniques, their usage, and the specific environments they have been designed to work in.

Only then can you select the right tool suitable for your specific needs and utilize the immense benefits offered by the advanced data analytics.

BIG DATA

A V KRISHNA MOHAN
DEPT. of CSE , SIT

Introduction to big data

The 21st century is characterized by the rapid advancement in the field of information technology[IT].

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 process a number of skills to create their place or position and to sustain in the market.

One of the most crucial of these skills is an understanding of and the ability to utilize the immense potential of Information Technology.

According to the information technology Association of America , Information Technology that is ,IT, is defined as the study ,design, development ,application , implementation , support or management of Computer Based information systems.

Need of Big Data

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 this big data generated through various sources to run their businesses.

They analyze the data to understood and interpret market trends , study customer behavior and to 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 data sets that cannot be managed efficiently by common database management systems. These data sets range from terabytes to exabytes. Mobile phones , credit cards , radio frequency identification devices that is RFID and social networking platforms create use amounts of data that may reside unutilized at unknown servers for many years ,

However with the evolution of big data that this big data can be accessed and analyzed on a regular basis to generate Useful information.

In this chapter we introduce you to the big data , the big Buzz word of the IT industry and its growing importance in almost every sector of human existence be it education ,health ,science, technology, defense , lifestyle etc..

Hadoop distributed architecture file system

Due Its complexity big data is stored in distributed architecture file system.

Hadoop by Apache is widely used for storing and managing big data.

Analyzing big data is a challenging task as it involves large distributed file systems , which are fault tolerant , flexible and scalable.

According to survey 90% of the data in the world today has been created in the last two years only.

As big data is exponentially growing at an alarmingly high rate and it would be a waste of time and effort if we just capture and store only , instead it can be analyzed & put to some logical use.

The process of capturing or collecting big data is known as *datafication.*

Big data is 'datafied' so that it can be used productively.

SCENARIO – related to Big Data

Big data cannot be made useful by simply organizing it , rather , the data's usefulness lies in determining what we can do with it.

To Extract meaningful value from big data , you need optimal processing power , analytical capabilities and skills.

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 common examples with his team to enhance their knowledge.

Real world examples of big data

1) Consumer product companies and retail organizations are observing data on social media websites such as Facebook and Twitter. These social media sites help them to analyze customer behavior , preferences and product perception. Accordingly the companies can line up their upcoming products to gain more profits. This phenomenon is also known as social media Analytics.

2)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 soon wastes money and replacing it too late , triggers an expensive work stoppage.

3) Advertising and Marketing Agencies are tracking social media to understand responsiveness to campaigns , promotions and other advertising mediums.

4) 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.

5) 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.

Structuring Big Data

Structuring of data is nothing but a process of arranging the available data in a manner such 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 of news articles that I come across ?
- How do I choose a book of the millions of books available on my favourite sites are stores ?
- How do I keep myself updated about new events ,sports ,inventions and discoveries taking place across the globe ?

Structuring Big Data

The answers to all of the above question can be found by information processing system.

These systems can analyse and the structure a large amount of data specifically for you on the basis of what you are searching , what you looked at and for how long you remained at a particular page of website ,

Thus scanning and presenting you with the customized information as per your behaviour and habits.

In other words structuring data helps in understanding user behaviour , requirements and preferences to make personalized recommendation for every individual.

When a user regularly visits or purchase from online shopping sites such as eBay or Amazon , each time he or she logs in ,the system can present a recommended list of products that may interest the user on the basis of his/her our earlier purchase or searches,

Thus presenting a specially customised recommendations set for every user.

This is the power of Big Data Analytics.

Structuring Big Data

Today various sources generate a variety of data such as images , text , audio etc..

All such different types of data can be structured

only if it is stored and organised 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 :

1) Internal sources such as organisational or enterprise data

2) External sources such as social data.

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 forms 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 • Product and sales data • Generally OLTP and operational data	This data (current data in the operational systems) 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.

Figure 1.2 shows the various parts of Big Data. It is obvious from Figure 1.3 that in Big Data, we deal with data storage, distributed system, data mining, and many other things.

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

Figure 1.3 shows the various parts of Big Data. It is obvious from Figure 1.3 that in Big Data, we deal with data storage, distributed system, data mining, and many other things.

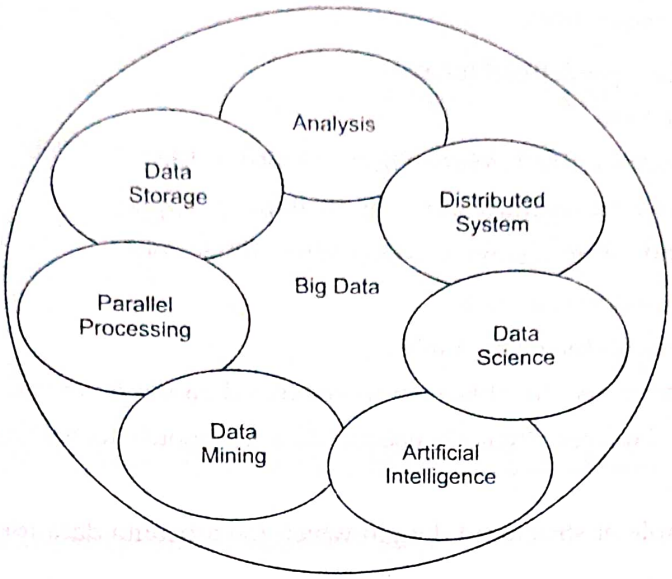


Figure 1.3: Concepts of Big Data.

On the basis of the data received from the sources mentioned in Table 1.3, 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, approximately 70% to 80% of data is in unstructured form. Figure 1.4 illustrates the types of data that comprise Big Data:

Figure 1.4: Types of Big Data

Let us discuss these types in detail in the following sections.

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 can be used to take various organizational decisions.

Structured data:

- Is organized data in a predefined format
- Is stored in tabular form
- 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 comma separated values (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

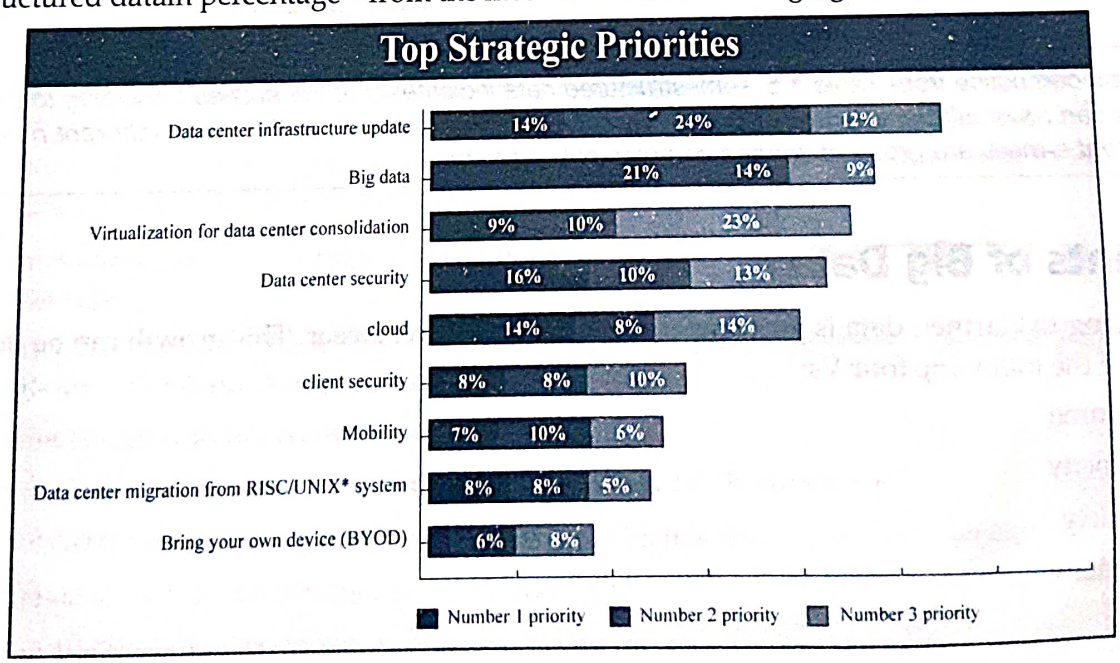
We all know that nowadays CCTV cameras are installed in almost every supermarket, and its footage is thoroughly analyzed by the management for various purposes. Some focus points of the analysis include the routes customers take to navigate through the store, customer behavior during a bottleneck, such 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 1.5 shows the result of a survey conducted to ascertain the challenges associated with unstructured datain percentage—from the most to the least challenging IT areas:



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

Figure 1.5: 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 contains 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

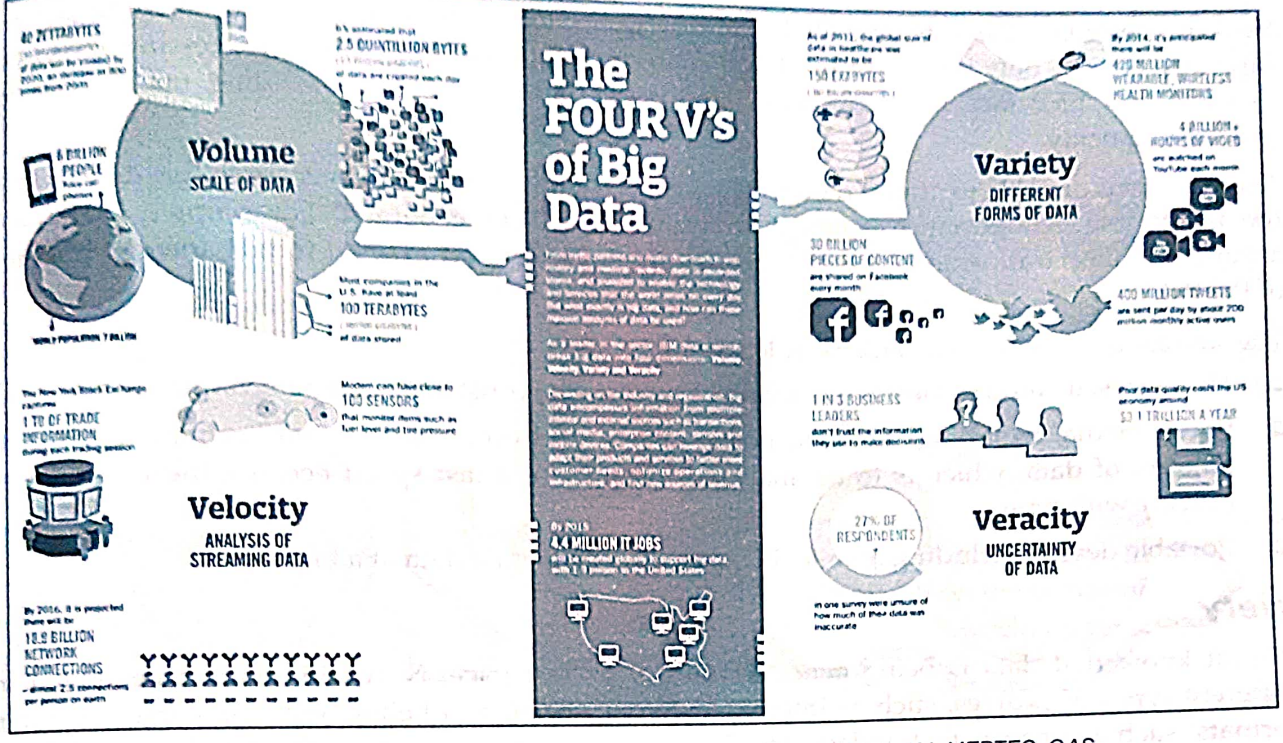
As you can notice from Table 1.5, semi-structured data indicates that the entities belonging to the same class can have different attributes even if they are grouped together. In this case, different names and different e-mails are grouped under a common column name.

Elements of Big Data

According to Gartner, data is growing at the rate of 59% every year. This growth can be depicted in terms of the following four Vs:

- Volume
- Velocity
- Variety
- Veracity

Figure 1.6 explains the four Vs of Big Data with examples:



Source: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Figure 1.6: Four Vs of Big Data

Volume

Volume is the amount of data generated by organizations or individuals. Today, the volume of data in most organizations is approaching exabytes. Some experts predict the volume of data to reach zettabytes in the coming years. Organizations are doing their best to handle this ever-increasing volume of data. For example, according to IBM, over 2.7 zetabytes of data is present in the digital universe today. Every minute, over 571 new websites are being created. IDC estimates that by 2020, online business transactions will reach up to 450 billion per day.

The Internet alone generates a huge amount of data. The followings figures help us to get an idea of the Internet traffic:

- Internet has around 14.3 trillion live Web pages, and 48 billion Web pages are indexed by Google Inc.; 14 billion Web pages are indexed by Microsoft Bing.
- Internet has around 672 exabytes of accessible data.
- Total world-wide Internet traffic in the year 2013 was 43,639 petabytes.
- Over 9,00,000 servers are owned by Google Inc., which is the largest in the world.
- Total data stored on the Internet is over 1 yottabyte.

Even by underestimation, the total data stored on the Internet, including images, videos, audio, etc., has crossed 1 yottabyte. The exact size of the Internet will never be known!

Source: Global Internet usages facts and statistics of 2013

Velocity

Velocity describes the rate at which data is generated, captured, and shared. Enterprises can capitalize on data only if it is captured and shared in real time. Information processing systems such as CRM and ERP face problems associated with data, which keeps adding up but cannot be processed quickly.

These systems are able to attend data in batches every few hours; however, even this time lag causes the data to lose its importance as new data is constantly being generated. For example, eBay analyzes around 5 million transactions per day in real time to detect and prevent frauds arising from the use of PayPal.

The sources of high velocity data include the following:

- IT devices, including routers, switches, firewalls, etc., constantly generate valuable data.
- Social media, including facebook posts, tweets, and other social media activities, create huge amount of data, which is to be analyzed instantly at a fast speed because the value degrades quickly with time.
- Portable device, including mobile, PDA, etc., also generate data at a high speed.

Variety

We all know that data is being generated at a very fast pace. Now, this data is generated from different types of sources, such as internal, external, social, and behavioral, and comes in different formats, such as images, text, videos, etc. Even a single source can generate data in varied formats, for example, GPS and social networking sites, such as Facebook, produce data of all types, including text, images, videos, etc. Figure 1.7 shows the various types of data:

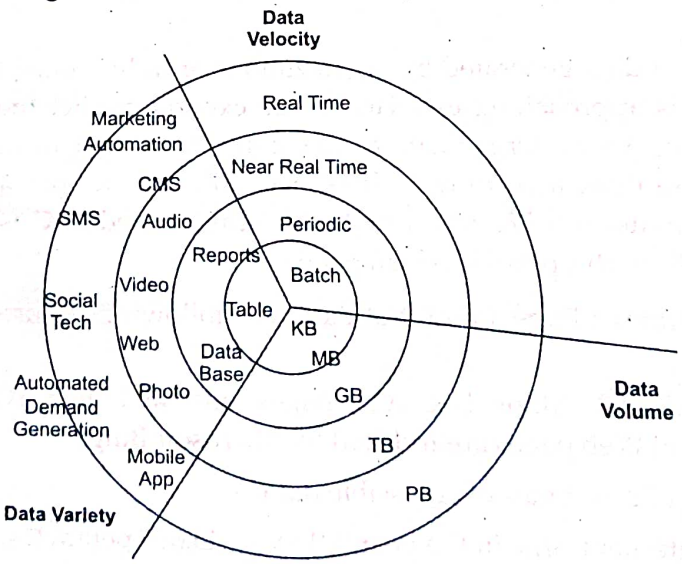


Figure 1.7: Various Types of Data

Veracity

Veracity generally refers to the uncertainty of data, i.e., whether the obtained data is correct or consistent. Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis. Data when processed becomes

information; however, a lot of effort goes in processing the data. Big Data, especially in the unstructured and semi-structured forms, is messy in nature, and it takes a good amount of time and expertise to clean that data and make it suitable for analysis.

Big Data Analytics

Big data analytics reformed the ways to conduct business in many ways, such as it improves, decisions making, business process management, etc. Business analytics uses the data and different other techniques like information technology, features of statistics, quantitive methods, and different models to provide results. There are three main types of business analytics: descriptive analytics, predictive analytics, and prescriptive analytics. The conventional database systems are not in a postion to process Big data defined by the four Vs: volume, variety, velocity, and veracity. Big data also affects the analytical process and technologies used for analytics.

There are mainly three types of analytics:

- **Descriptive Analytics**—Descriptive analytics is the most prevalent form of analytics, and it serves as a base for advanced analytics. It answers the question 'What happened in the business?' Descriptive analytics analyses a database to provide information on the trends of past or current business events that can help managers, planners, leaders, etc. to develop a road map for future actions. Descriptive analytics performs an in-depth analysis of data to reveal details such as frequency of events, operation costs, and the underlying reason for failures. It helps in identifying the root cause of the problem.
- **Predictive Analytics**—Predictive analytics is about understanding and predicting the future and answers the question 'What could happen?' by using statistical models and different forecast techniques. It predicts the near future probabilities and trends and helps in what-if analysis. In predictive analytics, we use statistics, data mining techniques, and machine learning to analyze the future. Figure 1.8 shows the steps involved in predictive analytics:

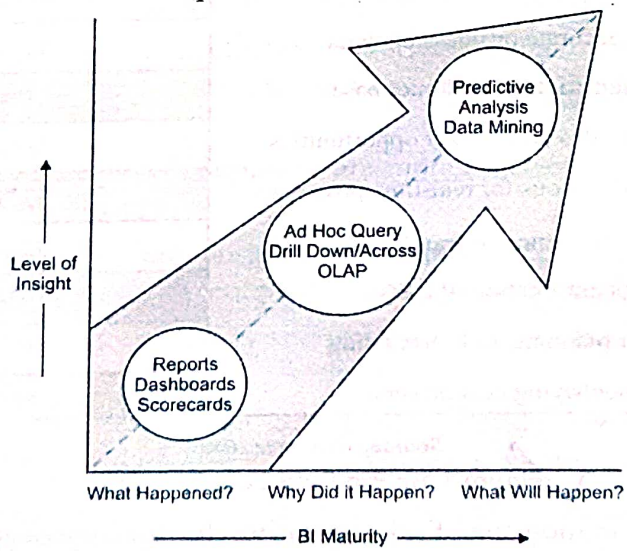


Figure 1.8: Predictive Analytics

- **Prescriptive Analytics**—Prescriptive analysis answers 'What should we do?,' on the basis of complex data obtained from descriptive and predictive analyses. By using the optimization

technique, prescriptive analytics determines the finest substitute to minimize or maximize some equitable finance, marketing, and many other areas. For example, if we have to find the best way of shipping goods from a factory to a destination, to minimize costs, we will use the prescriptive analytics. Figure 1.9 shows a diagrammatic representation of the stages involved in the prescriptive analytics:

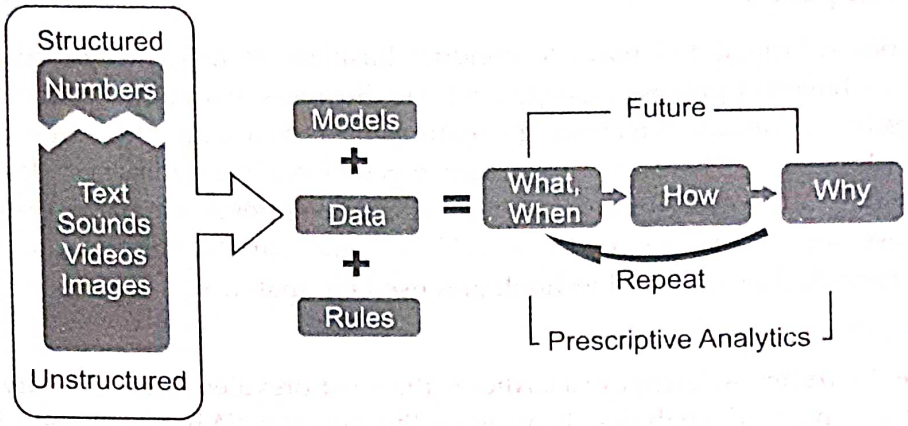


Figure 1.9: Prescriptive Analytics

Data, which is available in abundance, can be streamlined for growth and expansion in technology as well as business. When data is analyzed successfully, it can become the answer to one of the most important questions: how can businesses acquire more customers and gain business insight? The key to this problem lies in being able to source, link, understand, and analyze data.

Figure 1.10 highlights the proportion of business areas that have benefited by using Big Data:

Big Data Analytics Benefit	Proportion of Businesses Reporting Benefit (%)
Better social influences marketing	61%
More accurate business insights	45%
Segmentation of customer base	41%
Identifying sales and market opportunities	38%
Automated decisions for real-time processes	37%
Detection of fraud	33%
Quantification of risks	30%
Better planning and forecasting	29%
Identifying cost drivers	29%

Source: TDWI July 2013

Figure 1.10: Big Data Benefit Areas

Let us understand some common analytical approaches that businesses apply to use Big Data.

Table 1.6 describes various analytical approaches typically associated with Big Data:

Table 1.6: Analytical Approaches

Approach	Possible Evaluations
Predictive Analysis	• How can a business use the available data for predictive and real-time analysis across its different domains? • How can a business avail benefits from the unstructured enterprise data? • How can a business leverage upon new types of data such as sentiment data, social media, clickstream, and multimedia?

Behavioral Analysis	• How will a business leverage complex data in order to create new models for:
	  • Driving business outcomes
	  • Decreasing business costs
	  • Driving innovation in business strategy
	  • Improving overall customer satisfaction
	  • Converting an audience to a customer

Data Interpretation	• What new business analyses can be estimated from the available data? • Which data should be analyzed for new product innovation?

Advantages of Big Data Analytics

According to Atul Butte, Stanford, "Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world." So, the real power of Big Data lies in its analysis. Processing, studying, and implementing the conclusions derived from the analysis of Big Data help you to collect accurate data, take timely and more informed strategic decisions, target the right set of audience and customers, increase benefits, and reduce wastage and costs.

The right analysis of the available data can improve major business processes in various ways. For example, in a manufacturing unit, data analytics can improve the functioning of the following processes:

- **Procurement**—To find out which suppliers are more efficient and cost-effective in delivering products on time
- **Product Development**—To draw insights on innovative product and service formats and designs for enhancing the development process and coming up with demanded products
- **Manufacturing**—To identify machinery and process variations that may be indicators of quality problems
- **Distribution**—To enhance supply chain activities and standardize optimal inventory levels vis-à-vis various external factors such as weather, holidays, economy, etc.
- **Marketing**—To identify which marketing campaigns will be the most effective in driving and engaging customers and understanding customer behaviors and channel behaviors
- **Price Management**—To optimize prices based on the analysis of external factors

- **Merchandising**—To improve merchandise breakdown on the basis of current buying patterns and increase inventory levels and product interest insights on the basis of the analysis of various customer behaviors
- **Sales**—To optimize assignment of sales resources and accounts, product mix, and other operations
- **Store Operations**—To adjust inventory levels on the basis of predicted buying patterns, study of demographics, weather, key events, and other factors
- **Human Resources**—To find out the characteristics and behaviors of successful and effective employees, as well as other employee insights for managing talent better

Every business and industry today is affected by and benefitted from Big Data analytics in multiple ways.

A closer look at some specific industries will help you to understand the application of Big Data in these sectors.

Transportation

Big Data has greatly improved transportation services. The data containing traffic information is analyzed to identify traffic jam areas. Suitable steps can then be taken, on the basis of this analysis, to keep the traffic moving in such areas. Distributed sensors are installed in handheld devices, on the roads and on vehicles to provide real-time traffic information. This information is analyzed and disseminated to commuters and also to the traffic control authority.

Education

Big Data has transformed the modern-day education processes through innovative approaches, such as e-learning for teachers to analyze the students' ability to comprehend and thus impart education effectively in accordance with each student's needs. The analysis is done by studying the responses to questions, recording the time consumed in attempting those questions, and analyzing other behavioral signals of the students. Big Data also assists in analyzing the requirements and finding easy and innovative ways of imparting education, especially distance learning over vast geographical areas.

Travel

The travel industry also uses Big Data to conduct business. It maintains complete details of all the customer records that are then analyzed to determine certain behavioral patterns in customers. For example, in the airline industry, Big Data is analyzed for identifying personal preferences or spotting which passengers like to have window seats for short-haul flights and aisle seats for long-haul flights. This helps airlines to offer the similar seats to customers when they make a fresh booking with the airways.

Big Data also helps airlinesto track customers who regularly fly between specific routes so that they can make the right cross-sell and up-sell offers. Some airlines also apply analytics to pricing, inventory, and advertising for improving customer experiences, leading to more customer satisfaction, and hence, more business. Some airlines even go to the length of evaluating customers who tend to miss their flights. They try to help such customers by delaying the flights or booking them on another flight.

Government

Big Data has come to play an important role in almost all the undertaking and processes of government.

According to the UK free market,"the UK government could save up to £33 billion a year by using public Big Data more effectively." Analysis of Big Data promotes clarity and transparency in various government processes and helps in:

- Taking timely and informed decisions about various issues
- Identifying flaws and loopholes in processes and taking preventive or corrective measures on time
- Assessing the areas of improvement in various sectors, such as education, health, defense, and research
- Using budgets more judiciously and reducing unnecessary wastage and costs
- Preventing fraudulent practices in various sectors

Healthcare

In healthcare, the pharmacy and medical device companies use Big Data to improve their research and development practices, while health insurance companies use it to determine patient-specific treatment therapy modes that promise the best results. Big Data also helps researchers to work towards eliminating healthcare-related challenges before they become real problems. Big Data helps doctors to analyze the requirement and medical history of every patient and provide individualistic services to them, depending on their medical condition.

Telecom

The mobile revolution and the Internet usage on mobile phones have led to a tremendous increase in the amount of data generated in the telecom sector. Managing this huge pool of data has almost become a challenge for the telecom industry. For example, in Europe, there is a compulsion on the telecom companies to keep data of their customers for at least six months and maximum up to two years. Now, all this collection, storage, and maintenance of data would just be a waste of time and resources unless we could derive any significant benefits from this data. Big Data analytics allows telecom industries to utilize this data for extracting meaningful information that could be used to gain crucial business insights that help industries in enhancing their performance, improving customer services, maintaining their hold on the market, and generating more business opportunities.

Consumer Goods Industry

Consumer goods companies generate huge volumes of data in varied formats from different sources, such as transactions, billing details, feedback forms, etc. This data needs to be organized and analyzed in a systemic manner in order to derive any meaningful information from it. For example, the data generated from the Point-of-Sale (POS) systems provides significant real-time information about customers' preferences, current market trends, the increase and decrease in demand of different products at different regions, etc. This information helps organizations to predict any possible fluctuations in prices of goods and make purchases accordingly. It also helps marketing teams in taking suitable actions rapidly if there is a deviation in the expected sales of a product, thus,

preventing any further losses to the company. Therefore, we can say that Big Data analytics allows organizations to gain better business insights and take informed and timely decisions.

Aviation Industry

Big Data analytics also plays a significant role in the commercial aviation industry. Like other industries, the aviation industry also maintains a detailed record of all their customers that includes their personal information, flying preferences, and other trends and patterns. The organization analyzes this data to improve their customer services, and thus brand image. In addition, every aircraft generates a significant amount of data during operation. This data is then analyzed for enhancing operational efficiencies, identifying the parts that require repairs, and taking any necessary constructive or preventive measures on time.

Careers in Big Data

Now that you know that Big Data is really BIG in today's world, you can well understand that so are the opportunities associated with it. The market today needs plenty of talented and qualified people who can use their expertise to help organizations deal with Big Data.

Qualified and experienced Big Data professionals must have a blend of technical expertise, creative and analytical thinking, and communication skills to be able to effectively collate, clean, analyze, and present information extracted from Big Data.

Most jobs in Big Data are from companies that can be categorized into the following four broad buckets:

- Big Data technology drivers, e.g., Google, IBM, Salesforce
- Big Data product companies, e.g., Oracle
- Big Data services companies, e.g., EMC
- Big Data analytics companies, e.g., Splunk

Figure 1.11 shows the logos of some companies that hire Big Data professionals:

Source: Glassdoor Report October 2011

Figure 1.11: Companies Hiring Big Data Professionals

As shown in Figure 1.11, companies such as Google, Salesforce, and Apple offer various types of opportunities to Big Data professionals. These companies deal in various domains such as retail, manufacturing, information, finance, and consumer electronics. The hiring of Big Data experts in these domains, as per Big Data Analytics 2014 report, is shown in Figure 1.12:

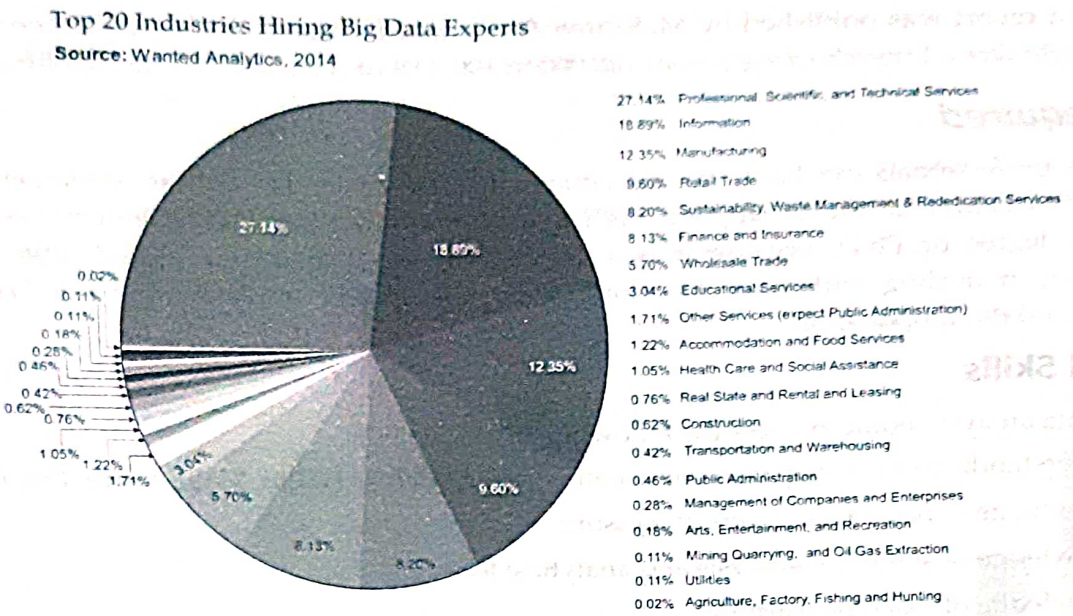


Figure 1.12: Top 20 Industries Hiring Big Data Experts

The most common job titles in Big Data include:

- Big Data analyst
- Data scientist
- Big Data developer
- Big Data administrator
- Big Data engineer

Figure 1.13 illustrates the roles of these profiles:

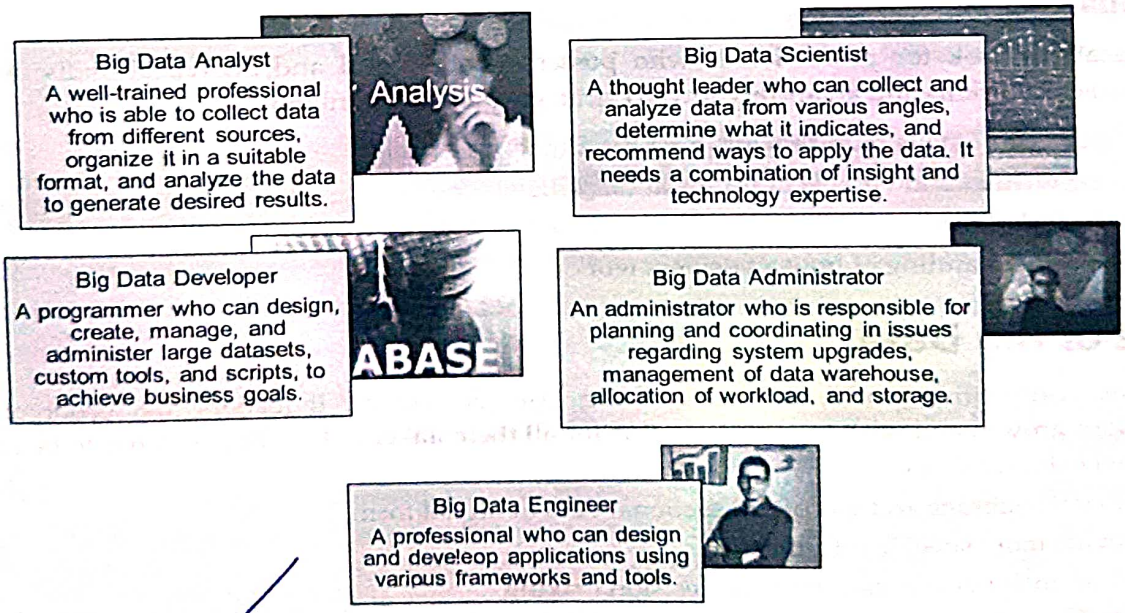


Figure 1.13: Role of Different JobTitles in Data Analytics

In 2011, a report was published by McKinsey & Co. that indicated that by 2018, the United States alone might face a huge shortage (about 140,000 to 190,000) of data analytics professionals.

Skills Required

Big Data professionals can have various educational backgrounds, such as econometrics, physics, biostatistics, computer science, applied mathematics, or engineering. Data scientists mostly possess a master's degree or Ph.D. because it is a senior position and often achieved after considerable experience in dealing with data. Developers generally prefer implementing Big Data by using Hadoop and its components.

Technical Skills

A Big Data analyst should possess the following technical skills:

- Understanding of Hadoop ecosystem components, such as HDFS, MapReduce, Pig, Hive, etc.
- Knowledge of natural language processing
- Knowledge of statistical analysis and analytical tools
- Knowledge of machine learning
- Knowledge of conceptual and predictive modeling

A Big Data developer should possess the following skills:

- Programming skills in Java, Hadoop, Hive, HBase, and HQL
- Understanding of HDFS and MapReduce
- Knowledge of ZooKeeper, Flume, and Sqoop

These skills can be acquired with proper training and practice. This book familiarizes you with the technical skills required by a Big Data analyst and Big Data developer.

Soft Skills

Organizations look for professionals who possess good logical and analytical skills, with good communication skills and an affinity toward strategic business thinking.

The preferred soft skills requirements for a Big Data professional are:

- Strong written and verbal communication skills
- Analytical ability
- Basic understanding of how a business works

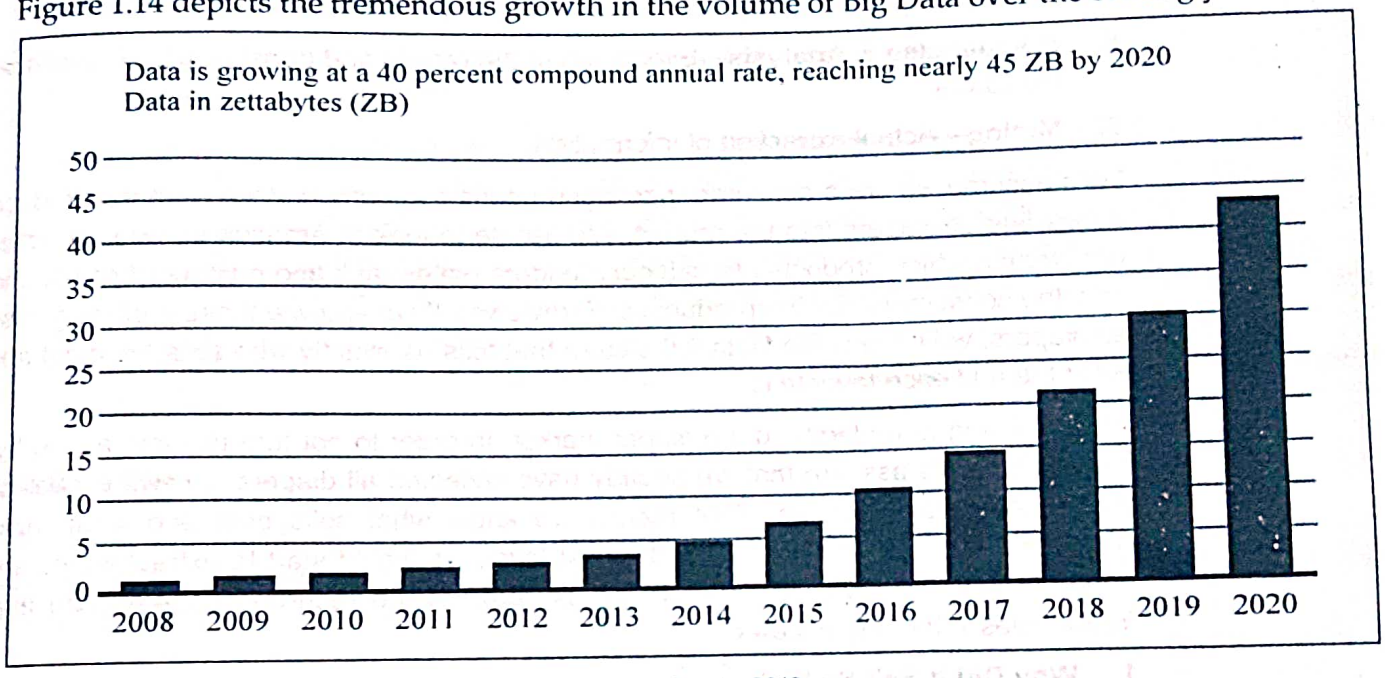
Future of Big Data

In today competitive world, the need of Big Data is evident. If leaders and economies want exemplary growth and wish to generate value for all their stakeholders, Big Data has to be embraced and used extensively to:

- Allow the storage and use of transactional data in digital form
- Provide more specific information
- Refine analytics that can improve decision making
- Classify customers for providing customized products and services based on buying patterns

Most organizations today consider data and information to be their most valuable and differentiated asset. By analyzing this data effectively, organizations worldwide are now finding new ways to compete and emerge as leaders in their fields to improve decision making and enhance their productivity and performance. At the same time, the volume and variety of data is also increasing at an immense rate every day. The global phenomena of using Big Data to gain business value and competitive advantage will only continue to grow as will the opportunities associated with it.

Figure 1.14 depicts the tremendous growth in the volume of Big Data over the coming years:



Source: Oracle, 2012

Figure 1.14: Growth Pattern of Data

Research conducted by MGI and McKinsey's Business Technology Office suggests that the use of Big Data is most likely to become a key basis of competition for individual firms for success and growth and strengthening consumer surplus, production growth, and innovation.

EXHIBIT 2: The Future of Big Data—Moving from Big Data 1.0 to 2.0

The future of Big Data is not about numeric data points but instead about asking the deeper questions and finding out why consumers make the decisions they do.

Today, clients often ask about the future of big data and what the next step is; how can we leverage data on an even deeper level in order to extract meaningful consumer insights that go beyond where we are now? Most of the standard answers are around the ability to get data and insights in real time and from more devices than ever. It's time we move beyond structured data and into the prime time of text analytics.

For us, the easiest way to get started with Big Data 2.0 is to focus on the unstructured data we collect every day. This can be reviews, customer support emails, community forums, or even your own CRM system. The simplest way to look at this data is through a process called text analytics.

Text analytics is a fairly straightforward process that breaks out like this:

1. **Acquisition**—Collecting and aggregating the raw data you want to analyze
2. **Transforming & Preprocessing**—Cleaning and formatting the data to make it easier to read
3. **Enrichment**—Enhancing the data by adding additional data points
4. **Processing**—Performing specific analyses and classifications on the data
5. **Frequencies & Analysis**—Evaluation of the results and translation into numerical indicators
6. **Mining**—Actual extraction of information

Let's assume we are a consumer packaged goods company and we want to introduce a new line of diapers into the market. We decide to look at Amazon in order to better understand which products are category leaders (sales rank and number of sales) and how the consumers like the product itself (reviews). If we analyze these metrics across all diapers, we have a Big Data 1.0 picture that tells us exactly who sells the most and what the audience favorite is.

We are trying to understand the diaper market. In order to not turn this into a step-by-step guide, let's assume that we already have collected all diapers reviews as well as their qualitative indicators. That means we know what sells best and what ranks best/worst. In order to take this to the next level, we would start to extract words and phrases from the reviews. This will tell us some of the recurring patterns and their frequencies within the reviews.

1. **Why Did It Sell So Well**

   When I looked at the reviews of the top-selling product, I found that the most mentioned terms across the majority of the helpful reviews were "price," "special," and "value." This tells us that people did not buy it because of its quality or features but because of its pricing. So, when we are launching our product, we want to look at this one for price/value guidance instead of features.

2. **Why People Did Not Like It**

   This one was very revealing. The brand with the most negative reviews had an extremely high frequency around the terms "tape," "stick," "stay closed," and "open." After a few reads, I discovered that consumers had no issues with the usual key features on a diaper such as "absorbency," "leakage," or "softness" but actually had issues with the tape on the side of the diaper, and the fact that it kept opening. The amount of negative reviews overall that mentioned these issues makes us believe that this is a feature that brands don't talk about but consumers care about. Therefore, we would recommend testing ads that address this issue.

3. **Smart Filtering**

   One interesting issue we came across is the fact that a lot of the negative reviews were not actually about the product but rather focused on shipping, stock level, and packaging concerns. By tagging and removing these from the set, we are able to evaluate product level in order to focus on product-related concerns. If we were to list our diaper on Amazon, we would recommend adding a shipping and stock

level guarantee prominently in the copy—a competitive advantage that speaks directly to consumer concerns.

4. **What Do They Want**

From an R&D perspective, this insight is worth gold. By evaluating reviews that have terms like "I wish," "hope," or "they should " we are able to detect common features consumers are looking for when thinking about diapers. These are great insights that address the constantly changing need of the consumers. We can feed these product feature-specific insights to our R&D team as well as our copywriters.

As you can see, when analyzing the diaper category, Big Data 2.0 yielded insights beyond binary performance indicators. We could see the crowd favorites but did not (yet) know the "why" behind purchases or understand the positive or negative reviews until our text analytics exercise. There are countless consumer insights to be mined from textual, unstructured data that give us the voice of the consumer, their motivations, and a deeper understanding of their purchasing behavior.

Source: *http://www.clickz.com/clickz/column/2377230/the-future-of-big-data-big-data-20*

Summary

This chapter, introduced you to Big Data—the big buzzword of today's IT industry. The chapter discussed some common features and sources of Big Data. Next, you learned about the evolution of Big Data along with its various types, namely, structured, unstructured, and semi-structured. The chapter further discussed the four Vs of Big Data, i.e., Volume, Velocity, Variety, and Veracity. You also learned about the use of Big Data in various domains, such as transportation, education, travel, government, and healthcare. The chapter also familiarized you with professional opportunities available in the career path of Big Data and the technical as well as soft skills required to enter this domain.

Quick Revise

Multiple-Choice Questions

Q1. Which of the following is not a characteristic of Big Data?

a. Volume
b. Variability
c. Variety
d. Velocity

Ans. The correct option is b.

Q2. Who among the following do you think would be able to deal with the growing number of data sources efficiently?

a. Business developer
b. Data scientist
c. Sales executive
d. Web designer

Ans. The correct option is b.

Q3. Which one of the following is not an example of external datasources?

a. Data from CRM
b. Data from Web logs
c. Data from government sources
d. Data from market surveys

Ans. The correct option is a.

Q4. Which of the following does not belong to the traditional database technology?

a. RDBMS
b. DBMS
c. Flat files
d. NoSQL

Ans. The correct option is d.

Q5. If a Big Data analyst were to analyze data from a database of call logs provided by a telecom service provider, which element of Big Data would he be dealing with?

a. Volume
b. Variable
c. Variety
d. Velocity

Ans. The correct option is a.

Q6. Some people call this data as "structured but not relational." Which data are we talking about?

a. Structured data
b. Unstructured data
c. Semi-structured data
d. Mixed data

Ans. The correct option is c.

Q7. The data generated from a GPS satellite and Web logs is classified as _____.

a. Structured data
b. Unstructured data
c. Both structured and unstructured data
d. Semi-structured data

Ans. The correct option is d.

Q8. The data being captured can be in any form or structure. Which characteristic of Big Data are we talking about?

a. Volume
b. Velocity
c. Variety
d. Value

Ans. The correct option is c.

Subjective Questions

Q1. List and discuss the four elements of Big Data.

Ans. Big Data primarily consists of the following four elements:

- **Volume**—Volume is the amount of data generated by organizations or individuals. Today, the volume of data in most organizations is approaching around exabytes. Some experts predict the volume of data to reach zetabytes in the coming years. Organizations are doing their best to handle this ever-increasing volume of data. For example, Google Inc. processes around 20 petabytes of data, and Twitter feeds generate around 8 terabytes of data every day.

- **Velocity**—Velocity describes the rate at which data is generated, captured, and shared. Enterprises can capitalize on data only if it is captured and shared in real time. Information processing systems such as CRM and Enterprise Resource Planning (ERP) face problems associated with data, which keeps adding up but cannot be processed quickly. These systems are able to attend data in batches every few hours; however, even this time lag causes the data to lose its importance as new data is constantly being generated. For example, eBay analyzes around 5 million transactions per day in real time to detect and prevent frauds arising from the use of PayPal.
- **Variety**—We all know that data is being generated at a very fast pace. Now, this data is generated from different types of sources, such as internal, external, social, and behavioral, and comes in different formats, such as images, text, videos, etc. Even a single source can generate data in varied formats; for example, GPS and social networking sites, such as Facebook, produce data of all types, including text, images, videos, etc.
- **Veracity**—Veracity generally refers to the uncertainty of data, i.e., whether the obtained data is correct or consistent. Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis. Data when processed becomes information; however, a lot of effort goes in processing the data. Big Data, especially in the unstructured and semi-structured forms, is messy in nature, and it takes a good amount of time and expertise to clean that data and make it suitable for analysis.

Q2. As an HR manager of a company providing Big Data solutions to clients, what characteristics would you look for while recruiting a potential candidate for the position of a data analyst?

Ans. A Big Data analyst should be a well-trained professional who is able to collect data from different sources, organize it in a suitable form, and analyze it to generate the desired results. A Big Data analyst should have the following technical and soft skills:

Technical Skills:

- Understanding of Hadoop, Hive, and MapReduce
- Knowledge of natural language processing
- Knowledge of statistical analysis and analytical tools
- Knowledge of conceptual and predictive modeling

Soft Skills:

- Strong written and verbal communication skills
- Analytical ability
- Basic understanding of how a business works

Q3. You are planning the marketing strategy for a new product in your company. Identify and list some limitations of structured data related to this work.

Ans. Structured data has certain limitations associated with it, when it comes to product marketing and advertising.

Some of these limitations are:

- Marketing strategies do not provide space for predefined rules
- There is hardly any definite relation between product sales and marketing strategies
- Structured data cannot find complex correlation patterns
- Behavioral analysis is not possible from structured data

Exploring the use of Big Data in Business Context

A V KRISHNA MOHAN

DEPT. of CSE , SIT

Objectives:

The main objectives of studying this chapter is you know :

- Use of big data in social networking
- Use of big data in preventing fraudulent activities
- Use of big data in detecting fraudulent activities in Insurance sector
- Use of big data in Retail Industry

Introduction

Almost all organisations collect and collate relevant data in various forms such as

Customers feedback [**customer's wants/preferences OR likes & dislikes**]

Inputs from retailers and suppliers

Current market Trends

This **big data** is analysed and **the information derived** is used by the management to **take major organisational decisions.**

An organisation generally has to spend huge amounts to collect data , for example , **customer surveys or market research reports , require a significant amount of investment by an organisation.**

The cost of collecting data goes on escalating /increasing as an organisation keeps on collecting more data.

The continuously increasing cost decreases the value of the collected data.

In other words collecting and maintaining a pool of data is just a waste of resources [time and money] , unless any logical conclusions and business insights can be derived from it.

This is where Big Data Analytics comes into picture.

Introduction

This chapter provides an in-depth look at various questions arrived in the use of Big Data in Business Context.

How big data influences businesses in today's world ?

How big data and different methods of Data Analytics are used in real time and why ?

How can large volumes of data are used to get better business insights /

How can the big data obtained be used to form better business Strategies and thereby help in scalability and profitability of the organisation.

In other words business insights gained from Big Data Analytics help the organisation to

- reduce their cycle time
- fulfil orders quickly
- Cut excess inventory
- Improve forecast accuracy and customer services

Introduction

This chapter focus on how big data is actually used in Real world to get better business insights and explore future Business expansion and also to detect and prevent frauds in different industries.

Scenario :

Consider again the same scenario as we discussed in the first chapter. Mr Smith , the data analyst of the Argon Technology plans to do some research on the use and effects of Big Data Analytics on the business sector.

All types of businesses give due importance to their customer's feedback , So Mr Smith decided to collect data for his research from social networking sites.

Use of big data in social networking

Some popular social networking sites are Twitter , Facebook and LinkedIn.

These social networking sites are also called as social media.

The main focus is to analyse the effects of big data generated from social media on different business industries.

Social network data refers to the data generated from people socializing on social media that is , In social networking sites , we find different people constantly adding and updating comments.

All these activities generate large amounts of data , analysing and mining such large volumes of data , show business trends with respect to wants and preferences and likes & dislikes of a wide audience.

This data can be segregated on the basis of different age groups , locations and genders for the purpose of analysis.

Based on the information extracted , organisation design products and services specific to People's need.

Figure 2.1 shows the social network data generated daily through various social media:

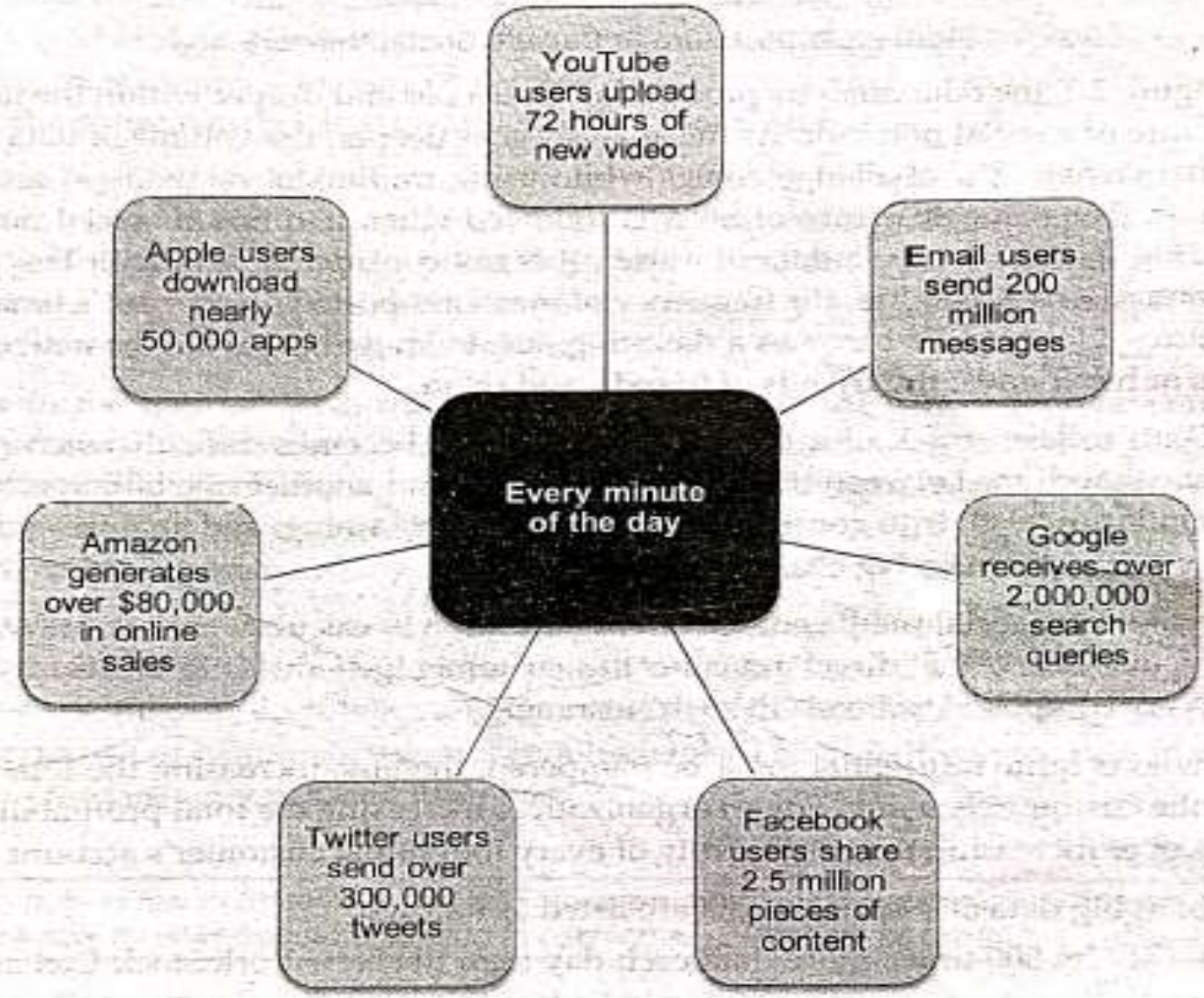


Figure 2.1: Social Network Data Generated Every Minute of the Day

Social Network Analysis

Social Network Analysis [SNA] is the analysis performed on the data obtained from social media.

Let us understand the importance of social network data with the help of an example such as

Mobile Network Operator [MNO].

The data captured by MNO in a day such as the mobile phone calls , text messages and other related details , of all its customers is very huge in volume. This type of data is used daily for different purposes.

An MNO does not simply need to record and analyse the calls of a customer but the entire network calls related to that customer. The company must study the data of the people whom the customer called and also of the people in the customers network , that is , who called back the customer. Such a network is called as a Social Network.

The data derived from social media enables an organisation to calculate the total revenue a customer can influence instead of the direct revenue the customer generates. Because of his this advantage , organisations are compelled to invest in such customers.

For an organisation, increasing the total profitability of a network takes priority over increasing the profitability of every individual customers account.

Social Network Analysis

Some facts about the big data and social media are listed as follows :

Facebook collects 500 times more data each day than the New York stock exchange.

Twitter produces 12 times more data each day than the New York stock exchange.

Social media analytics is now a days used for online reputation management, crisis management , lead generation and brand check to measure campaigning reports and much more.

The following are the areas in which decision-making processes are influenced by social networking data:

Business Intelligence

Marketing

Product design and Development

Business intelligence

Business intelligence is a data analysis process to convert a raw data set to meaningful information by using different techniques and tools for boosting business performance.

Business intelligence allows a company to collect ,store , access and analyse data for adding value to decision making.

The data generated from different social media is analysed to gain important business insights.

Social customer relationship management[CRM] data is the latest catch phrase used these days to describe this type of data.

Such a data analysis helps in changing the perspective of an organisation while valuing its customers.

Instead of valuing a single customer , organisations can now calculate the value of the entire network that is influenced by that customer.

Business intelligence

Consider an example of a mobile service provider that has a low- value customer. The customer has subscribed for a simple call plan and fails to provide any additional revenue as profit.

As per the traditional method of evaluating the profitability of a customer , if the customer is not satisfied with the services and if he or she wants to leave the company , generally has no problems to let the customer to go as he or she is providing low revenue.

However, with the help of a social network analysis ,SNA , the organization can now identify that some connections in the customer's network make a large number of calls and text messages and have a large network of friends.

With such analyses the organisation might take an altogether different decision and might start valuing that customer more. This implies that the influence of a customer is very important to the organisation.

Studies have shown that when a user of a calling network leaves , the others in the network often follow suit.

The mobile service provider can now think of investing in the customer to retain him or her.

Link Analysis & advantages / benefits of Social Network Analysis :

Social Network Analysis can also help in law enforcement and anti-terrorism efforts as it is possible to identify trouble groups or people who are directly or indirectly connected to each other. Such type of analysis is called link analysis.

Preceding examples and illustrations we can derive the following business insights or advantages / benefits of Social Network Analysis :

Social Network Analysis can help provide new contexts in which decisions are taken on the basis of available data and not on opinions.

Social Network Analysis allows organisations to shift goals from maximizing individual account profitability to maximizing the profitability of the customer's network.

Social Network Analysis helps organisations to identify highly connected customers. It also helps in identifying when , where and how to align and focus marketing efforts in building a better brand image.

Social Network Analysis enables organisations to lure[motivate] highly connected customers by offering them free trials of their services and soliciting[request] their feedback for the betterment of their products and services.

Social Network Analysis assist organisation by encouraging internal customers to become more active on corporate social networking sites and provide comments and opinions on various products and services.

Telemetry

Some organisation reward their influential customers with discounts and offers and those customers, in turn keep on spreading a positive brand image of the organisation.

The gaming industry also makes use of social networking data to develop business intelligence. Analysis of social network data or link analysis helps in tracing elementary data about gaming, including the information about who is playing which game , who plays with whom , for how much time they remain online to play the game , changing patterns of playing among people or groups and much more.

The term telemetry is popular in video gaming industry and refers to capture of in-game activities. Telemetry is analogous to a weblog analyser tool and captures activities performed by the player while they navigate through the game.

Telemetry also helps in finding the preferred partner of for a player for a particular game. Players are often segmented on the basis of their individual playing style.

Use of big data in business context

A V KRISHNA MOHAN

DEPT. of CSE , SIT

Marketing

In today's competitive in scenario, marketers aim to deliver what consumers want by using interactive communication across digital channels such as email , mobile , social and the web.

These channels in turn generate the social media data required to provide insights based upon the brand preferences of a target audience.

Conducting social network analysis on this data can generate very useful and meaningful business insights , that may help the marketers / organizations to take timely and informed decisions.

Let’s now take a case study to understand the importance of social media data on marketing.

Marketing [case study]

CASELET :

Walmart has acquired a social media analytics company named Kosmix and created at Walmart labs, a division that analyses social media data to understand retail Trends.

The director of the product management at Walmart Labs , Tracy Chu , said the the following in a blog post :

Walmart Labs is working on ways , to help Walmart to interpret social media to predict Trends and learn more about what customers want.

They are mining social media sources like Twitter and Facebook to find useful marketing insights.

One of the key responsibilities of this division is to monitor public domain conversations and then position Walmart products accordingly.

Walmart labs has been tracking social charts for analyzing the trends in various categories such as holiday Toys , online games and mobile commerce.

Walmart labs also launched Shopycat , a Facebook app , which analyses data from social networks and accordingly makes gift recommendations as per your friends likes and dislikes.

Product design and development

Business organizations / company's design and develop their products ,based on the outcome of the sentiment analysis performed on the social media data / customer feedback from social media.

To be able to measure sentiments more meticulously is of Great Value while designing a product or service.

By listening to what customers want or by understanding where the gap in the offering enable organizations to make the right decisions in the direction of their product design and development.

In this way social network data can help the organization to improve product design and development services , ensuring that the customers ultimately get the best products and services.

Sentiment analysis refers to a computer programming technique to analyze human emotions, attitudes and views across popular social networks including Facebook ,Twitter and other blogs.

This technique requires analytical skills as well as advanced computing applications.

Businesses research organizations and marketing professionals across the globe use sentiment analysis to identify , analyze and measure customer's behavior and online trends.

Most organizations today simply rely on the number of likes , tweets and comments , instead of actually studying the quality of the sentiments expressed in the conversations.

Use of big data in preventing fraudulent activities

SCENARIO :

While going through the feedbacks collected from social networking sites , Mr Smith noticed that almost all types of businesses have suffered on account of fraudulent activities and many of the business organizations are now turning to Big Data Analytics to solve this problem.

Fraud :

Fraud can be defined as the false representaion of facts , leading to concealment or distortion of the truth.

Frauds can be committed by both words and conduct and intended to deceive the other party, generally to gain financial advantage.

Frauds can occur frequently in Financial Institutions such as banks and insurance companies.

Because of these fraudulent activities , online retailers such as Amazon and eBay tend to Incur huge expenses and losses.

Common types of financial frauds

The following are the some of the common types of financial frauds :

1) Credit card fraud :

This type of fraud is quite common these days and is related to the use of credit card facilities.

In online shopping transaction , the online retailer cannot see the authentic user of the card and therefore the valid owner of the card cannot be verified.

It is quite it likely that a fake or a stolen credit card is used in the transaction.

2) Exchange or return policy fraud :

An online retailer always has a policy allowing the customer to exchange or return the goods and sometimes people take the advantage of this policy. These people buy a product online ,use it and then returned it back as they are not satisfied with the product with some reasons.

Such a fraud can be averted by charging a restocking fee on the returned goods and also staying cautious of such customers , who are known to commit such frauds.

Common types of financial frauds

3) Personal information fraud :

In this type of fraud , people obtain the login information of a customer and then log-in to the customer's account , purchase a product online and then change the delivery address to a different location. The actual customer keeps calling the retailer to refund the amount as he or she has not made the transaction.

Preventing fraudulent activities using Big Data Analytics

An organization analyses big data to prevent fraudulent activities as follows :

1) Keep track of and process huge volumes of data
2) Differentiate between real and fraudulent entries.
3) Identify the new methods of fraud and add them to the list of fraud prevention methods.
4) Verify whether a product has actually been delivered to the valid recipient.
5) Determine the location of the customer and the time when the product was actually delivered.

Fraud detection in real time

Big data also helps to detect frauds in real time.

It compares live transactions with different data sources to validate the authenticity of online transactions.

For example , in an online transaction , big data would compare the incoming IP address with the geo-data received from the customer's smart phone apps.

A valid match between the Two confirms the authenticity of the transaction.

Big data examine the entire historical data to track suspicious patterns of the customer order.

These patterns are then used to create checks for avoiding real-time fraud.

Big data analysis is performed in real-time by retailers to know the actual time , when the products were delivered to the customer.

Costly products often have sensors attached to them that transmit their location information. when such products are delivered to the customer , the streaming data obtained from the sensors, provides location information to the retailer, thereby , preventing frauds.

Fraud detection in real time [Caselet]

Visa uses a powerful fraud management system and the company has reported the identification and the prevention of potential fraud opportunities that could have cost the company a loss of around 2 billion dollars.

Fraud management system was based on a big data Technology known as Massively Parallel Processing [MPP] database.

In order to detect and prevent frauds , the system analyzes each transaction from 500 different aspects and returns with effective results.

Visually Analyzing Fraud

Image analytics is another emerging field that can help to detect frauds.

It refers to the process of analyzing image data with the help of digital processing of the image.

Examples include the use of barcodes and QR codes.

Some other examples include Complex solutions such as facial recognition and position and movement analysis.