Big Data, Cloud and Analytics

Block

1

INTRODUCTION AND APPLICATIONS OF BIG DATA

Editorial Team

Prof. R. Prasad IFHE (Deemed-to-be-University), Hyderabad	Dr. Sanjay Fuloria IFHE (Deemed-to-be-University), Hyderabad
Dr. Sindhuja IFHE (Deemed-to-be-University), Hyderabad	Dr. Nasina Jigeesh IFHE (Deemed-to-be-University), Hyderabad

Content Development Team

Prof. Anirudh Prakhya IFHE (Deemed-to-be-University), Hyderabad	Dr. Y. V. Subrahmanyam IFHE(Deemed-to-be-University), Hyderabad
Prof. R. Muthukumar IFHE (Deemed-to-be-University), Hyderabad	Prof. V. Srinivasa Murthy IFHE (Deemed-to-be-University), Hyderabad
Dr. K. Veena IFHE (Deemed-to-be-University), Hyderabad	Prof. KV Murali Mohan IFHE (Deemed-to-be-University), Hyderabad
Prof. M. Venkata Dharma Kumar IFHE (Deemed-to-be-University), Hyderabad	

Proofreading, Language Editing and Layout Team

Prof. Jayashree Murthy IFHE (Deemed-to-be-University), Hyderabad	Mr. K. Venkateswarlu IFHE (Deemed-to-be-University), Hyderabad
Ms. C. Sridevi IFHE (Deemed-to-be-University), Hyderabad	

Ref. No. BDCA-SLM-IFHE – 122022B1

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

Our E-mail id: cwfeedback@icfaiuniversity.in

Centre for Distance and Online Education (CDOE)
The ICFAI Foundation for Higher Education
(Deemed-to-be-University Under Section 3 of UGC Act, 1956)
Donthanapally, Shankarapalli Road, Hyderabad- 501203

COURSE INTRODUCTION

With digitization everywhere, data professionals have their contribution in all kinds of industry verticals. These are finance, manufacturing, information technology, communications, retail, logistics, and automobiles. Big data is used for gaining competitive advantage and making data-oriented decisions, i.e., approach of making decisions on the data analysis, than on intuition alone. Many real world big data business examples can be found in operation: Streamlined media streaming, Discovering consumer shopping habits, Predictive inventory ordering, Personalized marketing, Personalized health plans for cancer patients, Fuel optimization tools for the transportation industry, Monitoring health conditions through data from wearables, Real-time data monitoring and cyber security protocols, and Live road mapping for autonomous vehicles. Thus study of big data gives wider view of various business application and uses.

The domain of Big Data, Cloud and Analytics can greatly enhance analytical skills and reasoning. The decision can be on possible future trends, providing an advantage to be the first to market and also be competitive. The field of data analytics is changing business best practices in all types of industries from manufacturing to marketing. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more informed business decisions. Data analysts use a variety of tools and technologies to gather all sorts of data, like statistics about how much time users spend on a website, demographic information about customers or traffic patterns in a city. Descriptive analytics reveal what happened in the past. Diagnostic analytics answer why something happened. Predictive analytics tell what will probably happen in the future. Prescriptive analytics show what actions should be taken to make progress or avoid problems. Some of the corporate examples that used big data and data analytical tools and techniques include: Increasing the quality of medical care, Fighting climate change in local communities, Revealing trends for research institutions, Stopping hackers in their tracks, Serving customers with useful products, Driving marketing campaigns for businesses, Promoting smart energy usage for utility companies, Improving the insurance industry, and Creating manufacturer warranties that make sense.

The course covers the topics in five blocks:

Block - 1: Introduction and Applications of Big Data

Block - 2: Cloud Computing and Big Data Technologies

Block - 3: Business Analytics

Block - 4: Managing Talent for Big Data Analytics

Block - 5: Data Privacy and Analytics in Various Business Areas

Big data is being generated (structured as well unstructured) in all service organizations and other business areas. Especially in the areas of advertising; marketing as well healthcare, big data generates huge amounts of structures, semi structured and unstructured data. Thus Block-1 Introduction and Applications of Big Data covers overall introduction to big data, its importance, big data generation in marketing and advertising as well in various healthcare applications.

To work with big data and get benefits, the focus needs to move to know about big data technologies, various methods, how to use big data for decision making in business, focus on handling unstructured data storage, usage, analysis and big data in information management at organizations. Thus Block-2 on Cloud Computing and Big Data Technologies covers Hadoop, open source technologies, BI related analytics, decision making using big data, handling unstructured data.

Having large volumes of collected data will go waste, if crucial decisions could not be arrived at from the primary data. Data analytics is the science of analyzing raw data in order to make predictive analysis and conclusions about the information. Thus, data analytics is a vital element in various business areas like predictive analysis and decision making, which a learner needs to get prepared. Thus business analytics is essential knowledge to acquire. Block-3 on Business Analytics covers complete aspects of business analytics spanning over: introduction, models, and applications.

A data scientist's main role is to organize and analyze large portions of data through custom-designed analytical software, in order to provide stakeholders with findings that they can use to make informed business decisions. A data scientist typically has various roles and functions. More so working with big data, data scientist can help build many predictive solutions. The learner needs to play the data scientist role on need basis. The Block-4 on Managing Talent for Big Data Analytics covers completely about data scientists, role, use of mathematics and computer science, culture of decision science and play a significant role at business decisions.

Analytics role in various business functions like: HR, top management decision making, marketing intelligence, is a very useful knowledge to the learner. Current day practices need also to focus on data privacy and business ethics for a socially responsible business activity. The Block-5 on Data Privacy and Analytics in Various Business Areas covers HR analytics, Data analytics and allied analysis methods, business and market intelligence as well data privacy and ethics.

This edition has added a large number of contemporary examples and deleted old examples and exhibits.

BLOCK 1: INTRODUCTION AND APPLICATIONS OF BIG DATA

Big data is being generated (structured as well unstructured) in all service organizations and allied areas. Especially in the areas of advertising and marketing as well healthcare, big data helps arrive at many predictive analytical decisions. Thus Block 1- Introduction and Applications of Big Data, covers overall introduction to big data, its importance, big data generation in marketing and advertising as well in various healthcare applications. By virtue of social media proliferation, varieties of social media being used by millions of citizens, podcasts, you tubes, maps being shared across, generate multiple types of data, and the importance of big data has grown. While one aspect is the collection storage and analysis, the other aspect is, use it for further research, business decisions and predictive analytics in different verticals. This block deals with units on, what is big data? Why big data is important, big data in marketing and advertising details and big data in healthcare.

Unit-1 Before working with big data, the learner needs to get full knowledge of *What is Big Data*? It impresses the learner why it’s different than conventional data collections and analysis. Thus, this unit on what is big data; deals with - Introduction to big data, Concepts of big data; utilities in general and corporate life, and Key Trends in big data.

Unit-2 Data collection through various transactions had been a common approach, and present day social media usage, you-tube usage, podcasts, maps are generating very unstructured data. This unit on *Why is Big Data Important?*; examines why big data needs an appreciation from learner, about how such big data is feasible to be captured and what constitutes big data and covers various natures of data being generated in business and social platforms, what are unstructured data, and the growing magnitude of unstructured data.

Unit-3 One major business area where big data is generated includes marketing and advertising and allied effects and learner needs to analyze the data from these business areas to capture big data effects and analytics. This unit on *Big Data in Marketing and Advertising*; covers big data generation in marketing and advertising, the impact of present day digitization and the digital marketing perspective and database markets, how big data paves way for new thought of marketing, fraud; risk and big data, credit risk management, algorithmic trading, advertising and big data using consumer products as a doorway.

Unit-4 another socially useful business area is healthcare and learner needs to link the big data generation and use in healthcare for the benefit of the society as a whole. The unit on *Big Data in Healthcare* discovers big data generation and the advances in healthcare, advantages and disadvantages, paving way for new frontiers in healthcare, and healthcare consumer products.

Unit 1

What is Big Data?

Structure

"It is a capital mistake to theorize before one has data."

- Sherlock Holmes

1.1 Introduction

If a company truly wants to understand what's going on in its chosen market, it needs data.

The volume of data in our world has been detonating. Companies generate trillions of bytes of information about their suppliers, customers and operations. Lots of networked sensors are being entrenched in the physical world in automobiles and devices such as mobile phones. These sensors sense, produce and communicate data. People with smartphones, persons on social network sites and multimedia will continue to fuel exponential growth of data. Large pool of data that can be seized, transferred, accumulated, stored and scrutinized is called big data. Now, big data is a part of every sector and influences the global economy. Like other areas such as human capital and hard assets, much of modern economic activity, revolution and growth merely couldn't happen without data.

Big data is the buzz word of current information and technology world. It gained popularity in the first decade of 21st century. Big data is the data which has been accumulated and shared by many individuals and organizations. This is because of technological improvement and availability of equipment to store and share the information. As storage cost of data is inexpensive, every individual/ organization stores and shares huge amount of data every day. This creates voluminous amount of data which grow exponentially every day. The data is stored for future utilization to improve decision making. Since the data grow exponentially, there is a need to find useful information from this 'Big Data'.

But bigger data consequently require different approaches, tools and techniques to solve new problems and to solve old problems in a better way. Like many other technologies, big data can bring dramatic cost reduction and substantial improvements in time required to perform the computing task.

The previous decade's successful web startups are leading examples of how big data was used as an enabler of innovative products and services. For example, by relating a large number of signals from a user's activities and those of their contacts, Facebook has been able to expertise a highly custom-made user experience and generate a new kind of advertising business.

In this unit, we will explore big data, the reasons and history of its emergence, and what makes big data different from normal data.

1.2 Objectives

By the end of this unit, you should be able to:

- Explain what is big data
- Describe the emergence of big data
- Outline how big data differs from other normal data

1.3 What is Big Data? - Definition

Big data can be defined in many ways.

Big data can be described as, "Big volumes of unstructured, semi-structured and structured data, which has the inherent potential to be mined for important information" [Structured data is highly specific and is stored in a predefined format, where unstructured data is a conglomeration of many varied types of data that are stored in their native formats. Semi-structured data is a data type that contains semantic tags, but does not conform to the structure associated with typical relational databases].

Big data can be defined as "voluminous data that can be scrutinized computationally to expose patterns, trends, associations, relating to human interactions and behavior."

Big data is a buzzword used to designate a massive volume of structured and unstructured data that are extremely large and tough to be processed using customary database and software techniques available.

Big data is a wide term that works with large and complex data sets which cannot be handled by traditional data processing applications and distributed databases are required to process it.

Big data is a high "variety and velocity" and high-volume information resource. Big data mandates useful, advanced methods of information. It facilitates processing that facilitate better insight, decision making and process automation.

Big data is the term progressively used to designate the process of applying grave computing power to seriously gigantic and often extremely multifarious sets of information.

When the data which is in groups are large and heavy, they are called big data. It is difficult for the traditional database management to capture, store and share such a large and heavy data.

Big data refers to the immense amounts of data collected over time that are tough to analyze and manage using common database management tools.

Big data includes e-mail messages, business transactions, and surveillance videos, photos and activity logs generated by computers or smart machines. Big data also comprises unstructured text posted on blogs, social media and on the Web.

Any discussion about big data would be incomplete without a reference to its five fundamental characteristics, usually called 5V's:

- Volume denotes that the data has huge volume
- Variety refers to variety of type and source
- Velocity denotes the speeds of generation
- Veracity means that the data is authentic
- Value mentions that it adds value to the users

Example: Tesla Exploits Big Data Analytics to Become one of the most Valuable Auto Companies in a Short Time

Tesla was well known for its self-driving cars. Leveraging Big Data collected from on board sensors in the cars and from the environment, it was being processed by its servers in the Cloud in real time and it provided guidance to the vehicle even before the driver came to know. The unstructured and semi structured data from sensors was analysed using machine learning models to continuously add new automation features and improve safety of the vehicle and its occupants.

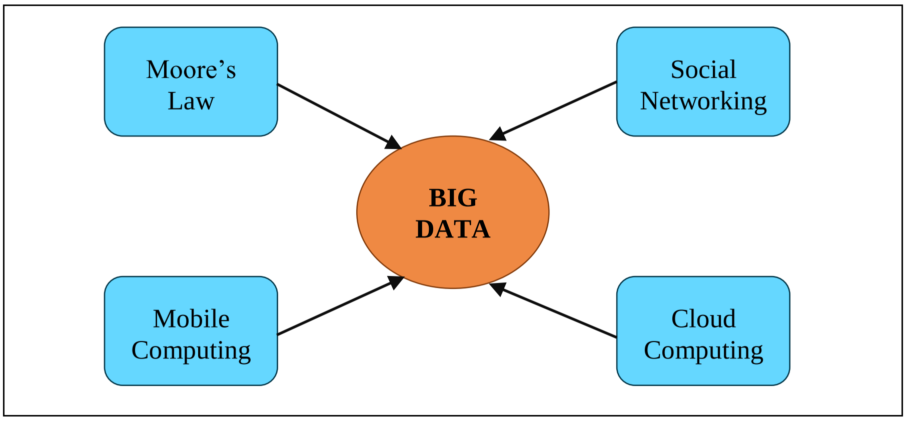
Source: https://digital.hbs.edu/platform-digit/submission/tesla-a-data-driven-future/, March 23, 2021, Accessed on 03/08/2022

1.4 Reasons for Big Data Emergence

Big data is not an overnight miracle. It has many roots and branches. Big data has emerged from firms that have been producing and managing tons of transactional data over the years. There are many reasons for the emergence of big-data. They are:

1. **Comfort of Computing Power:** Big data is the outcome of four major global trends: Moore's Law (technology always gets cheaper and computing power doubles in every two years), mobile computing (tablet or smart phone we use to communicate), social networking (Twitter, Facebook, LinkedIn, etc.), and cloud computing (don't even have to own computing infrastructure). It is shown in Figure 1.1.

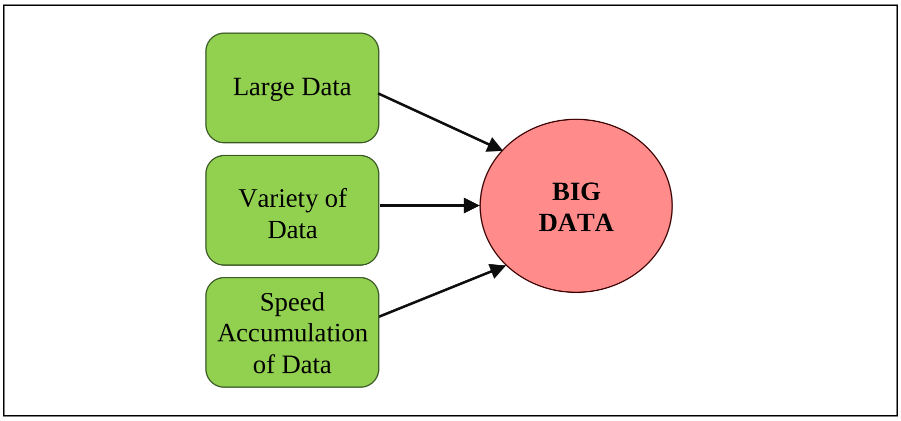
Figure 1.1: Enablers of Big Data



Source: ICFAI Research Center

2. **Availability of Data:** Volumes of transactional data are stored for decades in big firms. But computerization has resulted in fabrication of variety of data (text, image, video and audio) with speed (velocity) (Figure 1.2). This has made the conventional data management redundant.

Figure 1.2: Big Data



Source: ICFAI Research Center

3. **Emergence of Advanced Computing Techniques:** Conventional analytics, software and hardware technologies, data management and data processing tools, commodity hardware and open-source technology are merged to provide a solution to this problem. This enables IT professionals and business executives to efficiently handle big volumes of data.

Example: The Rise of Big Data Technologies and why it Matters

Big data happens when input is more than what can be processed with data management systems. Google is an American multinational technology company which focuses on search engine technology, online advertising, cloud computing, computer software, quantum computing, e-commerce, artificial intelligence, and consumer electronics. Google proposed Google file system, a technology for indexing and managing mounting data. A key tenet to the idea was using more low-cost machines to accomplish big tasks more efficiently and inexpensively than the hardware on a central server.

Source: https://fusionalliance.com/the-rise-of-big-data-technologies-and-why-it-matters/, March 24, 2021, Accessed on 09/26/2022

1.5 Evolution of Big Data

Big data is the emerging topic in the business functional activities. The roots of big data run deep. Certain milestones of tracking the historic improvement on how data has been collected, stored, managed, and analyzed are given below.

1926 - Nikola Tesla has foreseen human capability to access and investigate enormous amounts of data in the future.

1940’s-1950’ – Electronic computing was invented to make high speed calculations.

1944 - Fremont Rider, the librarian from Wesleyan University, quantified that the libraries in US are doubling in size every 16 years.

1949 - Father of Information, Claude Shannon did research on photographic data and punch cards which are big storage capacity items. The Library of Congress was the biggest item in Shannon’s list with 100 trillion bytes of data.

1960 - Data warehousing became inexpensive.

1996 - R.J.T. Morris and B.J. Truskowski stated that Digital data storage became more economic than paper storage.

1997 - The term “Big Data” was introduced by Michael Cox and David Ellsworth of NASA’s Ames Research Centre, for the first time in a paper published on Visualization. They highlighted the problems handling hefty unstructured data sets using the current traditional computing systems.

1998 - Carlo Strozzi developed ‘NoSQL’, an open source relational database.

1999 - Information was measured by computer storage terms like GB, TB, PB, etc. (Hal R. Varian and Peter Lyman at UC Berkeley)

According to the study in 1999 titled “How Much Information?” the world had created roughly 1.5 Exabyte’s of Information.

1999 - The term Internet of Things (IoT) was introduced for the first time by Kevin Ashton in Procter & Gamble’s business presentation

2001 - Gartner Analyst Doug Laney in a research article, defined the three dimensions of big data - 3Vs.

2004 - Wal-Mart claimed to have the leading data warehouse with 500 terabytes storage.

2005 - Doug Cutting and Mike Cafarella developed the tiny toy elephant Hadoop to manage the big data blast from the web.

2006 - International Data Corporation (IDC) did the first study of estimating the amount of information growth. Data creation of 161 exabytes (EB) of data was estimated for the year 2006. This was expected to go up by 6 times in 4 years.

2007 - Institute of Advanced Analytics, North Carolina State University offered the first Master’s degree in Analytics.

2008 - Bret Swanson and George Glider, forecasted the US IP traffic to touch 1 zettabyte by the end of 2015. Another prediction was that the US internet usage will be 50 times more than it was in 2006.

2008 - World’s CPUs can process 9.57 trillion gigabytes of data.

2008 - According to Global Information Industry Centre survey in 2008, a regular user ingests 34 gigabytes of information on an average and 100,500 words in a day.

2008 - In a single day, Google processed 20 petabytes of data.

2009 - eBay storage totaled eight petabytes (which is equal to 104 years of HD-TV video).

2009 - McKinsey report projected that, a US company with 1000 employees on an average stores data more than 200 TB.

2011 - The Yahoo warehouse amounted to 170 petabytes (which is equal to 8.5 times of all hard disk drives created in 1995).

2011 - It is computed that 1.8 zettabytes of data was generated from the beginning of civilization to the year 2003. In 2011, 2 days was enough to produce 1.8 zettabytes of data.

2011 - A McKinsey report on big data predicted the scarcity of analytics professionals talent in US by the year 2018. US alone needed 1.5 million analysts, 190,000 skilled professionals and 140,000 managers with deep analytical skills.

2011 - ‘Watson’, IBM’s supercomputer analyzed 4TB of data in seconds, equivalent to 200 million pages

2011 - Ventana Research Survey identified that 94% of Hadoop users use big data analytics on exponential amounts of unstructured data which was probably not feasible earlier.

2012 - Obama administration initiated big data research and development initiative. This comprised of 84 big data programs to tackle the increasing influx of data problems faced by the government. It is also interesting to recall that big data analysis played a key role in Obama’s 2012 re-election promotion. US government earned an amount of $200 million through big data research projects.

2013 - 4.4 zettabytes (4.4x 1021) of data were made by the Digital Universe Study. It appraised that the volume of big data is anticipated to grow exponentially to 44 zeta bytes by end of 2020.

2013 - Gartner forecast that rise in spending on big data analytics by various organizations would be to the tune of 72%.

2013 - EMC study predicted that only 35% of the data created will have semantic value by 2020.

According to Forrester, U.S. business-to-business (B2B) ecommerce transactions are expected to reach $1.8 trillion by 2023. This would account for 17% of all B2B sales in the country

An IDC Digital Universe study estimated that amount of digital data created per year will be 35 zettabytes by 2020. The study predicted that cloud computing will play an integral role in managing data growth. Other observations are that by 2021, 35 billion IoT devices will be installed around the world and the number of connected devices in 2020 would hit 50 billion mark.

2015 - Research assessments recommend that 2.5 quintillion (i.e. 2.5 followed by a total of 18 0’s) bytes of data is generated every day.

2015 - Google became the largest big data company in the world which stored 10 billion gigabytes of data and analyzed around 3.5 billion queries every day.

2015 - Amazon became the company with the maximum number of servers. 1,400,000 servers in various data centers, 152 million customers store 1,000,000,000 gigabytes of big data created at Amazon.

2018 – More than 2.5 quintillion bytes of data was created every day. People used 3.1 million gigabytes of internet data, and mined 1.25 new bitcoins every minute.

2021 – Astonishingly Amazon, Microsoft, Google, Facebook together store 1,200 petabytes of information. These industry leaders continue to increase the amount of data on their websites.

2025 Estimate – The world would have 75 billion Internet of Things (IoT) devices.
Globally, the amount of data generated each day would be 463 exabytes

Opportunities in big data industry for computer programmers, entrepreneurs, investors, and other IT professionals are increasing on a very large scale.

Activity 1.1

Anjali has been hired as the chief technology manager in an online learning firm. She finds piles of data associated with each learner and the way they pick up learning. She wants to customize and offer dynamic learning to each learner based on his/her style and type of learning. Further, she wants to help predict apt careers for each learner.

What can Anjali make use of to understand each learner`s need? What can be the reasons for the emergence of such an amount of data with Anjali?

Answer:

Check Your Progress - 1

1. What is 'big data'?
   a. Huge data of text
   b. Data which is big in nature
   c. Data with noise
   d. Large Data which needs more computing power and advanced techniques to process it
   e. Movie data
2. What does Moore's law tell?
   a. Computing power decreases
   b. Computing power increases with time
   c. Hardware cost increases
   d. Hardware cost decreases
   e. Technology gets cheaper and computing power doubles in every two years
3. Which of the following is not a social network?
   a. Twitter .
   b. Facebook
   c. Google Plus
   d. Amazon
   e. LinkedIn
4. Which of the following is an open source relational database?
   a. NoSQL
   b. Oracle
   c. DB2

d. SAS
e. SQL Plus

5. Hadoop is developed to manage which of the following?
    a. Sensor data
    b. Small amount of data
    c. The big data blast from the web
    d. Hard disk
    e. Semantic data

1.6 Evolution of Data Systems

Having discussed about data, evolution and volumes, one can realize that to process the data, data management systems are used. There is a massive change in the data management systems also. According to Misha Ghose, the evolution of data systems is categorized into three stages they are:

1. **Dependent (Early Days):** Users of data systems were not able to understand the analytical platform and were dependent on solution / database providers to solve their problem and were not able to define their business needs.
2. **Independent (Recent Years):** Users of data systems understood what an analytical platform was and worked collaboratively with IT to describe the business needs and solve it.
3. **Interdependent (Big Data Era):** This is the interactional stage marked by collaboration among various companies in order to meet the business requirements. This stage occurs when the database is very large, having variety and is accumulated with speed. Organization may need mining of data for interpretation and efficient decision. In this stage, the organizations look to collaborate with more external factors to achieve the result.

Big data era actually takes the organization from independence to interdependence for evolution of data.

Example: The Possibilities of Data and Computing - 180 Years in the Making

IBM started its journey with data and innovation in 1911 with punch card and other office automation machines. Since then, the data and technology that they fostered has been used to innovate Apollo landing on the moon, launch one of the first commercially available personal computers, and create Deep Blue – the first chess “machine” to win a match against a reigning human chess champion – and computers that could do natural language manipulation and cognition, such as Watson, featured on the gameshow ‘Jeopardy’. IBM has also taken the impressive approach to democratise quantum computing, and artificial intelligence.

Source: https://www.dnb.com/perspectives/master-data/the-evolution-of-data.html, February 2, 2022. Accessed on 09/26/2022

1.7 Flood of Mythic "Start-Up" Proportions

Companies like Google, Facebook, LinkedIn, eBay and others rely on skills of data scientist to overcome the barriers of traditional system. They leverage new technologies and methodologies to analyze the massive data. It is the data which drives their business. Today, even start-ups are applying technology to solve big data issues.

According to McKinsey Global Institute's recent study, business organizations (mostly financial institutions, health care institutions, FMCG and retail stores, meteorological sectors, etc.) capture trillions of bytes of data about their suppliers, customers, and operations using digital systems after digitization of companies. Networked sensors embedded in automobiles, mobile phones, and other electronic products are persistently sensing, creating, storing and communicating data. This results in 40% of volume of data growth every year. And it also specifies that 15 out of 17 US sectors already acquired enormous data.

1.8 How does Big Data Differ from Normal Data?

How does big data differ from the normal data?. Big data normally refers to the tremendous data which cannot be processed by the traditional data management systems and methods. New approaches are needed to store and process the high volumes of data. Data stored grow in size with time. The technology advances. This calls for different thinking to select the right information and pattern from the vast ocean of available data. It is here that the concept of big data gains relevance. Big data deals with not only the volume, but also the variety of data stored. E.g. text, image, video and audio and also structured, semi-structured and unstructured. Different types of data require different approach, processing time and speed. Methods used and the speed required to acquire the data (online transactions, embedded sensors, etc.) also matter.

Example: China Eastern Airlines Relies on Big Data Analytics to understand Customer Preferences and Provide More Customized Services to its Passengers

China Eastern Airlines is one of the major airlines in the world. It deployed Big Data analytics and machine learning to obtain a better view of customer travel preferences and provide more accurate services based on this. The company had recently introduced a new loyalty program which was "revenue based" rather than "mileage based" as was the norm in the airlines industry. The customer was offered other services apart from air travel by partnering with companies in other services.

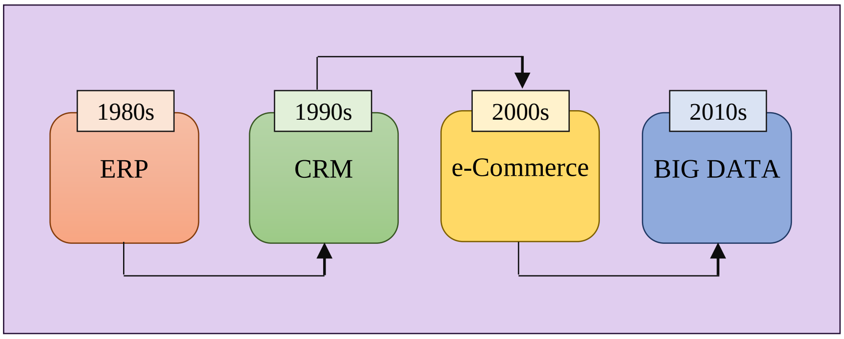
Source: https://www.businesswire.com/news/home/20210928006185/en/China-Eastern-Airlines-Released-New-Loyalty-Program-Globally%21, September 28, 2021, Accessed on 04/08/2022

1.9 Big Data - Why now?

What is the need for big data technology? Why do companies need to spend for it? These are queries which normally people ask. Figure 1.3 depicts the technological development in recent past which affected businesses. Whenever any new technology arises, business organizations spent on it to acquire it and face failure later. When Customer Relationship Management (CRM) emerged as the new concept in 1990s, many organizations spent for CRM suites and tools, but failed to attain the expected results. The reason was that the project required a harmonized involvement of process, people and technology which was lacking.

If the process, people and technology are not aligned together, the project fails. To overcome that, the organizations have to analyze and answer the questions like what is the usage of big data analytics? Whether its output/results will be used in organizational decision making? How far it will be useful, and in what way? If big data analytics could improve the business, how much can we spend for that?

Figure 1.3: Recent Technological Development Timeline



Source: ICFAI Research Center

Example: Accuweather Deploys Big Data Technologies to Provide 30 Billion Api Request Every Day from 2 Billion Users

Accuweather is ~~was~~ a global weather media company. It collected weather data from various locations on a real time basis and some 2 billion users daily accessed this data through APIs (Application Program Interface) to generate useful content to individual mobile users. The company received 30 billion requests every day and Big Data Technologies were deployed to handle such huge volume of data growing on a real time basis. High speed delivery of data to end users is ~~was~~ critical to enable them to gear up for dangerous weather situations like storms.

Source: https://cloud.google.com/customers/accuweather, 2022, Accessed on 04/08/2022

1.10 A Convergence of Key Trends

Different things have converged to shape big data analytics. Companies stored large information previously also. Initially, they were storing data on tapes, which are slow to access. Nowadays, technologies like Hadoop, can be used to extract and access the data in a faster manner. Even though, tremendous amount of data is stored, Hadoop can extract the required information from it.

Another reason is the availability of cheaper hardware. Hard disks and storage devices are very economical. So we can store any amount of data and it is possible to analyze the data quickly with low cost. Thus, convergence of more trends like less expensive and faster hardware, and technologies like Hadoop made big data analytics easier. In the last decade, data analytics was prohibitive due to the high cost, but now it has become affordable.

Previously, data was analyzed to answer general queries. But now data can be used to answer specific questions related to an individual. E.g., an insurance company can suggest a policy depending on the customer`s age. The large data stored in an organization can be used to compare different things, e.g. the insurance companies can find out what are the preferences of individual customers? What are the preferences of a particular age group? How can products be combined to give a combo offer to increase sales of slower moving items? In short, the huge data stored can be used to take very important decisions.

Example: Etsy (US Based Retail Major in Handmade Products) Develops Machine Learning Models to Improve its Product Search Based on Style Preference

Etsy operates a marketplace connecting around 40 billion buyers with some 4 billion sellers across the globe. The platform is exclusively for handmade and vintage products. Since the company deals with handmade products, most of its 60 million products are one of a kind and as such normal search engines fail to give correct results. So, the company data science team created machine learning models to go through text and visual data and arrive at some 42 styles into which the products can be categorised. This improved the search by many times.

Source: https://digital.hbs.edu/platform-digit/submission/etsy-building-an-algorithm-with-an-eye-for-fashion/, MAR 24, 2021, Accessed on 04/08/2022

Activity 1.2

You work as a Senior IT Manager in a private sector bank. You deal with lakhs of transactions in a day. You want to detect fraudulent transactions before it affects your organization. You wish to make use of the big data generated through various transactions made by customers. Explore how the convergence of key trends can be useful for your purpose.

Check Your Progress - 2

6. Which is the largest big data company in the world?
    - a. Yahoo
    - b. Amazon
    - c. Google
    - d. Gartner
    - e. IBM
7. Which Company has maximum number of servers as of now (2015)?
    - a. Yahoo
    - b. Amazon
    - c. Google
    - d. Gartner
    - e. IBM
8. Who introduced the term ‘big data’?
    - a. R.J.T. Morris and B.J. Truskowski
    - b. McKinsey
    - c. Obama
    - d. Gartner
    - e. Michael Cox and David Ellsworth
9. Who introduced ‘NoSQL’?
    - a. R.J.T. Morris and B.J. Truskowski
    - b. McKinsey
    - c. Doug Cutting and Mike Cafarella
    - d. Carlo Strozzi
    - e. Michael Cox and David Ellsworth
10. Who developed ‘Hadoop’?
    - a. R.J.T. Morris and B.J. Truskowski
    - b. McKinsey
    - c. Doug Cutting and Mike Cafarella
    - d. Carlo Strozzi
    - e. Michael Cox and David Ellsworth

1.11 Summary

- The unit explained what big data is and then how it is emerging as a new technology.
- It then discussed the evolution of big data concept and the evolution of data systems.
- Then we found how it is different from normal data which is processed by traditional systems and approaches.
- Then we continued with the reasons for the importance of big data, and discussed about quickly deriving business value from a range of emerging data sources, including location data generated by smartphones and other roaming devices, public information available online, social media data, and data from sensors embedded in cars, buildings and other objects.,.
- The unit concluded that various factors converged and enabled big data.

1.12 Glossary

B2C: When information about customer is available but scattered around files that is difficult to use, it is known as 'silo' effect.

BIG Data: Enormous amount of data which grow by business transactions and require advance techniques and approaches to manage and process it.

Cloud Computing: Practice of using a network of remote servers accommodated on the Internet to store, manage, and process data, rather than a local server or a personal computer on demand per usage cost.

Data Base (Management) System: The database (management) system is a computer application, which interact with the users and other applications, to store, manage and retrieve data.

EMV: EMV is a technical standard for smart payment cards and for payment terminals and automated teller machines that can accept them.

Hadoop: Hadoop is an open source solution which helps storage and process of large unstructured data sets. One such source of unstructured data is Electronic Mail. Large amount of ASCII files, jpg images and video attachments are transferred using electronic mails. The e-mails use SMTP (Simple Mail Transfer Protocol) for small mailing services.

IDC: International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets.

Mobile Computing: Technology that allows transmission of data (text, image, voice and video) via a computer or any other wireless enabled device without having to be connected to a fixed physical link.

Moore's Law: Computing power doubles approximately every two years.

Petabyte (PB): A petabyte is a measure of memory or storage capacity and is 2 to the 50th power bytes or, in decimal, approximately a thousand terabytes.

Predictive Modeling: Estimating purchasing and spending patterns of a customer to predict about his future purchasing behavior.

Quintillion: a thousand raised to the power of six (10^{18}).

Social Networking: The use of dedicated websites and applications to interact with other users, or to find people with similar interests to one's own.

Terabyte (TB): A terabyte is a measure of computer storage capacity that is 2 to the 40th power, or approximately a trillion bytes which is equal to 1024 GB (Gigabyte).

1.13 Self-Assessment Test

1. What is 'big data'?
2. Explain the reasons behind 'big data' emergence.
3. List out the important milestones in big data evolution.
4. Explain the evolution of data base systems.
5. Explain the difference between big data and the normal data.
6. Which made big data analytics possible and affordable?

1.14 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

1.15 Answers to Check Your Progress Questions

1. **(d) Large data which need more computing power and advanced techniques to process it**

   ‘Big data’ is the large data which needs more computing power and advanced techniques to process it.

2. **(e) Technology gets cheaper and computing power doubles every two years**

   Moore’s law tells that technology gets cheaper and computing power doubles in every two years.

3. **(d) Amazon**

   Amazon is a service provider.

4. **(a) NoSQL**

   NoSQL is a query language.

5. **(c) Manage the big data blast from the web**

6. **(c) Google**

   Google is the largest big data company.

7. **(b) Amazon**

   Amazon has maximum number of servers (2015).

8. **(e) Michael Cox and David Ellsworth**

   Michael Cox and David Ellsworth introduced the term big data.

9. **(d) Carlo Strozzi**

   Carlo Strozzi. introduced NoSQL.

10. **(c) Doug Cutting and Mike Cafarella**

    Doug Cutting and Mike Cafarella developed Hadoop.

Unit 2

Why is Big Data Important?

Structure

"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway"

- Geoffrey Moore, management consultant
and author of Crossing the Chasm

2.1 Introduction

Data analytics provides insights which the data tells. These insights provide competitive advantage to companies in the marketplace.

In the previous unit, we were introduced to big data, its definition and how it differs from normal data. Big data is relatively big in volume and its processing is not possible with conventional systems with normal capability. Normally companies and institutions which are computerized or are maintaining manual data, store the data for future decision making. This is true in case of banking sector, other financial sectors including stock brokerage companies, investment firms, retail businesses, medical institutions and hospitals, government agencies, meteorological departments, aerospace analytics, air-travel forecasting, research & development, air-traffic control, etc. These institutions store a lot of data. Their data grows in volumes within a short span.

The data stored should be utilized for future decision making, otherwise the efforts and money spent for storing the huge data become waste. After globalization, the companies have spread across the world. Hence to take any effective decision, the entire data, collected and stored from various units across the world need to be analyzed. This data is very large and cannot be processed by the conventional systems. This is a typical example of big data and big data analytics. Though big data has now emerged as a new technology, its roots date back to 1960s, when data base management systems were used to store data.

In this unit, we will explore more about big data, the characteristics of big data, wider variety of data used in big data analytics, structured and unstructured data and also why it is important to have big data analytics.

2.2 Objectives

By the end of this unit, you should be able to:

- Define big data in depth.
- Label the varieties and types of data.
- Recognize why big data and its analytics is important.

2.3 Why is Big Data Important?

Nowadays, like human capital and hard assets, data also is an essential factor of every sector. Every activity, be it production, service or marketing, needs data. Without data, most of the modern economic activities cannot take place. Data is the basis of growth for individual firms, improving productivity and generating substantial value for the world economy by plummeting waste and aggregating the quality of products and services.

According to McKinsey & Company’s Business Technology Office and McKinsey Global Institute (MGI), the complete volume of data produced, stored, and excavated for insights has become frugally relevant to businesses, consumers and the government.

The history of previous trends in IT investment and its impact on productivity and competitiveness strongly recommends that big data & its analytics can have a similar power. All companies must consider big data and analytics to create value earnestly if they want to compete.

There are four different approaches used in analytics, to react or pro-act for a business situation. They are,

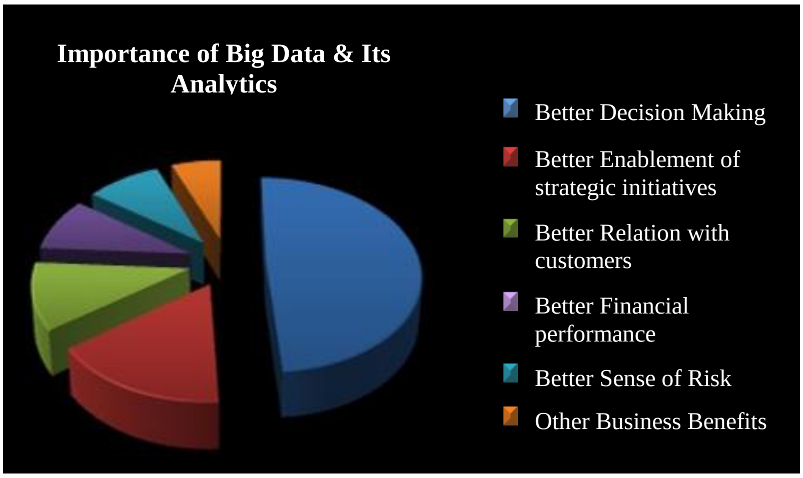
1. **Business Intelligence:** Business Intelligence (BI) provides ad hoc reports, standard business reports, alerts and notifications based on analytics, in the reactive category.
2. **Big data BI:** When reporting tweaks from vast data sets, this is termed big data BI. But judgments based on these two approaches are still conservative.

3. **Big analytics:** Creating proactive, forward-looking resolutions requires proactive big analytics like predictive modeling, optimization, forecasting and statistical analysis. These allow us to identify spot weaknesses and trends, or define conditions for making decisions. This might be proactive. But it cannot be performed on big data because traditional storage and processing are not sufficient for big data analytics.

4. **Big data analytics:** Using big data analytics, we can mine only the pertinent information from data stored in TBs (terabytes), PBs (petabytes) and EBs (exabytes), and analyze it to improve business decisions for the future. It is proactive, with sound knowledge and deep insight and can be used to meet the future requirements.

   Big data analytics is important in business. Big data leads to better decision making opportunities, utilization of strategies in a better way, better CRM, better financial performance and other business benefits. The importance of big data and its analytics is shown in Figure 2.1.

Figure 2.1: Importance of Big Data & Its Analytics



Source: https://medium.com/@pdvekariya1/10-reasons-why-big-data-analytics-is-the-best-career-move-526015c17cd9

Example: First Tennessee Bank Deploys Big Data Analytics to Reduce Marketing Costs by 20 Percent

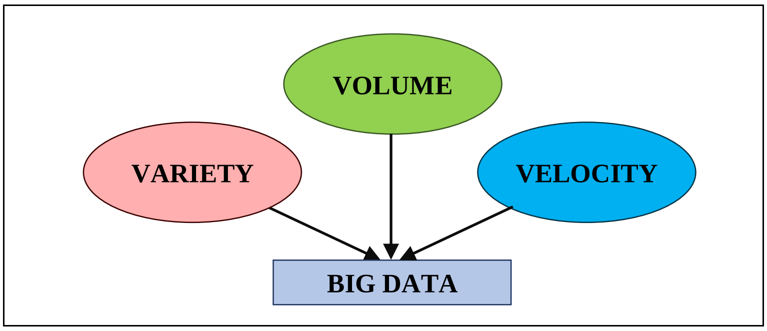
The First Tennessee Bank used predictive analytics to understand customer spending patterns and customized its offerings and attempting upselling and cross selling. The bank could offer different customized offers to customers which resulted in 600% ROI (Return on Investment). Predictive analytics helped the bank to achieve 3.1% increase in customer response and the marketing costs came down by 20%

Source: Big Data in Banking – Leapfrogging into Digital Banking Era - TechnoCapt.com, 02-July-2022, Accessed on 06/08/22.

2.4 Characteristics of Big Data

Big data cannot be defined using its size or volume alone. It can be measured or defined according to the context. For small companies, a data of 10 TB size is big data whereas big organizations and global establishments treat 500 TB or more as big data. Government organizations consider a petabyte or more of data as big data. Small datasets, though do not need much storage space, are deliberated as big data because they are principally complex in nature for analysis, and applications. Fifteen years ago, Gartner devised the "three V's" to describe the incipient big data revolution. These three Vs are considered as characteristics of big data (shown in Figure 2.2).

Figure 2.2: Three V's of Big Data



Source: ICFAI Research Center

1. **Volume:** This defines the size of data used for analytics. Normally data in MBs and GBs are not considered as big data. To be termed as big, the data should be in terabytes, petabytes and exabytes. Volume of the data is also measured by the number of attributes and dimensions used for prediction and also depending upon how many instances of such dimensions are available, e.g., number of transactions, events occurred or historic data available. Normally, analytics uses small datasets called "Training Data" to create the predictive/classification models. These models are used in future to predict the new or strange data. These predictions may go wrong due to restrictions in the size of training sets used. By eliminating the data volume constriction and using larger data sets, enterprises can ascertain indirect patterns that can lead to targeted micro-decisions or upsurge the exactitude of the predictive models.

2. **Variety:** Variety designates diverse formats of data that do not lend themselves to be stored in regular Relational Database Management System (RDBMS). These comprise an extended list of data such as emails, documents, social media text messages, still images, graphs, audio, video, machine-generated output data from sensors, RFID tags, cell phone GPS signals, machine logs, DNA analysis devices and many more. These types of data are categorized as semi-structured or unstructured. In fact, normally 90% or more data are of these formats in any organization.

Variety is also used to describe data that can be acquired from various sources, both from outside and from inside of the company, which can be used to yield new and appreciated insights, which were not previously available.

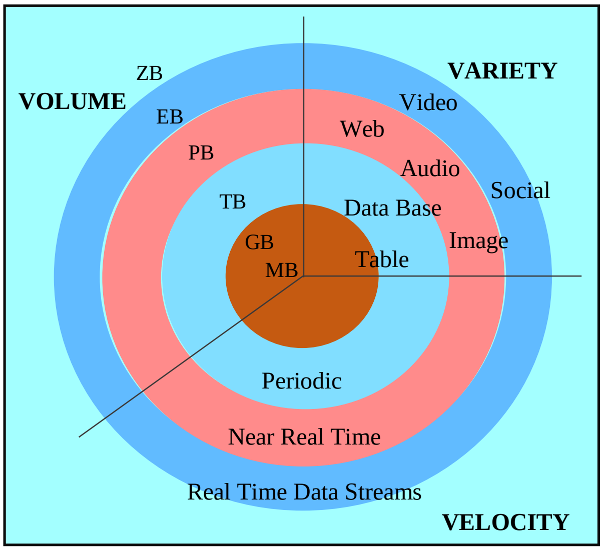
There is a connection between variety and volume. Unstructured data is increasing faster than structured data. According to Gartner's estimate unstructured data doubles every 3 months.

3. **Velocity:** There are two types of data used for analysis. They are data-in-motion and data-at-rest. Velocity describes what type of data it is. Data-in-motion, for example, is the data acquired from a sensor (in motion) or the web log history acquired from a web site about page visits and clicks by each visitor (at rest). Completeness and consistency of fast-moving streams of data are of concern. At the same time, matching them to specific outcome of events, based on their variety is also considered. It incorporates the latency or timeliness of data being seized at a rate or with a lag time that makes it beneficial.

   Velocity is also used to describe how long the data remain valuable/useful for analysis. Is it perpetual or relevant only for a particular point of time? The next dimension of velocity describes the speed at which the data must be retrieved and stored. It specifies whether the business needs near real responses or real time responses. This may mean that data is processed as streams to make rapid, real-time decisions or it may be that weekly/monthly batch processing is done to produce more judicious decisions.

   The characteristic of big data is depicted in Figure 2.3.

Figure 2.3: Characteristics of Big Data



Source: ICFAI Research Center

Apart from these characteristics, there are four other characteristics that have popped up in the literature occasionally. These characteristics share the definitional limits of the 4th V - 'Value' of big data used.

Veracity: What is the background of the data used? Does it come from a consistent source? Does it have the accuracy and the completeness of data.

Variability: There are numerous possible meanings for variability. Is the data consistent in terms of accessibility or interval of reporting? Does it exactly represent the event reported? Whether the data contain many extreme values or outliers? Whether the data determine what to do with these values and whether they comprise a new and significant signal or do they present noisy data?

Viscosity: This is used to pronounce the lag time or latency in the data comparative to the event described.

Virality: It describes the rate at which the data spreads or how frequently it is picked up by other users or events.

Handling the four Vs helps organizations excerpt the value of big data. The three utilities of the 4 V's are the followings:

1. **Informed intuition:** Envisaging possible future happenings and deciding course of actions which could be successful.
2. **Intelligence:** Observing the happenings in real time and shaping the actions to be taken.
3. **Insight:** Appraising things that have happened and defining the actions to eradicate the loss. Of late big data characteristics are discussed based on five factors or 5V's.

Twitter uses Big Data Analytics on Real Time Data to Detect Fake News

Twitter was dealing with massive volume of data from user activity. This data represented both "volume" "velocity". In 2020 Twitter used analytics on steaming data to identify, classify and take appropriate action on fake news

Source: (99+) The 8 Best Examples of Real-Time Data Analytics | LinkedIn, 08-Jun-2021 & https://bernardmarr.com/the-8-best-examples-of-real-time-data-analytics/, 28-May-2021, Accessed on 06/08/2022

Activity 2.1

Ringer Retails, a retail grocery shop chain across Hyderabad and Secunderabad wants to make use of data generated from its everyday transactions to understand which products sells better, which brand has more demand, inventory stocking levels, customers' spending habits and so on. Explore the characteristics of big data set to be generated by the retailer and explain the reasons.

Answer:

Check Your Progress - 1

1. What is ‘Volume’?

   a. It defines the size of data

   b. It describes the format of data used

   c. It describes the quality of data.

   d. It describes the speed of data retrieval

   e. It describes the accuracy, completeness and how fast it spreads

2. What is meant by ‘Variety’?

   a. It defines the size of Data

   b. It describes the format of data used or from where the data is obtained

   c. It describes the quality of data

   d. It describes the speed of data retrieval

   e. It describes the accuracy, completeness and how fast it spreads

3. What is meant by ‘Value’?

   a. It defines the size of data

   b. It describes the format of data used

   c. It describes the quantity of data

   d. It describes the speed of data retrieval

   e. It describes the accuracy, completeness of data and how fast the data spreads

4. What do you mean by ‘Business Intelligence’?

   a. It provides ad hoc and/or standard business reports based on analytics, a reactive strategy

   b. Reporting tweaks from vast data sets

   c. Creating proactive, forward-looking resolutions using predictive modeling, forecasting and statistical analysis

   d. Mining information from traditional data stores

   e. Mining pertinent information from huge data stores/other resources and analyzing it to improve future business decisions, a proactive strategy

5. What do you mean by 'Big Data Analytics'?
   a. It provides ad hoc and/or standard business reports based on analytics
   b. Reporting tweaks from vast data sets
   c. Creating proactive, forward-looking resolutions using predictive modeling, forecasting and statistical analysis
   d. Mining information from traditional data stores
   e. Mining pertinent information from huge data stores/other resources and analyzing it to improve future business decisions, a proactive strategy

2.5 A Wider Variety of Data

The variety of data sources used for data acquisition in organizations continues to increase. Conventionally, data for analytical processing could be acquired from internal information systems, like CRM (Customer Relationship Management), MIS (Management information system), ERP (Enterprise Resource Planning) and EIS (Executive Information System) applications. However, to improve the knowledge and awareness and take better decisions, the data could be taken from wide variety of other information sources like:

1. Internet data (i.e., web logs, social media, clickstream and social networking links)
2. Primary research data (i.e., market surveys, observations, experiments, feedback analysis of customers)
3. Secondary research data (i.e., market researches, competitive and marketplace data, consumer data, business data)
4. Data about Locality (i.e., data from GPS, mobile device data, geospatial data)
5. Image & Video data (i.e., video, surveillance, satellite image,)
6. Data from the Supply Chain (i.e., vendor catalogs and pricing, EDI (Electronic Data Interchange) from the supply chain, quality information)
7. Device data (i.e., Wireless sensors, radio frequency devices, programmable logic controllers, telemetry)

The wide variety of data leads to complications in consuming the data into data storage. It also complicates the transformation and processing of the data using traditional analytical approaches.

Example: UPS (the Logistics Major) Deploys IOT Sensor Based Big Data Analytics Solution to Cut Down Delivery Times to Less than 3 Days for 90 Percent Customers

UPS had always been investing in technology solutions to serve its customers even during the recent pandemic. With all the constraints, the technology solutions could enable the logistics company to maintain the pre pandemic delivery timelines.

Contd….

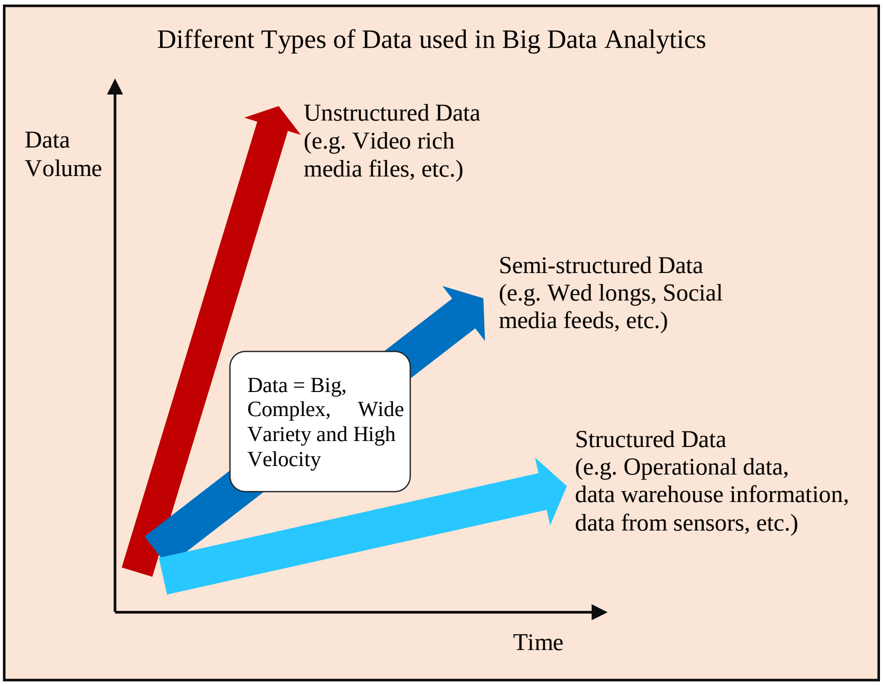
The company made the hubs as data analytics centres whereby nearly 2 million packages were tracked using a network of IOT sensors. The whole exercise resulted in bringing down the delivery time to less than 3 days for majority of deliveries.

Source: https://www.aidataanalytics.network/data-science-ai/articles/digital-transformation-and-business-resilience-in-action-a-look-into-upss-visionary-approach-to-automation-ai-and-analytics, 01/08/2021, Accessed on 06/08/2022

2.6 Different Types of Data used in Big Data Analytics

There are three different types of data used in big data analytics as depicted in Figure 2.4. The figure illustrates the growth of data in volume related to time.

Figure 2.4: Different Types of Data used in Big Data Analytics


Different Types of Data used in Big Data Analytics
Data Volume
Unstructured Data (e.g. Video rich media files, etc.)
Semi-structured Data (e.g. Wed longs, Social media feeds, etc.)
Data = Big, Complex, Wide Variety and High Velocity
Structured Data (e.g. Operational data, data warehouse information, data from sensors, etc.)
Time

Source: ICFAI Research Center

1. **Structured data:** This type of data can be stored and managed by relational data base management systems which could be represented in row and column format. The data consistency allows it to retort to simple queries to attain usable information, based on operational needs of the organization. Example includes organization’s transactional data which can be stored in tables and databases.

2. **Semi-structured data:** This type of data does not adapt to fixed or explicit schema like row-column format. The data is naturally self-describing and comprises of tags or other markers to impose hierarchies within the data. Examples include social media feeds and weblogs.

3. **Unstructured data:** This type of data cannot be stored into relational tables or databases for analysis or querying. This data could take more space to store and cannot be confined with fixed size. Examples include images, graphs, audio and video files.

Example: Airbnb Uses Social Media Analytics to Give Innovative Offers to Customers

Airbnb, the disruptor in the hospitality industry was severely affected as travel and stays were restrictive. But the company used Big Data analytics to process unstructured and semi structured information generated by users on the Social Media platforms of the Airbnb brand. They noticed the customer preferences and trends and came up with campaigns like #GoNear. This campaign encouraged local travel within 300 miles. This has opened new opportunities for the company to utilize its properties and generate revenues

Source: https://sproutsocial.com/insights/social-spotlight-airbnb/, November 16, 2020, Accessed on 07/08/2022.

2.7 The Expanding Universe of Unstructured Data

Unstructured data normally does not fit into any explicit structure or relational data model. It is more complex. Nevertheless we cannot omit unstructured data, since it has many takeaways. Some of them are as follows:

1. According to Gartner’s estimate, unstructured data doubles every three months.
2. Most of the new data is unstructured, it is about 95% of the total data used for analytics and only 5% of data is structured.
3. Unstructured data grows exponentially, but structured data grows in linear manner (Figure 2.4).
4. Unstructured data is massively underutilized.
5. Unstructured data is not totally waste or noisy. It has its value. We need to find the technological partner to extract the value from the unstructured data. i.e., we need to find the mechanism to find the signal out of the noisy information.

Example: Bank of America uses Big Data Analytics to get Actionable Insights from Customer Posts on Social Media

Bank of America used Big Data analytics to analyse customer social media posts data to identify service problems of customers before they could damage the bank reputation.

Contd….

The bank used Data mining to analyze around 40,000 social media comments of customers. Large part of the comments pointed to unverified rumors about purchase limits. This could have led to customer loss. Since they found such things early, the damage could be contained

Source: https://www.techerati.com/features-hub/opinions/how-big-data-has-revolutionised-the-finance-industry/, 1st July, 2021, Accessed on 07/08/2022.

2.8 Big Data Analytics

The upward demands for data volume, variety and velocity to analyze the data have placed accumulative demands on software technologies. The demand is on the rise for computing platforms as well, to tackle and manage the competitiveness of the business organizations in the global marketplace.

From a business perspective, we need to learn how to:

- Use big data analytics to get competitive advantage for the business enterprise and use the core competencies.
- Get benefit from new technology capabilities and also utilize existing technology assets.
- Facilitate the applicable organizational change to obtain fact-based decisions.
- Provide earlier and superior results by approving and capitalizing on the new technologies occurring in the global market place.

Following Exhibit 2.2 talks about the importance of big data.

Exhibit 2.2: Big Data Analytics – Transportation application

Transportation

Maps to apps. That's the nutshell version of how navigation has been transformed by technology, with the vast majority of smartphone users relying on their devices for directions. And those directions are courtesy of big data — relevant information (on traffic patterns, for example) gleaned from government agencies, satellite images and other sources.

But big data doesn't just affect how people move, it affects how everything moves — including packages, planes and cars. Packages have tracking numbers (data!). Planes analyze data to (among many other things) increase fuel efficiency and predict maintenance issues. And cars, via onboard sensors and IoT connectivity, collect and transmit so much data that the autonomous driving revolution might be closer than we think.

Here are some examples of big data in motion.

Source: https://builtin.com/big-data/big-data-examples-applications. July 3, 2019-Updated: April 6, 2020

Big data analytics uses wide variety of techniques (as listed in Figure 2.5).

Figure 2.5: Three Phases in Big Data

Phase-I	Phase-II	Phase-III
DBMS-based, structured content: 1. RDBMS & Data Warehousing 2. Extract Transfer Load 3. Online Analytical Processing 4. Dashboards & Scorecards 5. Data Mining & Statistical Analysis	Web based, unstructured content 1. Information retrieval and extraction 2. Opinion mining 3. Question answering 4. Web analytics and web intelligence 5. Social media analytics 6. Social network analysis 7. Spatial-temporal analysis	Mobile and sensor based content 1. Location-aware analysis 2. Person-centered analysis 3. Context-relevant analysis 4. Mobile visualization 5. Human-computer interaction

Source: Banu, & Yakub, Md. (2020, October). EVOLUTION OF BIG DATA AND TOOLS FOR BIG DATA ANALYTICS. ResearchGate. Retrieved December 13, 2022, from https://www.researchgate.net/publication/345573305_EVOLUTION_OF_BIG_DATA_AND_TOOLS_FOR_BIG_DATA_ANALYTICS

These variety of techniques enable:

1. **Deeper insights.** It is used to provide insights into 'ALL' as a whole or a category – all the individuals, all the events, all the products, all the transactions, etc., instead of looking at classes, segments or regions.
2. **Broader insights.** In globally connected economy, it is difficult to operate the business within constantly changing environments. Our conventional plans may go wrong because of estimation or approximation. Big data analytics considers all the data even from new data sources, to understand the intricate, changing, and interrelated circumstances to produce more precise insights.
3. **Frictionless actions.** Increased consistency and accuracy will permit deeper and broader insights to be automated into efficient actions.

2.9 Setting the Tone at the Top

Dr. Usama Fayyad, one of the great minds in big data analytics, explains how big data analytics is different from traditional analytics by giving an astronomy example. Just as an astronomer needs best competent telescope to get a deeper insight into a star or a galaxy, the business organizations need to use the big data analytics to get deeper and broader insights to solve business problems. For that, it needs to pick the information of interest in dynamic way (whereas traditional analytical approaches analyze with static information). The information of interest may change over time, or depending on the problem it may be surrounded with noisy data. Big data analytics could find it from the gigantic collection of data and could utilize it for solving solutions. Finding the key attributes from the collection of data to draw certain conclusions is the specialty of big data analytics.

Activity 2.2

Metro cities nowadays use surveillance cameras to track the residents. These cameras are positioned at railway station, bus stand, road crossings and other public places. In the process it generates tones of images, videos, call records and so on. What different types of data can you identify in the big data generated by the surveillance cameras?

Answer:

Check Your Progress - 2

6. What do you mean by 'Big Analytics'?
   a. It provides ad hoc and/or standard business reports based on analytics, a reactive strategy
   b. Reporting tweaks from vast data sets
   c. Creating proactive, forward-looking resolutions using predictive modeling, forecasting and statistical analysis
   d. Mining information from traditional data stores
   e. Mining pertinent information from huge data stores/other resources and analyzing it to improve future business decisions, a proactive strategy

7. In three I's, 'Intelligence' refers to which of the following?
   a. Planning for future
   b. Envisaging possible future happenings and deciding course of actions
   c. Taking appropriate action for past happenings
   d. Observing the happenings in real time and shaping the actions to be taken
   e. Appraising things that have happened and defining the actions to eradicate the loss

8. In three I's, 'Informed Intuition' refers to which of the following?
   a. Planning for present
   b. Envisaging possible future happenings and deciding course of actions
   c. Taking appropriate action for past happenings
   d. Observing happenings in real time and shaping the actions to be taken
   e. Appraising things that have happened and defining the actions to eradicate the loss

9. Which of the following is the example for ‘Unstructured Data’?
   a. Employee Table
   b. Sensor data
   c. Google
   d. Video obtained from ‘Monitoring’ system of shop floor
   e. Web log about number of products clicked on your e-commerce web-site
10. Which of the following is the example for ‘Semi-structured Data’?
   a. Employee Table
   b. Sensor data
   c. Google
   d. Video obtained from ‘Monitoring’ system of shop floor
   e. Web log file containing number of products clicked by the customers on your e-commerce web-site

2.10 Summary

- Explanation for why big data analytics is important by describing the different reactive and proactive analytical approaches used in organizations to take appropriate actions.
- Clarification on big data by intensifying the details of its characteristics – the 4 ‘Vs and also how these 4 ‘Vs are used to get three I’s in the business organizations.
- Discussion on the wider variety of data used in big data analytics and the different types of data used for big data analytics.
- More details on ‘unstructured data’ which occupies more volume of data needed for analytics.
- The need for big data analytics and variety of techniques used in big data analytics are briefed.
- Conclusion by stating that big data analytics is not analyzing the entire data (by differentiating it from traditional analytics), but finding the interesting patterns using key attributes which can be used to draw certain significant conclusions.

2.11 Glossary

3 Vs of Big Data: Volume, Variety and Velocity.

Big Data Analytics: Mining pertinent information from huge data stores/other resources and analyze it to improve future business decisions, a proactive strategy.

Big Data: Enormous amount of data which grow by business transactions (which have more unstructured data) and require advance techniques and approaches to manage and process it.

Business Intelligence: It provides ad hoc and/or standard business reports based on traditional analytics, a reactive strategy.

Data Base (Management) System: The database (management) system is a computer application, which interacts with the users and other applications, to store, manage and retrieve the structured data.

Data Stream: It is a sequence of digitally encoded coherent signals (data packets) used to transmit or receive information that is in the process of being transmitted.

EB: An EB or Exabyte is actually 1,152,921,504,606,846,976 bytes (2^{60}) bytes, as the measurement of internal computer memory is based on a base 2, or binary, number system.

GB: A gigabyte (GB) is a measure of computer data storage capacity that is roughly equivalent to 1 billion bytes. A gigabyte is two to the 30th power or 1,073,741,824 bytes (2^{30} B).

MB: A megabyte is a measure of computer data storage capacity that is roughly equivalent to 1048576bytes (2^{20} B).

Real-time Data: Real-time data denotes information that is delivered immediately after collection. There is no delay in the timeliness of the information provided. Real-time data is often used for navigation or tracking.

Semi-Structured Data: Data is a form of structured data that does not conform to the formal structure of data models associated with relational databases, but however contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.

Structured Data: Data that could be stored in row-column format and which can be of any specific data type which has limited size (e.g. Integer, float, etc.). It could be readily searchable by simple, straightforward search engine algorithms or other search operations.

Unstructured Data: This refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

Web Log File: Web log file is a file that records events that occur in a website.

ZB: A zettabyte (ZB) is a unit of digital information storage used to denote the size of data. It is equivalent to 1024 exabytes.

2.12 Self-Assessment Test

1. Why is ‘big data’ important? Explain.
2. Explain in detail the characteristics of ‘Big Data’.
3. List out the variety of data used in big data analytics and give examples.

4. Differentiate between ‘Unstructured’ ‘Semi-structured’ and ‘Structured Data’.

5. Explain the different types of data used for big data analytics.

6. What are the varieties of techniques used in big data analytics?

2.13 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.

2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.

3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.

5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition

6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

2.14 Answers to Check Your Progress Questions

1. (a) It defines the size of data.

This defines the size of data used for analytics. Normally data in MBs and GBs are not considered as big data. To be termed as big, the data should be in terabytes, petabytes and exabytes.

2. (b) It describes the format of data used or from where the data obtained

Variety designates diverse formats of data that do not lend themselves to be stored in regular Relational Database Management System (RDBMS).

Variety is also used to describe data that can be acquired from various sources, both from outside and from inside of the company, which can be used to yield new and appreciated insights, which were not previously available.

3. (e) **It describes the accuracy, completeness of data and the speed of data**

Characteristics that share the definitional limits of the 4th V - 'Value' of big data used.

- Veracity: What is the background of the data used? Does it come from a consistent source? Does it have the accuracy and the completeness of data.
- Variability: There are numerous possible meanings for variability. Is the data consistent in terms of accessibility or interval of reporting?
- Viscosity: This is used to pronounce the lag time or latency in the data comparative to the event described.
- Virality: It describes the rate at which the data spreads or how frequently it is picked up by other users or events.

Handling the four Vs helps organizations excerpt the value of big data.

4. (a) **It provides ad hoc and/or standard business reports based on analytics, a reactive strategy**

Business Intelligence (BI) provides ad hoc reports, standard business reports, alerts and notifications based on analytics, in the reactive category.

5. (e) **Mining pertinent information from huge data stores/other resources and analyze it to improve future business decisions, a proactive strategy**

Using big data analytics, we can mine only the pertinent information from data stored in TBs (terabytes), PBs (petabytes) and EBs (exabytes), and analyze it to improve business decisions for the future. It is proactive, with sound knowledge and deep insight and can be used to meet the future requirements.

6. (c) **Creating proactive, forward-looking resolutions using predictive modeling, forecasting and statistical analysis**

Creating proactive, forward-looking resolutions requires proactive big analytics like predictive modeling, optimization, forecasting and statistical analysis. These allow us to identify spot weaknesses and trends, or define conditions for making decisions. This might be proactive. But it cannot be performed on big data because traditional storage and processing are not sufficient for big data analytics.

7. (d) **Observing the happenings in real time and shaping the actions to be taken.**

8. (b) **Envisaging possible future happenings and deciding course of actions.**

9. (d) **Video obtained from 'Monitoring' system of shop floor**

   This type of data cannot be stored into relational tables or databases for analysis or querying. This data could take more space to store and cannot be confined with fixed size. Examples include images, graphs, audio and video files.

10. (e) **Web log file containing number of products clicked by the customers on your e-commerce web-site**

    This type of data does not adapt to fixed or explicit schema like row-column format. The data is naturally self-describing and comprises tags or other markers to impose hierarchies within the data. Examples include social media feeds and weblogs.

Unit 3

Big Data in Marketing and Advertising

Structure

"Information is the oil of the 21st century, and analytics is the combustion engine"

- Peter Sondergaard, Senior Vice President, Gartner

3.1 Introduction

In the previous unit, we realized the importance of big data and also got introduced to big data analytics. Technological advances have brought great changes in the entire world in the past decades. It is the era of instant communication and big data. Every aspect of life, including product consumption and purchase behavior, reflects these advances. Marketing managers always used data to create their marketing campaigns. In the past, the amount of data was limited because collecting the data was very difficult. But with the advent of the internet and other technologies, the scenario has changed. Now market managers and marketers are using big data, collecting it from numerous online channels and other ways. Marketers have found a way to boost their marketing insights, from social media sites, forums, blogs, reviews and many more.

In the past decades, marketers collected data with a lot of difficulty. Such data were utilized to create advertisements to target broad demographics. In the present situation, big data platforms are helping companies to manage and analyze their gigantic data, to create advertisements directed towards individuals at various segments. Previously, evaluating the expected effect and quality of marketing campaigns was cumbersome. Today, it can be easily done using big data simulations in a virtual marketplace. It greatly reduces failure costs involved in marketing and advertisement.

Big data helps marketers to understand the past and present behaviors of customers without difficulty. In addition, it helps to predict the accurate need, preferences, and demands of the customer, easily. Currently, both large and small companies play on the same field, both having access to big data and its associated technologies. The ability to retain and analyze enormous amounts of unstructured and structured data is helping digital advertisers. Now they are able to ascertain new relationships, spot budding trends and patterns and achieve actionable insights that provide a competitive advantage.

This unit explores more about big data by viewing its applications in marketing and advertising areas of business.

3.2 Objectives

By the end of this unit, you should be able to:

- Explain more insights of big data
- Relate how big data can be used in marketing and advertisement to improve business

3.3 Digital Marketing and the Non-Line World

Digital Marketing - Digital marketing refers to the collective term used for various channels of promotion through on-line platforms. This concept is becoming increasingly relevant and important, with more and more organizations beginning to have online presence, either through their own websites or through dedicated company pages on social media networks. The way organizations communicate with customers and the way customers interact with organizations have changed drastically with the advent of social media networks. Hence, a coherent strategy involving both offline and online marketing has to be used by businesses, to interact with their customers and understand their needs.

Non-line World - Prior to the advent of internet, companies operated with only offline marketing. With the rapid spread of internet, they turned to online/ digital marketing. However, companies need to realize that it is not offline or online commerce that encompasses marketing, but a combination of both the forms is the ideal approach to marketing. Non-line world, the term coined by

Avinash Kaushik (author, and Digital Marketing evangelist for Google) emphasizes that it is neither the online world nor the offline world that organizations need to operate, but it is the non-line world that they must aspire to operate, and eventually move to. Customers often move between online and offline worlds, for example, mobile apps are used to download or get information. Similarly, bar code scanning could be used to upload information from offline to online world.

According to Google, two factors are having an effect on offline and online shopping: 1. Consumers feeling more comfortable with online shopping, and 2. Their reduced fear of online frauds.

Nowadays, consumers are progressively going to a physical store to become more accomplished about a product – by touching it, playing with it and seeing how it works actually. Then they go online to compare its price with other shops. Either they purchase it online or go back to the store and make their purchase there. This leads to non-line shopping or the non-line world. It presents a new prospect for digital marketers who are already visualizing a world in which the offline and online mix are interchangeably used. That world is likely to include some form of "non-line"

Google's digital marketing evangelist explained that big data warehouse tactics do not work well in the online world, and to be successful online, one needs to embrace multiplicity. Here multiplicity refers to requirement of multiple skills in the decision-making team, multiple types of data (consumer data, click stream data, competitive intelligence data, etc.) and multiple tools. Many companies fail to make smart decisions in the online world because they cannot embrace multiplicity.

To make good decisions in online world, you should know how to use variety of tools to fetch multiple types of data together at right time and make high speed decisions. Digital marketing has an impact of what happens in the "offline" real world, together with what happens in the online world. This is because of the existence of "non-line" world as mentioned above. The customers move sinuously between online & offline worlds. When a customer is in the offline world, he uses his mobile phone or laptop to access information from the online world and vice versa. Nowadays, most of the newspaper publishers and other e-commerce websites struggle to select the contents which might be of interest to their customers. The available analytical tools work based on the current data. This is a drawback.

Don't Abdicate Relationships

Nowadays companies are collecting and assessing data about their customers through different social media platforms like Facebook. Hitherto, the companies were connected with the customers only through the retailers and there was no direct relationship with the customer. Now any company can have a direct relationship with the customers, no matter how far they are from a customer. Direct relationship with the customer enables the firm to take critical business decisions (such as new product development) much easier and faster.

Is IT Losing Control of Web Analytics?

Nowadays most of the industries are losing their customer data by sending the customer details to large consultants for processing. The purpose of these consultants is to collect, create and crush data (Puking). Big data available for a company will completely renovate its ability to understand the efficiency of its marketing. Companies can also hold employees responsible for the millions of dollars being spent. Big data helps companies understand their competitor's behavior.

Example: Walgreens's Digital Marketing Strategy Recognizes that Customers Operate in the "Non - Line World"

Walgreens has a digital marketing strategy in place which assumes that customer experience has to be seamless across both in-store and online. In fact, the company believes customers who use in-store and mobile channels are 6x times more valuable than the customers who use only in-store. The idea is to give seamless exceptional customer experience which ever channel the customer chooses.

Source: https://www.globalconveniencestorefocus.co.uk/features/walgreens-loves-loyalty/, 9th June 2020, Accessed on 08/08/2022

3.4 Database Marketers, Pioneers of Big Data

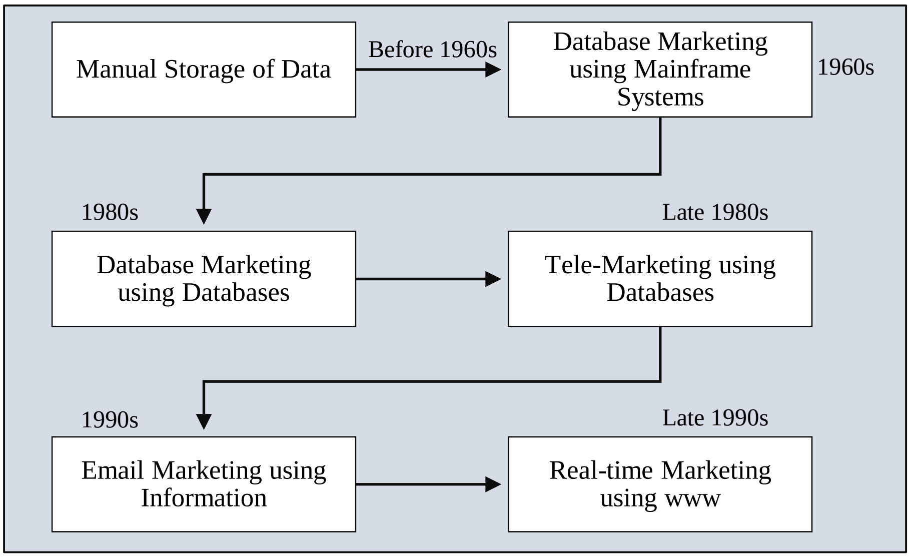
Database marketing deals with: 1. Storing the details of individual customers, 2. Using that information to understand customer better and 3. Communicating with some of those customers effectively, to drive business value. By 1960s, database marketing started by storing the customer information in mainframe systems and using them to communicate with the customers. Later information about customers was stored in multiple systems due to storage and other constraints. Collecting the details from different systems became difficult. Then companies started developing software that could remove duplicate customer information from the systems (de-duping).

In 1980s, database concepts were introduced and databases have since been used to store the customer details making it simple and easy to access the customer details. Reports driven from the databases gave deeper insight into purchasing preferences and buying habits of the customers. This enabled the marketers to decide which kind of direct mail marketing campaigns produced the most responses and which customer segments were more likely to respond. Later, telemarketing was introduced to use these customer databases to make direct calls to customers through call centers, for promotional or sales activities.

In 1990s, when email entered the picture, the marketers started using Internet and World Wide Web to improve their marketing. Once the dot-com boom hastened, marketers adopted new technologies to start the company websites. This led to "real-time" marketing. Marketing companies started collecting more and more details of the customers. Different software companies started developing solutions. Later, most of the companies developed these solutions in-house. The main benefit of in-house approach is that it enabled them to receive lot of different data from different places. They started to see detailed transactional data; detailed data on a product, e.g., in banks, detailed data about distinct transactions within bank accounts were seen. Companies started collecting information from customer care centers. These data created a better picture of the individual, besides the mix of products and services they purchased.

In the late 1990s, with the emergence of marketing automation software, proactive communication became the hot trend. It enabled communication at right time called "time around communication". As technology progressed to fascinate greater volumes of data, the costs of data environments came down, and companies started collecting more transactional data. Entire growth in database marketing (shown in Figure 3.1) is the cause for big data generation, which led to big data analytics.

Figure 3.1: Growth in Database Marketing


Manual Storage of Data
Before 1960s
Database Marketing using Mainframe Systems
1960s
1980s
Late 1980s
Database Marketing using Databases
Tele-Marketing using Databases
1990s
Late 1990s
Email Marketing using Information
Real-time Marketing using www

Source: ICFAI Resource Center

Today, most of the companies have the ability to store and analyze data gathered from every exploration on their websites. By combining that exact data from their internal sources with unidentified data from external sources, companies can predict customer's behavior with surprising accuracy.

3.4.1 Key Techniques in Database Marketing

Find below various techniques in database marketing.

1. **Classification Techniques** – logistic regression and classification trees (classifying customers into defined groups).
   - From the prospective universe, customers who are likely to respond to marketing communication. (Target them first after analyzing them for potential profitability).
   - Customers likely to commit fraud, and type of loan applicants who are likely to default.
   - Customers expected to leave the service or switch the product (Make attempts by offering exciting deals and discounts to retain them).
2. **Prediction Techniques** – Similar to classification, but rather than simply classifying into a group prediction, it aims to answer precisely a particular value.
3. **Association Rules** – By analyzing large databases of customer transactions, marketers can infer what products are purchased together; this could be very well used in arranging products in a physical store or a super market. E-commerce companies could use this to give suggestions for a customer searching or purchasing a particular product. In short, 'association rules' help identify what goes with what.
4. **Data Visualization** – Visualizing data helps to grasp and comprehend huge information very quickly. Histograms, pie charts, scatter plots, correlation matrices, and box plots etc. are some examples of commonly used data visualization techniques.

Example: Walmart Uses Data Mining with Association Rules to Figure out Seven-Fold Increase in Sales of Strawberry Pop-Tarts Before Hurricanes

Walmart, the retail major uses data mining on a big way to understand customer profiles and preferences and rearranged its shelfs accordingly both in store and online. In one study, the retailer used data mining using association rules and found out that the sales of strawberry pop-tarts went up by seven times before a hurricane. With this insight, the company stocked more units of that specific pop-tart across the outlets along the hurricane path.

Source: https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109, 9th June 2020, Accessed on 08/08/2022

3.5 Big Data and the New School of Marketing

New school of marketing defines the way customers to have changed their behavior of collecting their requirements. The customers decide which marketing messages they can receive, when, from where, and from whom. They desire interactive communication across various digital power channels like: email, social media, mobile, and the web.

Consumers Have Changed. So Must Marketers

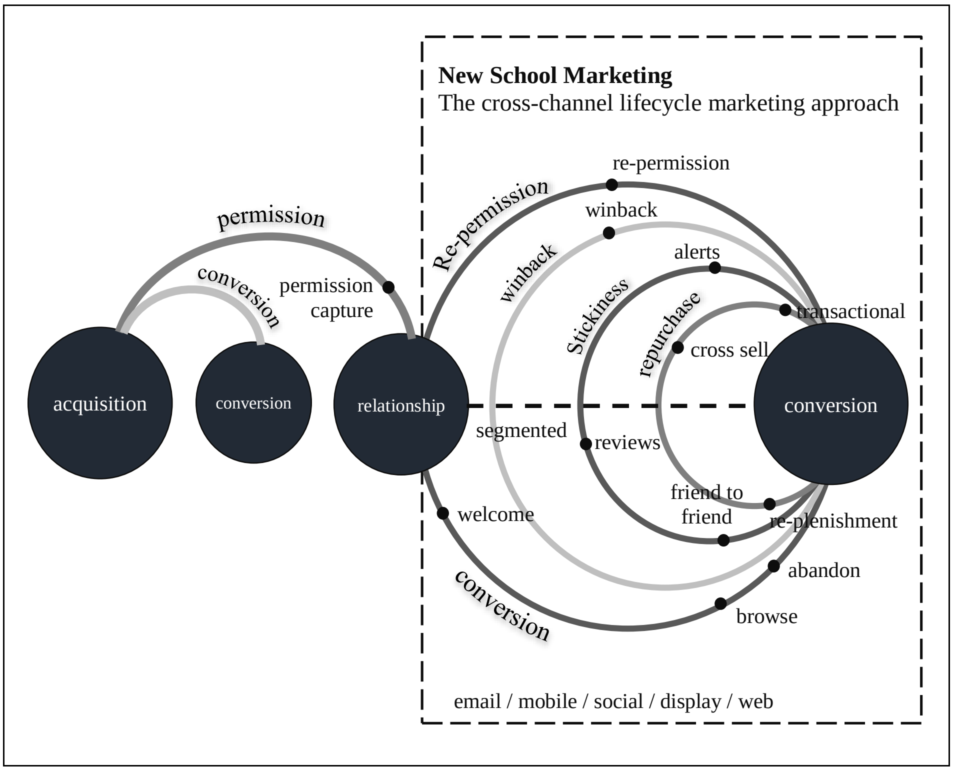
Lifestyle of customers has changed. They are picking information from cross-channels. They are at various moods at a given moment. So the marketer has to change themselves to adopt an appropriate marketing strategy at any moment. They should be able to deliver the information at multiple channels in time. In addition, they need to study the customer's holistic data profiles, including their preference data and channel response, area of influence/ social footprint, etc.

Consumers have changed in the way they perceive traditional form of advertising. TV commercials, mass emails and advertisements on newspaper may not be as effective as they once used to be. Consumers have shifted to digital lifestyle and the marketing messages have to be tailored to suit them. Consumers became informed and are unpredictable than ever. They need relevant and interactive communication across digital channels, emails, mobile messages, social networks, and websites. Marketing and advertising managers must realize this change and embrace new and innovative methods to target consumers and draw their attention towards products and services, offered.

Example **-** An online newspaper publisher could collect and analyze data on the amount of time spent by viewers on each of the sections and the type of news that is usually read. This helps them to design the news display in a better fashion on their website with a mix of news from various fields that are usually preferred by its customers. This is displayed prominently in the front page of the website and rest of the sections are appropriately placed.

Cross-Channel Lifecycle Marketing starts with the approval of the customer, followed by call information and then preferences of customer for multiple channels. In this approach, marketers need to have the customer information systems and applicable integrated marketing. This helps in having absolute understanding of customers through affirmed preferences and practical behavior. They can automate and optimize their processes and programs. Figure 3.2 depicts various loops involved in the Cross-Channel Lifecycle Marketing approach between relationship and conversion, namely, stickiness, win-back, repurchase, and re-permission.

Figure 3.2. New School of Marketing



Source: Minelli, M., Chambers, M. and Dhiraj, A. Big Data: Big Analytics. Wiley India, 2015

Social and Affiliate Marketing

Word-of-mouth marketing was the most powerful marketing before introduction of the internet. In 1996, William J. Tobin, the founder of PC Flowers & Gifts introduced the concept of affiliate marketing. This is also referred to as performance marketing on the Internet. In this approach, the affiliate brings business to the company and is rewarded for that. Industry analysts estimate affiliate marketing as $3 billion. Affiliates operate through behind the screen channels, which most of the customers are unaware of. In 2012, after vast usage of social media networks, the barriers of affiliate marketing were totally removed. This has brought unimaginable changes in the marketing world.

Empowering Marketing with Social Intelligence

Increasing and extensive use of social media led to the user-generated content, which is "big data". Enormous data is created and the volume is expected to increase exponentially in a continuous manner. Millions of status updates, blog posts, photographs and videos are generated every second among social network users. Successful organizations need to carve out the appropriate information for their company's business. They need to scrutinize and, understand it on a constant basis, and in real time to predict the likely future behavior of customer.

Example: ASOS (The UK Based ecommerce fashion brand) Uses "Social Intelligence" to Understand the Different Preferences of Customers in UK and USA and Plan Marketing Accordingly

The UK based fashion brand has major operations in UK and USA. It is active on social media with the same marketing strategy for both the markets. But it wanted to understand the differences better and plan suitable marketing plans for them. It has used "social intelligence" technique (part of new school of marketing) for this. It found UK customers, mostly students, active on social media later in the day. The US customers were more interested in celebrities and events. So separate strategies were deployed for social media in the two markets.

Source: https://lmd.lk/social-intelligence-2/ , 31st March, 2022, Accessed on 08/08/2022

Activity 3.1

You are hired as a marketing manager in a software enterprise. You understand the power of social media and how it holds the key for success to spread positive message or spread a rumor amongst masses. How do you intent to use big data to empower marketing with social intelligence for your software organization?

Answer:

Check Your Progress - 1

1. Gathering information online and off-line before making a purchase is known as which of the following?

    a. Online Shopping

    b. Offline Shopping

    c. Show-casing

    d. E-Commerce

    e. Non-line Shopping

2. Online Analytics involves which of the following?

    a. Consumer Behavior data

    b. Click Stream data

    c. Competitive intelligence data

    d. Offline information

    e. Consumer data, click stream data, competitive intelligence data, etc.

3. What is Social Media Marketing?
   a. Word-of-mouth Marketing
   b. Efforts to create content that attracts attention and encourages readers to share it across their social networks
   c. Affiliate Marketing
   d. Offline Marketing
   e. Marketing by making phone calls
4. What is meant by ‘Data Puking’?
   a. Presenting all details available about the customer
   b. Presenting online data transaction of the customer
   c. Collecting transactional data
   d. It is to collect, create, crush data and present the data for analytical purpose
   e. All the above
5. When did database Marketing started using mainframe systems?
   a. Before 1960s
   b. In 1960s
   c. In late 1990s
   d. In 1980s
   e. In 1990s
6. When did real-time marketing started using www?
   a. Before 1960s
   b. In 1960s
   c. In late 1990s
   d. In 1980s
   e. In 1990s

3.6 Fraud and Big Data

Fraud is a deliberate attempt made for personal advantage, or to cause damage to another Individual. Common and well-known forms of frauds are credit card fraud and insurance fraud. Social media and mobile phones are the frontlines for fraud. Most of the customers use their birthday information, their high school name, their phone number or their pet’s name as their personal information to verify identity. These details are available in their profile in the public websites, or the individuals share this information through social media. Big data technology provides an optimal technology solution to detect fraud based on three Vs, namely, high volume (customer records and transactions for years), high velocity (social media information and dynamic transactions), high variety (unstructured data such as call center conversations and customer emails, as well as structured data like transactional data). Different approaches are used to detect fraud using big data.

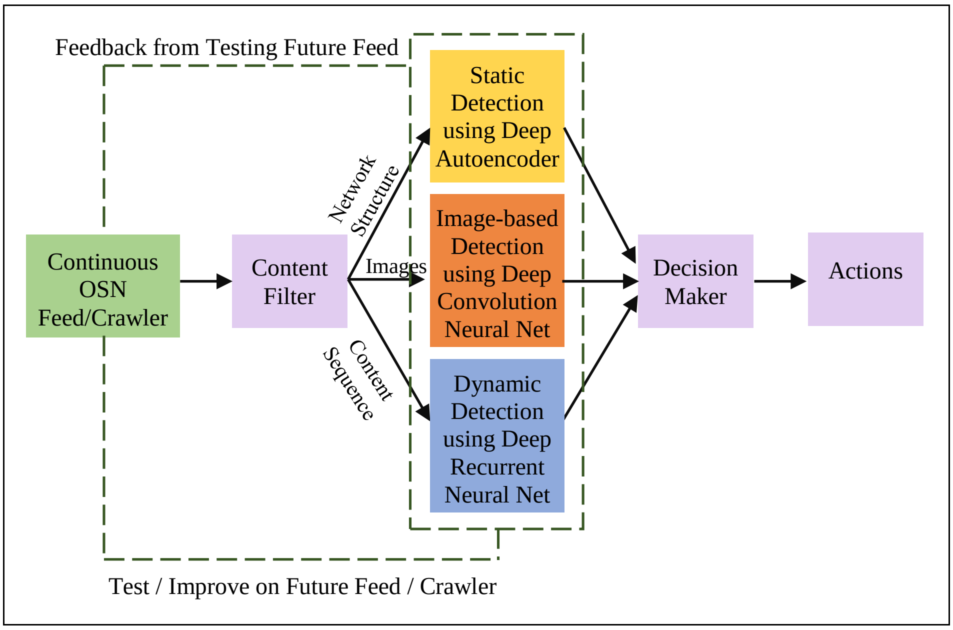
3.6.1 Elastic Search

It is an open-source tool. It is based on Apache Lucene. It can search any kind of documents at near real-time. You can use this tool to work on huge historical data-sets in conjunction with real-time data to pick the identified deviations in payment patterns. Once the transactional data has been processed, the defined query identifies new transactions both in unstructured and structured data that have up stretched profiles. This provides scalability to the event processing framework, and allows specific doubtful transactions to be enriched with added unstructured information. This ability can reduce false positives.

3.6.2 Social Network Analysis (SNA)

It is an extensive analysis of social media networks. SNA could expose every individual involved, from original criminals to their companions. It also helps to understand their relations and activities to identify a 'bust out'. SNA lets the company to proactively look through large amounts of data to show relationships via links and nodes. The SNA tool encompasses a mixture approach of analytical methods. This includes statistical methods, organizational business rules, network linkage analysis and pattern analysis, to reveal the large amount of data to find relationships via links. Link analysis is used to look for clusters and how those clusters link to other clusters. Civic records such as fore-closures, judgments, address change frequency, criminal records, and bankruptcies are all data sources that can be assimilated into a model. Using the mixture approach, the company can rate the transactions. If the rating is high, it indicates that the transaction is fraudulent. This process is depicted in Figure 3.3.

Figure 3.3: Fraud Detection using Social Network Analysis


Feedback from Testing Future Feed
Continuous OSN Feed/Crawler
Content Filter
Network Structure
Images
Content Sequence
Static Detection using Deep Autoencoder
Image-based Detection using Deep Convolution Neural Net
Dynamic Detection using Deep Recurrent Neural Net
Decision Maker
Actions
Test / Improve on Future Feed / Crawler

Source: https://tuansinung.net/wp-content/uploads/2017/02/Final_Exam_Diagram.png

3.6.3 Detecting and Preventing of Fraud

Data is collected from disparate sources in real time which is then aggregated and is processed in real time to identify potential threats. For example, credit card transactions are monitored in real time to identify patterns and inconsistencies in usage. Any inconsistency would set an alert that enables companies to act quickly and prevent potential threats.

Traditionally, the approach of detecting fraud has been to look out for factors like unusual logins from different places and times, besides bad IP addresses. Also, some of the companies have traditionally been using statistical analyses to identify fraudulent activity. Statistical techniques, however, use sampling of data and do not take the entire data for processing. This allows for one or more frauds going undetected, the cost of which may be very high. Identification of low incidence events (very rare events) raising an alert on suspicious activity or claim is generally used as a method for detecting fraud.

3.6.4 Big Data for Fraud Detection

With the advent of big data, this approach is changing and sophisticated solutions are being incorporated to look at various factors that set fraud alerts. Also, the disadvantage of not operating on all the available data as in the case of statistical analysis is eliminated. Advanced analytic solutions that are fast and powerful to process real time activity and identify potential risks are increasingly being deployed by banks and financial institutions. The following three benefits make the usage of big data technology almost indispensable for financial organizations, when it comes to detecting fraud.

- **Enormous volume:** Operate on all data, ability to handle very large data sets.
- **High speed:** Near real-time processing using sophisticated technology.
- **Type of data:** It is the ability to process and analyze unstructured data apart from the structured ones, which makes big data solutions very effective.

Several companies already use advanced analytics to identify the pattern of normal sales. Once this is known, any outlier or a significant deviation from this usual pattern could be flagged. Immediate action could be taken to prevent huge loss arising from fraudulent activity, including impersonation and identity theft. For instance, if the normal spending patterns are known and analyzed regularly, any significant 'spend amount' could immediately be flagged.

Example: Wells Fargo used Machine Learning to Internal, to Locate and Respond to Fraud Attacks in Real-Time

Wells Fargo with close to 2 trillion-dollar assets serves one third households and one tenth of small businesses in USA. The bank partnered with FICO to analyze both internal data and third-party data to detect frauds and protect customer assets while offering very good customer experience.

Contd….

The multi layered machine learning approach has drastically reduced false positives. The bank's focus was to detect and thwart frauds from continuously changing threats with minimal disruption to excellent customer service.

Source: https://www.businesswire.com/news/home/20220316005321/en/Wells-Fargo-Enhances-Fraud-Protection-and-Customer-Experience-Using-FICO-Solutions, March 16, 2022, Accessed on 08/08/2022

3.7 Risk and Big Data

Risk Management would not exist without advanced data analytics. There are two common types of risk management, namely market risk management and credit risk management. Another type, operational risk management, is not that common. The risk professionals may avoid risk, or reduce the negative effect of risk, or reduce the probability of occurrence of negative risk, or accept some or all potential benefits.

3.7.1 Market Risk Analytics

This analytics helps in understanding the decrease of probability due to the change in interest rates, foreign exchange rates, stock prices and product prices. For example, a question of an investor – 'Should we vend this stock, if the price falls another 13 percent?' Here the prior stock prices of the portfolio can be analyzed to make a decision.

3.7.2 Credit Risk Analytics

It focuses on historical credit behavior. It is an analytical tool to forecast the possibility that a borrower will default on a given type of debt he has applied for. For example, "Is this person expected to default on his Rs.5,00,000 loan?". In this case the customer's past transactions and repayments could be analyzed or if he/she is a new customer, the similar demographic customer profiles could be analyzed to make a decision.

Example: Master Card Deploys Ai Based Analytics to Identify Loan Default Risk and thus Ease Lending Process in Asia Pacific

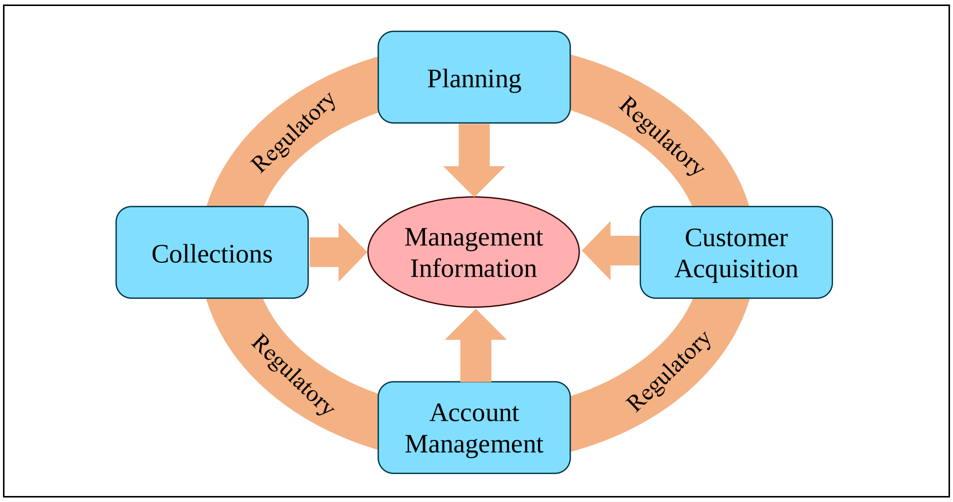
Mastercard recognized that lending is cumbersome for the small businesses in Asia Pacific region as the process of assessing creditworthiness was complex. The company tied up with Singapore based AI tool company Eureka to help its customers in Asia Pacific region with data-based analytics to establish credit worthiness and make loans available to them more easily. The solution used credit score and location checks by using telco data.

Source https://www.yahoo.com/video/mastercard-ma-ties-ease-credit-152003224.html, December 15, 2021, Accessed on 08/08/2022

3.8 Credit Risk Management

Credit risk management deals with minimizing the loss and maximizing the profit. Mostly the financial institutions, which have more number of customers, need to find out the profitable customers and retain them. Credit risk professionals need to analyze the complex and abundant customer data, in addition to other details of the customer from various other sources. Nowadays, these data are overwhelming and need good analytical tools to analyze them. Conventional scoring methods emphasize on predicting the likelihood of misbehavior or bankruptcy. When you add scoring solutions, those can help companies recognize the profitable potential of customers. Today, most of the companies out-source credit rating of customers to third party companies and depend on their results of credit scoring to rate their customer and retain them. In the future, credit risk management methods will leverage new data sources deriving from a big data generated from highly digital and mobile world. American product leader for Master Card Advisors, Risk & Marketing Solutions, indicates that the typical credit risk framework involves four critical parts, viz., planning, account management, collections and customer acquisition. Figure 3.4.

Figure 3.4: Credit Risk Framework



Source: Ori Peled.[from the book: Big data, big analytics: M Minelli, M Chambers, Ambiga Dhitaj, John Wiley and Sons]

Example: Zest Finance (Fintech start-up) Uses Customer Social Media Data to Assess Credit Risk Before Lending

Zest finance went beyond just credit scores to assess credit risk for borrowers. It used analytical tools to analyse Social Media data of the customer like location, people in his network, how transparent the customer was etc. to arrive at the credit score.

Source: https://content.accion.org/wp-content/uploads/2019/01/Risk-Management-Tool-Guide-3-Credit-Scoring.pdf Accessed on 08/08/2022

3.9 Big Data and Algorithmic Trading

Algorithmic trading is a highly refined set of processes or mathematics or trading rules used by computer program or software agents, to provide "insights" into specific situations to take automated decisions. Algorithmic trading is mainly used by investment bankers to determine orders for equities, interest rate and foreign exchange rates. It is also used to buy and sell commodities, fixed income instruments, and derivatives in real time speed. It determines risk and return of each potential trade. Thus it facilitates better decision-making to buy or sell. Algorithmic trading comprises of enormous number of transactions with multifaceted interdependent data. Data of every millisecond matters for taking a decision.

3.10 Advertising and Big Data

In the early days, advertising executives used influential resources to reach their audiences. These were billboard, radio, newspaper and ultimately the television. Their clients were also attentive and were anxious to get their messages through these channels. As time progressed, the industry matured. The advertising executives required to learn more about their spectators. It created mandate for firms such as Nielsen Media Research, to statistically measure different segments of the population viewing different television programs. This helped the publicists to pick the best option to place their ads on media. Nowadays, clever media planning and the presence of more and more data are needed to target their ads.

Reach, Resonance and Reaction

According to Randall Beard the global head of Nielsen's Global Head of Advertiser Solutions, big data is changing the way the advertisers address certain related needs for instance:

1. How much is to be spent?
2. How to distribute budget across all the marketing communication touch points?
3. How to enhance advertising efficiency against brand equity and return-on-investment (ROI), in real-time?

The Need to Act Quickly (Real-Time when possible)

Normally, advertisers have to understand the impact of advertisements quickly and need to take immediate actions.

Measurement Can Be Tricky

Immediate increase in on-the-spot sales is not the only measurement. Apart from that, other measurements could study the actual offline purchase behavior also.

Content Delivery Matters Too

Where (in which site) the content is delivered also matters. If the site is popular, the advertisement could succeed.

Optimization and Marketing-Mixed Modeling (MMM)

Marketing-mixed modeling aids advertisers to recognize the impact of other marketing activities and advertising on sales results. It is also used to identify the relative performance of different medium in which the advertisement was broadcast or propagated.

Today's advertising involves large volume of data at different velocity. It may be broadcasted automatically and involve high variety of data to analyze the impact of the advertisement. Therefore, big data analytics is a must for today's advertising.

Example: Hotels.com is Using Insights from Big Data Analytics to come up with Very Focussed Ad Campaigns

Hotels.com (hotel booking platform) has used AI based Big Data analytics to analyse over 50 lac #TravelGrags hashtags found in tweets linked to Instagram posts. This gave insights into what locations, what food, what activities, what tourist attractions the travellers were bragging about. Based on this, the company could run very focussed ad campaigns around those themes.

Source:https://datainsight.wbresearch.com/blog/how-hotelscom-is-using-ai-and-big-data-to-reveal-social-medias-top-travelbrags, 2022, Accessed on 08/08/2022

Activity 3.2

A major financial lending institute wants to roll out the ability to offer customers the option to supplement their credit score by providing the option to supplement their payment associated information from several other sources like utility services, mobile phone bills, and other bills paid. How do you think big data would play a role in assessing the credit score evaluation of the customers and ultimately credit risk management?

Answer:

Check Your Progress - 2

7. Credit Risk Management dealt with which of the following?

   a. Minimizing loss and maximizing profit

   b. Identifying potential customers

   c. Leveraging customer data from various sources

   d. Rating customers

   e. All the above

8. Which one of the following focuses on realizing the probability, leading to the analysis that the value of a portfolio will decrease due to the change in stock prices?

   a. Credit Risk Analysis

   b. Market Risk Analysis

   c. Operational Risk Analysis

   d. Personal Risk Analysis

   e. None of the above

9. Which of the following focuses on existing credit behaviors, to assess the possibility of a borrower defaulting on any type of debt applied for?

   a. Credit Risk Analysis

   b. Market Risk Analysis

   c. Operational Risk Analysis

   d. Personal Risk Analysis

   e. None of the above

10. What is ‘algorithmic trading’?

    a. It is a trading game played using cards

    b. It is a program to calculate the credit risk of a person

    c. It is a program used by investment bankers to determine the company’s profit

    d. It is a program used by investment bankers to find orders for equities, interest rate and foreign exchange rates, determine buy and sell commodities, fixed income instruments, and derivatives at striking speed

    e. None of the above

3.11 Using Consumer Products as a Doorway

Today’s success in business depends on how close you are with your customers and how much you have understood them. This can be achieved by observing and recording their movements and information. This involves high volume, variety, and velocity of data which ultimately is the big data and its analytics. Without that, success in business can never be imagined.

3.12 Summary

- The unit explained the usage of big data in marketing and business.
- This is done by explaining the customer behavior to buy items in the conventional and digital marketing.
- It also deals with the different ways of marketing in new school of marketing, the role of big data in fraud detection with different approaches, risk management through big data, algorithmic trading, and the relationship between advertising and big data is explained.
- The chapter concludes that without understanding the customer, the business will never succeed and that can be achieved using big data and analytics.

3.13 Glossary

Affiliate Marketing: It is a type of performance-based marketing in which a business rewards one or more affiliates for each visitor or customer brought by the affiliate's own marketing efforts.

Algorithmic Trading: Algorithmic trading (automated trading, black-box trading, or simply algo-trading) is the process of using computers, programmed to follow a defined set of instructions, for placing a trade in order to generate profits at a speed and frequency that is impossible for a human trader**.**

Database Marketing: Database marketing is a form of direct marketing using databases of customers or potential customers to generate personalized communications in order to promote a product or service for marketing purposes. The method of communication can be any addressable medium, as in direct marketing.

De-duping: De-dupe stands for de-duplication and is defined as optimizing data storage by eliminating duplicate copies of data.

Elastic Search: It is an open source search engine built on top of Apache Lucene and released under an Apache license. It is Java-based, and can search and index document files in diverse formats.

Link Analysis: Link analysis is a data analysis technique used in network theory that is used to evaluate the relationships or connections between network nodes. These relationships can be between various types of objects (nodes), including people, organizations and even transactions.

Marketing Automation: It refers to software platforms designed to automate repetitive tasks in critical areas such as campaign management.

Non-line World: Consumer uses both on-line or offline interchangeably to explore more about a product or to buy the product.

Off-Line Marketing: Offline marketing is an obvious choice for any business with a local customer base. Offline marketing strategies utilize offline media channels to create awareness of a company’s products and services. These campaigns can include radio and print advertising.

Online Marketing: Internet marketing, or online marketing, refers to advertising and marketing efforts that use the Web and email to drive direct sales via electronic commerce, in addition to sales leads from web sites or emails.

Social Marketing: Itis an approach used to develop activities aimed at using social networks for changing or maintaining people's behavior. It is also used for the benefit of individuals and society as a whole.

Social Media Marketing: Social media marketing is the process of gaining website traffic or attention through social media sites. Social media marketing programs usually center on efforts to create content that attracts attention, and encourages readers to share it across their social networks.

Social Network Analysis: Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties or edges (relationships or interactions) that connect them.

3.14 Self-Assessment Test

1. What is ‘Non-line’ world of customer?
2. Explain the growth in database marketing.
3. What is the role of big data in fraud detection? Explain the different approaches used in it.
4. How big data can be used in risk management?
5. Explain the use of big data in advertising.
6. What is meant by algorithmic trading? Explain.

3.15 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.

3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.

5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition

6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

3.16 Answers to Check Your Progress Questions

1. (e) Non-line shopping

Consumers are progressively going to a physical store to become more accomplished about a product – by touching it, playing with it and seeing how it works actually. Then they go online to compare its price with other shops. Either they purchase it online or go back to the store and make their purchase there. This leads to non-line shopping

2. (e) Consumer data, click stream data, competitive intelligence data, etc.

Google's digital marketing evangelist explained that big data warehouse tactic does not work well in the online world, and to be successful online, one need to embrace multiplicity. Here multiplicity refers to requirement of multiple skills in the decision-making team, multiple types of data (consumer data, clickstream data, competitive intelligence data, etc.). Hence analytics need to address these.

3. (b) Efforts to create content that attracts attention and encourages readers to share it across their social networks

4. (d) It is to collect, create, crush data and present the data for analytical purpose

Nowadays most of the industries are losing their customer data by sending the customer details to large bureaucratic organizations. The purpose of these organizations is to collect, create, and crush data (Puking).

5. (b) In 1960s

By 1960s, database marketing started by storing the customer information in mainframe systems and using them to communicate with the customers.

6. (c) **In late 1990s**

In 1990s, when email entered the picture, the marketers started using Internet and world wide web to improve their marketing. Once the dot-com boom hastened, marketers adopted new technologies to start the company websites. This led to "real-time" marketing.

7. (e) **All the above**

Minimizing loss and maximizing profit, identifying potential customers, rating customers, leveraging customer data from various sources.

8. (b) **Market Risk Analysis**

This analytics helps in understanding the decrease of probability due to the change in interest rates, foreign exchange rates, stock prices and product prices. For example, a question of an investor – 'Should we vend this stock, if the price falls another 13 percent?' Here the prior stock prices of the portfolio can be analyzed to make a decision.

9. (a) **Credit Risk Analysis**

It focuses on historical credit behavior. It is an analytical tool to forecast the possibility that a borrower will default on a given type of debt he has applied for. For example, "Is this person expected to default on his Rs.5,00,000 loan?". In this case the customer's past transactions and repayments could be analyzed or if he/she is a new customer, the similar demographic customer profiles could be analyzed to make a decision.

10. (d) **It is a program used by investment bankers to determine orders for equities, interest rate and foreign exchange rates, buy and sell commodities, fixed income instruments and derivatives at striking speed**

Algorithmic trading (automated trading, black-box trading, or simply algo-trading) is the process of using computers programmed to follow a defined set of instructions for placing a trade in order to generate profits at a speed and frequency that is impossible for a human trader.

Unit 4

Big Data in Healthcare

Structure

> *"If Big Data is the new oil in healthcare, clinical & business intelligence is the refinery."*
>
> - Brendan Fitzgerald, HIMSS Analytics

4.1 Introduction

In the era of big data our healthcare organizations have learned that gathering data for the sake of gathering data yields little benefit. However, when data – as with crude oil - is refined the opportunities for use are multiplied.

In the previous unit, we have discussed about big data uses in marketing. With the increased use of electronic systems in our work, data gets generated and flows into the system every second. Such data can be of varied types, such as operational, administrative, financial and clinical. As organizations will have large number of systems and users. The data is trapped in various forms as input. It is seen that data flow is different in different domains and has a varied velocity. Whatever be the case, data needs to be processed and presented in a human-readable form to the user who can then use it based on his/her need. When the data is voluminous and it is difficult to process it using traditional soft wares and techniques, it is referred to as "big data". Big data is larger, more complex data set, especially from new data sources. These data sets are so voluminous that traditional data processing software just cannot manage them. But these massive volumes of data can be used to address business problems. This would not have been possible before.[1] Big data is an evolving term and commonly used to refer to large volumes of data (structured, unstructured or in any other form). Big data is complex and poses challenges to normal harnessing. In an article titled "Why healthcare may finally be ready for big data", published in the *Harvard Business Review* in 2014, Shah and Pathak point out that a huge amount of 2.5 quintillion terabytes of data was being generated every day in 2012 alone. The total amount of data created, captured, copied and consumed in the world is expected to reach 59 zettabytes in 2020.

Healthcare is a costly business. It accounts for a major share of the GDP in any country. In other words, healthcare consumes a major share of income of nations. The healthcare sector is at an inflection point and is poised for rapid changes. The healthcare industry, across the globe, is facing significant challenges around cost, quality care of the patient and accessibility of healthcare to the population. Harnessing big data is seen as one of the ways to be able to understand the status, need and possible solutions for the ever-increasing challenges in healthcare. This unit provides an overview of big data in healthcare, its applications and areas of impact.

In this unit, we will cover topics such as big data in healthcare – overview, challenges and benefits, leveraging strategies for big data, applications in the healthcare industry and source of innovation and new technologies.

4.2 Objectives

By the end of this unit, you should be able to:

- Describe what is big data and its uses in healthcare.
- Recognize some challenges of big data in healthcare.
- Discuss leveraging strategies for big data.
- Define applications of big data in healthcare industry.
- Relate to how big data is a source of innovation and new technologies.

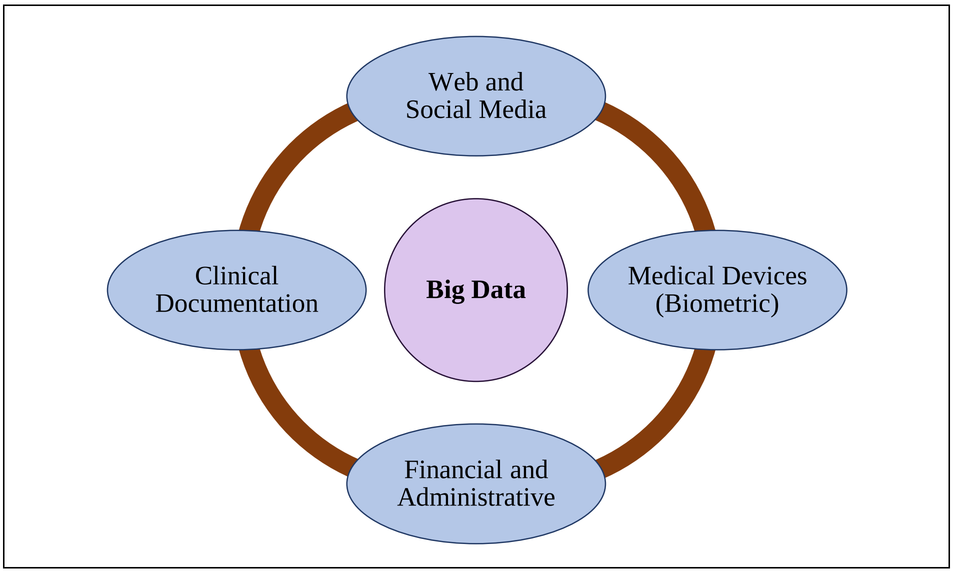
4.3 Big Data in Healthcare –An Overview

The need for analyzing healthcare data to improve delivered care and to meet quality measures is initiating revolution in healthcare. To heal the patients and stay in business, healthcare stakeholders such as the physicians, pharmaceutical companies and payers want to make quick decisions. To be able to do that, they need to rely on authentic and well-processed data presented to them in a manner that they can quickly see and understand. This can be achieved with the help of data analytics. Analytics is systematic computational analysis of data or statistics

[1] https://www.oracle.com/in/big-data/what-is-big-data.html

and is a process that deals with data mining. “Analytics” is used to describe statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen. Figure 4.1 illustrates big data.

Figure 4.1: Big Data - Illustration



Source: ICFAI Resource Centre

Healthcare Data are of Four Types:

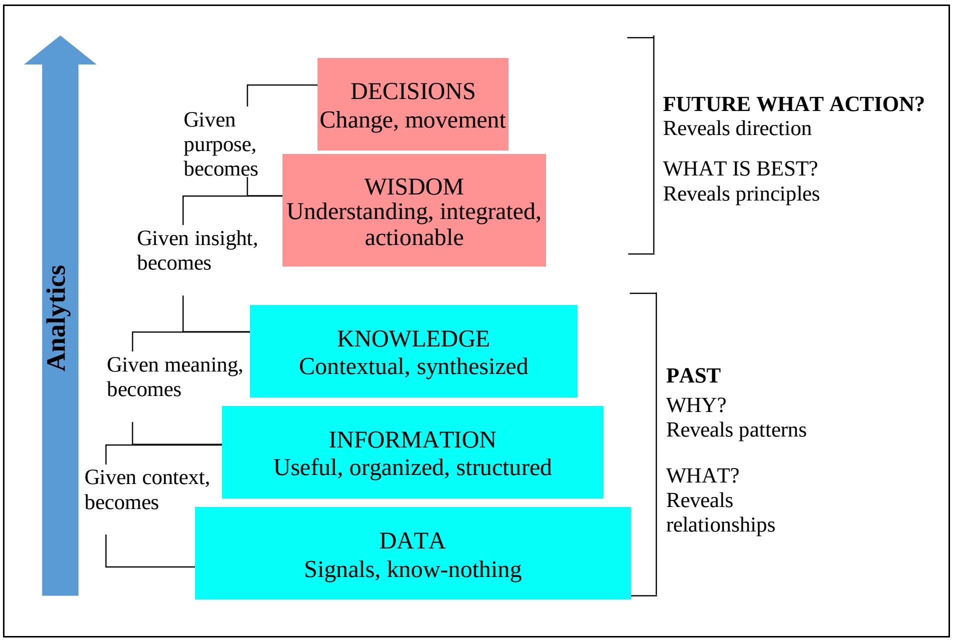
- **Clinical -** Data pertaining to the clinical facts of the patient such as vitals, labs, medications, blood pressure, pulse rate etc.
- **Operational-** Data related to operational aspects such as patient movements, human resources, material stocks etc.
- **Financial -** Data related to insurance and payment as money changes hands
- **Administrative -** Data related to registration, transaction and record keeping, usually during the delivery of a service such as data captured at the front office, surgical centers etc.

As is seen in the Figure 4.1, the big data comes in from:

- *Web and Social Media* - The use of this by patients and all the stakeholders generates a huge data related to the patient.
- *Medical Device (Biometric)* - The devices generate a continuous or periodic data from the user.
- *Financial/Administrative* - Related to payers and patient care administration process.
- *Clinical Documentation* - Documentation related to clinical parameters of the patients.

With analysis, capture, storage, processing and transfer in place, big data analytics holds promise to change the face of healthcare information and decision making. Needless to say, information privacy is embedded in any best practice around data management. The DIKW (Data - Information - Knowledge-Wisdom) cycle depicts the progression of data based on connectedness and understanding. Figure 4.2 presents a step wise move from data to decisions.

Figure 4.2: DIKW Cycle - Data - Information - Knowledge- Wisdom



Source: https://www.i-scoop.eu/big-data-action-value-context/dikw-model/

Example: Paris based Hospital Group Assistance Publique-Hôpitaux de Paris is using Big Data Analytics to Predict Number of Patient Admissions on an Hourly and Daily Basis

Estimating the number of staff to be deployed at any time in the hospital is a big challenge for hospital administration. Too many will add to unnecessary costs while too less will affect the patient service. The hospital administration used big data analytics to solve this problem. They used data mining techniques on historical data to predict the number of likely patients on an hourly basis and daily basis. This could help plan the staff deployment well and reduce costs and enhance patient satisfaction. 10-year data of admissions was analysed using “time series analysis techniques” and pattens in admission data were observed.

Source: https://www.datapine.com/blog/big-data-examples-in-healthcare/, June 2, 2022 Accessed on 09/08/2022

4.4 Challenges and Benefits

Big data is set to offer tremendous insight in healthcare management systems. With terabytes and petabytes of healthcare data pouring in the system today, it is not easy for the traditional infrastructure and architecture to handle it. There need to be new ways to process such volumes of data.

Need of the healthcare industry is to generate:

- Adhoc reports and other report formats as and when needed.
- Requests for intelligent data analysis for clinical decision support.
- Self-service access to information for self-management.
- Information that is understandable and sharable.
- Integrating clinical trials data with the clinical data for newer patterns in patient care management.

Challenges

Big data poses multiple challenges to the system as listed below:

- Varied types of data like:
  - Insurance claims.
  - Administrative data.
  - Clinical notes within the medical record.
  - Images from patient scans.
  - Lab reports.
  - Conversations about health in social media.
  - Information from medical devices.
- Disparate data sources - Data is trapped in multiple systems which may or may not “talk” to each other for data integration to proceed.
- Integrations – Due to lack of standardization of healthcare data, integration becomes challenging. Poor data standardization can result in incomplete and inaccurate data collection, patient matching issues and slower organizational work flows. For healthcare organizations to deliver the best, most efficient and high quality of care, they must need accurate and up to the minute patient information and data.
- Multiple users and consumers of the data, such as:
  - Providers.
  - Payers.
  - Employers.
  - Disease management companies and wellness facilities.
  - Patients and care givers.
  - Life science/Pharmaceutical companies.

- Meeting the need of speed - To find the relevant and granular data quickly can be a challenge, given the volume and variety of the data.
- Understanding the data - Fallout of the unstructured data and data types. It is required to make some structure and sense of the input, to be able to present it in a meaningful way.
- Data ownership - Custodian of certain forms of data, how and who should use that information. This is important from the data use and security standpoint too.

  Benefits

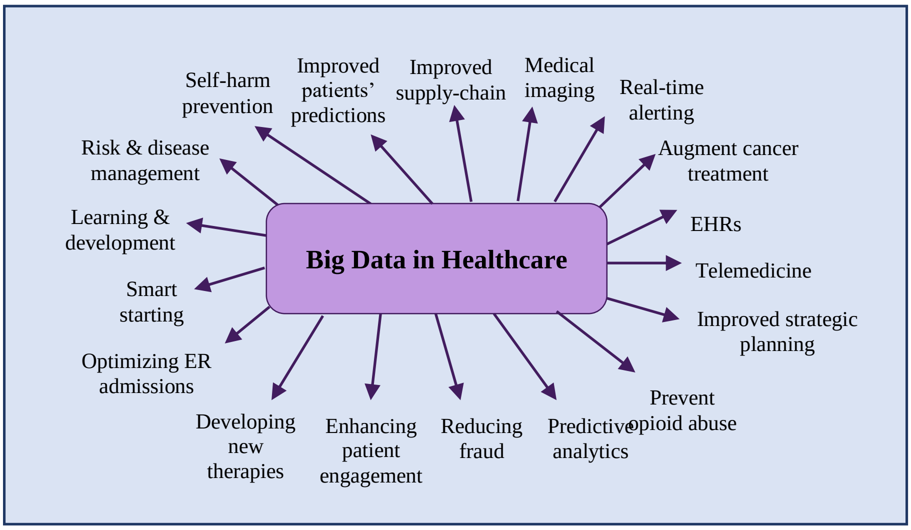
Benefits of big data are vast, given its potential to churn the data and come up with innumerable possibilities of combinations like:

- Applying advanced analysis and computation to modify existing strategies or create new ones.
- Leveraging heterogeneous datasets (web of things technology data are acquired from different sorts of distributed sensors and applications. The data types vary from structured data such as table data, semi-structured like extensible markup language (XML) or resource description framework (RDF), and unstructured data like images and videos. This is possible using predictive analytics where we use various types of data (administrative, clinical, financial, environmental, etc.) and integrate them to come up with better predictions of health and target the right interventions to the right patients.
- Improve both quality and efficiency in healthcare in multiple delivery areas such as:
  - Predicting re-admissions and adverse events (An event, preventable or non-preventable, that caused harm to a patient as a result of medical care. This includes never events, hospital-acquired conditions, events that required life-sustaining intervention and events that caused prolonged hospital stays, permanent harm or death.)
  - Treatment optimizations.
  - Identify gaps in care.
  - Assess medication/treatment compliance.
  - Identification of in-need population early and direct resources to such segment of population.
- Reduction on cost of health care: Using augmentation techniques to harness existing healthcare data with big data/at-risk strategies such as in-memory technologies and advanced analytics.

- Faster and better decision making: For example, using natural language processing technologies to understand customer satisfaction and time to intervene.
- Creation of new products and services: Driven by changing needs, emerging scenarios and enabled by generation of new knowledge, new products and services are to be created to overcome existing healthcare challenges.
- Detection of frauds: Reduced submissions of improper, erroneous or fraudulent claims using data analytics.

Figure 4.3 presents uses of big data in healthcare.

Figure 4.3: Uses of Big Data



Source: https://www.datapine.com/blog/big-data-examples-in-healthcare/

Example: Kaiser Permanente (Healthcare company) has Overcome the Challenge of Health Data Sharing by Using Big Data Analytics and Saved One Billion Dollars

Kaiser Permanente (a healthcare company) has been deploying Big Data Analytics on data stored by the company for a large number of years. The group has resolved the issue of problem in data sharing across hospitals by implementing a program called “HealthConnect” through which it makes it easier to use EHR (Electronic Health Records) across his group hospitals easily. This well-designed integrated system has resulted in better outcomes for cardiovascular diseases while saving $1 billion due to reduced appointment costs and lab test costs.

Source: https://www.datapine.com/blog/big-data-examples-in-healthcare/, June 2, 2022 Accessed on 09/08/2022

4.5 Leveraging Strategies for Big Data

Big data is to be used by healthcare to better the services. It is up to healthcare to leverage what is coming in and is existing. This enables them to come up with strategies to consume the data in a meaningful way so that the data serves its purpose in imparting better experience in healthcare. Table 4.1 presents strategies for leveraging big data for healthcare:

Table 4.1: Strategies for Leveraging Big Data

Strategy	Details
Engage stakeholders	Engaging critical stakeholders such as patients and providers will help get their buy-in and decreases their resistance.
Data governance framework	Framework that is well structured to manage enterprise-wide data governance. According to Data Governance Institute, framework is defined as "logical structure for classifying, organizing and communicating complex activities involved in making decisions about, and taking action, on enterprise data".
Integrate data analytics in training	Introducing the data culture at the initiation stage will help it root better into the overall system. Hospitals are encouraged to modify their training courses to include information on understanding how big data tools add value to overall healthcare performance. This will encourage users to use the data.
Ease of use of data consumption	Creating easy to use systems, such as dashboards, for the users to see the data as they use it mostly for clinical decision support. The data needs to be accurately processed and easily presented for the users to be able to make quick decisions.
Quality metrics and big data	Quality metrics drive healthcare quality. Analytics and quality improvement teams must work together to integrate analytics into the various quality-improvement methodologies.
Encourage in-house teams over commercial products	Encourage in-house teams to work on the challenges and build products or services. Inflexibility of commercial products can be an obstacle to the objective of big data analytics.

Source: ICFAI Research Center

Example: Apollo Hospitals Partners with Microsoft to use Big Data Analytics (Machine Learning) to Develop an India Specific Heart Risk Score to Better Predict Cancer Risk Among General Public

The Apollo Hospital group has always been trying to find better methods to predict diseases like heart attack and as such help the health care professionals to plan treatment plans. The hospital chain analysed some 4 lac patient data collected over seven years using AI based machine learning model approach. The model included additional risk factors compared to traditional factors and so the predictions were more accurate. The model provided a user-friendly tool for the doctors. Also, they developed a platform where patients can themselves find their risk without a detailed health check-up.

Source: https://www.broadbandcommission.org/Documents/working-groups/AIinHealth_Report.pdf, September 2020, Accessed on 09/08/2022

Check Your Progress - 1

1. Which of the following generates a continuous or periodic data from the user?
   a. Clinical
   b. Operational
   c. Medical Device (Biometric)
   d. Financial
   e. Administrative
2. Which need(s) of the healthcare industry can be translated as drivers for big data?
   a. Information that is understandable and sharable
   b. Quick generation of reports
   c. Intelligent data analysis for decision making
   d. Self-service access to information
   e. Information that is understandable and sharable, Quick generation of reports, Intelligent data analysis for decision making, Self-service access to information
3. What is a type(s) of healthcare data?
   a. Administrative
   b. Clinical
   c. Financial
   d. Operational
   e. Administrative, clinical, Financial, Operational

4. Which of these is a challenge of big data?
   a. Adhoc reports
   b. Intelligent Data Analysis
   c. Access to Information for Self Management
   d. Information that is Understandable and Shareable
   e. Disparate Data Sources
5. What is a type(s) of strategy(ies) to leverage big data in healthcare?
   a. Integrate quality metrics and analytics
   b. Encourage users to use the input data
   c. Integrate analytics into training
   d. Make the data consumption easy and understandable
   e. Integrate quality metrics and analytics, Encourage users to use the input data, Integrate analytics into training, Make the data consumption easy and understandable

Activity 4.1

Sunshine hospital is a large treating facility with a varied range of specialties and departments. This hospital is a tertiary care center for the region and has a large draining population to cater to. Being a center for the highest care, Sunshine has a huge inflow of patients. The hospital has the right and skilled staff. All its departments are equipped with a gamut of healthcare systems to trap and track the patients.

Sunshine hospital faces the problem of data deluge and plans to involve a vendor to help it with harnessing and analyzing data for its use. The senior management of the hospital wishes to have analytics run on the big data to improve existing services and come up with new service lines. Classify healthcare data into types (at least three) giving 2 examples of each type.

Answer:

4.6 Applications in Healthcare Industry

In spite of the challenges that emerging areas of the big data face, there is a lot of potential in it. Big data analytics is able to help the healthcare industry in many ways. A few examples of the application of big data in the healthcare industry are as follows:

- **Cancer care -** Compare millions of cell codes quickly to find cord blood matches from across the world for cancer patients in need of a stem-cell transplant.
- **Disease surveillance -** Monitoring disease trends and tracking infectious disease outbreaks for readiness of the system.
- **Payer analytics -** Consumers using comprehensive customer databases to gain deep understanding of the health plan. This helps to improve service offerings and create new products and service lines.
- **Clinical decision support applications -** Helps physicians of all strata from learners to specialists, to find out the optimal assessment and plan for a given patient, from millions of records of similar patients worldwide.
- **Monitor patients inwards -** To alert and predict the onset of nosocomial (hospital acquired) infections before the symptoms were noticed.
- **Prediction of likely outcomes -** In case of chronic diseases such as diabetes by extensive computing, based on patient's longitudinal health data, management protocols, and association with a particular physician.
- **Population health management -** Identify the unique needs and interventions for patient sub-populations through predictive modelling, by collecting and analysing data for at-risk groups

Example: Turquoise Health (A Health Care Startup) has Created a Platform which Ensures Complete Price Transparency Prior to Care Leading to Avoiding Friction Among Patients, Payers, and Providers

Turquoise Health (a health care startup) provided technology-based solutions for the patients to benefit from the price transparency from the payers and providers. The company got machine-readable data from providers and the patients got a comparative picture of price for specific treatment. The hospital also benefitted because the payments (already fixed) were realized in real time as and when a service was given. The company also created a platform for contracts among providers, payers, and patients. All three parties had certainty about the prices and payment prior to the treatment. This avoided friction among the three stakeholders.

Source: https://www.mobihealthnews.com/news/turquoise-health-scores-20m-price-transparency-platform, May 11, 2022, Accessed on 09/08/2022

4.7 Source of Innovation and New Technologies

Big data has opened new frontiers in healthcare. With the US federal government bill on HITECH act, encouraging transparency and use of electronic health records by the providers, the ground is fertile for innovation for creating new knowledge, new services and new products.

The input and approach for such innovations and technologies could be as follows:

- Opening and release of data from the pharmaceutical companies which have been aggregating data.
- Making the payer and provider data available to integrating mechanisms.
- Sharing the personal health details of patients who continue to input data in their records.
- Collecting, saving, retrieving, and analyzing information from multiple sources.
- Providing different forms of data to new models of healthcare such as Accountable Care Organizations (ACO). These models are based on “risk-sharing” and “fee-for-outcomes” tenets. Hence they need different form of data to work on.
- Aggregating big data into algorithms to form evidence-based care plans is fast gaining momentum.
- Visualizing easy to understand data display that would enable making quick decisions using relevant technologies.

Example: BenevolentAI is using Natural Language Processing to Analyse Large Medical Data to See if any Already Approved Drugs May Help in Blocking COVID Replication

BenevolentAI is a pharma company at the forefront of a revolution in drug discovery and development. It combined advanced AI and machine learning with cutting edge science to decipher complex disease biology and discover new drugs. The company used natural language processing to scan through large historical medical information and also scientific literature to find if any already approved drugs can block COVID virus replication. It succeeded, and it found six such drugs so far.

Source: https://www.broadbandcommission.org/Documents/working-groups/AIinHealth_Report.pdf, September 2020, Accessed on 09/08/2022

Check Your Progress - 2

6. Which of the following helps consumers using comprehensive customer databases to gain deep understanding of the health plan?
   a. Cancer Care
   b. Disease Surveillance
   c. Payer Analytics
   d. Clinical Decision Support Applications
   e. Monitor Patient Inwards

7. What is/are some example(s) of the applications of big data in healthcare?
   a. Cancer/Chronic disease care
   b. Treatment optimizations
   c. Risk documentation and predictive analytics of population
   d. Payers use data to analyze their customer base
   e. Cancer/Chronic disease care, Treatment optimizations, Risk documentation and predictive analytics of population, Payers use data to analyze their customer base

8. Which of the following helps physicians of all strata to find out the optimal assessment and plan for a given patient?
   a. Payer Analytics
   b. Clinical Decision Support Applications
   c. Monitor Patient Inwards
   d. Prediction of Likely Outcomes
   e. Population Health Management

9. What is an examples of approach to innovation and technologies in big data?
   a. Pharmaceutical companies data availability
   b. Regulation opening up clinical data to data scientists
   c. New models of care need new data type
   d. New data visualization techniques
   e. Pharmaceutical companies data availability, Regulation opening up clinical data to data scientists, New models of care need new data type, New data visualization techniques

10. Which of the following is a source of innovation and new technology?
   a. HITECH Act
   b. Population Health Management
   c. Cancer Care
   d. Disease Surveillance
   e. Prediction of Likely Outcomes

Activity 4.2

A data scientist was employed by CARE hospitals to understand patient’s data that was in the electronic systems in the hospitals for the past 10 years. CARE hospitals wanted to mine the data to relate to care of chronic diseases such as asthma, diabetes, heart failure and myocardial infarction to find out gaps in care and devise novel services to address this issue.

Imagine you are the data scientist appointed by the CARE hospitals. Suggest to the management at least three ways to leverage the existing and incoming data. For example, come up with the types of data that will be trapped to analyse a chronic disease, how these data will be used and the right data integration.

Answer:

4.8 Summary

- Big data is defined as ‘a term that describes large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information’.
- Healthcare data is a big data that can be used to churn out new information about the patients and current status of healthcare.
- The data of innumerable patients are available across multiple categories such as social media, clinical, administrative, operational and pharmaceuticals, besides the data from devices flowing into multiple healthcare systems, every second.
- Authentic data with intelligent analytics form a platform on which evidence-based medicine can run.
- Big data with its 5 “Vs”—Volume, velocity, variety, veracity and value—have helped healthcare define new goals regarding the standard of care, needs of populations and sub-populations, and personalized medicine.
- Big data seems to be the next revolution. It has been at the center of the discussions of all information technology seminars. But it comes with its set of challenges related to security, data fragmentation and lack of framework.
- The healthcare industry, though a recent entrant in this game, is fast to catch up and soar to new heights on ways to leveraging big data quickly.
- In this unit, we have covered topics such as leveraging strategies for big data, applications in the healthcare industry, and source of innovation and new technologies.

4.9 Glossary

Accountable Care Organization (ACO): Care models that include physicians, clinics, hospitals, and other healthcare providers, who come together voluntarily, to give coordinated high-quality care to their medicare patients.

Big Data: Large volumes of data flowing into the healthcare system that poses challenges to be analyzed by traditional infrastructure or systems.

Clinical Decision Support (CDS): Guidelines that are generally presented in IT systems to provide physicians, allied staff, and patients with knowledge and person-specific information, organized and intelligently filtered to enhance health and healthcare.

Emerging Areas in Healthcare: Newer areas that are evolving to address cost, patient care, quality, healthcare accessibility, and challenges around healthcare.

Health Information Technology for Economic and Clinical Health (HITECH): Provides the U.S. Department of Health and Human Services (HHS) with the authority to establish programs through healthcare IT.

Population Health: According to Kindig and Stoddart, population health is ‘Health outcomes of a group of individuals, including the distribution of such outcomes within the group of population’.

4.10 Self-Assessment Test

1. What is big data and how is it different from normal data and its lifecycle/management?
2. Explain a few challenges that the healthcare industry is facing and the need for a big data revolution.
3. Name three benefits and three challenges of big data.
4. Name three strategies that the healthcare users should know to be able to leverage big data.
5. Big data has many applications in healthcare. Mention any three applications in the healthcare scenario, justifying the need for big data.

4.11 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

4.12 Answers to Check Your Progress Questions

1. (c) Medical Device (Biometric)

The devices generate a continuous or periodic data from the user.

2. (e) Information that is understandable and sharable, quick generation of reports, intelligent data analysis for decision making, Self-service access to information

Data and order sets, guidelines, care plans, alerts and notifications are all types of Clinical Decision Tools. Information that is understandable and sharable, quick generation of reports and self-service access to information are all needs of the healthcare industry that are drivers for big data.

3. (e) Administrative, clinical, financial, operational

Clinical, financial, administrative, and operational are by far the most types of data in healthcare.

4. (e) Disparate Data Sources

Data is trapped in multiple systems which may or may not “talk” to each other for data integration to proceed.

5. (e) Integrate quality metrics and analytics, encourage users to use the input data, integrate analytics into training, make the data consumption easy and understandable

Integrating quality metrics in analytics, encouraging users to use analytics, integrating analytics into training and making data consumption easy and understandable are also strategies to leverage big data.

6. (c) Payer Analytics

Consumers using comprehensive customer databases to gain deep understanding of the health plan. This helps to improve service offerings and create new products and service lines.

7. (e) Cancer/chronic disease care, treatment optimizations, risk documentation and predictive analytics of population, payers use data to analyze their customer base

Cancer care, care plans for chronic diseases, treatment optimizations, risks documentation and predictive analytics, and payers using data for their customer analysis, are all examples of big data applications in healthcare.

8. (b) Clinical Decision Support Applications

Helps physicians of all strata from learners to specialists, to find out the optimal assessment and plan for a given patient, from millions of records of similar patients worldwide.

9. (a) Pharmaceutical companies data availability

Pharmaceutical companies data availability, regulation opening up clinical data to data scientists, new models of care need new data type, new data visualization techniques, data from pharmaceutical companies, clinical data, new models of care needed for new data and new data visualization techniques are all examples of innovations in big data in healthcare.

10. (a) HITECH Act

Big data has opened new frontiers in healthcare. With the US federal government bill on HITECH act, encouraging transparency and use of electronic health records by the providers, the ground is fertile for innovation for creating new knowledge, new services and new products.

Big Data, Cloud and Analytics

Course Structure

Block 1: Introduction and Applications of Big Data	
Unit 1	What is Big Data?
Unit 2	Why Big Data is Important?
Unit 3	Big Data in Marketing & Advertising
Unit 4	Big Data in Healthcare
Block 2: Cloud Computing and Big Data Technologies	
Unit 5	Big Data and Cloud Technologies
Unit 6	Big Data Technologies and Terminologies
Unit 7	Cloud Computing and Big Data Management for Decision Making
Unit 8	Handling Unstructured Data
Unit 9	Information Management
Block 3: Business Analytics	
Unit 10	Analytics in Database Marketing
Unit 11	Business Analytics Techniques
Unit 12	Data Visualization and Modelling
Block 4: Managing Talent for Big Data Analytics	
Unit 13	Talent Management-I
Unit 14	Talent Management-II
Block 5: Data Privacy and Analytics in Various Business Areas	
Unit 15	HR Analytics in HR Planning
Unit 16	Data Analytics for Top Management Decision Making
Unit 17	Business and Marketing Intelligence Using Analytics
Unit 18	Data Privacy and Ethics