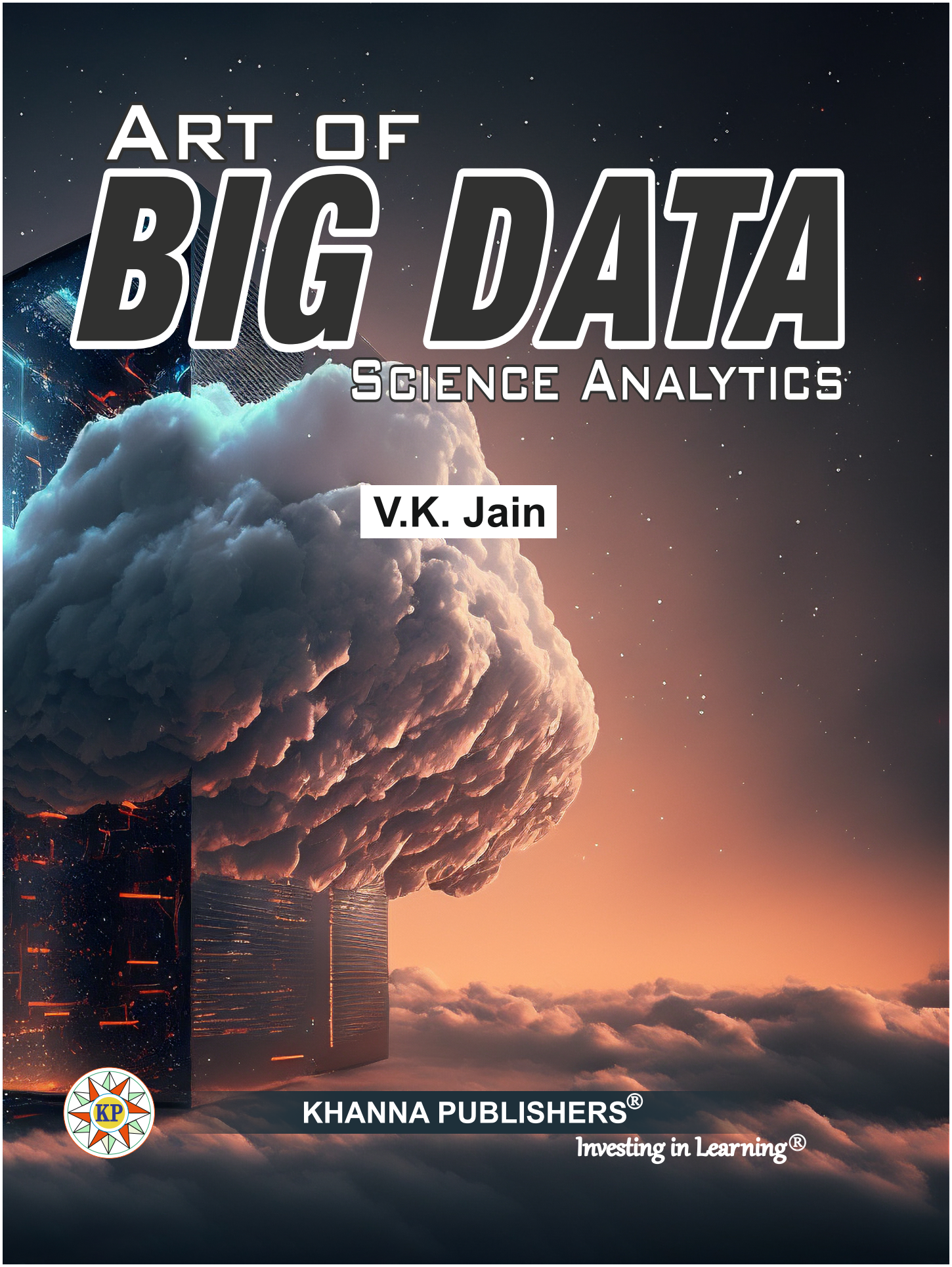
ART OF BIG DATA

SCIENCE ANALYTICS

V.K. Jain



KHANNA PUBLISHERS®

Investing in Learning®

Art of Big Data Science Analytics

Er. V.K. Jain
B.E., M.Tech., FIE, FIETE,
Life Member ISHRAE, NOIDA

KHANNA PUBLISHERS®
Investing in Learning®
Operational Office:
4575/15, Onkar House, Opp. Happy School,
Ground Floor, Daryaganj, New Delhi 110002
Phones : 011-45033819 *Mob*. 09811541460
E-mail : contactus@khannapublishers.in
website : khannapublishers.in

Published by :

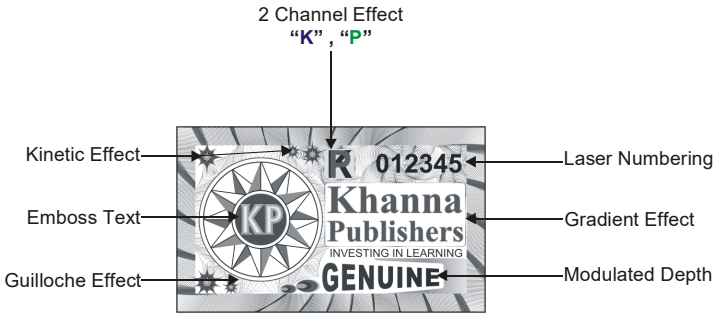
Romesh Chander Khanna & Vineet Khanna
for KHANNA PUBLISHERS
2-B, Nath Market, Nai Sarak
Delhi- 110 006 (India)

Website : www.khannapublishers.in

Hologram & Description

To all readers of our books, to prevent yourself from being defrauded by piraters, please make sure that there is an Hologram on the cover of our books with the below specifications. If you find any book without Hologram and Description, please mail us at **contactus@khannapublishers.in**

Thanking you



ISBN No. : 978-93-92549-51-9

***First Edition :* 2024**

Preface

Big data analytics is the often complex process of examining big data to uncover information such as hidden patterns, correlations, market trends and customer preferences that can help organizations make informed business decisions.

Big Data Analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

There's no single technology that encompasses big data analytics. Of course, there's advanced analytics that can be applied to big data, but in reality several types of technology work together to help you get the most value from your information.

Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media much of it generated in real time and at a very large scale.

Machine learning, a specific subset of AI that trains a machine how to learn, makes it possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks.

The rise of cloud storage and the demand for remote, cloud-based access has transformed the role of the enterprise data architect. Instead of simply managing an in-house database, architects now need to handle a host of new challenges related to data integrity, security, integration and analysis.

There is no single book available that covers all topics like Big Data and Hadoop, Data Management. Data Mining and Ware Housing, Machine Learning In-memory analytics. Predictive, Prescriptive and Diagnostic or Inferential Analytics, Visual Text Analytics at one place. Each of these topics require a book to cover the full text-ware but there are many disciplines of Science and Technology which need only consolidated knowledge to be conferred to learner in the subjects. This book is down to earth effort in that direction. It was very tough task for me to present the matter in very concise manner and which is commensurate with latest practices and state of affairs.

I am very thankful to the publisher of this book to push forward the publication of this book in very short time during Covid-19 period, especially to Kratu Khanna and Akshita Khanna for taking very keen interest in editing. I am very much indebted to Late Shri Ramesh Khanna ji who gave me the chance of publishing my first engineering book and really took active interest in promoting and establishing me as author. While composing the book, I must have taken help from hundreds of sources, in hard and soft formats, and therefore express my gratitude to authors and presenters of them. I also express my deep sense of gratitude to At last I can't forget the amount of support provided to me by my better half Heera Jain and other members of my family.

Er. V K Jain
B.E., M. Tech., FIE, FIETE,
Life Member ISHRAE
NOIDA

Contents

1

CHAPTER

Big Data and Data Science

1.1. INFORMATION EXPLOSION

The biggest innovation of the second decade of twenty first century was the phenomenon of Expansion and movement of information from hands of elite society to hands of masses.

There are now more than 5.16 billion people around the world using the internet. Well over half of the world's population is now online,

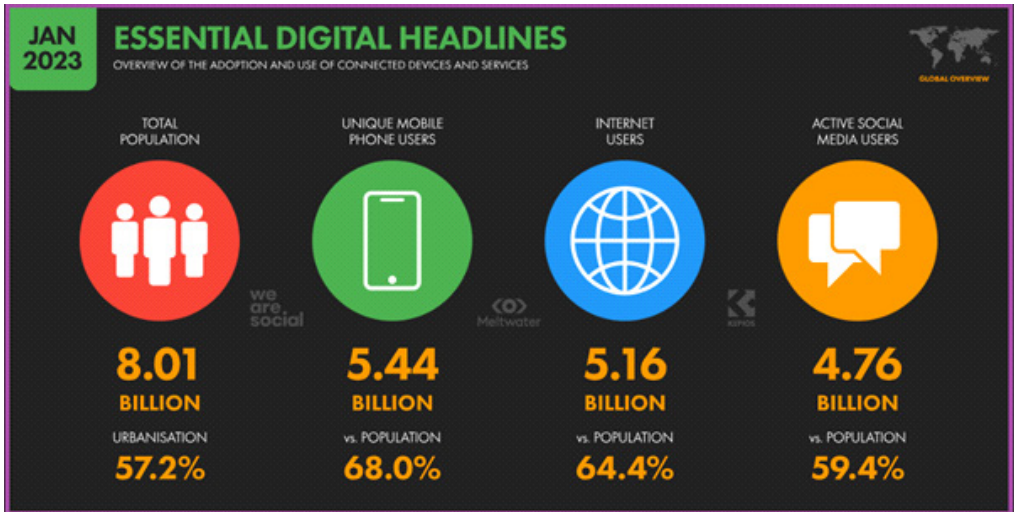


Fig. 1.1. Display of Impact of Digitisation 2023 (reported by DataReportal–Digital 2023: Global Overview Report web site)

- There are **5.11 billion** unique mobile users in the world today, up 100 million (2 percent) in the past year.
- There are **4.39 billion** internet users in 2019, an increase of 366 million (9 percent) versus January 2018.
- There are **3.48 billion** social media users in 2019, with the worldwide total growing by 288 million (9 percent) since this time last year.
- **3.26 billion** people use social media on mobile devices in January 2019, with a growth of 297 million new users representing a year-on-year increase of more than 10 percent.

By 2020, there will be around 40 trillion gigabytes of data (40 zettabytes).

Social media and digitisation has completely revolutionised the world. The various popular

social media content providers are :

The world's population passed 8 billion on 15 November 2022, and has reached 8.01 billion at the start of 2023. Just over 57 percent of the world's population now lives in urban areas.

A total of 5.44 billion people use mobile phones in early 2023, equating to 68 percent of the total global population. Unique mobile users have increased by just over 3 percent during the past year, with 168 million new users over the past 12 months.

There are 5.16 billion internet users in the world today, meaning that 64.4 percent of the world's total population is now online. Data show that the global internet user total increased by 1.9 percent over the past 12 months, but delays in data reporting mean that actual growth will likely be higher than this figure suggests.

1.1.1. Social Media Platforms: Monthly Active Users

If we rank platforms by monthly active users–which offers perhaps the most consistent basis for comparison – the latest "official" data suggest that Facebook still comes out top at a worldwide level.

Figures published in Meta's Q3 2022 investor earnings report show that the platform now has 2.958 billion monthly active users (MAU), which equates to almost 37 percent of the world's total population.

Meanwhile, YouTube's latest "official" statement indicates that the platform has "over 2 billion monthly logged-in users", but figures published in the company's own advertising resources suggest that the platform now attracts more than 2.5 billion users each month.

Instagram has consolidated its position amongst the top social media platforms since our October 2022 report too, with the company recently announcing that it has 2 billion monthly active users.

That puts the platform in similar territory to stablemate WhatsApp, although it's worth noting that Meta now reports WhatsApp attracts 2 billion active users per day, so its monthly user figure is likely even higher.

WeChat rounds out the top five, with Tencent's most recent investor earning announcement revealing that the platform now has more than 1.3 billion monthly active users.

However, Kepios analysis indicates that users in China still account for the vast majority of WeChat's global user base.

At a worldwide level, data.ai intelligence reveals that YouTube has the greatest number of active users of any mobile app, not just social media apps.

Facebook ranks second in terms of the social apps in this overall ranking, but it's important to stress that data.ai's figures show that Facebook's active users have continued to grow over the past twelve months.

WhatsApp is the third "social" app in this ranking, with the latest MAU figures putting it just behind Facebook.

Meta stablemates Instagram and Messenger claim the remaining "social" places in the top 10 apps by MAU, and it's worth highlighting that Google and Meta account for all of these top 10 apps.

Perhaps unsurprisingly, the hugely popular short-video platform now ranks sixth at a worldwide level, although it's worth noting that these rankings don't include users of TikTok's sister app, Douyin.

India is the country with the most active Facebook users (290 million), the US is second (190 million), and Indonesia third (140 million).

Telegram comes in seventh, ahead of Twitter in eighth place, while Snapchat and Pinterest round out the top ten social media apps by monthly active users.

India has the biggest YouTube audience size. As of April 2022, there were 467 million subscribers, followed by the United States with 247 million and Indonesia with 139 million. Facebook Messenger – 1.3 Billion Active Users

TikTok is one of the fastest growing social media apps, especially among younger users. Globally, TikTok reached 1 billion monthly active users in September 2021. To put this growth into perspective, in July 2020, it was "only" 689 million monthly active users.

As of January 2022, Pinterest was ranked as the 14th most-used social network across the globe based on global active users.

As of Q1 2022, the Snapchat app had approximately 332 million daily active users

According to LinkedIn's own statistics, LinkedIn is the world's largest professional network, with more than 875 million users in over 200 countries and territories across the world.

1.2. BIG DATA

Big Data is a phrase used to mean a massive volume of both structured and unstructured **data** that is so **large,** it is difficult to process it using traditional database methods and software techniques. In most enterprise scenarios the volume of **data** is too **big** or it moves too fast or it exceeds current processing capacity.

The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Although the term "big data" have been gaining momentum in current decade, the act of gathering and storing large amounts of information for eventual analysis is ages old. Humanity tried to gather the trusted data in form of Wikipedia. Currently, the English Wikipedia alone has over 5,904,484 articles of any length, and the combined Wikipedias for all other languages greatly exceed the English Wikipedia in size, giving more than 27 billion words in 40 million articles in 293 languages. The English Wikipedia alone has over 3.6 billion words, over 60 times as many as the next largest English-language encyclopedia,

The World Data Centre for Climate (WDCC) is the largest database in the world. The WDCC claims having data worth 220 terabytes readily accessible on the web including information on climate research and anticipated climatic trends, as well as 110 terabytes (or 24,500 DVD's) worth of climate simulation data.

If you do a Google search on big data (13.08.18), **over** 6,29,00,00,000 results search results are returned in 0.61 seconds in my PC, showing the importance of involvement of this technology. Big Data refers to technologies and initiatives that involve data that is too diverse, fast-changing or massive for conventional technologies, skills and infrastructure to address efficiently.

Over the past decade, major web companies like Google, Amazon and Facebook pioneered businesses built on monetizing massive data volumes. In the process, they invented new paradigms not only for extracting value from data, but also for managing data and compute resources from data center design, to hardware, to software, to application provisioning. Governments and even Google can detect and track the emergence of disease outbreaks via

social media signals. Oil and gas companies can take the output of sensors in their drilling equipment to make more efficient and safer drilling decisions. In the same way that the mission to the moon spawned a wave of innovation across multiple industries, Big Data has pushed information technology a quantum leap forward.

1.2.1. How Big is the Canvas of Big Data?

Big data does not have always to be big (*i.e.*, data in range of peta/exabytes). Even 50 GB can be said as big data if the structure is too complex for a normal RDBMS to store. What is small data? Small data is simple data structures, e.g. numbers (be it monetary, integers, fractions or floating points), strings (names, description, types), dates, times, and all the data we used to know in the last 30 years of data warehousing history.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.

The Information size is calculated as below (to be multiplied by 1.024 times, as actually 1 KB = 1024 = 2^8) :

1000 (10^3) KB	1 kilobyte	
1000 (10^6) MB	1 megabyte	(1 million byte)
1000 (10^9) GB	1 gigabyte	(1 billion byte)
1000 (10^{12}) TB	1 terabyte	(1 trillion byte)
1000 (10^{15}) PB	1 petabyte	
1000 (10^{18}) EB	1 exabyte	
1000 (10^{21}) ZB	1 zettabyte	
1000 (10^{24}) YB	1 yottabyte	

As per IBM 2.5 petabytes is Memory capacity of the human brain. A processor with a 64-bit address bus can address 18 exabytes of memory. Vedic Mathematics thought of 10^{60} as ultimate that in Sanskrit known as Mahogh. As per IDC, Digital universe is doubling in size every two years, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes.

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. Internet of Things (IoT) is making evry such things possible. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data are created. One question for large enterprises is determining who should own big data initiatives that affect the entire organization. IDC and EMC project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010.

1.2.2. Examples of Big Data

Artificial Intelligence (AI), mobile, social and Internet of Things (IoT) are driving data complexity, new forms and sources of data. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and

unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale.

What we can do, however, is gain a sense of just how much data the average organization has to store and analyze today. Toward that end, here are some metrics that help put hard numbers on the scale of Big Data today:

- Analysts predict that by 2020, there will be 5,200 gigabytes of data on every person in the world.
- Amazon sells 600 items per second.
- On average, each person who uses email receives 88 emails per day and send 34. That adds up to more than 200 billion emails each day.
- Master Card processes 74 billion transactions per year.
- 2–40 exabytes of storage capacity will be needed by 2025 just for the human genomes
- *eBay.com* uses two data warehouses at 7.5 *peta bytes* and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.
- *Amazon.com* handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- *Facebook* handles 50 billion photos from its user base.
- *Google* was handling roughly 100 billion searches per month as of August 2012.
- *Oracle NoSQL Database* has been tested to past the 1M ops/sec mark with 8 shards and proceeded to hit 1.2M ops/sec with 10 shards.

1.3. THE 5 V's OF BIG DATA

Big data is often characterised by the 5 **V's of Big Data** . Gartner was the first to put forward the concept of the term 3 V of big data which now stands modified as the five V's: Volume, *Velocity, Variety, Value, and Veracity*. The 5 V's, signifies the unique features of big data:

1. **Volume:** This signifies large amounts of data. Typically when people discuss big data volumes, they discuss peta-bytes. But in reality, most real-life big data implementations are still in the 10's to 100's of terabytes range which is still a lot of data.
2. **Velocity:** Velocity in the context of big data refers to the speed of data acquisition and processing. Big data technologies provide horsepower that accelerates these processes, thereby making data provisioning and usage faster, too.
3. **Variety:** This refers to the evolving types and growing sources of data, including semi-structured and unstructured data. An example of semi-structured data might be an e-mail message, where the date might be in a common, structured format, but the e-mail text itself is more unstructured. An example of unstructured data would be notes that a customer support representative might type in, free-form, about a customer's trouble ticket. Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

   Today's data is unstructured. In fact, 80% of all the world's data fits into this category, including photos, video sequences, social media updates, etc. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

4. **Value :** Talking about value means we are referring to the worth of the data being extracted. We can extract endless amounts of data but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data. The most important part of embarking on a big data initiative is to understand the costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is reaped can be monetized.
5. **Veracity :** Veracity is the quality or trustworthiness and accuracy of the data. For example, think about the data available on Wikipedia that is an amulgamation of data provided by millions of users worldwide. And think of the all the Twitter posts depending upon one's ego and alliance, having a lot of conflicts, hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content. Although Wikipedia makes a lot of efforts to verify the data, it is often biased to western world. Gleaning loads and loads of data is of no use if the quality or trustworthiness is not accurate. Another good example of this relates to the use of Global Positioning System (GPS) data that depends upon the accuracy of measuring equipment. Often the GPS will "drift" off course as you peruse through an urban area. Satellite signals are lost as they bounce off tall buildings or other structures. When this happens, location data has to be fused with another data source like road data, or data from an accelerometer to provide accurate data.

1.4. BIG DATA HANDLING PROCESS: DATA MINING, DATA WARE HOUSING, DATA LAKES AND DATA MARTING

Big data is extracted or collected from various soft sources *i.e.*, mined and stored in data lakes and then sent to data warehouse and ultimately marketed to business market by and to various business houses and Government department with slight variations here and there.

1.4.1. Data Mining

Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence(AI) and statistical.

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency.

More extensive data mining techniques were needed to get resorted and resolved, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 30 or 40 million records of detailed customer information, knowing that two million of them live in one location is not enough. You want to know whether those two million are a particular age group and their average earnings so that you can target your customer needs better.

As per IBM these business-driven needs changed simple data retrieval and statistics into more complex data mining. The business problem drives an examination of the data that helps to build a model to describe the information that ultimately leads to the creation of the resulting report.

For example, IBM SPSS®,(we shall read this in chapter 13 on Machine Learning), which has its roots in statistical and survey analysis, can build effective predictive models by looking

at past trends and building accurate forecasts. IBM InfoSphere® Warehouse provides data sourcing, pre-processing, mining, and analysis information in a single package, which allows you to take information from the source database straight to the final report output.

1.4.2. Data Lakes and Data Warehouses

Data lakes and Data warehouses are two different types of data storage *repository*, but with many differences. The data lake stores *raw structured* and *unstructured data* in whatever form the data source provides. It does not require prior knowledge of the analyses you think you want to perform.

The data warehouse integrates data from different sources and suits business reporting.

Data warehouse stores data in files or folders, in hierarchical manner where as data lake uses a flat architecture to store data.

Data warehouse is a core component of business intelligence, the data warehouse is a central repository of integrated data from one or more disparate sources, and it's used for reporting and data analysis. When the board makes a strategic decision on its future, or a call center agent reviews a customer's profile the data is typically being sourced from a data warehouse.

Characteristics Features of Data warehouse concept:

1. Holds multiple subject areas
2. Holds very detailed information
3. Works to integrate all data sources
4. Does not necessarily use a dimensional model but feeds dimensional models.

Data mart is a special category of data warehouse that often holds only one subject area-for example, *Finance*, or *Sales*. It may hold more summarized data and concentrates on integrating information from a given subject area or set of source systems.

The following are the reasons for creating a data mart

1. Easy access to frequently needed data
2. Creates collective view by a group of users
3. Improves end-user response time
4. Ease of creation
5. Lower cost than implementing a full data warehouse
6. Potential users are more clearly defined than in a full data warehouse
7. Contains only business essential data and is less cluttered.

1.5. DIFFERENCE BETWEEN DATA WAREHOUSING AND BIG DATA TECHNOLOGY

Big data uses Hadoop ecosystem to extract useful data from dynamic social and technical data present on Internet while data warehouse is a static repository of raw data first naturally collected in data lake, then sorted systematically and modelled into useful database.

The diagram below illustrates how the data warehouse and big data environments can come together in an integrated and very complementary way. In this scenario, the Hadoop

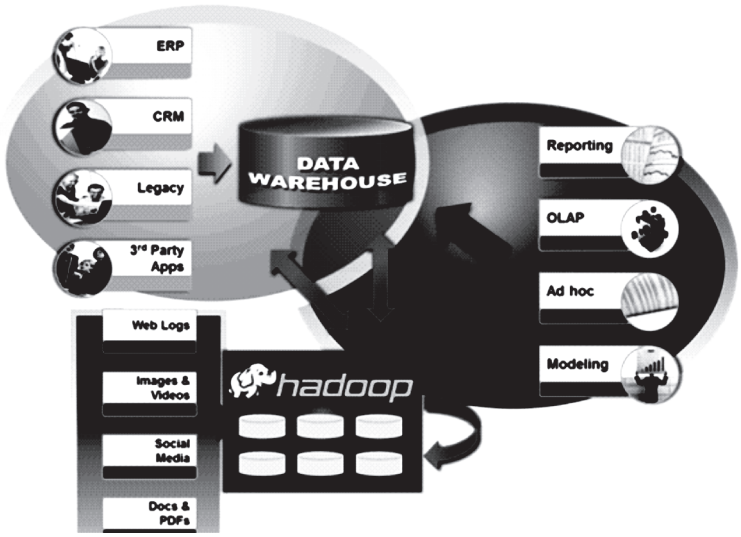


Fig. 1.2. Difference Between Big data and Data Warehousing

system can perform quickly. For instance, a high-tech company might decide to pull data from its social networking site and combine it with data from the data warehouse to update a customer's social network circle of friends. The environment might also use Hadoop to quickly "score" that person's social influence. Then that data will be provisioned back to the data warehouse so that, say, a campaign manager can view that person's influence score and re-segment him (or her) accordingly.

The example here is one of many potential uses for Hadoop and the data warehouse working together. The point to make here is that each system is doing what it's best designed to do. In the case of Hadoop, it's processing large amounts of social networking data quickly and in parallel. In the case of the data warehouse, it's availing that data to business users, knowledge workers, or data scientists, who are using that data—and other data as well—to make business decisions.

While there are exceptions to every rule, big data and data warehouse technologies are optimized for different purposes. Again, the goal is to use these solutions for what they were designed to do. In other words, use the best tool for the job.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in better and faster decisions.

BUSINESS REQUIREMENT	BIG DATA	DATA WAREHOUSE
Discovery of unexplored business questions	●	●
Clean, consistent, high-quality data	◒	●

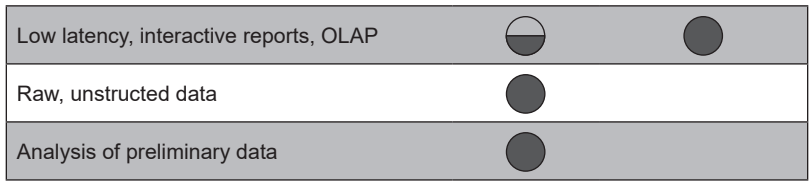


Fig. 1.3. Comparing Business Requirement for Big Data and Datawarehousing

1.6. BIG DATA TYPES

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. Lately, the term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."

Data sets grow rapidly, in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.

Big Data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data[1]. Big data 'size' is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data [2]. Big data requires a set of techniques and technologies with new forms of integration to reveal *insights* from datasets that are diverse, complex, and of a massive scale[3].

Following are examples of the three types:

1. Structured Data : Relational data.
2. Semi Structured Data : XML data.
3. Meta Data
4. Unstructured data : Word, PDF, Text, Media Logs.

1.6.1. Sources of Unstructured Big Data

Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.

Here are some examples of machine-generated unstructured data:

1. **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
2. **Scientific data:** This includes seismic imagery, atmospheric data, and high energy physics.
3. **Photographs and video:** This includes security, surveillance, and traffic video.
4. **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic seismic profiles.

The following list shows a few examples of human-generated unstructured data:

1. **Text internal to your company:** Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
2. **Social media data:** This data is generated from the social media platforms such *as YouTube, Facebook, Twitter, LinkedIn, and Flickr.*
3. **Mobile data:** This includes data such as text messages and location information.
4. **Website content:** This comes from any site delivering unstructured content, like ***YouTube, Flickr, or Instagram, Wikipedia.***

Of the useful data, IDC estimates that in 2013 perhaps 5% was especially valuable, or "target rich". That percentage should more than double by 2020 as enterprises take advantage of new Big Data and analytics technologies and new data sources, and apply them to new parts of the organization.

1.6.2. Semi-structured Data

Semi-structured data can contain both the forms of data. We can see semi-structured data as a strcutured in form but it is actually not defined with e.g., a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.

```
Personal data stored in a XML file
<rec><name>Amitabh Bajaj</name><sex>Male</sex><age>49</age></rec>
<rec><name>JyotsnAgarwal</name><sex>Female</sex><age>47</age></rec>
<rec><name>Anurag Jain</name><sex>Male</sex><age>44</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Web pages are generated in scripting of HTML which is also an example of **Semi-structured data.**

1.6.3. Meta Data

Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data which can make tracking and working with specific data easier.

There are three main types of metadata:

- **Descriptive metadata** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- **Structural metadata** indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- **Administrative metadata** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data, two that are sometimes listed as separate metadata types are:
  1. Rights management metadata, which deals with intellectual property rights, and
  2. Preservation metadata, which contains information needed to archive and preserve a resource.

1.6.4. Structured Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, now days, we are foreseeing issues when size of such data grows to a huge extent, typical sizes are being in the zettabyte. rage of multiple.

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employer_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushi Roy	Male	Admin	500000
7500	Subhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

1.7. DATA SCIENCE

Data science, also known as *data-driven science*, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. We shall learn about data mining in short in this unit itself. In fact it is emerging as convergence of various knowledge domains for effective utilisations of various analysis methods for better output of experts in their activities (Fig. 1.4).

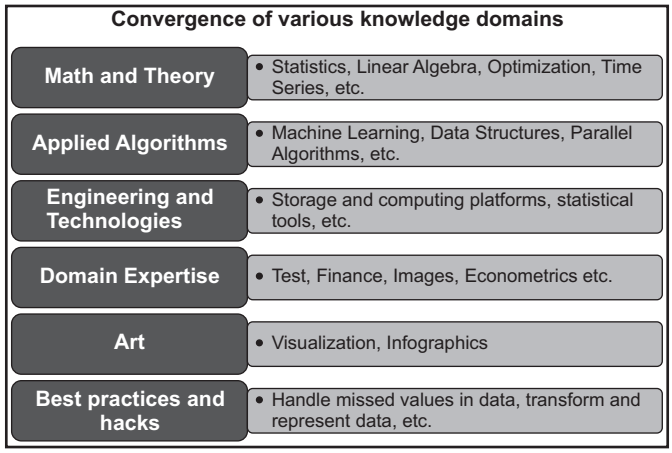


Fig. 1.4. Data science as convergence of various knowledge domains

As such Data Science is one of the recent fields combining *big data, unstructured data* and *combination of statistics* and *analytics* and *business intelligence*. It is a new field that has emerged within the field of Data Management providing understanding of correlation between structured and unstructured data. More accurately, Data Science is the discipline of using quantitative methods from **statistics** and **mathematics** along with **technology** (computers and software) to develop algorithms designed to discover patterns, predict outcomes, and find optimal solutions to complex problems. Nowadays, data scientists are in great demand as they can transform unstructured data into actionable insights, helpful for businesses.

ART OF BIG DATA SCIENCE ANALYTICS

About the Book

This version of the book on the analytics of big data illustrates various emerging technologies available on various distinct platforms. The reader will find valuable and substantial information on the following- Data warehousing and mining technologies dealing with Big Data, Hadoop Ecosystem, Tableau data visualization software, a platform on Google Cloud. Two Apache open-source distributed products have also been discussed. Kafka—an event streaming platform, Storm—a real-time computation system and the recent data platform on Google Cloud, used by whatsapp *i.e. Qubole* are also explained. Alongwith these, this book provides information on two data warehousing applications-*Presto* (including PrestoDB and PrestoSQL/Trino) and Teradata Enterprise access for Hadoop.

Following applications also find place in the book:
SAP HANA, a platform for ERP (Enterprise Resource Planning) software, *InfoSphere BigInsights* 2.1.2, Amazon Elastic MapReduce Hadoop Distribution, Hewlett Packard Enterprise's (HPE), a big data platform- *Vertica* Statistical Analysis System (SAS) *viz.* one of the best tools for creating statistical modelling used by data analysts.

Contents

KHANNA PUBLISHERS®
ISO 9001:2015
4575/15, Onkar House, Opp. Happy School,
Ground Floor, Daryaganj, New Delhi-110002
Phones: 011-45033819, 9811541460
E-mail: contactus@khannapublishers.in

Website:
www.khannapublishers.in

9 789392 549519