SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY

TOPICS

UNIT - I Introduction to Big Data: Big Data and its Importance – Four V's of Big Data – Drivers for Big Data – Introduction to Big Data Analytics – Big Data Analytics applications.

Introduction to Big Data:

What is Big Data?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time.

It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

What is an Example of Big Data?

Following are some of the Big Data examples-

The New York Stock Exchange is an example of Big Data that generates about *one terabyte* of new trade data per day.

Social Media

The statistic shows that *500+terabytes* of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

A single Jet engine can generate *10+terabytes* of data in *30 minutes* of flight time. With many thousand flights per day, generation of data reaches up to many *Petabytes.*

Types of data:

The digital data is divided into three types

1.**unstructured data:** This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.

Forexample:

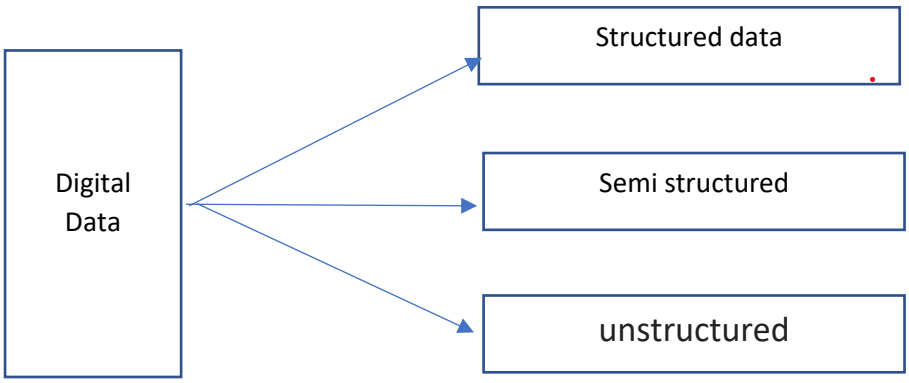
chatrooms,memos,power point presentation,images,videos,letters,researches,white papers,body of an email.

Semi-structured data:

This is the data which does not conform to a data model but has some structure.how ever it is not in the form which can be used easily by a computer program.

Examples:E-mails,XML,markup languages,HTML

Structured data: This is the data which is in the organized form. i.e in the form of rows and columns and can be easily used by a computer programs. Relationships exist between the entities of data, such as classes and their objects.Data stored in data bases is an example of structured data.



Structured data:

The data when it conforms to the schema/structure we say it is structured data.

- **Structured data** is generally tabular data that is represented by columns and rows in a database.
- Databases that hold tables in this form are called ***relational databases***.
- The mathematical term *"relation"* specify to a formed set of data held as a table.
- In structured data, all row in a table has the same set of columns.
- SQL (Structured Query Language) programming language used for structured data.

Table 1.1 A relation/table with rows and columns

	Column 1	Column 2	Column 3	Column 4
Row 1				

Table 1.2 Schema of an "Employee" table in a RDBMS such as Oracle

Column Name	Data Type	Constraints
EmpNo	Varchar(10)	PRIMARY KEY
EmpName	Varchar(50)	
Designation	Varchar(25)	NOT NULL
DeptNo	Varchar(5)	
ContactNo	Varchar(10)	NOT NULL

Table 1.3 Sample records in the "Employee" table

EmpNo	EmpName	Designation	DeptNo	ContactNo
E101	Allen	Software Engineer	D1	0999999999
E102	Simon	Consultant	D1	0777777777

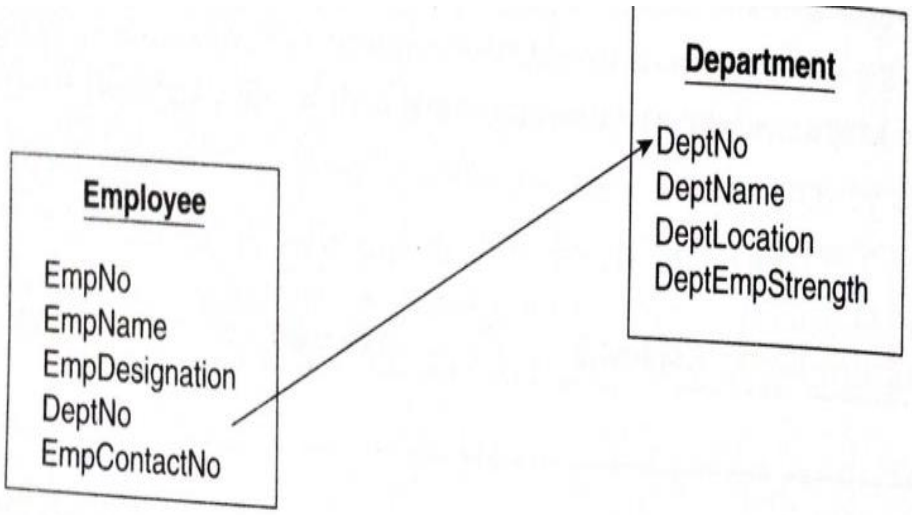


Figure 1.3 Relationship between "Employee" and "Department" tables.

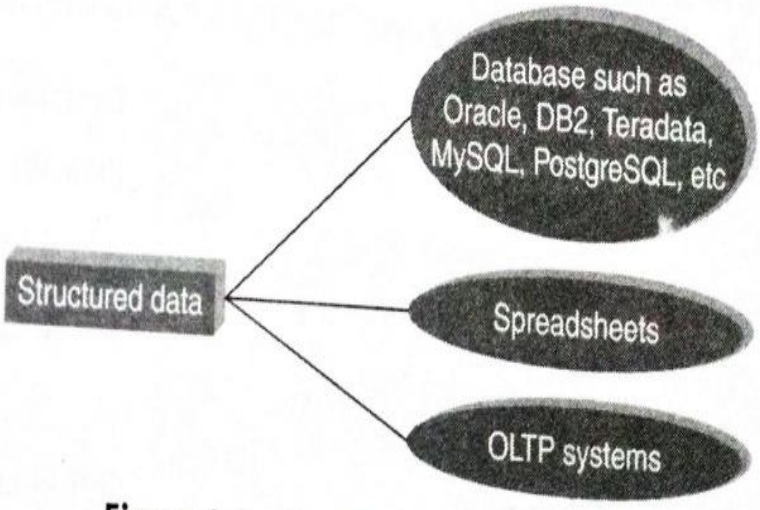


Figure 1.4 Sources of structured data.

1.1.1.2 Ease of Working with Structured Data

Structured data provides the ease of working with it. Refer Figure 1.5. The ease is with respect to the following:

1. **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
2. **Security:** How does one ensure the security of information? There are available staunch encryption and tokenization solutions to warrant the security of information throughout its lifecycle. Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.
3. **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.
4. **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server (increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.).
5. **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction. Given next is a quick explanation of the ACID properties:
   - ***Atomicity:*** A transaction is atomic, means that either it happens in its entirety or none of it at all.
   - ***Consistency:*** The database moves from one consistent state to another consistent state. In other words, if the same piece of information is stored at two or more places, they are in complete agreement.
   - ***Isolation:*** The resource allocation to the transaction happens such that the transaction gets the impression that it is the only transaction happening in isolation.
   - ***Durability:*** All changes made to the database during a transaction are permanent and that accounts for the durability of the transaction.

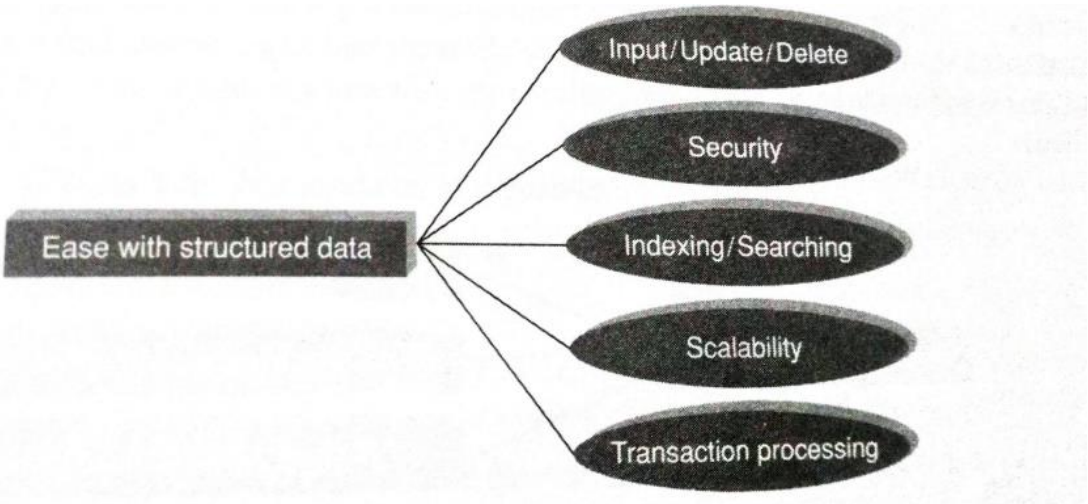


Figure 1.5 Ease of working with structured data.

Sources of structured data:

The sources from where the data is generated is RDBMS,oracle,DB2,Microsoft Sql server,Teradata,Mysql and OLTP systems.

Semi structured data:

The semi structured data is also referred to as self describing tags.

It uses tags to segregate the semantic elements.

Sources of semi structured data:

The sources of semi structured are

XML-Extensible mark up language

JSON-Java script object Notation.

An example of HTML as follows

<html>

<head>

<title>place your title here

</head>

<body bgcolor=”FFFFFF”

</html>

Sample JSON document

{

_id:9

Booktitle:Fundamentals of Business Analyticts’

Author Name:”Seema acharya”

Publisher:”Wiley India”

Year of Publication:”2011”

}

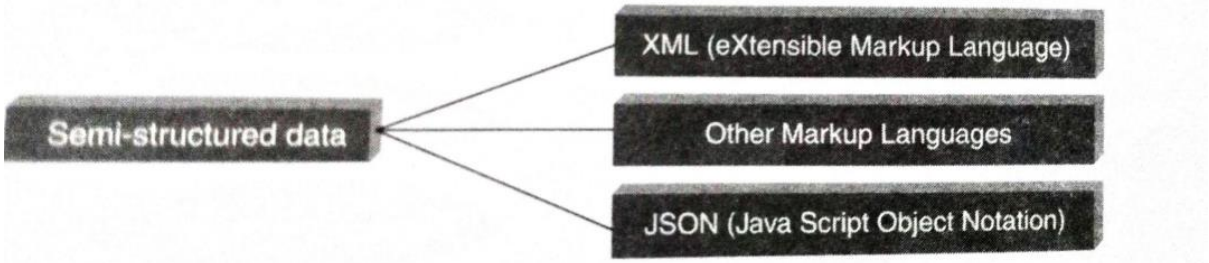


Figure 1.7 Sources of semi-structured data.

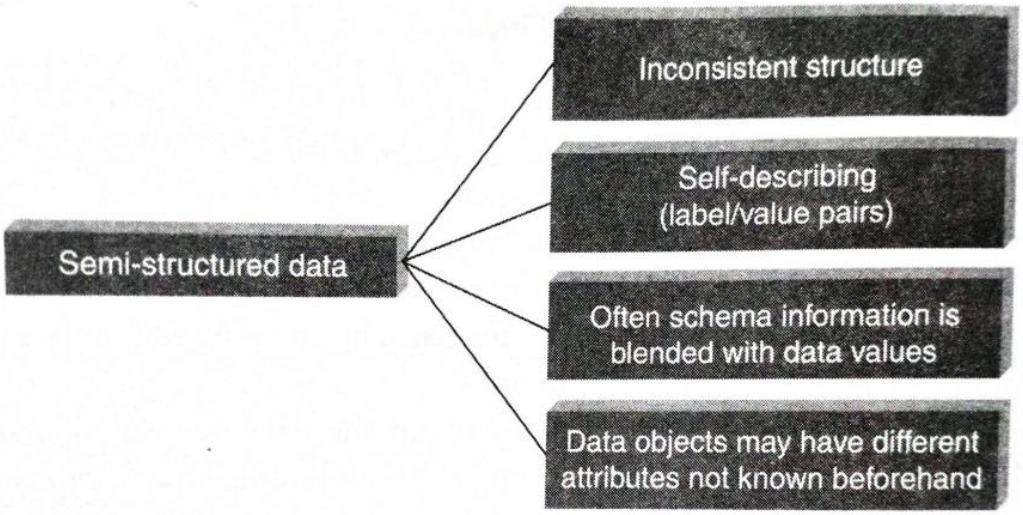


Figure 1.6 Characteristics of semi-structured data.

Unstructured data:

Unstructured data does not conform to any predefined model.

Sources of Unstructured data:

Table 1.4 Few examples of disparate unstructured data

Twitter message	Feeling miffed ☹. Victim of twishing.
Facebook post	LOL. C ya. BFN
Log files	127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I; Nav)"
Email	Hey Joan, possible to send across the first cut on the Hadoop chapter by Friday EOD or maybe we can meet up over a cup of coffee. Best regards, Tom

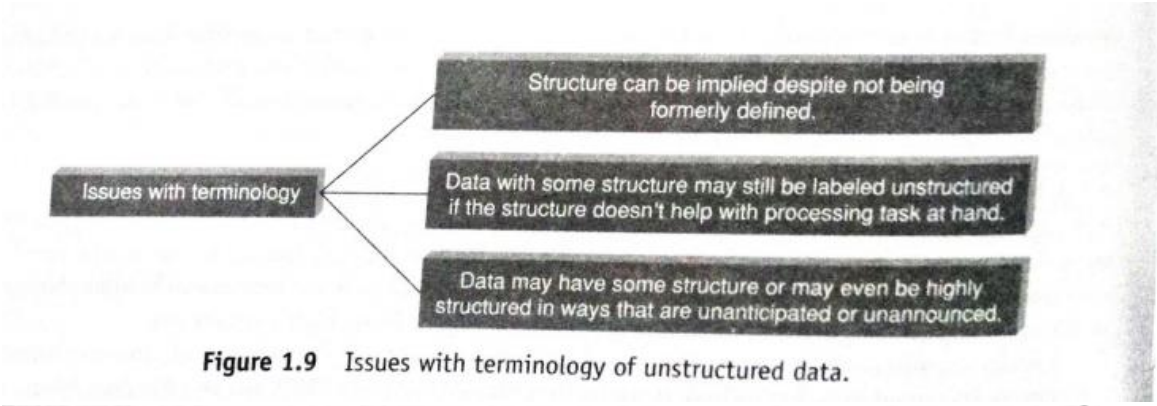


Figure 1.9 Issues with terminology of unstructured data.

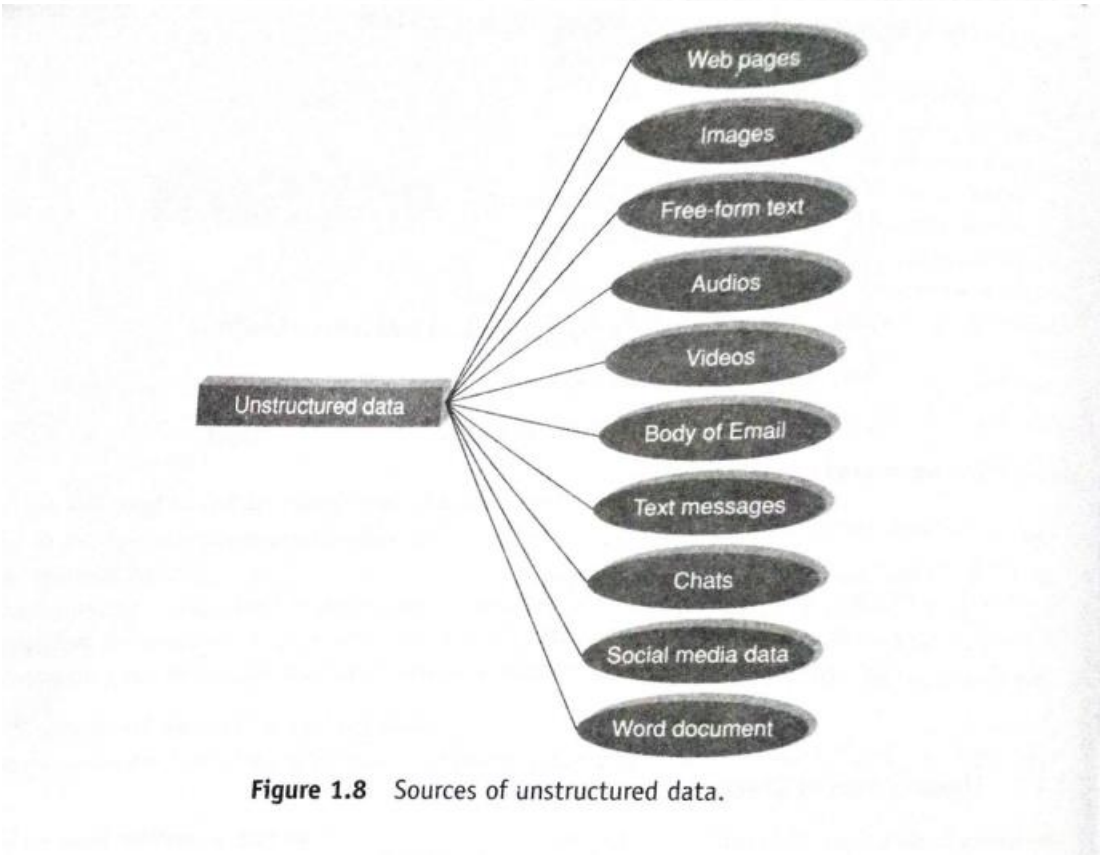


Figure 1.8 Sources of unstructured data.

How to deal with unstructured data:

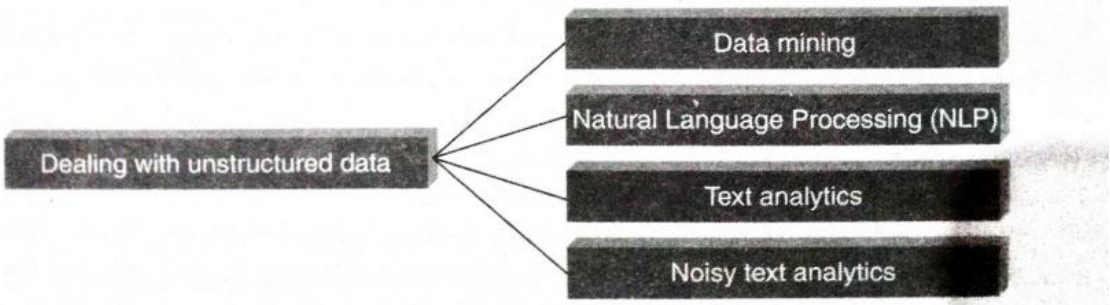


Figure 1.11 Dealing with unstructured data.

The following techniques are used to find patterns in or interpret unstructured data:

1. **Data mining:** First, we deal with large data sets. Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables. It is the analysis step of the "knowledge discovery in databases" process.

   Few popular data mining algorithms are as follows:

   - ***Association rule mining:*** It is also called "market basket analysis" or "affinity analysis". It is used to determine "What goes with what?" It is about when you buy a product, what is the other product that you are likely to purchase with it. For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.
   - ***Regression analysis:*** It helps to predict the relationship between two variables. The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.

- ***Collaborative filtering:*** It is about predicting a user's preference or preferences based on the preferences of a group of users. For example, take a look at Table 1.5. We are looking at predicting whether User 4 will prefer to learn using videos or is a textual learner depending on one or a couple of his or her known preferences. We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

2. **Text analytics or text mining:** Compared to the structured data stored in relational databases, text is largely unstructured, amorphous, and difficult to deal with algorithmically. Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text. It includes tasks such as text categorization, text clustering, sentiment analysis, concept/entity extraction, etc.
3. **Natural language processing (NLP):** It is related to the area of human computer interaction. It is about enabling computers to understand human or natural language input.
4. **Noisy text analytics:** It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc. The noisy unstructured data usually comprises one or more of the following: Spelling mistakes, abbreviations, acronyms, non-standard words, missing punctuation, missing letter case, filler words such as "uh", "um", etc.
5. **Manual tagging with metadata:** This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
6. **Part-of-speech tagging:** It is also called POS or POST or grammatical tagging. It is the process of reading text and tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective", etc.
7. **Unstructured Information Management Architecture (UIMA):** It is an open source platform from IBM. It is used for real-time content analytics. It is about processing text and other unstructured data to find latent meaning and relevant relationship buried therein. Read up more on UIMA at the link: http://www.ibm.com/developerworks/data/downloads/uima/

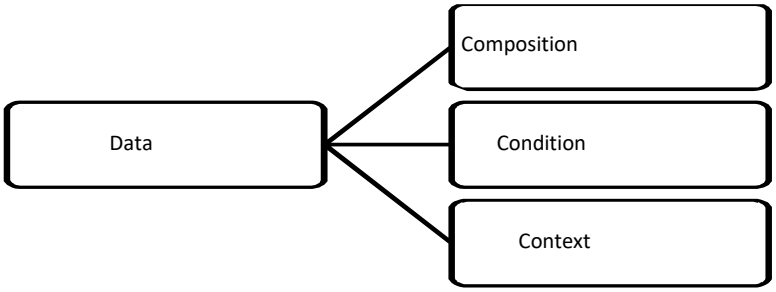
Characteristics of Data:

1.Composition: The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

2.Condition: The condition of data deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"

3.Context: The context of data deals with "Where has this data been generated?" "Whywas this data generated?" How sensitive is this data?" "What are the events associated with this data?" and so on.

Small data (data as it existed prior to the big data revolution) is about certainty. It is about known data sources; it is about no major changes to the composition or context of data.



Evolution of Big Data:

1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IOT) have led to anonslaught of structured, unstructured, and multimedia data. Refer Table 1.1.

	DATA GENERATIONAND STORAGE	DATA UTILZATION	DATA DRIVEN
COMPLEX AND UNSTRUCTURED			Structured data, Unstructured data, Multimedia data
COMPLEX AND RELATIONAL		Relational databases: Data-intensive applications	
PRIMITIVE AND STRUCTURED	Mainframes: Basic data storage		
	1970 and before	**Relational (1980 and 1990s)**	**2000 and beyond**

Big Data and its Importance

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in

its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

Big Data importance doesn't revolve around the amount of data a company has but lies in the fact that how the company utilizes the gathered data.

Every company uses its collected data in its own way. More effectively the company uses its data, more rapidly it grows.

By analysing the big data pools effectively the companies can get answers to :

Cost Savings :

o Some tools of Big Data like Hadoop can bring cost advantages to business when large amounts of data are to be stored.

o These tools help in identifying more efficient ways of doing business.

Time Reductions :

o The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately.

o This helps us to make quick decisions based on the learnings.

Understand the market conditions :

o By analyzing big data we can get a better understanding of current market conditions.

o For example: By analyzing customers' purchasing behaviours, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

Control online reputation :

o Big data tools can do sentiment analysis.

o Therefore, you can get feedback about who is saying what about your company.

o If you want to monitor and improve the online presence of your business, then big data tools can help in all this.

Using Big Data Analytics to Boost Customer Acquisition(purchase) and Retention :

o The customer is the most important asset any business depends on.

o No single business can claim success without first having to establish a solid customer base.

o If a business is slow to learn what customers are looking for, then it is very likely to deliver poor quality products.

o The use of big data allows businesses to observe various customer-related patterns and trends.

Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights :

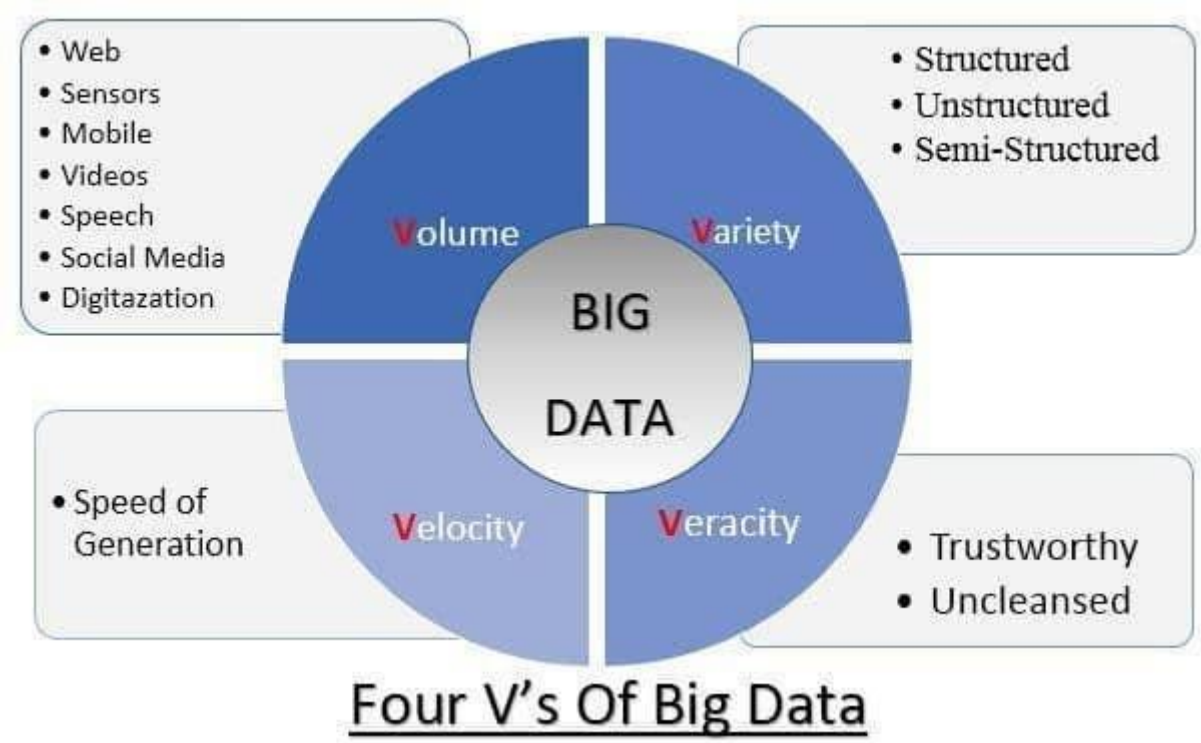
o Big data analytics can help change all business operations.

o Like the ability to match customer expectations, changing company's product line, etc.

o And ensuring that the marketing campaigns are powerful

Four v's Of Big Data

In recent years, Big Data was defined by the "*3Vs*" but now there is "6*Vs*" of Big Data which are also termed as the characteristics of Big Data as follows:



Four V's Of Big Data

1. Volume:

Table 2.2 Growth of data

Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

is unstructured data; a CCTV coverage, à weather forecast report is unstructured data too. Refer Figure 2.7 for the sources of big data.

1. **Typical internal data sources:** Data present within an organization's firewall. It is as follows:
   - ***Data storage:*** File systems, SQL (RDBMSs – Oracle, MS SQL Server, DB2, MySQL, PostgreSQL, etc.), NoSQL (MongoDB, Cassandra, etc.), and so on.
   - ***Archives:*** Archives of scanned documents, paper archives, customer correspondence records, patients' health records, students' admission records, students' assessment records, and so on.

2. **External data sources:** Data residing outside an organization's firewall. It is as follows:
   - ***Public Web:*** Wikipedia, weather, regulatory, compliance, census, etc.
3. **Both (internal + external data sources)**
   - ***Sensor data:*** Car sensors, smart electric meters, office buildings, air conditioning units, refrigerators, and so on.
   - ***Machine log data:*** Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
   - ***Social media:*** Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
   - ***Business apps:*** ERP, CRM, HR, Google Docs, and so on.
   - ***Media:*** Audio, Video, Image, Podcast, etc.
   - ***Docs:*** Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exa Bytes of data.

2. Velocity:

2.5.2 Velocity

We have moved from the days of batch processing (remember our payroll applications) to real-time processing.

Batch → Periodic → Near real time → Real-time processing

- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like 'velocity'.

- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

3. Variety:

- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.
    - **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
    - **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.
    - **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 6V's.

6. Variability:

- How fast or, available data that extent is the structure of your data is changing ?.
- How often does the meaning or shape of your data!to ?change?.
- example : if you are eating same ice-cream daily and the taste just keep changing.

Drivers for Big Data

Big Data has quickly risen to become one of the most desired topics in the industry.

The main business drivers for such rising demand for Big Data Analytics are :

1. The digitization of society
2. The drop in technology costs
3. Connectivity through cloud computing
4. Increased knowledge about data science
5. Social media applications
6. The rise of Internet-of-Things(IoT)

Example: A number of companies that have Big Data at the core of their strategy like :

Apple, Amazon, Facebook and Netflix have become very successful at the beginning of the 21st century.

1. The digitization of society

Big Data is largely consumer driven and consumer oriented. Most of the data in the world is generated by consumers, who are nowadays 'always-on'.

Most people now spend 4-6 hours per day consuming and generating data through a variety of devices and (social) applications.

With every click, swipe or message, new data is created in a database somewhere around the world.

Because everyone now has a smartphone in their pocket, the data creation sums to incomprehensible amounts.

Some studies estimate that 60% of data was generated within the last two years, which is a good indication of the rate with which society has digitized.

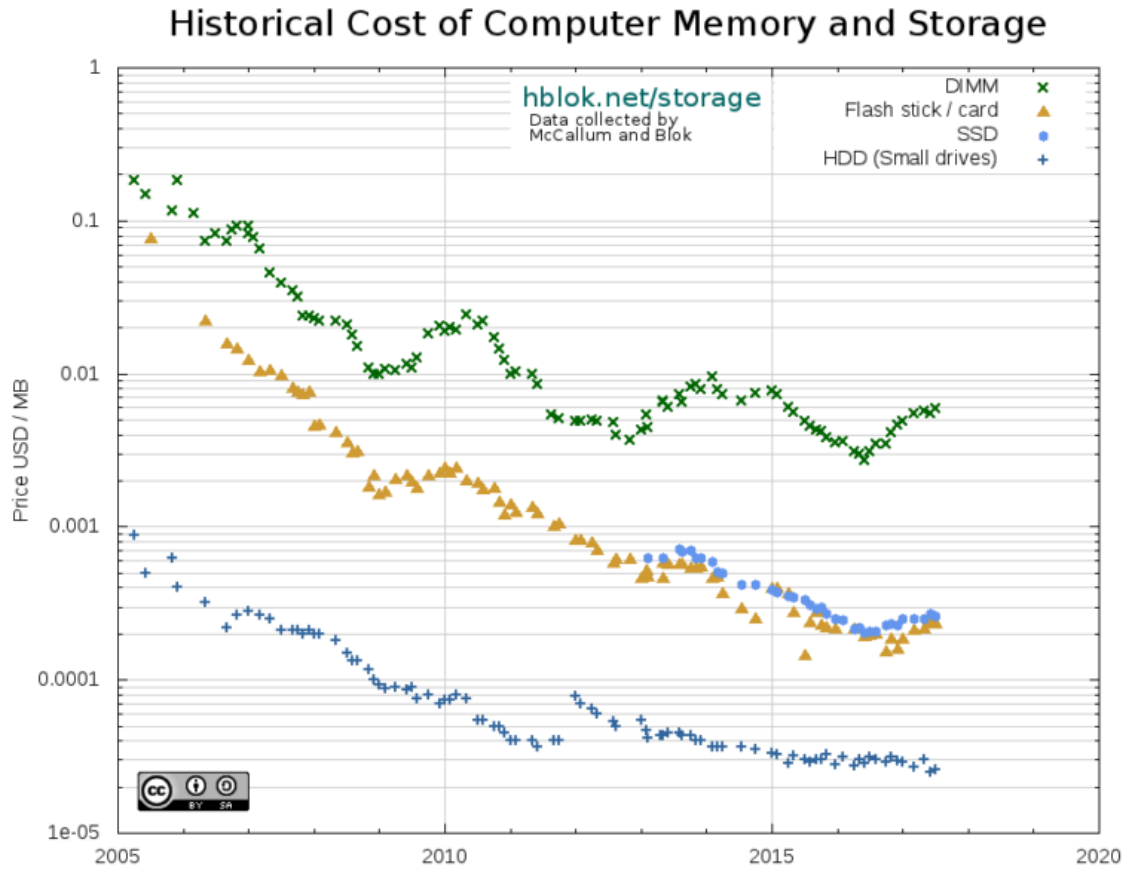
2. The drop in technology costs

Technology related to collecting and processing massive quantities of diverse (high variety) data has become increasingly more affordable.

The costs of data storage and processors keep declining, making it possible for small businesses and individuals to become involved with Big Data.

For storage capacity, the often-cited Moore's Law still holds that the storage density (and therefore capacity) still doubles every two years.

The plummeting of technology costs has been depicted in the figure below.



Besides the plummeting of the storage costs, a second key contributing factor to the affordability of Big Data has been the development of open source Big Data software frameworks.

The most popular software framework (nowadays considered the standard for Big Data) is Apache Hadoop for distributed storage and processing.

Due to the high availability of these software frameworks in open sources, it has become increasingly inexpensive to start Big Data projects in organizations.

3. Connectivity through cloud computing

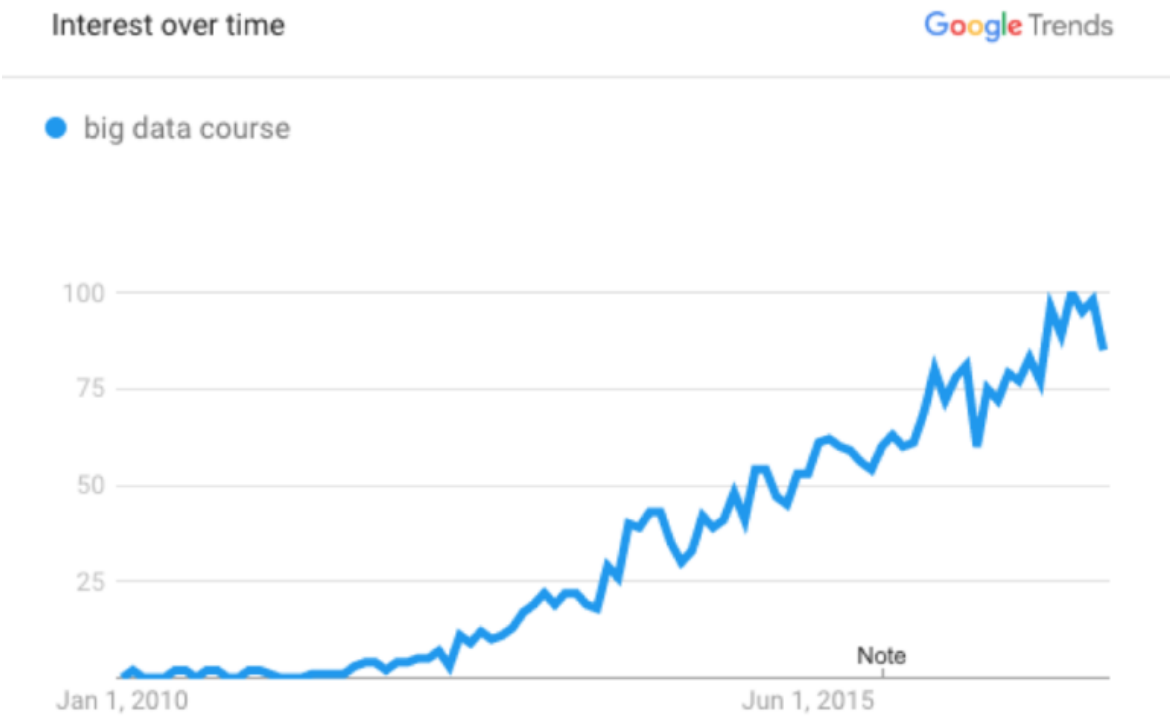
Cloud computing environments (where data is remotely stored in distributed storage systems) have made it possible to quickly scale up or scale down IT infrastructure and facilitate a pay-as-you-go model.

This means that organizations that want to process massive quantities of data (and thus have large storage and processing requirements) do not have to invest in large quantities of IT infrastructure.

Instead, they can license the storage and processing capacity they need and only pay for the amounts they actually used. As a result, most of Big Data solutions leverage the possibilities of cloud computing to deliver their solutions to enterprises.

4. Increased knowledge about data science

In the last decade, the term data science and data scientist have become tremendously popular. In October 2012, Harvard Business Review called the data scientist "sexiest job of the 21st century" and many other publications have featured this new job role in recent years. The demand for data scientist (and similar job titles) has increased tremendously and many people have actively become engaged in the domain of data science.



As a result, the knowledge and education about data science has greatly professionalized and more information becomes available every day. While statistics and data analysis mostly remained an academic field previously, it is quickly becoming a popular subject among students and the working population.

5. Social media applications

Everyone understands the impact that social media has on daily life. However, in the study of Big Data, social media plays a role of paramount importance. Not only because of the sheer volume of data that is produced everyday through platforms such as Twitter, Facebook, LinkedIn and Instagram, but also because social media provides nearly real-time data about human behavior.

Social media data provides insights into the behaviors, preferences and opinions of 'the public' on a scale that has never been known before. Due to this, it is immensely valuable to anyone

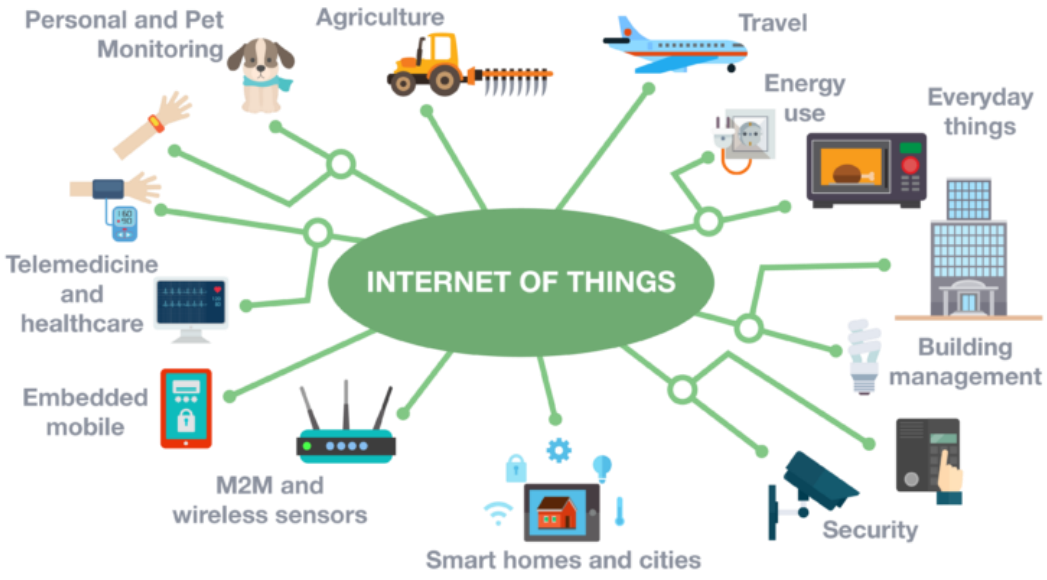
who is able to derive meaning from these large quantities of data. Social media data can be used to identify customer preferences for product development, target new customers for future purchases, or even target potential voters in elections. Social media data might even be considered one of the most important business drivers of Big Data.

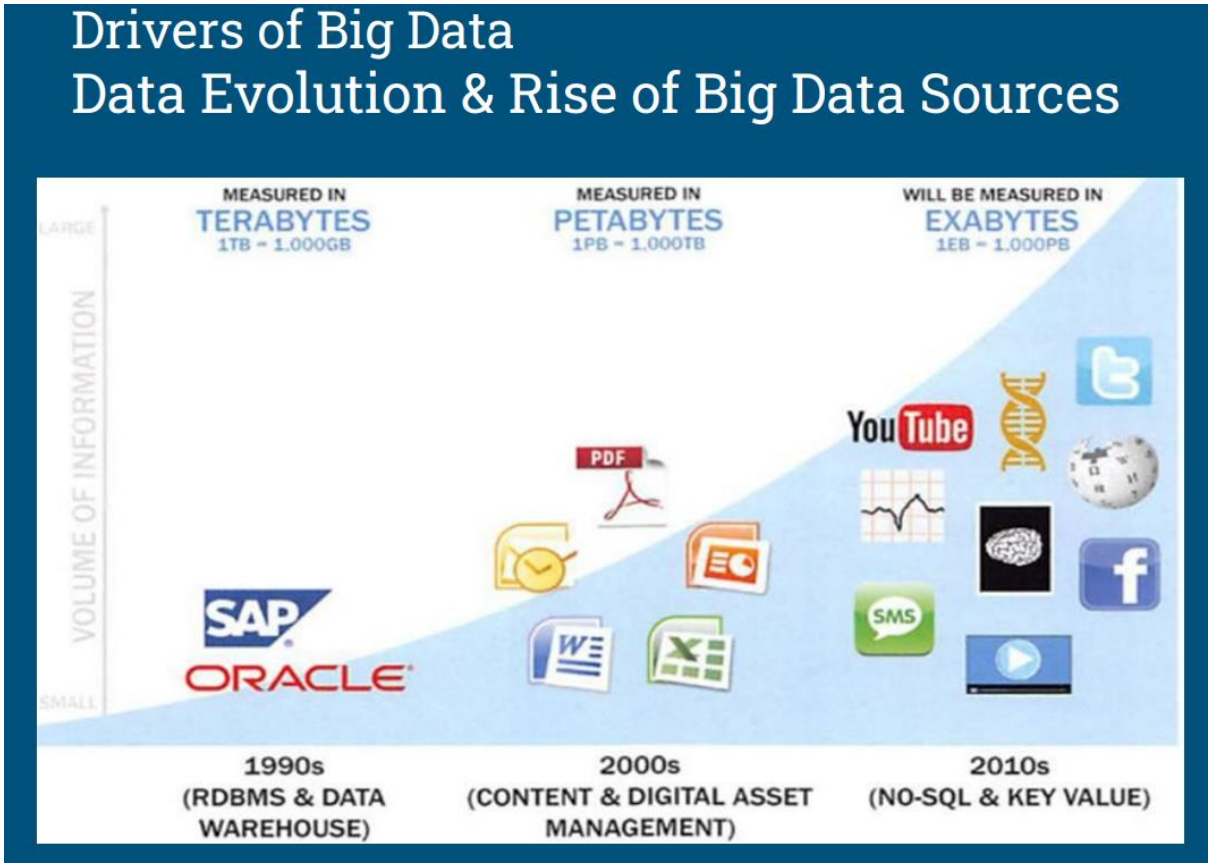
6. The upcoming internet of things (IoT)

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and network connectivity which enables these objects to connect and exchange data.

It is increasingly gaining popularity as consumer goods providers start including ‘smart’ sensors in household appliances. Whereas the average household in 2010 had around 10 devices that connected to the internet, this number is expected to rise to 50 per household by 2020.

Examples of these devices include thermostats, smoke detectors, televisions, audio systems and even smart refrigerators.





● Medical information, such as diagnostic imaging

● Photos and video footage uploaded to the World Wide Web

● Video surveillance, such as the thousands of video cameras across a city

● Mobile devices, which provide geospatial location data of the users

● Metadata about text messages, phone calls, and application usage on smart phones

● Smart devices, which provide sensor-based collection of information from smart

● Non traditional IT devices, including the use of RFID readers,

GPS navigation systems, and seismic processing.

These are the multiple sources where the data can be generated from multiple sources.

Challenges of Big Data

The major challenges associated with big data are as follows –

- Capturing data
- Curation
- Storage
- Searching
- Sharing
- Transfer
- Analysis
- Presentation

1.Data is growing in an exponential rate.Most of the data have been generated in the last 2-3 years.

Introduction to Big Data Analyticts

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging.

As this new digital economy continues to evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing to test the outcomes of machine learning techniques.

Other vendors offer added value by repackaging open source tools in a simpler way and bringing the tools to market. Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

Data devices and the "Sensornet" gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data.

For example, consider someone playing an online video game through a PC, game console, or smartphone. In this case, the video game provider captures data about the skill and levels attained by the player. Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can fine-tune the difficulty of the game, suggest other related games that would most likely interest the user, and offer additional equipment and enhancements for the character based on the user's age, gender, and interests. This information may get stored locally or uploaded to the game provider's cloud to analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical profiles of specific kinds of users.

Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location. This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.

Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of products purchased together. Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.

Data collectors include sample entities that collect data from the device and users.

Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content

Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips

Data aggregators make sense of the data collected from the various entities from the "SensorNet" or the "Internet of

Things." These organizations compile data from the devices and usage patterns collected by government agencies, retail stores, and websites. In turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

Data users and buyers

These groups directly benefit from the data collected and aggregated by others within the data value chain.

Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit. To provide input for this analysis, retail banks may purchase data from a data aggregator. This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodeling projects.

Obtaining data from these various sources and aggregators will enable a more targeted marketing campaign, which would have been more challenging before Big Data due to the lack of information or high-performing technologies.

Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analyzing related blogs and online comments. Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects first and how it moves, based on which geographic areas are tweeting about it or discussing it via social media.

A field to analyze and to extract information about the big data involved in the business or the data world so that proper conclusions can be made is called big data Analytics.

These conclusions can be used to predict the future or to forecast the business.

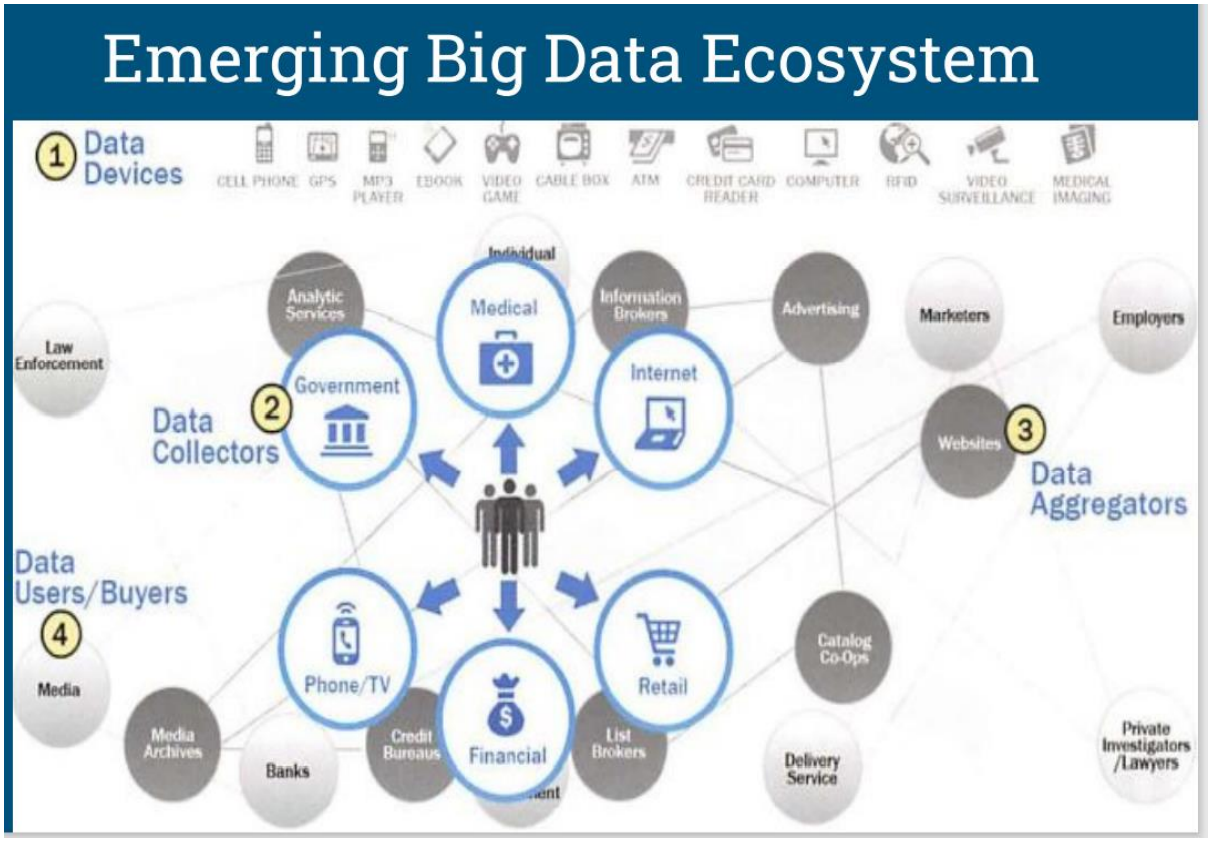
Also, this helps in creating a trend about the past.

Skilled professionals in statistics and engineering with domain knowledge are needed in the analysis of big data as the data is huge, and analysis needs proper determination and skillset.

This data is more complex that it cannot be dealt with with traditional methods of analysis.

We produce a massive amount of data each day, whether we know about it or not. Every click on the internet, every bank transaction, every video we watch on YouTube, every email we send, every like on our Instagram post makes up data for tech companies.

With such a massive amount of data being collected, it only makes sense for companies to use this data to understand their customers and their behavior better. This is the reason why the popularity of Data Science has grown manifold over the last few years.



A big data platform is a type of IT solution that combines the features and capabilities of several big data applications and utilities within a single solution, this is then used further for managing as well as analyzing Big Data.

It focuses on providing its users with efficient analytics tools for massive datasets.

The users of such platforms can custom build applications according to their use case like to calculate customer loyalty (E-Commerce user case), and so on.

Goal: The main goal of a Big Data Platform is to achieve: Scalability, Availability, Performance, and Security.

Example: Some of the most commonly used Big Data Platforms are :

- Hadoop Delta Lake Migration Platform

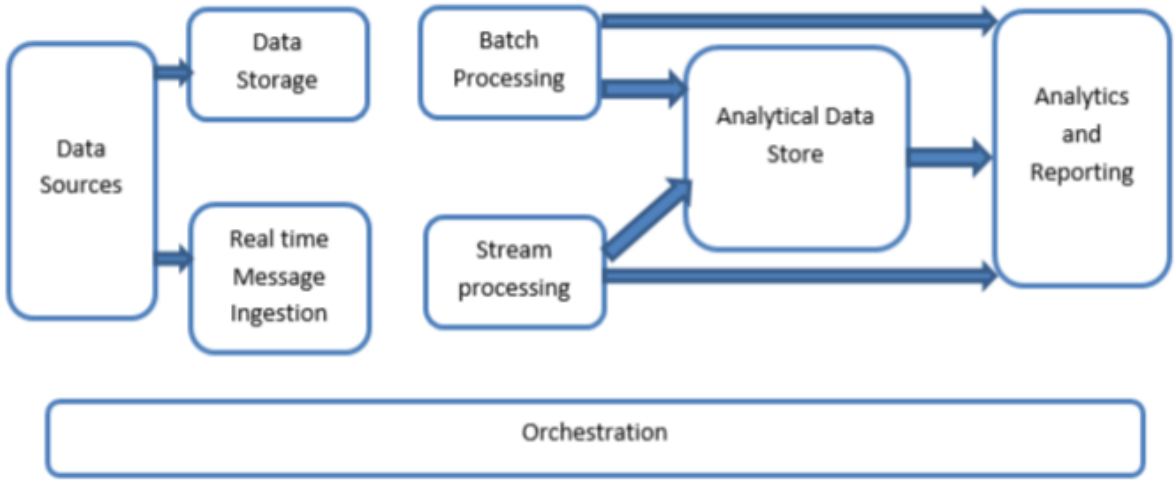
- Data Catalog Platform
- Data Ingestion Platform
- IoT Analytics Platform
- Drivers for Big Data
- Big Data has quickly risen to become one of the most desired topics in the industry.
- The main business drivers for such rising demand for Big Data Analytics are :

1. The digitization of society
2. The drop in technology costs
3. Connectivity through cloud computing
4. Increased knowledge about data science
5. Social media applications
6. The rise of Internet-of-Things(IoT)

Example: A number of companies that have Big Data at the core of their strategy like :

Apple, Amazon, Facebook and Netflix have become very successful at the beginning of the 21st century.

- **Big Data Architecture :**
- Big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems.



- The big data architectures include the following components:
- **Data sources**: All big data solutions start with one or more data sources.

- Example,
- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT devices.
- **Data storage**: Data for batch processing operations is stored in a distributed file store that can hold high volumes of large files in various formats (also called data lake).
- Example,
- Azure Data Lake Store or blob containers in Azure Storage.
- **Batch processing:** Since the data sets are so large, therefore a big data solution must process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis.
- **Real-time message ingestion:** If a solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing.
- **Stream processing:** After capturing real-time messages, the solution must process them by filtering, aggregating, and preparing the data for analysis. The processed stream data is then written to an output sink. We can use open-source Apache streaming technologies like Storm and Spark Streaming for this.
- **Analytical data store:** Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. Example: Azure Synapse Analytics provides a managed service for large-scale, cloud-based data warehousing.
- **Analysis and reporting:** The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyze the data, the architecture may include a data modelling layer. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts.
- **Orchestration:** Most big data solutions consist of repeated data processing operations, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report. To automate these

workflows, we can use an orchestration technology such as Azure Data Factory.

Applications of Big Data

In today's world big data have several applications, some of them are listed below :

Tracking Customer Spending Habit, Shopping Behavior :

In big retails stores, the management team has to keep data of customer's spending habits, shopping behaviour, most liked product, which product is being searched/sold most, based on that data, the production/collection rate of that product gets fixed.

Recommendation :

By tracking customer spending habits, shopping behaviour, big retail stores provide recommendations to the customers.

Smart Traffic System :

Data about the condition of the traffic of different roads, collected through cameras, GPS devices placed in the vehicle.

All such data are analyzed and jam-free or less jam way, less time taking ways are recommended.

One more profit is fuel consumption can be reduced.

Secure Air Traffic System :

At various places of flight, sensors are present.

These sensors capture data like the speed of flight, moisture, temperature, and other environmental conditions.

Based on such data analysis, an environmental parameter within flight is set up and varied.

By analyzing flight's machine-generated data, it can be estimated how long the machine can operate flawlessly and when it can be replaced/repaired.

Auto Driving Car :

In the various spots of the car camera, a sensor is placed that gathers data like the size of the surrounding car, obstacle, distance from those, etc.

These data are being analyzed, then various calculations are carried out.

These calculations help to take action automatically.

Virtual Personal Assistant Tool :

Big data analysis helps virtual personal assistant tools like Siri, Cortana and Google Assistant to provide the answer to the various questions asked by users. This tool tracks the location of the user, their local time, season, other data related to questions asked, etc.

Analyzing all such data provides an answer.

Example: Suppose one user asks “Do I need to take Umbrella?”The tool collects data like location of the user, season and weather condition at that location, then analyzes these data to conclude if there is a chance of raining, then provides the answer.

IoT :

Manufacturing companies install IOT sensors into machines to collect operational data.

Analyzing such data, it can be predicted how long a machine will work without any problem when it requires repair.

Thus, the cost to replace the whole machine can be saved.

Education Sector Energy Sector :

Online educational courses conducting organization utilize big data to search candidates interested in that course.

If someone searches for a YouTube tutorial video on a subject, then an online or offline course provider organization on that subject sends an ad online to that person about their course.

Media and Entertainment Sector :

Media and entertainment service providing company like Netflix,

Amazon Prime, Spotify do analysis on data collected from their users.

Data like what type of video, music users are watching, listening to most, how long users are spending on site, etc are collected and analyzed to set the next business strategy.

Traditional BI versus Big data

Here are some key differences between traditional BI and big data:

1. **Data sources**: Traditional BI typically relies on structured data from internal systems, while big data can come from a wide range of external and internal sources, including social media, IoT devices, and sensors.
2. **Data volume**: Traditional BI deals with relatively small amounts of data, while big data deals with extremely large and complex sets of data.
3. **Data variety**: Traditional BI typically deals with structured data, while big data can include structured, semi-structured, and unstructured data from various formats such as text, images, videos, and audio.
4. **Data processing**: Traditional BI relies on traditional data processing techniques such as SQL, while big data uses newer technologies such as Hadoop, NoSQL databases, and machine learning to process and analyze the data.
5. **Time frame**: Traditional BI focuses on historical data, while big data can also include real-time data streams.
6. **Scale:** Traditional BI is typically implemented in a small scale, while big data is implemented in large scale and distributed environments.

Traditional Data	Big Data
Traditional data is generated in enterprise level.	Big data is generated outside the enterprise level.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Traditional database system deals with structured data.	Big data system deals with structured, semi-structured,database, and unstructured data.
Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
Data integration is very easy.	Data integration is very difficult.

Traditional Data	Big Data
Normal system configuration is capable to process traditional data.	High system configuration is required to process big data.
The size of the data is very small.	The size is more than the traditional data size.
Traditional data base tools are required to perform any data base operation.	Special kind of data base tools are required to perform any databaseschema-based operation.
Normal functions can manipulate data.	Special kind of functions can manipulate data.
Its data model is strict schema based and it is static.	Its data model is a flat schema based and it is dynamic.
Traditional data is stable and inter relationship.	Big data is not stable and unknown relationship.
Traditional data is in manageable volume.	Big data is in huge volume which becomes unmanageable.
It is easy to manage and manipulate the data.	It is difficult to manage and manipulate the data.
Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc.	Its data sources includes social media, device data, sensor data, video, images, audio etc.

A typical data warehouse often includes the following elements:

- A relational database to store and manage data
- An extraction, loading, and transformation (ELT) solution for preparing the data for analysis
- Statistical analysis, reporting, and data mining capabilities
- Client analysis tools for visualizing and presenting data to business users

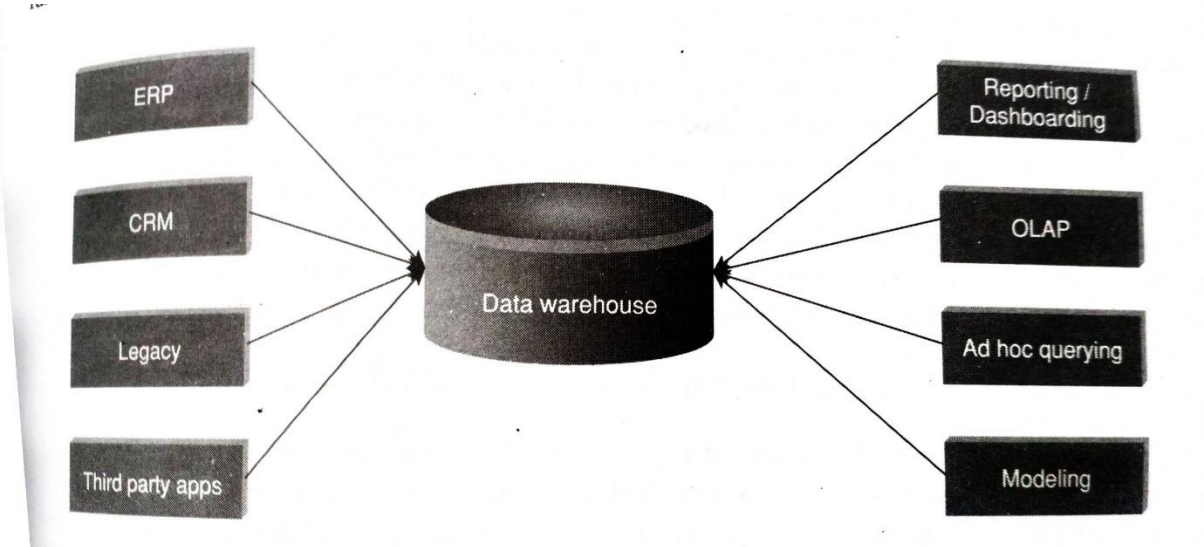


Figure 2.9 A typical data warehouse environment.

A typical Hadoop environment

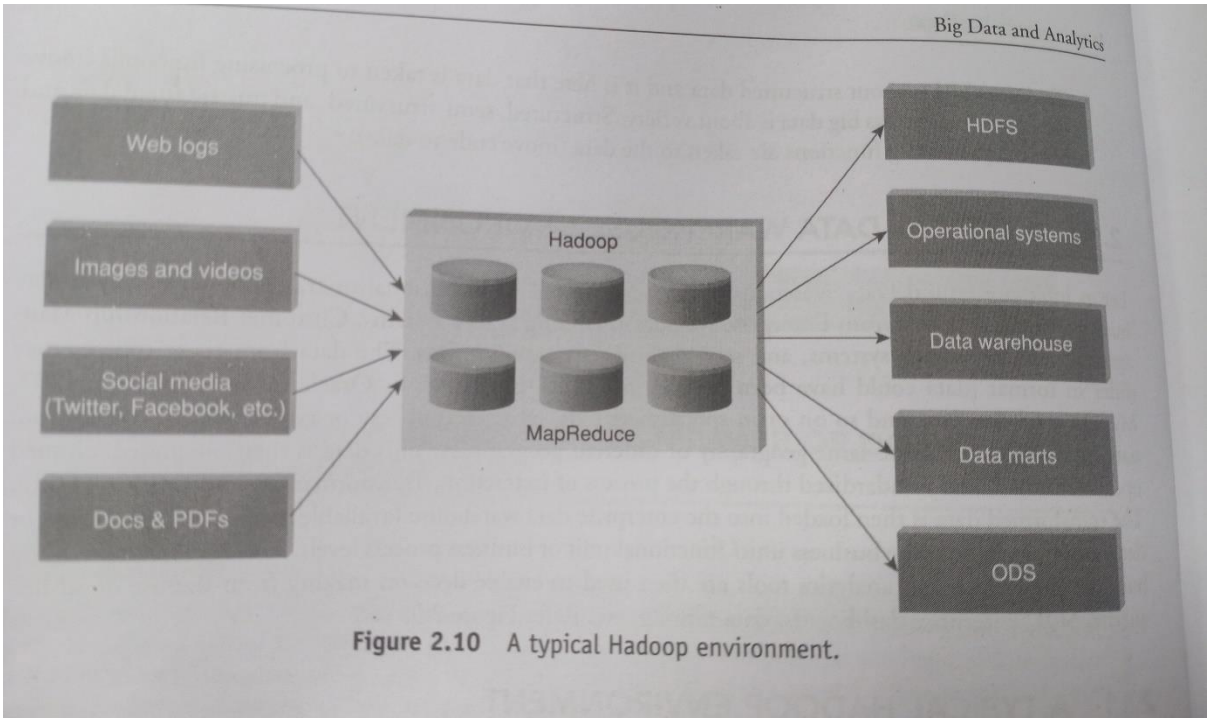


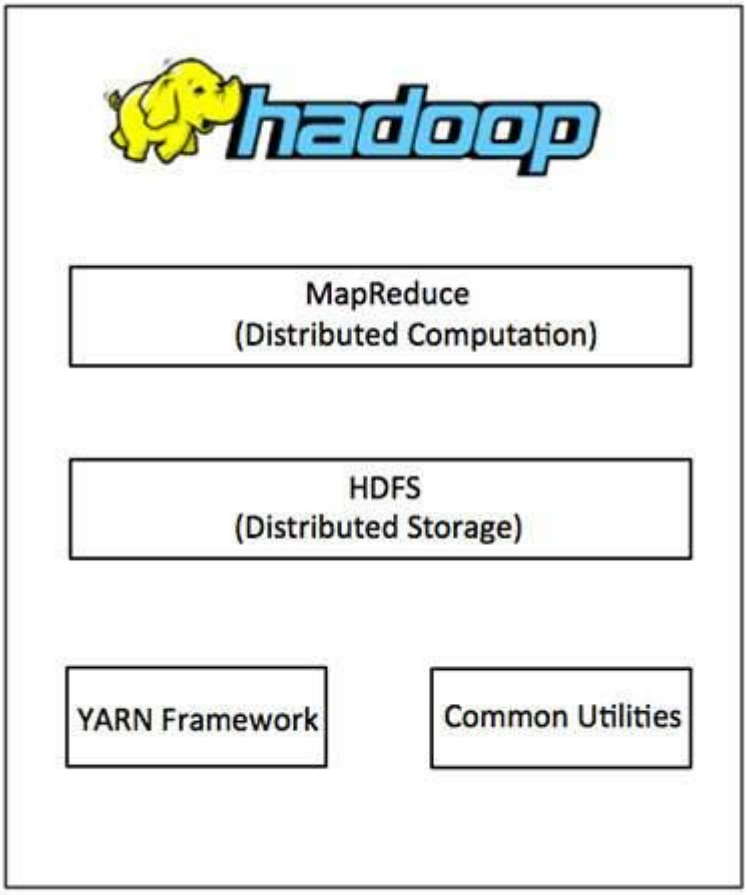
Figure 2.10 A typical Hadoop environment.

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

At its core, Hadoop has two major layers namely –

- Processing/Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed File System).



MapReduce

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However,

the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules –

- **Hadoop Common** – These are Java libraries and utilities required by other Hadoop modules.
- **Hadoop YARN** – This is a framework for job scheduling and cluster resource management.

How Does Hadoop Work?

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs –

- Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
- These files are then distributed across various cluster nodes for further processing.
- HDFS, being on top of the local file system, supervises the processing.
- Blocks are replicated for handling hardware failure.
- Checking that the code was executed successfully.
- Performing the sort that takes place between the map and reduce stages.
- Sending the sorted data to a certain computer.
- Writing the debugging logs for each job.

Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.

- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

A coexistence of data warehouse and Hadoop environment

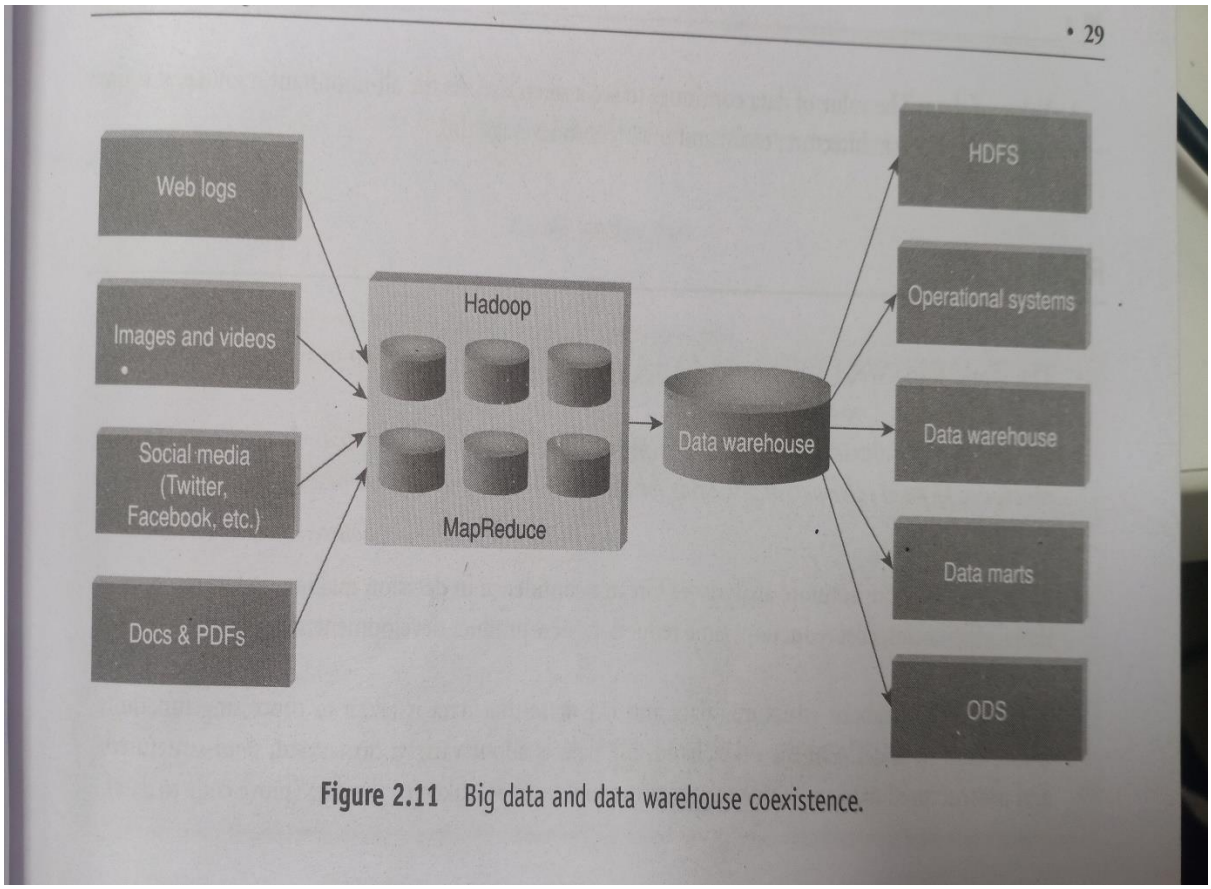


Figure 2.11 Big data and data warehouse coexistence.