INTERNET AND EMERGING TECHNOLOGIES

TOPIC: BIG DATA

TOPIC OUTLINE

- ✓ History of Big Data
- ✓ Definition
- ✓ Types of big data
- ✓ Dimensions/characteristics of big data
- ✓ Why big data is important?
- ✓ Sources of big data
- ✓ Tools for managing big data
- ✓ Applications of big data
- ✓ Challenges of big data
- ✓ Advantages of big data
- ✓ Disadvantages of big data

What is Big Data?

- The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.

The History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

What is Big Data?

Apache Hadoop defined big data as “datasets, which could not be captured, managed, and processed by general computers within an acceptable scope”. **McKinsey Global Institute** defined big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze“. **International Data Corporation (IDC)** defines “big data technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis”.

Academicians define big data as huge size of unstructured data produced by high-performance heterogeneous group of applications that spans from social network to scientific computing applications. The datasets range from a few hundred gigabytes to zetabytes that it is beyond the capacity of existing data management tools to capture, store, manage and analyze.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- ✓ It refers to a massive amount of data that keeps on growing exponentially with time.
- ✓ It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- ✓ It includes data mining, data storage, data analysis, data sharing, and data visualization.
- ✓ The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data

Big data analytic leads to more precise analysis thus helps to bring more accurate decision-making and better performance. Big data are collected either through structured or unstructured data sources (online or offline). Unstructured data can come from social media (Facebook, Instagram, Twitter posts, etc). While, structured data sources can come from internal database of organization. In businesses, both sources are used to understand the patterns of the customers. Indeed, an organization nowadays relies the fact that any data could be analyzed and used to reveal patterns of their customers. In other words, big data will help the organization to understand the behavior of their customers and use it to win a competition.

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.***

Apache Hadoop defined big data as “datasets, which could not be captured, managed, and processed by general computers within an acceptable scope”. **McKinsey Global Institute** defined big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze“. **International Data Corporation (IDC)** defines “big data technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis”. **Academicians** define big data as huge size of unstructured data produced by high-performance heterogeneous group of applications that spans from social network to scientific computing applications. The datasets range from a few hundred gigabytes to zetabytes that it is beyond the capacity of existing data management tools to capture, store, manage and analyze.

One definition about big data is that it refers to sets of data where their size is too advanced, too complex for typical data software tools to use their abilities for analysis or management etc.

Big data ranges from a few dozen terabytes to many petabytes of data. Furthermore, Big Data can be differentiated with 3 key differences i.e. the 3 V's. They are Volume, in which organizations collect data from different sources with different sizes of data, 2.5 Exabyte are created each day. Velocity, where many applications priorities the speed of data creation over its volume which allows real time information to be processed and Variety, where big data can take in different forms such as messages, updates and even pictures.

Big data is defined as a massive amount of data, very quickly in processing from many different forms to support decision making. Therefore, it is widely known with the volume, velocity and variety. Other definition is a massive volume of structured and unstructured data that is gathered and analyzed through new methods that can produce value for the organization.

Types of Big Data

Fig 1.12 Connects different types of data

a) Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

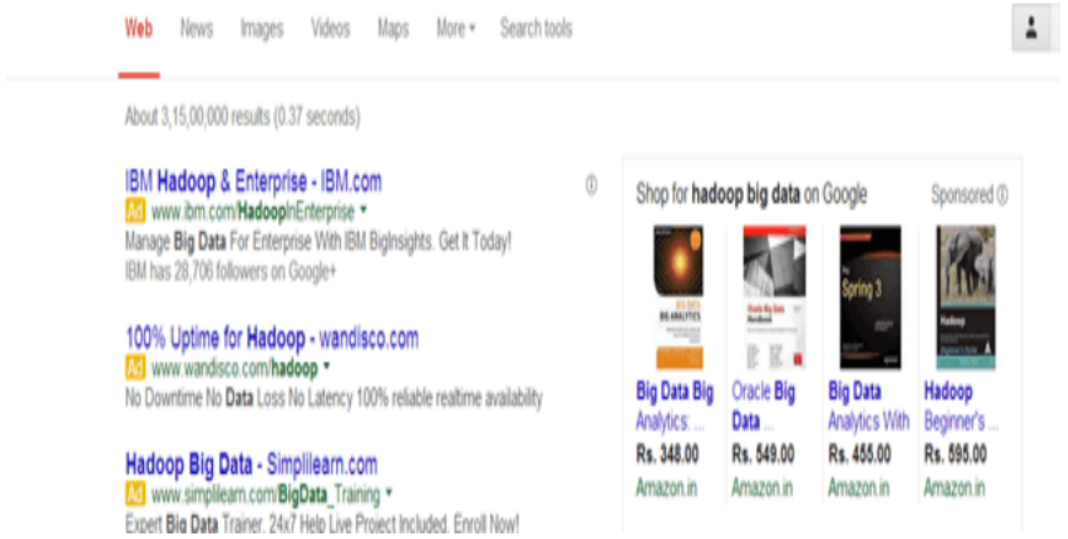
- **Examples Of Structured Data**

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

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

b) Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

The output returned by 'Google Search'



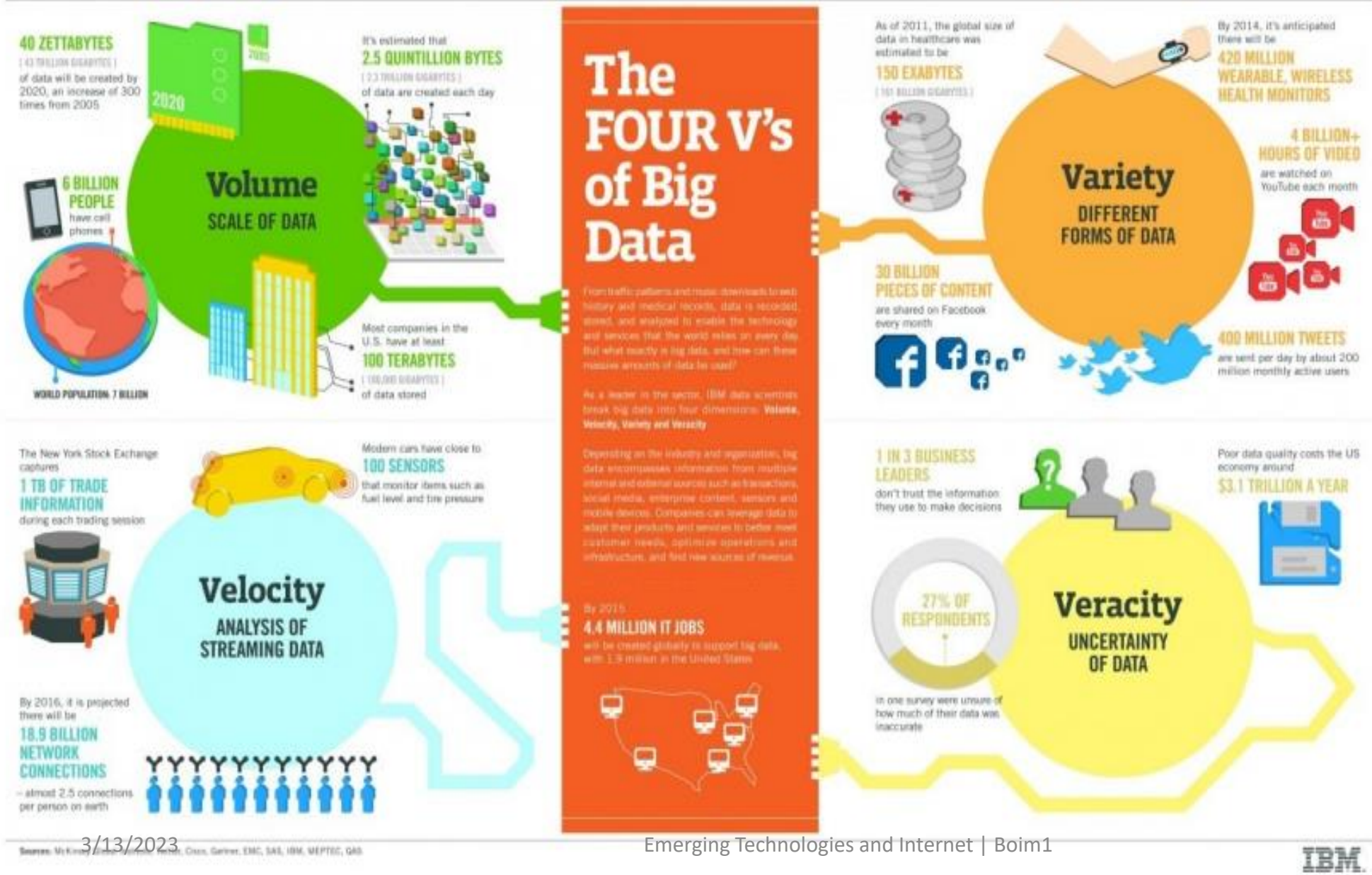
c) Semi-structured Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data.

- **Examples Of Semi-structured Data**

  Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</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>
```

Dimensions/Characteristics of Big Data



Dimensions/Characteristics of Big Data

Back in 2001, Gartner analyst Doug Laney listed the 3 ‘V’s of Big Data – Variety, Velocity, and Volume. These characteristics, isolated, are enough to know what big data is. Let’s look at them in depth:

a) Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data

b) Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

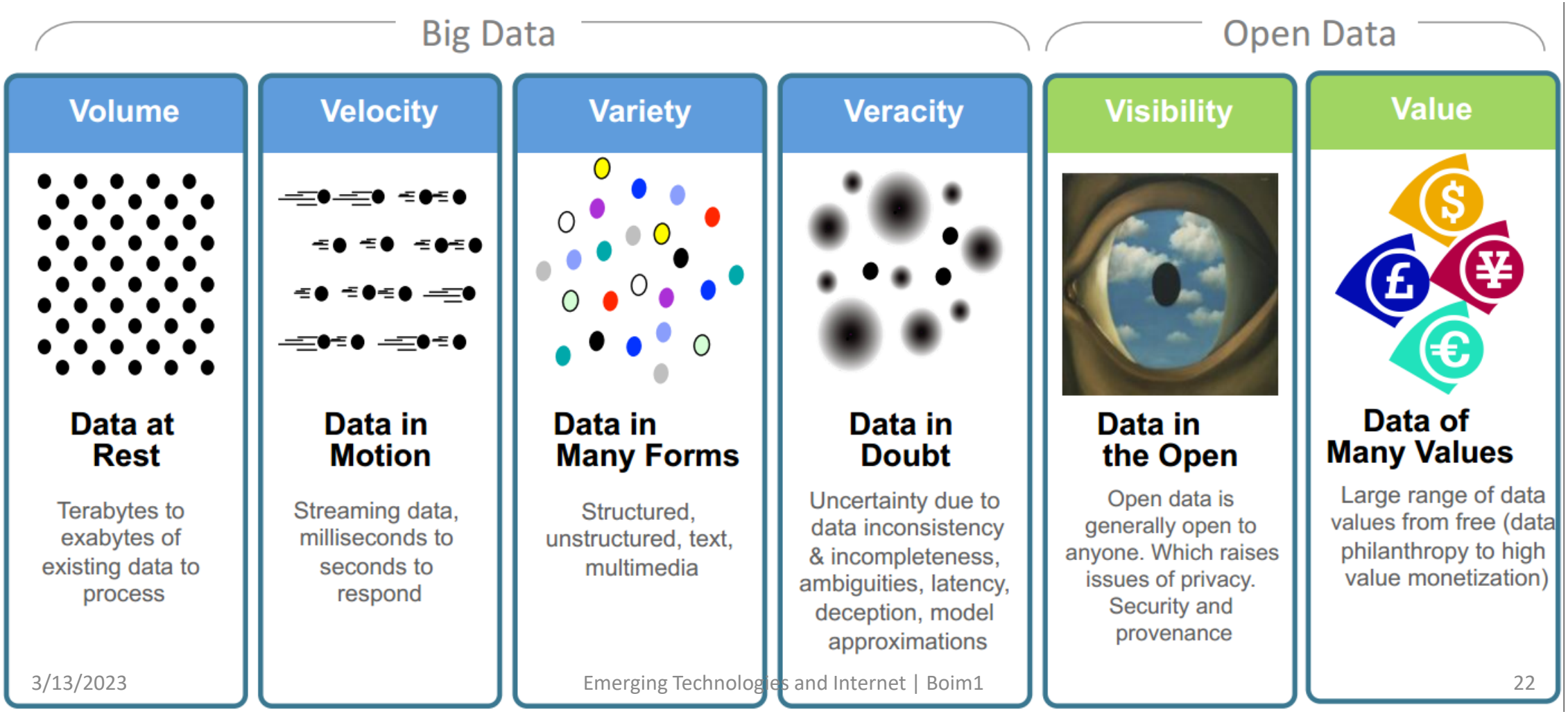
c) Volume is one of the characteristics of big data. We already know that Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus comes to the end of characteristics of big data.

Later, few more dimensions have been added, which are enumerated below:

d) Veracity: Coined by IBM, veracity refers to the unreliability associated with the data sources. For instance, sentiment analysis using social media data (Twitter, Facebook, etc.) is subject to uncertainty. There is a need to differentiate the reliable data from uncertain and imprecise data and manage the uncertainty associated with the data.

e) Variability: Variability and Complexity were added as additional dimensions by SAS. Often, inconsistency in the big data velocity leads to variation in flow rate of data, which is referred to as variability. Data are generated from various sources and there is an increasing complexity in managing data ranging from transactional data to big data. Data generated from different geographical locations have different semantics.

f) Low-Value density: Data in its original form is unusable. Data is analyzed to discover very high value (Sun and Heller, 2012). For example, logs from the website cannot be used in its initial form to obtain business value. It must be analyzed to predict the customer behavior.



Why is Big Data Important?

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:

1. Cost Savings: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.
2. Time Reductions: The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning

3. Understand the market conditions: By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, 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.

4. Control online reputation: Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

5. Using Big Data Analytics to Boost Customer Acquisition and Retention The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces.

If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.

6. Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights. Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.

7. Big Data Analytics As a Driver of Innovations and Product Development Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Sources of Big Data

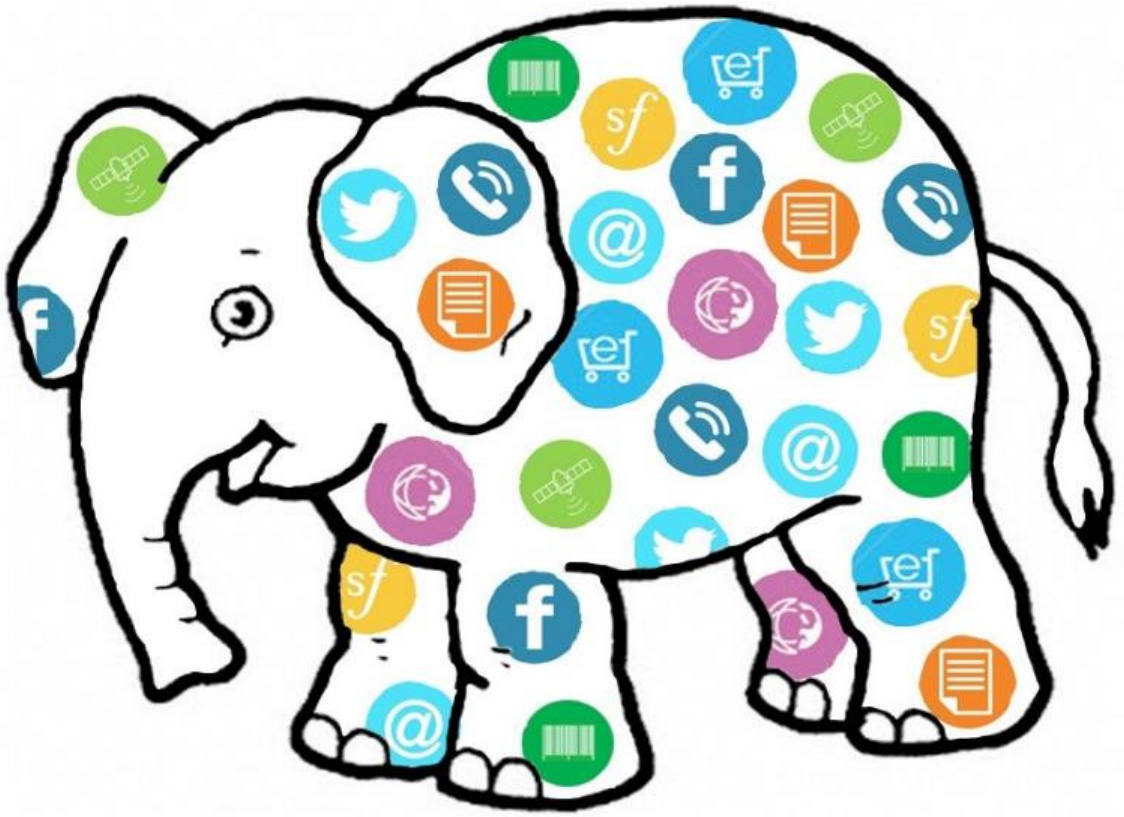


Fig. 1.5 Sources of Big Data

Table 1 summarizes the various types of data produced in different sectors.

Table 1. Different Sources of Data

Sector	Data Produced	Use
Astronomy	Movement of stars, satellites, etc.	To monitor the activities of asteroid bodies and satellites
Financial	News content via video, audio, twitter and news report	To make trading decisions
Healthcare	Electronic medical records and images	To aid in short-term public health monitoring and long-term epidemiological research programs
Internet of Things (IoT)	Sensor data	To monitor various activities in smart cities
Life Sciences	Gene sequences	To analyze genetic variations and potential treatment effectiveness
Media/Entertainment	Content and user viewing behavior	To capture more viewers
Social Media	Blog posts, tweets, social networking sites, log details	To analyze the customer behavior pattern
Telecommunications	Call Detail Records (CDR)	Customer churn management
Transportation, Logistics, Retail, Utilities	Sensor data generated from fleet transceivers, RFID tag readers and smart meters	To optimize operations
Video Surveillance	Recordings from CCTV to IPTV cameras and recording system	To analyze behavioral patterns for service enhancement and

Tools For Managing Big Data

Table 2: Big Data Capabilities and their Primary Technologies

Big Data Capability	Primary Technology	Features
Storage and management capability	Hadoop Distributed File System (HDFS)	Open source distributed file system, Runs on high-performance commodity hardware, Highly scalable storage and automatic data replication
Database capability	Oracle NoSQL	Dynamic and flexible schema design, Highly scalable multi-node, multiple data center, fault tolerant, ACID operations, High-performance key-value pair database
	Apache HBase	Automatic failover support between Region servers, Automatic and configurable sharding of tables
	Apache Cassandra	Fault tolerance capability for every node, Column indexes with the performance of log-structured updates and built-in caching
	Apache Hive	Query execution via MapReduce, Uses SQL-like language HiveQL, Easy ETL process either from HDFS or Apache HBase
Processing capability	MapReduce	Distribution of data workloads across thousands of nodes, Breaks problem into smaller sub-problems
	Apache Hadoop	Highly customizable infrastructure, Highly scalable parallel batch processing, Fault tolerant
Data integration capability	Oracle big data connectors, Oracle data integrator	Exports MapReduce results to RDBMS, Hadoop, and other targets, Includes a Graphical User Interface
Statistical analysis capability	R and Oracle R Enterprise	Programming language for statistical analysis

Big Data vs Data Warehouse

Big Data has become the reality of doing business for organizations today. There is a boom in the amount of structured as well as raw data that floods every organization daily. If this data is managed well, it can lead to powerful insights and quality decision making. Big data analytics is the process of examining large data sets containing a variety of data types to discover some knowledge in databases, to identify interesting patterns and establish relationships to solve problems, market trends, customer preferences, and other useful information. Companies and businesses that implement Big Data Analytics often reap several business benefits. Companies implement Big Data Analytics because they want to make more informed business decisions. A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources.

A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the Data Warehouse through the processes of extraction, transformation and loading (ETL tools). Data analysis tools, such as business intelligence software, access the data within the warehouse. A data warehouse stores current and historical data for the entire business and feeds BI and analytics. Data warehouses use a database server to pull in data from an organization's databases and have additional functionalities for data modeling, data lifecycle management, data source integration, and more. Modern data warehouses are designed to handle both structured and unstructured data, like videos, image files, and sensor data. Some leverage integrated analytics and in-memory database technology (which holds the data set in computer memory rather than in disk storage) to provide real-time access to trusted data and drive confident decision-making.

Without data warehousing, it's very difficult to combine data from heterogeneous sources, ensure it's in the right format for analytics, and get both a current and long-range view of data over time.

Benefits of data warehousing

A well-designed data warehouse is the foundation for any successful BI or analytics program. Its main job is to power the reports, dashboards, and analytical tools that have become indispensable to businesses today. A data warehouse provides the information for your data-driven decisions – and helps you make the right call on everything from new product development to inventory levels. There are many benefits of a data warehouse. Here are just a few:

•**Faster queries:** Data warehouses are built specifically for fast data retrieval and analysis. With a DW, you can very rapidly query large amounts of consolidated data with little to no support from IT.

•**Improved data quality:** Before being loaded into the DW, data cleansing cases are created by the system and entered in a worklist for further processing, ensuring data is transformed into a consistent format to support analytics – and decisions – based on high quality, accurate data.

•**Historical insight:** By storing rich historical data, a data warehouse lets decision-makers learn from past trends and challenges, make predictions, and drive continuous business improvement.

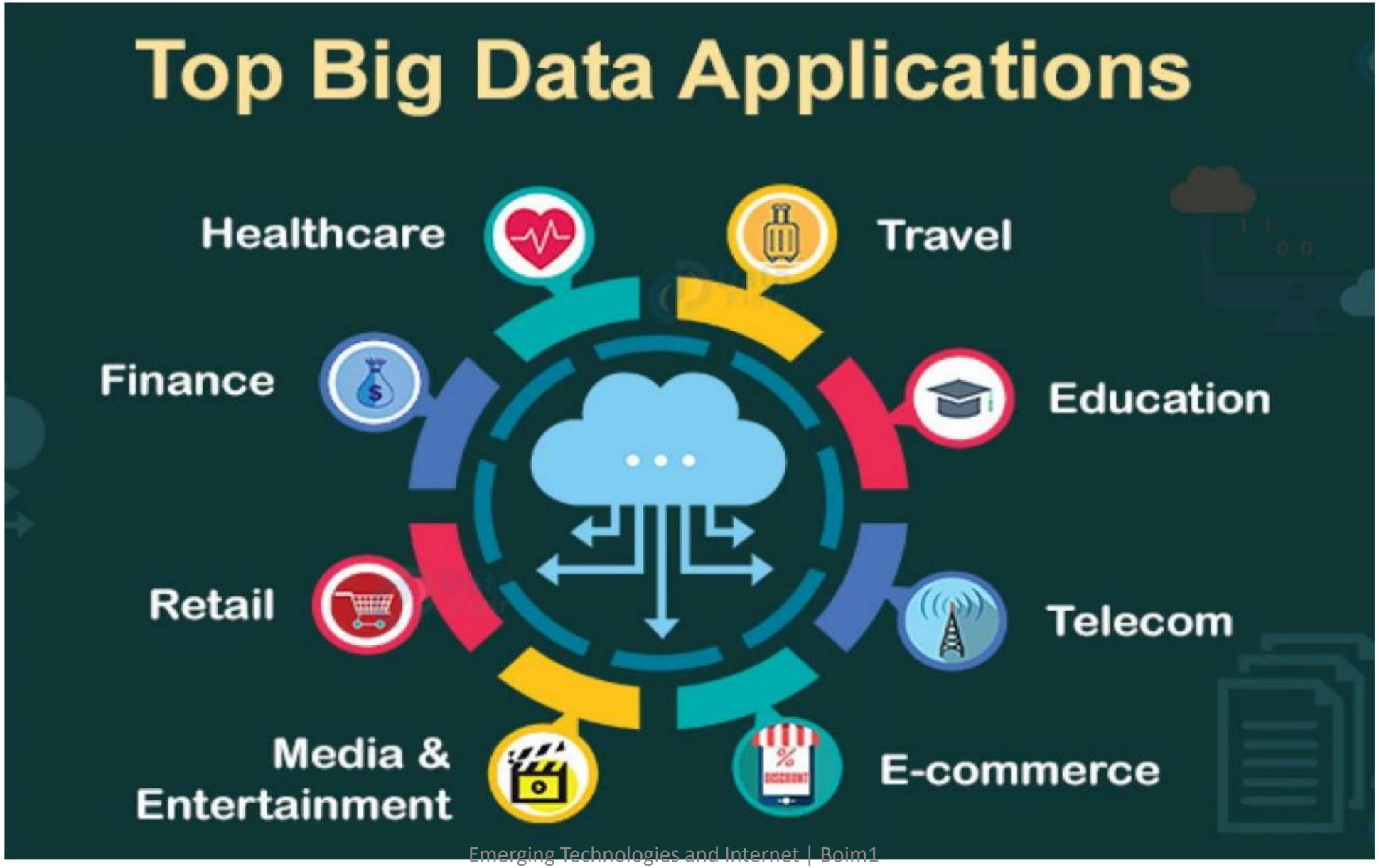
•**Better business analytics:** With data warehousing, decision-makers have access to data from multiple sources and no longer have to make decisions based on incomplete information.

Big Data Mining

Big data mining is referred to the collective data mining or extraction techniques that are performed on large sets /volume of data or the big data. Big data mining is primarily done to extract and retrieve desired information or pattern from humongous quantity of data.

This is usually performed on large quantity of unstructured data that is stored over time by an organization. Typically, big data mining works on data searching, refinement , extraction and comparison algorithms. Big data mining also requires support from underlying computing devices, specifically their processors and memory, for performing operations / queries on large amount of data. Big data mining techniques and processes are also used within big data analytics and business intelligence to deliver summarized targeted and relevant information, patterns and/or relationships between data, systems, processes and more.

Applications of Big Data



Healthcare

Data analysts obtain and analyze information from multiple sources to gain insights. The multiple sources are electronic patient record; clinical decision support system including medical imaging, physician's written notes and prescription, pharmacy and laboratories; clinical data; and machine generated sensor data. The integration of clinical, public health and behavioral data helps to develop a robust treatment system, which can reduce the cost and at the same time, improve the quality of treatment.

Telecommunication

In order to improve the customer experience, MSPs analyze a number of factors such as demographic data (gender, age, marital status, and language preferences), customer preferences, household structure and usage details (CDR, internet usage, value-added services (VAS)) to model the customer preferences and offer a relevant personalized service to them. This is known as targeted marketing, which improves the adoption of mobile services, reduces churn, thus, increasing the revenue of MSPs. For example, Ufone, a Pakistan-based MSP, reduced the churn rate by precisely marketing the customized offers to their customers. The company analyzes the CDR data to identify the call patterns to offer different plans to customers. The services are marketed to the customers through a call or text message. Their responses are recorded for further analysis.

Financial Firms

Currently, capital firms are using advanced technology to store huge volumes of data. But increasing data sources like Internet and Social media require them to adopt big data storage systems. Capital markets are using big data in preparation for regulations like EMIR, Solvency II, Basel II etc, anti-money laundering, fraud mitigation, pre-trade decision-support analytics including sentiment analysis, predictive analytics and data tagging to identify trades. The timeliness of finding value plays an important role in both investment banking and capital markets, hence, there is a need for real-time processing of data.

Retail

Evolution of e-commerce, online purchasing, social-network conversations and recently location specific smartphone interactions contribute to the volume and the quality of data for data-driven customization in retailing. Major retail stores might place CCTV not only to observe the instances of theft but also to track the flow of customers. It helps to observe the age group, gender and purchasing patterns of the customers during weekdays and weekends. Based on the purchasing patterns of the customers, retailers group their items using a well-known data mining technique called Market Basket Analysis, so that a customer buying bread and milk might purchase jam as well. This helps to decide on the placement of objects and decide on the prices. Nowadays, e-commerce firms use market basket analysis and recommender systems to segment and target the customers. They collect the click stream data, observe behavior and recommend products in the real time.

Analytics help the retail companies to manage their inventory. For example, Stage stores, one of the brand names of Stage Stores Inc. which operates in more 40 American states, used to analytics to forecast the order for different sizes of garments for different geographical regions.

Law Enforcement

Law enforcement officials try to predict the next crime location using past data i.e., type of crime, place and time; social media data; drone and smartphone tracking. Researchers at Rutgers University developed an app called RTM Dx to prevent crime and is being used by police department at Illinois, Texas, Arizona, New Jersey, Missouri and Colorado. With the help the app, the police department could measure the spatial correlation between the location of crime and features of the environment. A new technology called facial analytics that examines images of people without violating their privacy. Facial analytics is used to check child pornography. This saves the time of manual examination. Child pornography can be identified by integration of various technologies like Artemis and PhotoDNA by comparing files and image hashes with existing files to identify the subject as adult or child. It also identifies the cartoon based pornography.

Fraud and compliance

When it comes to security, it’s not just a few rogue hackers; you’re up against entire expert teams. Security landscapes and compliance requirements are constantly evolving. Big data helps you identify patterns in data that indicate fraud and aggregate large volumes of information to make regulatory reporting much faster.

Predictive maintenance

Factors that can predict mechanical failures may be deeply buried in structured data, such as the equipment year, make, and model, as well as in unstructured data that covers millions of log entries, senor data, error messages, and engine temperature., By analyzing these indications of potential issues before the problems happen, organizations can deploy maintenance more cost effectively and maximize parts and equipment uptime.

New Product Development

There is a huge risk associated with new product development. Enterprises can integrate both external sources, i.e., twitter and Facebook page and internal data sources, i.e., customer relationship management (CRM) systems to understand the customers' requirement for a new product, to gather ideas for new product and to understand the added feature included in a competitor's product. Proper analysis and planning during the development stage can minimize the risk associated with the product, increase the customer lifetime value and promote brand engagement. Ribbon UI in Microsoft 2007 was created by analyzing the customer data from previous releases of the product to identify the commonly used features and making intelligent decisions.

Banking

The investment worthiness of the customers can be analyzed using demographic details, behavioral data, and financial employment. The concept of cross-selling can be used here to target specific customer segments based on past buying behavior, demographic details, sentiment analysis along with CRM data.

Energy and Utilities

Consumption of water, gas and electricity can be measured using smart meters at regular intervals of one hour. During this interval, a huge amount of data is generated and analyzed to change the patterns of power usage. The real-time analysis reveals energy consumption pattern, instances of electricity thefts and price fluctuations.

Insurance

Personalized insurance plan is tailored for each customer using updated profiles of changes in wealth, customer risk, home asset value, and other data inputs. Recently, driving data of customers such as miles driven, routes driven, time of day, and braking abruptness are collected by the insurance companies by using sensors in their cars. Comparing individual driving pattern and driver risk with the statistical information available such as peak hours of drivers on the road develops a personalized insurance plan. This analysis of driver risk and policy gives a competitive advantage to the insurance companies.

Life Sciences:

Clinical research is a slow and expensive process, with trials failing for a variety of reasons. Advanced analytics, artificial intelligence (AI) and the Internet of Medical Things (IoMT) unlocks the potential of improving speed and efficiency at every stage of clinical research by delivering more intelligent, automated solutions.

Manufacturing:

For manufacturers, solving problems is nothing new. They wrestle with difficult problems on a daily basis - from complex supply chains, to motion applications, to labor constraints and equipment breakdowns. That's why big data analytics is essential in the manufacturing industry, as it has allowed competitive organizations to discover new cost saving opportunities and revenue opportunities.

Media and Entertainment

Media and Entertainment industry is all about art and employing Big Data in it is a sheer piece of art. Art and science are often considered to be the two completely contrasting domains but when employed together, they do make a deadly duo and Big Data's endeavors in the media industry are a perfect example of it. Viewers these days need content according to their choices only. Content that is relatively new to what they saw the previous time. Earlier the companies broadcasted the Ads randomly without any kind of analysis. But after the advent of Big Data analytics in the industry, companies now are aware of the kind of Ads that attracts a customer and the most appropriate time to broadcast it for seeking maximum attention. Customers are now the real heroes of the Media and entertainment industry - courtesy to Big Data and Analytics.

Travel Industry

While Big Data is spreading like wildfire and various industries have been cooking its food with it, the travel industry was a bit late to realize its worth. Better late than never though. Having a stress-free traveling experience is still like a daydream for many. And now Big Data's arrival is like a ray of hope, that will mark the departure of all the hindrances in our smooth traveling experience. Through Big Data and analytics, travel companies are now able to offer more customized traveling experience. They are now able to understand their customer's requirements in a much-enhanced way. From providing them with the best offers to be able to make suggestions in real-time, Big Data is certainly a perfect guide for any traveler. Big Data is gradually taking the window seat in the travel industry.

Automobile

Big Data has now taken complete control of the automobile industry and is driving it smoothly. Big Data is driving the automobile industry towards some unbelievable and never before results. The automobile industry is on a roll and Big Data is its wheels or I must say Big Data has given wings to it. Big Data has helped the automobile industry achieve things that were beyond our imaginations From analyzing the trends to understanding the supply chain management, from taking care of its customers to turning our wildest dream of connected cars a reality, Big Data is well and truly driving the automobile industry crazy.

Marketing

Marketing analytics helps the organizations to evaluate their marketing performance, to analyze the consumer behavior and their purchasing patterns, to analyze the marketing trends which would aid in modifying the marketing strategies like the positioning of advertisements in a webpage, implementation of dynamic pricing and offering personalized products.

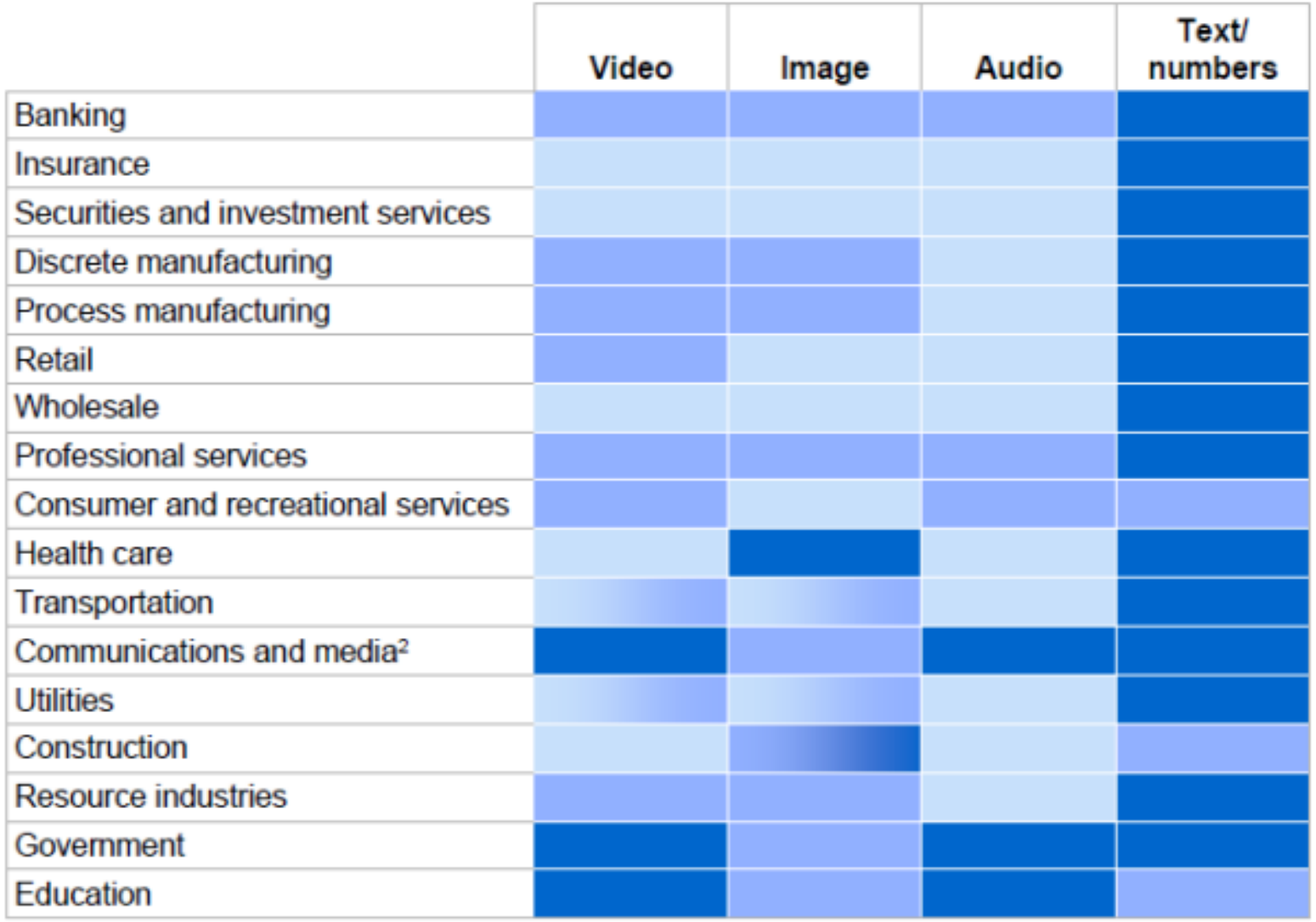
Education

With the advent of computerized course modules, it is possible to assess the academic performance real time. This helps to monitor the performance of the students after each module and give immediate feedback on their learning pattern. It also helps the teachers to assess their teaching pedagogy and modify based on the students‟ performance and needs. Dropout patterns, students requiring special attention and students who can handle challenging assignments can be predicted.

Other sectors

With increasing analytics skills among the various organizations, the advantage of big data analytics can be realized in sectors like construction and material sciences.

The type of data generated and stored varies by sector[1]



Penetration
- High
- Medium
- Low

1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.

SOURCE: McKinsey Global Institute analysis

Big Data Challenges

Scalability and Storage Issues:

The rate of increase in data is much faster than the existing processing systems. The storage systems are not capable enough to store these data. There is a need to develop a processing system that not only caters to today's needs but also future needs.

Timeliness of Analysis:

The value of the data decreases over time. Most of the applications like fraud detection in telecom, insurance and banking, require real time or near real time analysis of the transactional data.

Representation of Heterogeneous Data:

Data obtained from various sources are heterogeneous in nature. Unstructured data like Images, videos and social media data cannot be stored and processed using traditional tools like SQL. Smartphones now record and share images, audios and videos at an incredibly increasing rate, forcing our brains to process more. However, the process for representing images, audios and videos lacks efficient storage and processing.

Data Analytics System:

Traditional RDBMS are suitable only for structured data and they lack scalability and expandability. Though non-relational databases are used for processing unstructured data, but there exist problems with their performances. There is a need to design a system that combines the benefits of both relational and non-relational database systems to ensure flexibility.

Privacy and Security:

New devices and technologies like cloud computing provide a gateway to access and to store information for analysis. This integration of IT architectures will pose greater risks to data security and intellectual property. Access to personal information like buying preferences and call detail records will lead to increase in privacy concerns. Researchers have technical infrastructure to access the data from any data source including social networking sites, for future use whereas the users are unaware of the gains that can be generated from the information they posted. Big data researchers fail to understand the difference between privacy and convenience.

Not always better data:

Social media mining has attracted researchers. Twitter has become a new popular source. Twitter users do not represent the global population. Big data researchers should understand the difference between big data and whole data. The tweets containing references to pornography and spam are eliminated resulting in the inaccuracy of the topical frequency. There is a redundancy in number of twitter users and twitter accounts one account accessed by multiple people and multiple accounts created by single user. There are active users and passive users who just sign in to listen. There are two types of accounts public and protected or private.

Out of Context:

Data reduction is one of common ways to fit into a mathematical model. Retaining context during data abstraction is critical. Data which are out of context lose meaning and value. There is an obsession for „social graph‟ with the rise of social networking sites. Big data introduces two types of social networks: `articulated networks' and „behavioral network‟. Articulated networks are those resulting from specifying contacts through mediating technology. “Friends”, “Acquaintances” in Facebook, “Follow” is twitter and “Best Friends”, “Friends” and other circles in Google+ are examples of articulated network. Articulated networks are created to have separate group for friends, colleagues, friends of friends and filter the content that each group can view. Behavioral networks are obtained from social media interactions and communication patterns. But communication patterns necessarily need not reveal tie strength.

Digital Divide:

Gaining access to big data is one of the most important limitations. Data companies and social media companies have access to large social data. Few companies decide who can access data and to what extent. Few sell the right to access for a high fees while others offer a portion of data sets to researchers. This results in “Digital Divide" in the realm of big data: Big Data rich and Big Data poor.

Data errors:

With increase in the growth of information technology, huge amount of data is generated. With advent of cloud computing for storage and retrieval of data, there is a need to utilize the big data. Large datasets from internet sources are prone to errors and losses, hence unreliable. The source of the data should be understood to minimize the errors caused while using multiple datasets. The properties and limits of the dataset should be understood before analysis to avoid or explain the bias in the interpretation of data.

Advantages of using Big Data

1. Improved business processes
2. Fraud detection
3. Improved customer service
4. Better decision-making
5. Increased productivity
6. Reduce costs
7. Improved customer service
8. Increased revenue
9. Increased agility
10. Greater innovation
11. Faster speed to market

Disadvantages of Big Data

1. Privacy and security concerns
2. Need for technical expertise
3. Need for talent
4. Data quality
5. Need for cultural change
6. Compliance
7. Cybersecurity risks
8. Rapid change
9. Hardware needs
10. Costs
11. Difficulty integrating legacy systems