Big Data Analytics

by Jyoti Budhwar

What is Big Data?

- According to Gartner, the definition of Big Data – “Big data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”
- Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

- Big Data:- is a massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools.

- Today, there are millions of data sources that generate data at a very rapid rate. These data sources are present across the world. Some of the largest sources of data are social media platforms and networks. Let's use Facebook as an example—it generates more than 500 terabytes of data every day. This data includes pictures, videos, messages, and more.
- Data also exists in different formats, like structured data, semi-structured data, and unstructured data. For example, in a regular Excel sheet, data is classified as structured data—with a definite format. In contrast, emails fall under semi-structured, and your pictures and videos fall under unstructured data. All this data combined makes up Big Data.

There are certain basic tenets of 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.

Evaluation of Big data

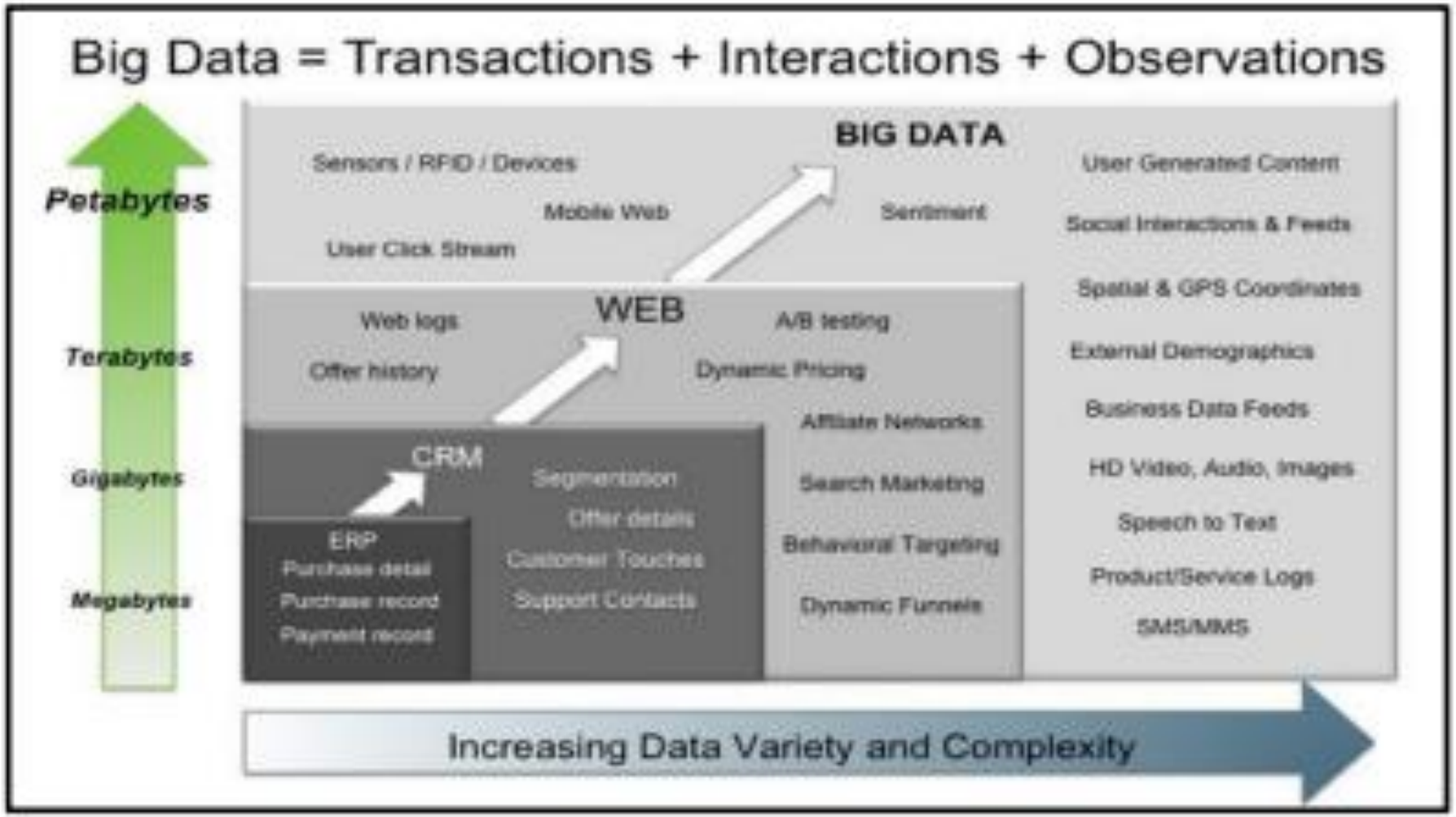
- Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.
- Around 2005, people began to realize just how much data 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.

	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**

Table 1.1 The 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 an onslaught of structured, unstructured, and multimedia data.


Big Data = Transactions + Interactions + Observations
Petabytes
Terabytes
Gigabytes
Megabytes
BIG DATA
Sensors / RFID / Devices
Mobile Web
User Click Stream
Sentiment
User Generated Content
Social Interactions & Feeds
Spatial & GPS Coordinates
External Demographics
Business Data Feeds
HD Video, Audio, Images
Speech to Text
Product/Service Logs
SMS/MMS
WEB
Web logs
Offer history
A/B testing
Dynamic Pricing
Affiliate Networks
Search Marketing
Behavioral Targeting
Dynamic Funnels
CRM
Segmentation
Offer details
Customer Touches
Support Contacts
ERP
Purchase detail
Purchase record
Payment record
Increasing Data Variety and Complexity

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Types of Digital Data

- Structured
- Unstructured
- Semi-structured

Structured:- 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.

- **Unstructured** :- 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.

- **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.

Characteristics of Big Data

- Back in 2001, Gartner analyst Doug Laney listed the 3 ‘V’s of Big Data – Variety, Velocity, and Volume.
- **Variety** :- 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.

- **Velocity** :- 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.
- **Volume** :- 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.

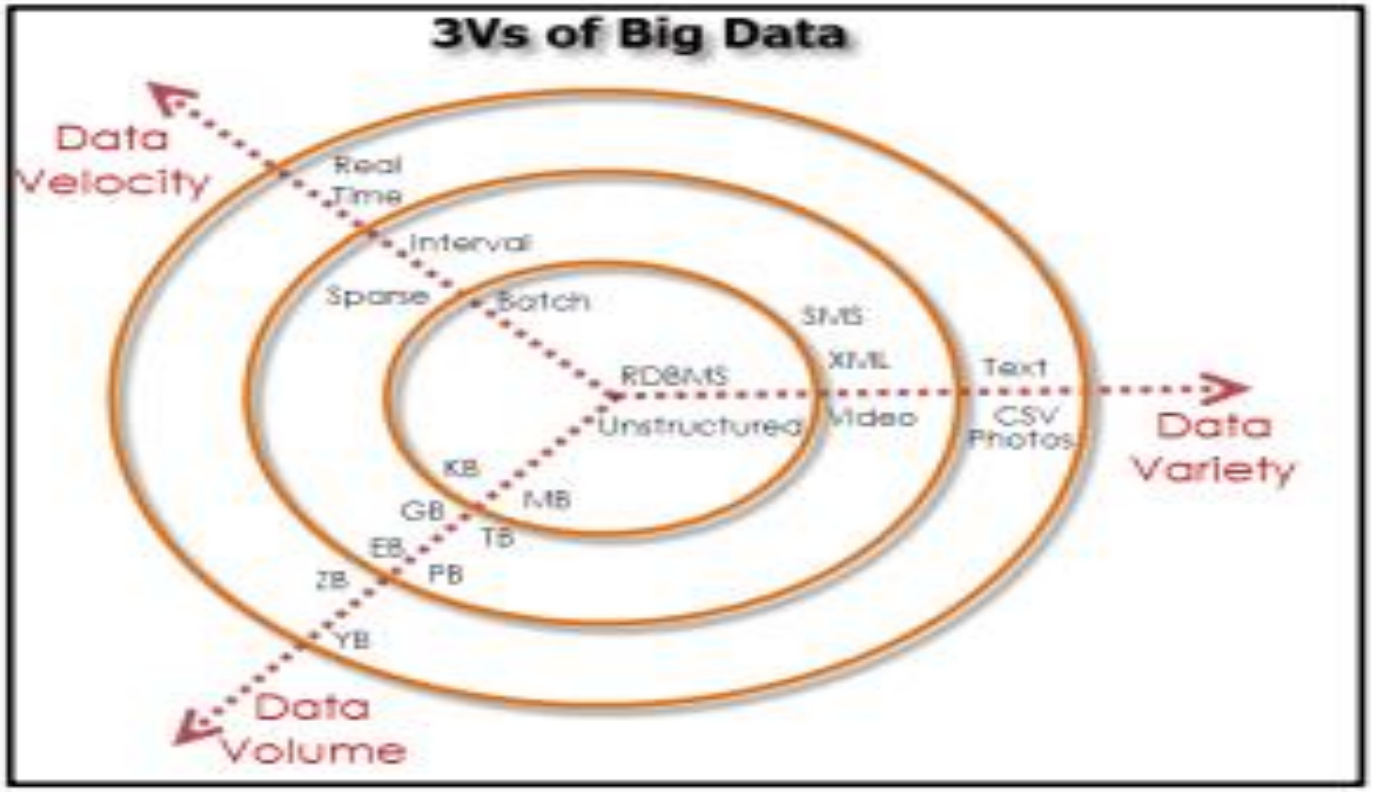
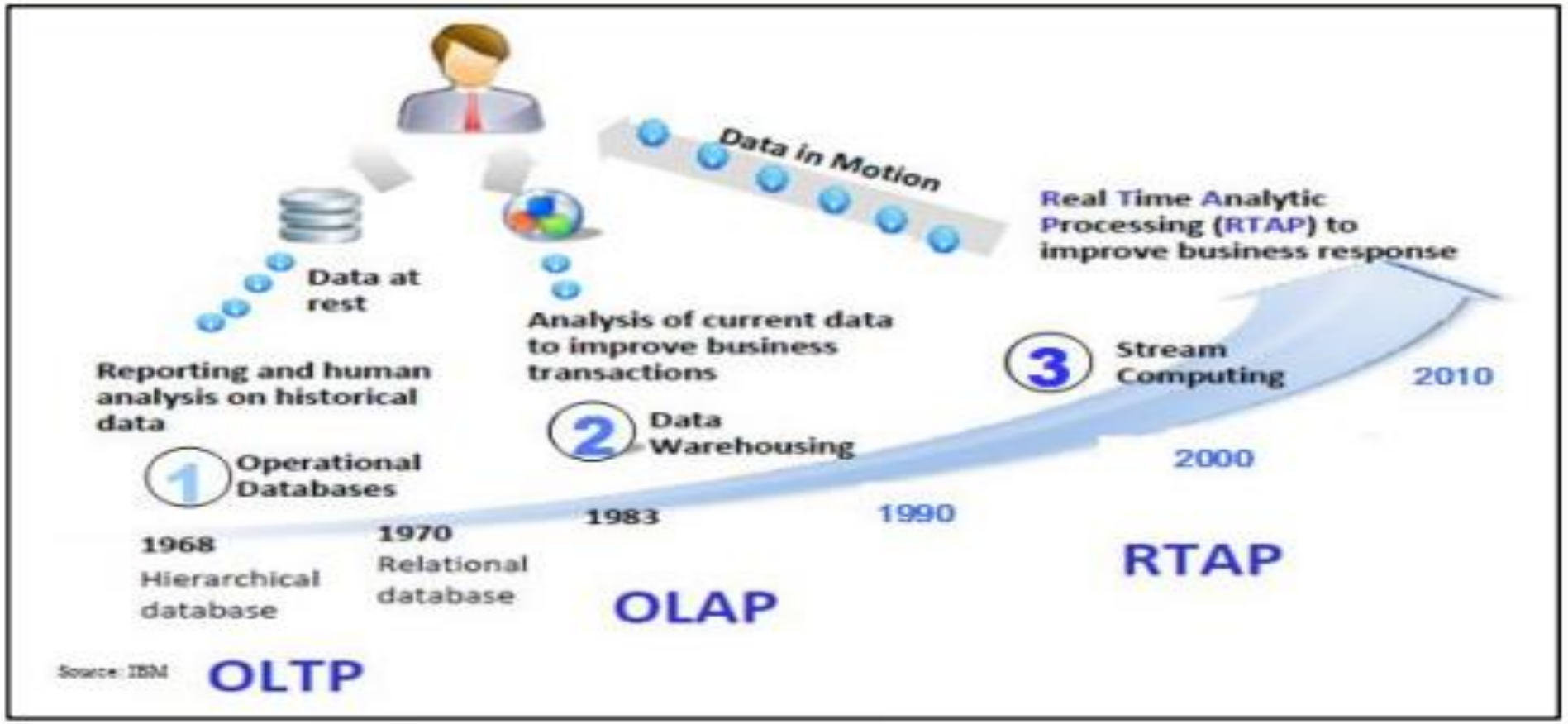


Figure : Characteristics of Big

Harnessing Big Data



- OLTP: Online Transaction Processing (DBMSs)
- OLAP: Online Analytical Processing (Data Warehousing)
- RTAP: Real-Time Analytics Processing (Big Data Architecture & technology)

Challenges with Big Data

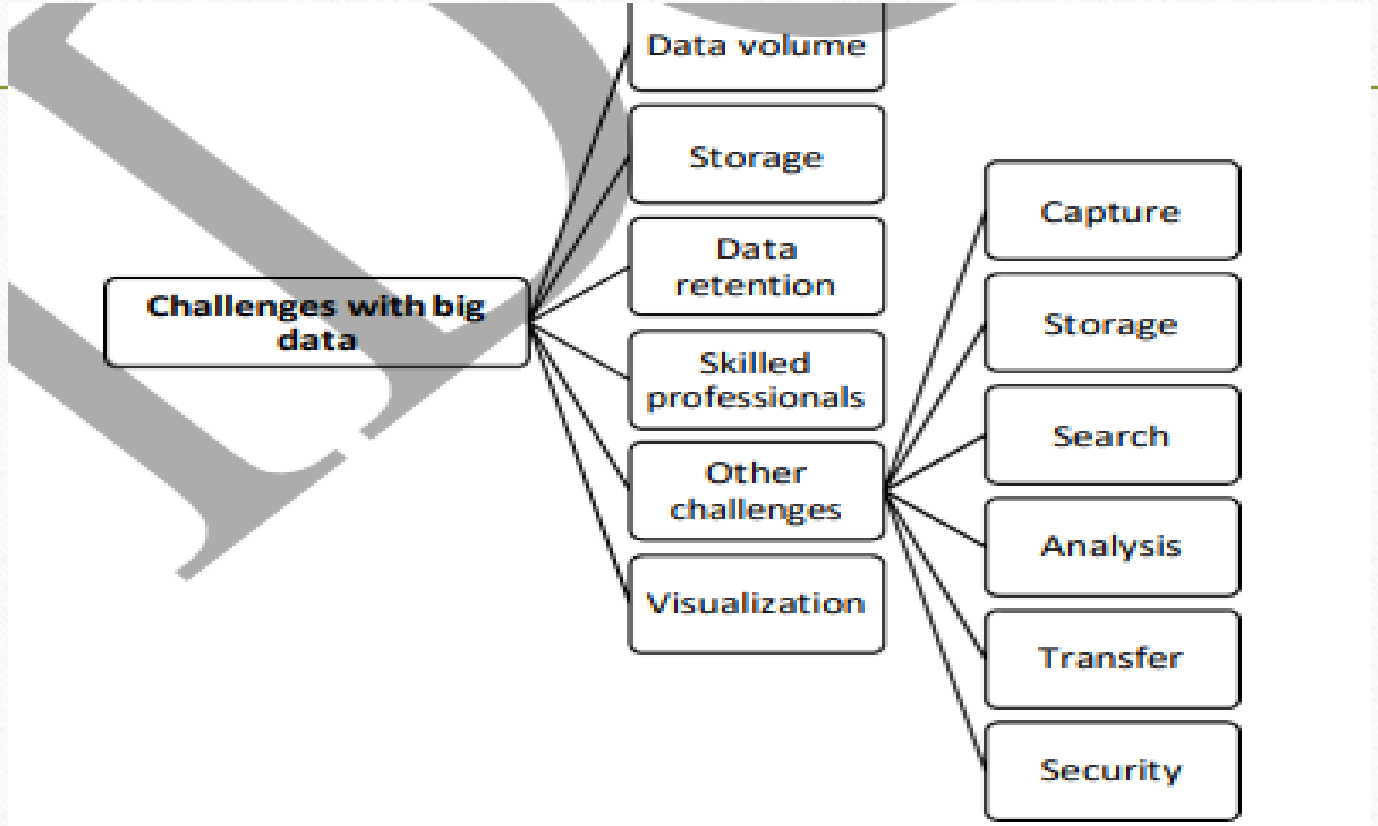


Figure 1.5. Challenges with big data

- **Data volume**: Data today is growing at an exponential rate. This high tide of data will continue to rise continuously. The key questions are –
- “will all this data be useful for analysis?”,
- “Do we work with all this data or subset of it?”
- **Storage**: Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity and easy upgrading / downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.
- **Data retention**: How long should one retain this data? Some data may require for log-term decision, but some data may quickly become irrelevant and obsolete.

- **Skilled professionals**: In order to develop, manage and run those applications that generate insights, organizations need professionals who possess a high-level proficiency in data sciences.
- **Other challenges**: Other challenges of big data are with respect to capture, storage, search, analysis, transfer and security of big data.
- **Visualization**: Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how big the data set should be for it to be considered bigdata. Data visualization(computer graphics) is becoming popular as a separate discipline. There are very few data visualization experts.

Why is Big Data Important?

- Cost Savings
- Time Reductions
- Understand the market conditions
- Control online reputation
- Using Big Data Analytics to Boost Customer Acquisition and Retention
- Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights
- Big Data Analytics As a Driver of Innovations and Product Development

Figure 1.6: Why big data?

What is Big Data Analytics?

- Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences.
- Big Data analytics provides various advantages—it can be used for better decision making, preventing fraudulent activities, among other things.

Uses and Examples of Big Data Analytics

- There are many different ways that Big Data analytics can be used in order to improve businesses and organizations. Here are some examples:
- Using analytics to understand customer behavior in order to optimize the customer experience
- Predicting future trends in order to make better business decisions
- Improving marketing campaigns by understanding what works and what doesn't
- Increasing operational efficiency by understanding where bottlenecks are and how to fix them
- Detecting fraud and other forms of misuse sooner
- These are just a few examples — the possibilities are really endless when it comes to Big Data analytics. It all depends on how you want to use it in order to improve your business.

The Lifecycle Phases of Big Data Analytics

- Stage 1 - Business case evaluation - The Big Data analytics lifecycle begins with a business case, which defines the reason and goal behind the analysis.
- Stage 2 - Identification of data - Here, a broad variety of data sources are identified.
- Stage 3 - Data filtering - All of the identified data from the previous stage is filtered here to remove corrupt data.
- Stage 4 - Data extraction - Data that is not compatible with the tool is extracted and then transformed into a compatible form.

- Stage 5 - Data aggregation - In this stage, data with the same fields across different datasets are integrated.
- Stage 6 - Data analysis - Data is evaluated using analytical and statistical tools to discover useful information.
- Stage 7 - Visualization of data - With tools like Tableau, Power BI, and QlikView, Big Data analysts can produce graphic visualizations of the analysis.
- Stage 8 - Final analysis result - This is the last step of the Big Data analytics lifecycle, where the final results of the analysis are made available to business stakeholders who will take action.

Different Types of Big Data Analytics

- 1. Descriptive Analytics
- 2. Diagnostic Analytics
- 3. Predictive Analytics
- 4. Prescriptive Analytics

- 1. Descriptive Analytics:-This summarizes past data into a form that people can easily read. This helps in creating reports, like a company's revenue, profit, sales, and so on. Also, it helps in the tabulation of social media metrics.
- 2. Diagnostic Analytics:-This is done to understand what caused a problem in the first place. Techniques like drill-down, data mining, and data recovery are all examples. Organizations use diagnostic analytics because they provide an in-depth insight into a particular problem.

- 3. Predictive Analytics:-This type of analytics looks into the historical and present data to make predictions of the future. Predictive analytics uses data mining, AI, and machine learning to analyze current data and make predictions about the future. It works on predicting customer trends, market trends, and so on.
- 4. Prescriptive Analytics:-This type of analytics prescribes the solution to a particular problem. Perspective analytics works with both descriptive and predictive analytics. Most of the time, it relies on AI and machine learning.

Big Data Industry Applications

- Ecommerce - Predicting customer trends and optimizing prices are a few of the ways e-commerce uses Big Data analytics
- Marketing - Big Data analytics helps to drive high ROI (Return on Investment) marketing campaigns, which result in improved sales
- Education - Used to develop new and improve existing courses based on market requirements
- Healthcare - With the help of a patient’s medical history, Big Data analytics is used to predict how likely they are to have health issues

- Media and entertainment - Used to understand the demand of shows, movies, songs, and more to deliver a personalized recommendation list to its users
- Banking - Customer income and spending patterns help to predict the likelihood of choosing various banking offers, like loans and credit cards
- Telecommunications - Used to forecast network capacity and improve customer experience
- Government - Big Data analytics helps governments in law enforcement, among other things

Big Data Analytics Tools

- Hadoop - helps in storing and analyzing data
- MongoDB - used on datasets that change frequently
- Talend - used for data integration and management
- Cassandra - a distributed database used to handle chunks of data
- Spark - used for real-time processing and analyzing large amounts of data
- STORM - an open-source real-time computational system
- Kafka - a distributed streaming platform that is used for fault-tolerant storage

- There are hundreds of ***data analytics tools*** out there in the market today but the selection of the right tool will depend upon your business ***NEED, GOALS, and VARIETY*** to get business in the right direction.
- **1. APACHE Hadoop:-** It's a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers. Hadoop also offers **cross-platform** support for its users. Today, it is the best ***big data analytic tool*** and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

- **Features of APACHE Hadoop**:- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.
- It works on small commodity hardware like JBOD or a bunch of disks.

- **2. Spark:-** APACHE Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is **capable of handling multi-petabytes of data** as well. Recently, Spark made a record of processing **100 terabytes of data in just 23 minutes** which broke the previous world record of **Hadoop (71 minutes)**. This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

- **Features of APACHE Spark:-**
- *Ease of use:* It allows users to run in their preferred language. (JAVA, Python, etc.)
- *Real-time Processing:* Spark can handle real-time streaming via Spark Streaming

- **3. Cassandra:-** APACHE Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It’s one of the **most popular tools for data analytics** and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. It is **capable of delivering thousands of operations every second** and can handle petabytes of resources with almost zero downtime

- **Features of APACHE Cassandra:**
- *Data Storage Flexibility:* It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- *Data Distribution System:* Easy to distribute data with the help of replicating data on multiple data centers.
- *Fast Processing:* Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
- *Fault-tolerance:* The moment, if any node fails, it will be replaced without any delay.

- **4. Qubole:-**It’s an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

- **<u>Features of Qubole</u>:**
- *Supports ETL process:* It allows companies to **migrate data from multiple sources in one place**.
- *Real-time Insight:* It monitors user's systems and allows them to view real-time insights
- *Predictive Analysis:* Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
- *Advanced Security System:* To protect users' data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.