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. And graph databases are becoming increasingly important as well, with their ability to display massive amounts of data in a way that makes analytics fast and comprehensive.

BIG DATA

Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

Types Of Big Data

The following are the types of Big Data:

1. Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period, developed technology in computer science has achieved greater success in developing techniques for working with such kinds of data (where the format is well known in advance) and also deriving value from it. However, nowadays, we are foreseeing issues when the size of such data grows to a huge extent; typical sizes are in the range of multiple zettabytes.

2. Unstructured

Any data with an unknown form or structure is classified as unstructured data. In addition to the huge size, unstructured data poses multiple challenges regarding its processing for deriving value out of it. A

typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Nowadays, organisations have a wealth of available data. Still, unfortunately, they don't know how to derive value from it since this data is in its raw form or unstructured format.

3. Semi-structured

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

Big data has advantages and disadvantages of its own, just like any other technology. There are times when the disadvantages of big data outweigh some of its advantages when it comes to practical applications. Therefore, before utilising big data, businesses must consider both its advantages and disadvantages. Let's talk about Big data's benefits and drawbacks after defining it.

Here are some big data examples that are helping transform organizations across every industry:

- **Tracking consumer behavior and shopping habits to deliver hyper-personalized retail product recommendations tailored to individual customers**
- **Monitoring payment patterns and analyzing them against historical customer activity to detect fraud in real time**
- **Combining data and information from every stage of an order's shipment journey with hyperlocal traffic insights to help fleet operators optimize last-mile delivery**
- **Using AI-powered technologies like natural language processing to analyze unstructured medical data (such as research reports, clinical notes, and lab results) to gain new insights for improved treatment development and enhanced patient care**
- **Using image data from cameras and sensors, as well as GPS data, to detect potholes and improve road maintenance in cities**
- **Analyzing public datasets of satellite imagery and geospatial datasets to visualize, monitor, measure, and predict the social and environmental impacts of supply chain operations**

Why Is big data important?

The importance of big data doesn't simply revolve around how much data you have. The value lies in how you use it. By taking data from any source and analyzing it, you can find answers that

1) streamline resource management

2) improve operational efficiencies

3) optimize product development

4) drive new revenue and growth opportunities

5) enable smart decision making. When combing big data with high-performance analytics,

you can accomplish business-related tasks such as:

- Determining root causes of failures, issues and defects in near-real time.
- Spotting anomalies faster and more accurately than the human eye.
- Improving patient outcomes by rapidly converting medical image data into insights.
- Recalculating entire risk portfolios in minutes.
- Sharpening deep learning models' ability to accurately classify and react to changing variables.
- Detecting fraudulent behavior before it affects your organization.

What are the 5 V's of Big Data?

Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity.

- **Volume:** the size and amounts of big data that companies manage and analyze
- **Value:** the most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data.
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time

- **Veracity:** the “truth” or accuracy of data and information assets, which often determines executive-level confidence

The additional characteristic of variability can also be considered:

- **Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases

Advantages of Big Data

1. Making wiser decisions

Businesses use big data to enhance B2B operations, advertising, and communication. Big data is primarily being used by many industries, such as travel, real estate, finance, and insurance, to enhance decision-making. Businesses can use big data to accurately predict what customers want and don't want, as well as their behavioural tendencies because it reveals more information in a usable format.

Big data provides business intelligence and cutting-edge analytical insights that help with decision-making. A company can get a more in-depth picture of its target market by collecting more customer data.

Business trends and behaviours are revealed by data-driven insights, which also help businesses compete and grow by enhancing their decision-making. Additionally, these insights help companies develop more specialised goods and services, strategies, and intelligent marketing campaigns to compete in their sector.

2. Cut back on the expense of business operations

According to surveys done by New Vantage and Syncsort (now Precisely), big data analytics has helped businesses significantly cut their costs. Big data is being used to cut costs, according to 66.7% of survey participants from New Vantage. Moreover, 59.4% of Syncsort survey participants stated that using big data tools improved operational efficiency and reduced costs. Do you know that Hadoop and Cloud-Based Analytics, two popular big data analytics tools, can help lower the cost of storing big data

3. Detection of Fraud

Financial companies especially use big data to identify fraud. To find anomalies and transaction patterns, data analysts use artificial intelligence and machine learning algorithms. These irregularities in transaction patterns show that something is out of place or that there is a mismatch, providing us with hints about potential fraud.

For credit unions, banks, and credit card companies, fraud detection is crucial for identifying account information, materials, or product access. By spotting frauds before they cause problems, any industry, including finance, can provide better customer service.

For instance, using big data analytics, banks and credit card companies can identify fraudulent purchases or credit cards that have been stolen even before the cardholder becomes aware of the issue.

4. A rise in productivity

A survey by Syncsort found that 59.9% of respondents said they were using big data analytics tools like Spark and Hadoop to boost productivity. They have been able to increase sales and improve customer retention as a result of this rise in productivity. Modern big data tools make it possible for data scientists and analysts to analyse a lot of data quickly and effectively, giving them an overview of more data.

They become more productive as a result of this. Additionally, big data analytics aids data scientists and analysts in learning more about themselves to figure out how to be more effective in their tasks and job responsibilities. As a result, investing in big data analytics gives businesses across all sectors a chance to stand out through improved productivity.

5. Enhanced customer support

As part of their marketing strategies, businesses must improve customer interactions. Since big data analytics give businesses access to more information, they can use that information to make more specialised, highly personalised offers to each individual customer as well as more targeted marketing campaigns.

Social media, email exchanges, customer CRM (customer relationship management) systems, and other major data sources are the main sources of big data. As a result, it provides businesses with access to a wealth of data about the needs, interests, and trends of their target market.

Big data also enables businesses better to comprehend the thoughts and feelings of their clients to provide them with more individualised goods and services. Providing a personalised experience can increase client satisfaction, strengthen bonds with clients, and, most importantly, foster loyalty.

6. Enhanced speed and agility

Increasing business agility is a big data benefit for competition. Big data analytics can assist businesses in becoming more innovative and adaptable in the marketplace. Large customer data sets can be analysed to help businesses gain insights ahead of the competition and more effectively address customer pain points.

Additionally, having a wealth of data at their disposal enables businesses to assess risks, enhance products and services, and improve communications. Additionally, big data assists businesses in strengthening their business tactics and strategies, which are crucial in coordinating their operations to support frequent and quick changes in the industry.

7. Greater innovation

Innovation is another common benefit of big data, and the NewVantage survey found that 11.6 per cent of executives are investing in analytics primarily as a means to innovate and disrupt their markets. They

reason that if they can glean insights that their competitors don't have, they may be able to get out ahead of the rest of the market with new products and services.

Disadvantages of Big Data

1. A talent gap

A study by AtScale found that for the past three years, the biggest challenge in this industry has been a lack of big data specialists and data scientists. Given that it requires a different skill set, big data analytics is currently beyond the scope of many IT professionals. Finding data scientists who are also knowledgeable about big data can be difficult.

Data scientists and big data specialists are two well-paid professions in the data science industry. As a result, hiring big data analysts can be very costly for businesses, particularly for start-ups. Some businesses must wait a long time to hire the necessary personnel to carry out their big data analytics tasks.

2. Security hazard

For big data analytics, businesses frequently collect sensitive data. These data need to be protected, and security risks can be detrimental if they are not properly maintained.

Additionally, having access to enormous data sets can attract the unwanted attention of hackers, and your company could become the target of a potential cyber-attack. You are aware that for many businesses today, data breaches are the biggest threat. Unless you take all necessary precautions, important information could be leaked to rivals, which is another risk associated with big data.

3. Adherence

Another disadvantage of big data is the requirement for legal compliance with governmental regulations. To store, handle, maintain, and process big data that contains sensitive or private information, a company must make sure that they adhere to all applicable laws and industry standards. As a result, managing data governance tasks, transmission, and storage will become more challenging as big data volumes grow.

4. High Cost

Given that it is a science that is constantly evolving and has as its goal the processing of ever-increasing amounts of data, only large companies can sustain the investment in the development of their Big Data techniques.

5. Data quality

Dealing with data quality issues was the main drawback of working with big data. Data scientists and analysts must ensure the data they are using is accurate, pertinent, and in the right format for analysis before they can use big data for analytics efforts.

This significantly slows down the reporting process, but if businesses don't address data quality problems, they may discover that the insights their analytics produce are useless or even harmful if used.

6. Rapid Change

The fact that technology is evolving quickly is another potential disadvantage of big data analytics. Businesses must deal with the possibility of spending money on one technology only to see something better emerge a few months later. This big data drawback was ranked fourth among all the potential difficulties by Syncsort respondents.

Big data management includes the following approaches

- Using a centralised interface or dashboard to monitor and ensure the availability of all big data resources
- Maintaining the database to get better outcomes.
- Monitoring big data analytics, big data reporting and other similar solutions and implementing them
- Efficient design and implementation of data cycle processes
- Control access and security of big data repositories
- Data visualization to reduce volume and improve big data operations
- Data visualization techniques allow multiple users to use it simultaneously.
- Capturing and storing data from all resources.

Technology Challenges with Big Data

Big data challenges include the storing, analyzing the extremely large and fast-growing data.

Some of the Big Data challenges are:

1. ***Sharing and Accessing Data:***
    - Perhaps the most frequent challenge in big data efforts is the inaccessibility of data sets from external sources.
    - Sharing data can cause substantial challenges.
    - It include the need for inter and intra- institutional legal documents.

- Accessing data from public repositories leads to multiple difficulties.
- It is necessary for the data to be available in an accurate, complete and timely manner because if data in the companies information system is to be used to make accurate decisions in time then it becomes necessary for data to be available in this manner.

2. ***Privacy and Security:***
    - It is another most important challenge with Big Data. This challenge includes sensitive, conceptual, technical as well as legal significance.
    - Most of the organizations are unable to maintain regular checks due to large amounts of data generation. However, it should be necessary to perform security checks and observation in real time because it is most beneficial.
    - There is some information of a person which when combined with external large data may lead to some facts of a person which may be secretive and he might not want the owner to know this information about that person.
    - Some of the organization collects information of the people in order to add value to their business. This is done by making insights into their lives that they're unaware of.

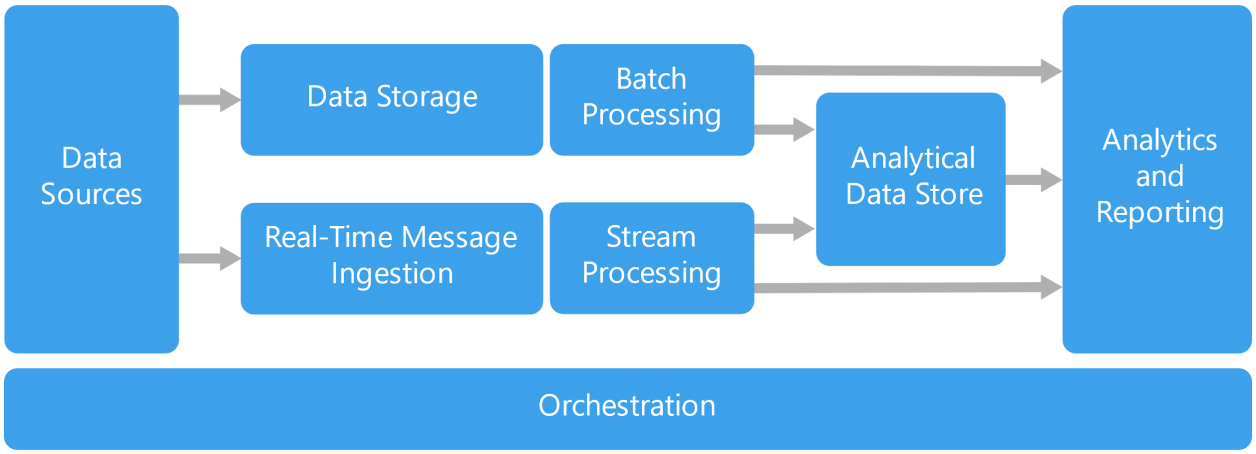
3. ***Analytical Challenges:***
    - There are some huge analytical challenges in big data which arise some main challenges questions like how to deal with a problem if data volume gets too large?
    - Or how to find out the important data points?
    - Or how to use data to the best advantage?
    - These large amount of data on which these type of analysis is to be done can be structured (organized data), semi-structured (Semi-organized data) or unstructured (unorganized data). There are two techniques through which decision making can be done:
        - Either incorporate massive data volumes in the analysis.
        - Or determine upfront which Big data is relevant.

4. ***Technical challenges:***
    - **Quality of data:**
        - When there is a collection of a large amount of data and storage of this data, it comes at a cost. Big companies, business leaders and IT leaders always want large data storage.

- For better results and conclusions, Big data rather than having irrelevant data, focuses on quality data storage.
- This further arise a question that how it can be ensured that data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not.

- **Fault tolerance:**
  - Fault tolerance is another technical challenge and fault tolerance computing is extremely hard, involving intricate algorithms.
  - Nowadays some of the new technologies like cloud computing and big data always intended that whenever the failure occurs the damage done should be within the acceptable threshold that is the whole task should not begin from the scratch.
- **Scalability:**
  - Big data projects can grow and evolve rapidly. The scalability issue of Big Data has lead towards cloud computing.
  - It leads to various challenges like how to run and execute various jobs so that goal of each workload can be achieved cost-effectively.
  - It also requires dealing with the system failures in an efficient manner. This leads to a big question again that what kinds of storage devices are to be used.

Big data architecture



A big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems.

Big data solutions typically involve one or more of the following types of workload:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

Most big data architectures include some or all of the following components:

- **Data sources**: All big data solutions start with one or more data sources. Examples include:
    - 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 typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a *data lake*. Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.
- **Batch processing**: Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files. Options include running U-SQL jobs in Azure Data Lake Analytics, using Hive, Pig, or custom Map/Reduce jobs in an HDInsight Hadoop cluster, or using Java, Scala, or Python programs in an HDInsight Spark cluster.
- **Real-time message ingestion**: If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are dropped into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. Options include Azure Event Hubs, Azure IoT Hubs, and Kafka.
- **Stream processing**: After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL queries that operate on unbounded streams. You can also use open source Apache streaming technologies like Spark Streaming in an HDInsight cluster.
- **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. The analytical

data store used to serve these queries can be a Kimball-style relational data warehouse, as seen in most traditional business intelligence (BI) solutions. Alternatively, the data could be presented through a low-latency NoSQL technology such as HBase, or an interactive Hive database that provides a metadata abstraction over data files in the distributed data store. Azure Synapse Analytics provides a managed service for large-scale, cloud-based data warehousing. HDInsight supports Interactive Hive, HBase, and Spark SQL, which can also be used to serve data for analysis.

- **Analysis and reporting**: Analytics refers to the systematic computational analysis of data or statistics. It involves applying various techniques such as statistical analysis, data mining, predictive modeling, and machine learning to interpret and make sense of data. The primary goal of analytics is to uncover patterns, generate insights, and support decision-making by predicting future trends and behaviors based on historical data.

  (Analytics ka matlab hai data ya statistics ka systematic analysis karna. Ismein alag-alag techniques ka use hota hai jaise ki statistical analysis, data mining, predictive modeling, aur machine learning. Analytics ka main objective hai data mein patterns ko samajhna, insights lena, aur future trends aur behaviors ko predict karna taaki better decisions liye ja sakein.)

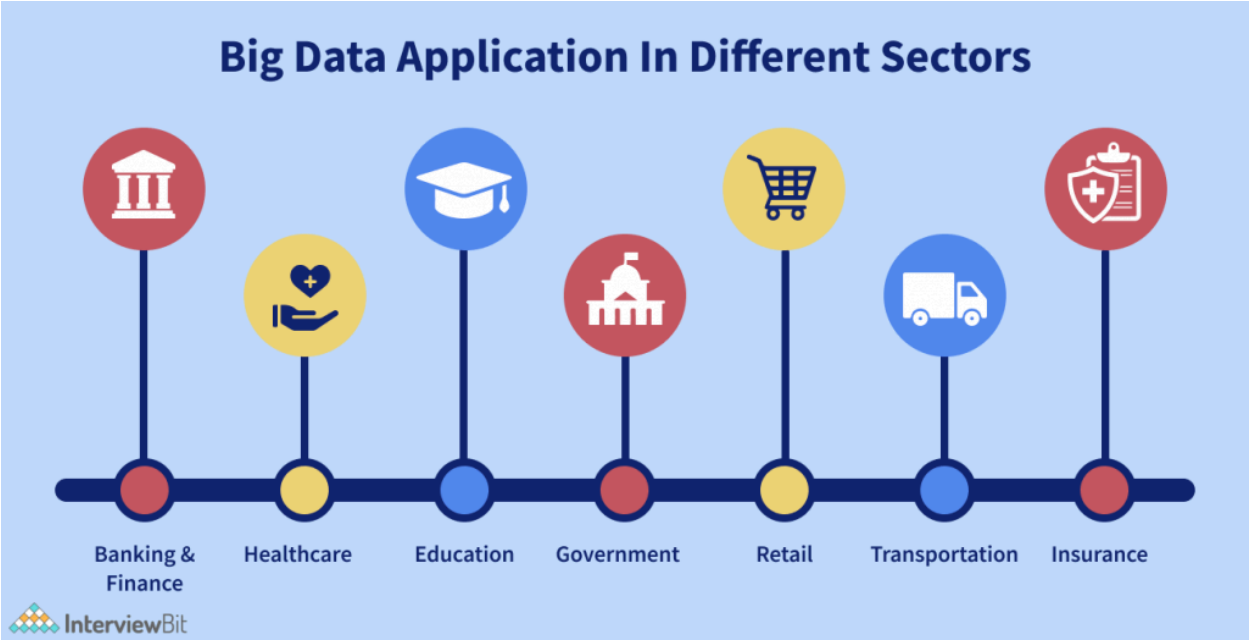
- Reporting is the process of organizing data into summaries to monitor how different areas of a business are performing. It typically involves the creation of static or dynamic reports, often in the form of tables, charts, and dashboards, that present data in an easily understandable format. The purpose of reporting is to provide a clear and concise view of past and present data to help stakeholders track performance and make informed decisions.

  (Reporting ka matlab hai data ko organize karke summaries mein present karna taaki business ke alag-alag areas ka performance track kiya ja sake. Reporting mein static ya dynamic reports banayi jaati hain, jismein tables, charts, aur dashboards ke through data ko easily samajhne layak format mein dikhaya jaata hai. Reporting ka purpose hai ek clear aur concise view dena past aur present data ka, taaki stakeholders ko performance track karne aur decisions lene mein help mile.)

- **Orchestration**: Orchestration in Big Data refers to the automated management, coordination, and arrangement of complex data processing tasks across different systems and environments. It involves the integration and scheduling of various data pipelines, ensuring that data flows smoothly from collection to processing, storage, and analysis without manual intervention. Orchestration tools manage the dependencies between tasks, handle failures, and optimize resource usage.

  (Orchestration ka matlab Big Data ke context mein hota hai data processing tasks ko automate karke manage karna, coordinate karna, aur arrange karna. Yeh process ensures karta hai ki data smoothly flow kare from collection to processing, storage, aur analysis tak, bina manual intervention ke. Orchestration tools alag-alag tasks ke beech dependencies manage karte hain, failures handle karte hain, aur resources ka efficient use karte hain.)

Applications of Big Data



1. Banking

Banking Application

Be it financial management or cash collection, big data has made banks more efficient for each industry. The technology's application has defeated the user's struggle, helping the bank to generate more revenue and their insights are more transparent and comprehensible than before. Varying from distinguishing fraud, analyzing and streamlining transaction processing, improving understanding of the users, perfecting trade execution, and promoting an exceptional user experience, Big Data extends a range of applications.

An interesting example of a company making use of Big Data efficiently in this sector is of Western Union. The company promotes an omnichannel approach that personalizes user experiences by processing more than 29 transactions per second and collecting all the data onto a common platform for the purpose of statistical modelling and predictive interpretation.

2. Education

When talking about the Education industry, the data garnered from the courses, students, faculty, and results is huge, the interpretation of which can bring forth insights useful for improving the operations and functioning of educational institutes. From promoting efficient learning, improving International recruiting for universities, supporting students in establishing career goals, decreasing university dropouts, promoting definite student evaluation, enhancing the decision-making process, and improving student results, Big Data has an indispensable role in this sector.

An outstanding example would be that of the University of Florida. The university uses IBM InfoSphere for obtaining, loading, and transporting data through various resources. The university employs IBM SPSS Modeler in the case of predictive analytics and data modelling, and IBM Cognos Analytics for examining and foretelling the performances of students.

Several variables varying from the student's grades, demographics, and economic background aid in measuring the assess dropout possibilities for the students. This helps the university in establishing its policies and promoting regular intervention for students on the verge of dropping out.

3. Media

The buzz for the conventional methods of consuming media is gradually fading away because the current strategies of consuming online content with the help of gadgets have become the latest trend. Since an immense amount of data is generated, big data has triumphantly made its way into this industry. Ranging from assisting to predicting what the audience needs, in the genre, music, and content as per their age group, to proposing them insights regarding customer churn, Big Data has made the lives of media houses much easier.

Another appropriate example of how big data plays a pivotal role in transforming media platforms would be Netflix. The technology not only impacts the series invested in by Netflix but also how the series is presented to their users. The search history and viewing history of the user, including the places where the user has paused the video, impacts everything from the personalized thumbnails to the shows we watch on the "Popular on Netflix" section.

4. Healthcare

Big Data has an essential role to play in improving modern healthcare operations. Technology has fully remodelled the healthcare sector By decreasing the cost of treatment, predicting epidemic outbreaks, dodging preventable diseases, improving life quality, prophesying the income obtained by daily patients to adjust staffing, adopting Electronic Health Records (EHRs), using real-time alerts to promote immediate care, utilizing health data for more efficient strategic planning, to decreasing frauds and flaws.

Mayo Clinic is a credible example of Big Data in healthcare. The platform embraces big-data analytics to help in identifying various conditions of patients and enhancing their life quality. This analytics can also identify at-risk patients and grant them more comprehensive health control and necessary medical intervention.

5. Agriculture

In the field of Agriculture, big data analytics drives smart farming and accurate agriculture operations, saving costs and unleashing new business possibilities. Some important areas where big data work involve meeting the food demand by providing farmers with information regarding the changes in weather, rainfall, and factors affecting crop yield, propelling smart and correct application of pesticides, management of equipment, guaranteeing supply chain productivity, etc.

A Bayer Group unit, Bayer Digital Farming, uses machine learning and artificial intelligence to identify weeds. Farmers share pictures of weeds in the app and then match them against a large Bayer database to detect the species. This app intervenes at the appropriate time, shielding the crops and improving yields.

6. Travel

Big Data plays an intrinsic role in shaping transportation in a more perfect and effective manner. Be it managing the revenue earned, maintaining the reputation gained, or following strategic marketing, Big Data has influenced this sector. It also helps in mapping out the route as per the requirements of the user, assisting in efficiently managing wait time, and identifying accident-prone areas to increase the safety level of traffic.

The perfect example of Big Data's use in the transportation industry would be Uber. The platform creates and uses a huge range of data on vehicles, drivers, locations, and the trip made by each vehicle, which is again tested and utilized for foretelling the demand, supply, accurate location of drivers, and trip fares.

7. Manufacturing

Thanks to Big Data, manufacturing is no longer an arduous manual process. Technology and Data analytics have succeeded in completely revolutionizing the manufacturing process. Big Data improves manufacturing, personalizing product design, guaranteeing accurate quality maintenance, overseeing the supply chain, and also evaluating to keep track of potential risks.

Rolls Royce is an interesting example of the importance of big data in this industry. The company uses Big data analytics for improving its design process, decrease its product development term, and magnify the production and quality of its products, at the same time reducing costs. The platform has also made its production processes easier by rectifying errors happening amidst the design process.

8. Government

Governments come across a huge level of data on an everyday basis, irrespective of the nation as they have to maintain various records and databases of their citizens, growth, geographical surveys, energy resources, etc. This data is needed to be reviewed and analyzed, thereby becoming an ally for the government in its operations. Primarily, the government utilizes this data in two areas, in its developmental plans and in the case of cybersecurity.

The Department of Homeland Security (DHS) is an important example of the government's utilization of Big Data. For ensuring security, the Department of Homeland Security (DHS) uses an intrusion-identifying system for sensors that has the capacity to investigate internet traffic both in and out of Federal systems other than recognizing attempts of malware and unsanctioned access.

9. Retail

Talking about retail, big data plays an important part in foretelling rising trends, targeting fitting customers at the relevant time, reducing marketing expenses, and improving the quality of customer service. From keeping a detailed view of each user and promoting personal engagement, enhancing pricing to acquire the best value from forthcoming trends, systemizing back-office operations, and improving customer services, Big Data gives a wide array of applications when it comes to Retail.

Amazon uses the Big Data obtained from consumers to grow its recommendation engine. The more Amazon learns about its consumers, the better it can foretell what the consumer long to buy. Knowing about consumer behaviour enables them to simplify the process and convince the consumer to purchase it, like recommending particular products instead of making the consumer browse through the entire catalogue.