Unit-1

Introduction to Data Science

What is Data Science

- The term “Data Science” has emerged only recently to specifically designate a new profession that is expected to make sense of the vast stores of big data. But making sense of data has a long history and has been discussed by scientists, statisticians, librarians, computer scientists, and others for years.
- Nowadays, Data Science as a business field is really complicated, and due to its remarkable popularity, there are numerous descriptions of data science
- However, in simple words, **data scientists just try to get insights from massive amounts of data that can help companies make smarter business decisions.** We also define Data Science as a methodology by which actionable insights can be inferred from data.
- Data science uses a wide array of data-oriented technologies, including **SQL**, **Python**, **R**, and **Hadoop**, etc. However, it also makes extensive use of statistical analysis, data visualization, distributed architecture, and more to extract meaning out of sets of data. The information extracted through data science applications is used to guide business processes and reach organizational goals.

A brief history of Data Science

- Data Science has revolutionized several different aspects of our world. Let's take a look then at when and where data science comes from.
- In 1962, John W. Tukey wrote in “The Future of Data Analysis” - The first milestone in the history of data science is globally recognized for the bright American mathematician John Tukey. The influence of John Tukey in statistical terms is enormous, but the most famous coinage attributed to him is related to computer science. In fact, it should be mentioned that he was the first to introduce the term "bit" as a contraction of "binary digit."
- In 1974, Peter Naur published the Concise Survey of Computer Methods, which surveyed data processing methods across a wide variety of applications. The term “data science” becomes clearer, as he puts his own definition on it: “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”
- In 1977, the International Association for Statistical Computing (IASC) was founded.
- In 1989, Gregory Piatetsky-Shapiro organized and chaired the first Knowledge Discovery in Databases (KDD) workshop.

A brief history of Data Science

- In 1994, BusinessWeek published a cover story on “Database Marketing.”
- In 1996, on the occasion of the conference of the International Federation of Classification Societies (IFCS), for the first time, the term “data science” is included in the title of the conference (“Data science, classification, and related methods”). In the same year, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth publish “From Data Mining to Knowledge Discovery in Databases.”
- In 1997, during his inaugural lecture as the H. C. Carver Chair in Statistics at the University of Michigan, Jeff Wu called for statistics to be renamed “data science” and statisticians to be renamed “data scientists.”

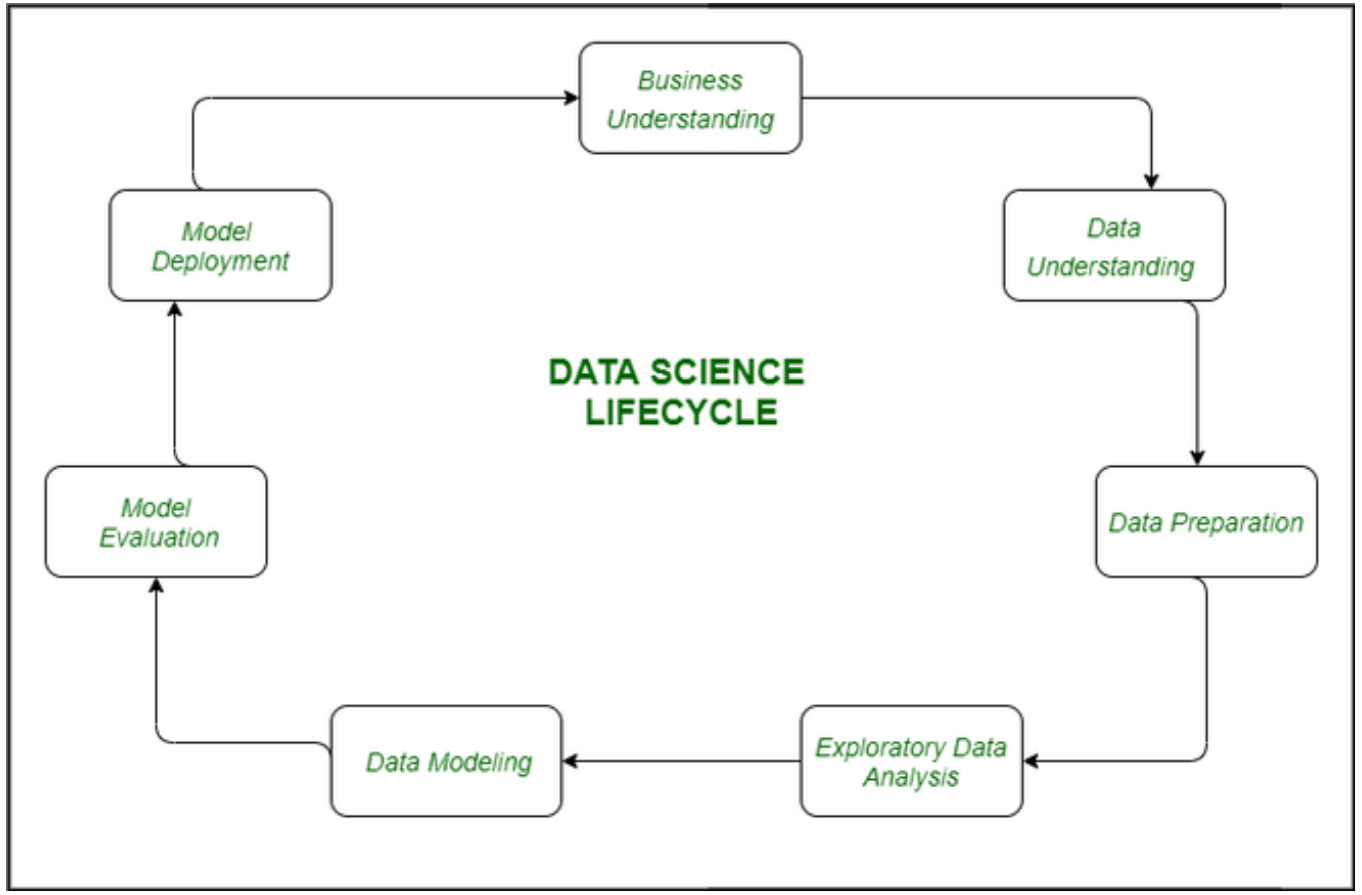
A brief history of Data Science

- Since the beginning of the 21st century, data stockpiles have expanded exponentially, largely thanks to advents in processing and storage that are both efficient and cost-effective at scale. The capability to collect, process, analyze, and display data and information in “real-time” give us an unprecedented opportunity to conduct a new form of knowledge discovery. To process this huge amount of data, Data Scientists need high performance also of a large portfolio of technologies to speed up tasks and data processing in a matter of seconds.
- Disruptive technologies like artificial intelligence, machine learning, and deep learning are nowadays available for Data Scientists thanks to powerful platforms available.

Data Science Life Cycle

- Data Science Lifecycle revolves around the use of machine learning and different analytical strategies to produce insights and predictions from information in order to acquire a commercial enterprise objective.
- The complete method includes a number of steps like data cleaning, preparation, modelling, model evaluation, etc.
- It is a lengthy procedure and may additionally take quite a few months to complete.
- So, it is very essential to have a generic structure to observe for each and every hassle at hand.
- The globally mentioned structure in fixing any analytical problem is referred to as a Cross Industry Standard Process for Data Mining or CRISP-DM framework.

Data Science Life Cycle



Data Science Life Cycle

- **1. Business Understanding:**
- The complete cycle revolves around the enterprise goal. What will you resolve if you do not longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

Data Science Life Cycle

- **2. Data Understanding:**
- After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information.
- This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

Data Science Life Cycle

- **3. Preparation of Data:**
- Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them.
- Constructing new data, derive new elements from present ones.
- Format the data into the preferred structure, eliminate undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

Data Science Life Cycle

- **4. Exploratory Data Analysis:**
- This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model.
- Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps.
- Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

Data Science Life Cycle

- **4. Exploratory Data Analysis:**
- This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model.
- Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps.
- Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

Data Science Life Cycle

- **5. Data Modeling:**
- Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output.
- This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem.
- After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them.
- We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

Data Science Life Cycle

- **5. Data Modeling:**
- Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output.
- This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem.
- After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them.
- We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

Data Science Life Cycle

- **6. Model Evaluation:**
- Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics.
- We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved.
- Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric.
- We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

Data Science Life Cycle

- **7. Model Deployment:** The model after a rigorous assessment is at the end deployed in the preferred structure and channel.
- This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully.
- If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste.
- For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

Applications of Data Science

1. In Search Engines

- The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

2. In Transport

- Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

3. In Finance

- Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company.
- Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

Applications of Data Science

- **4. In E-Commerce**
- E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.
- **5. In Health Care**
- In the Healthcare Industry data science act as a boon. Data Science is used for:
- Detecting Tumor.
- Drug discoveries.
- Medical Image Analysis.
- Virtual Medical Bots.
- Genetics and Genomics.
- Predictive Modeling for Diagnosis etc.

Applications of Data Science

- **6. Image Recognition**
- Currently, Data Science is also used in Image Recognition.
- **For Example,** When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.
- **7. Targeting Recommendation**

Targeting Recommendation is the most important application of Data Science.

Whatever the user searches on the Internet, he/she will see numerous posts everywhere.

This can be explained properly with an example: Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy offline. Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

Applications of Data Science

- **8. Airline Routing Planning**
- With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.
- **9. Data Science in Gaming**
- In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance. There are many games like Chess, EA Sports, etc. will use Data Science concepts.

Applications of Data Science

- **10. Medicine and Drug Development**
- The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.
- **11. In Delivery Logistics**
- Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

Applications of Data Science

- **12. Autocomplete**
- AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.

Natural Language Processing (NLP)

- Natural Language Processing or NLP in data science is the automatic manipulation of natural languages, like speech and text, by using software that helps computers observe, analyze, understand, and derive valuable meaning from natural or human-spoken languages.
- In other words, it is a branch of data science that focuses on training computers to process and interpret conversations in text format in a way humans do by listening.
- It is a field that is developing methodologies for filling the gap between Data Science and human languages.

Natural Language Processing (NLP)

- NLP applications are difficult and challenging during development as computers require humans to interact with them using programming languages like Java, Python, etc., which are structured and unambiguous. But human-spoken languages are ambiguous and change with regional or social change, so it becomes challenging to train computers to understand natural languages.

NLP use cases

- Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:
- **Spam detection:** You may not think of spam detection as an NLP solution, but the best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing.
- These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more.
- Spam detection is one of a handful of NLP problems that experts consider 'mostly solved' (although you may argue that this doesn’t match your email experience).

NLP use cases

- **Machine translation:** Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another.
- Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language. Machine translation tools are making good progress in terms of accuracy.

NLP use cases

- **Virtual agents and chatbots:** Virtual agents such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments.
- Chatbots perform the same magic in response to typed text entries. The best of these also learn to recognize contextual clues about human requests and use them to provide even better responses or options over time.

NLP use cases

- **Social media sentiment analysis:** NLP has become an essential business tool for uncovering hidden data insights from social media channels.
- Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

NLP use cases

- **Text summarization:** Text summarization uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text.
- The best text summarization applications use semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

Computer Vision

- Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world.
- Machines can accurately identify and locate objects then react to what they **“see”** using digital images from cameras, videos, and deep learning models.

How does computer vision work?

- Computer vision needs lots of data. It runs analyses of data over and over until it discerns distinctions and ultimately recognize images. For example, to train a computer to recognize automobile tires, it needs to be fed vast quantities of tire images and tire-related items to learn the differences and recognize a tire, especially one with no defects.
- Two essential technologies are used to accomplish this: a type of machine learning called deep learning and a convolutional neural network (CNN).

Computer vision applications

- IBM used computer vision to create My Moments for the 2018 Masters golf tournament. IBM Watson watched hundreds of hours of Masters footage and could identify the sights (and sounds) of significant shots. It curated these key moments and delivered them to fans as personalized highlight reels.
- Google Translate lets users point a smartphone camera at a sign in another language and almost immediately obtain a translation of the sign in their preferred language.
- The development of self-driving vehicles relies on computer vision to make sense of the visual input from a car’s cameras and other sensors. It’s essential to identify other cars, traffic signs, lane markers, pedestrians, bicycles and all of the other visual information encountered on the road.
- IBM is applying computer vision technology with partners like Verizon to bring intelligent AI to the edge, and to help automotive manufacturers identify quality defects before a vehicle leaves the factory.

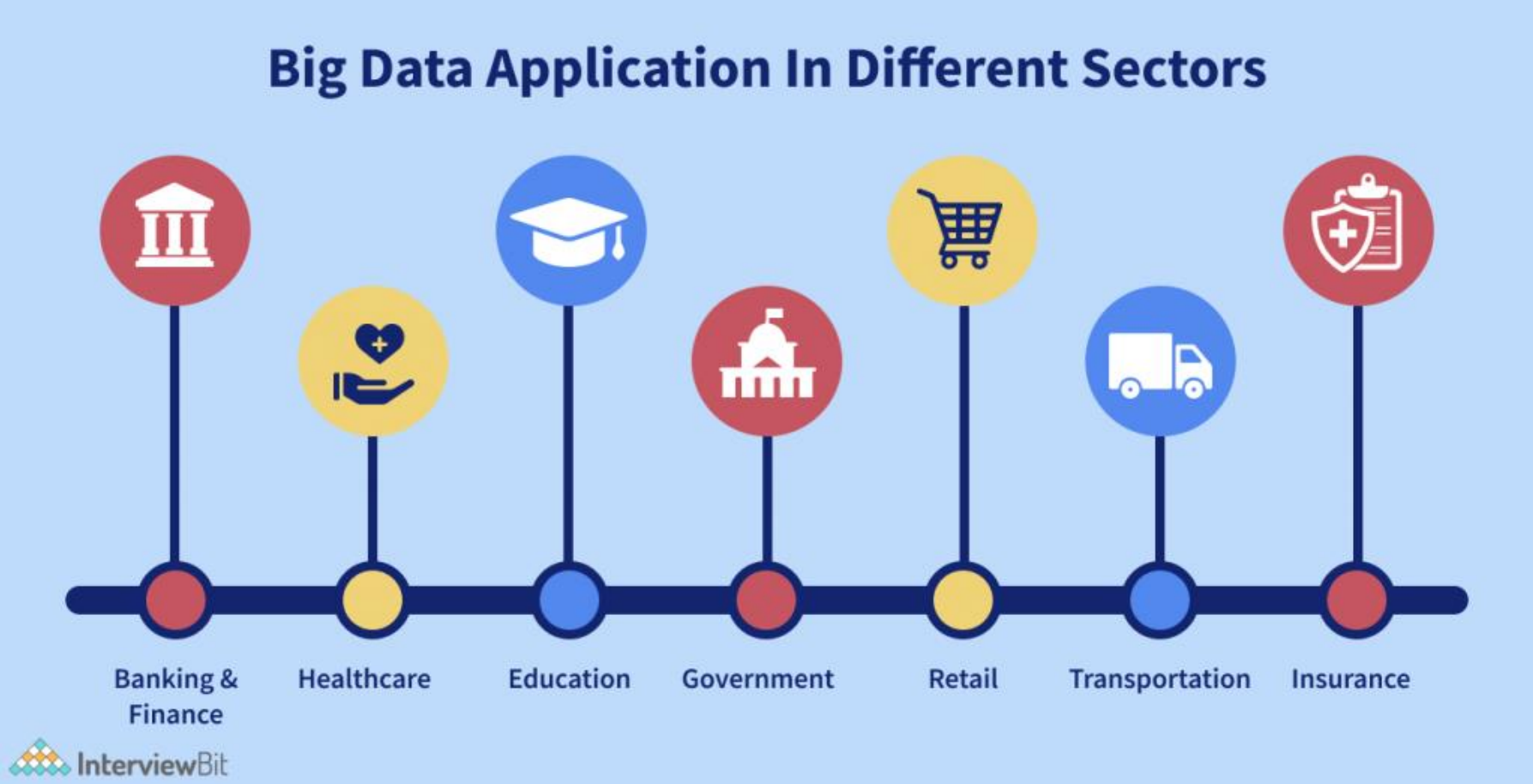
Big Data

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- Systems that process and store big data have become a common component of data management architectures in organizations, combined with tools that support big data analytics uses. Big data is often characterized by the three V's:
- the large *volume* of data in many environments;
- the wide *variety* of data types frequently stored in big data systems; and
- the *velocity* at which much of the data is generated, collected and processed.

Big Data

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- Systems that process and store big data have become a common component of data management architectures in organizations, combined with tools that support big data analytics uses. Big data is often characterized by the three V's:
- the large *volume* of data in many environments;
- the wide *variety* of data types frequently stored in big data systems; and
- the *velocity* at which much of the data is generated, collected and processed.

Applications of Big Data



Applications of Big Data

- **1. Banking**
- 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.

Applications of Big Data

- **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.

Applications of Big Data

- **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.

Applications of Big Data

- **4. Healthcare**
- Big Data has an essential role to play in improving modern healthcare operations.
- Technology has fully remodeled 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.

Applications of Big Data

- **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.

Applications of Big Data

- **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.

Applications of Big Data

- **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.

Issues in data science

- Data science is one of the most exciting fields at present that are empowering companies to enhance their business.
- With so much data constantly being produced by network servers, IoT sensors, official social media pages, databases, and company logs, it has to be handled and cannot be ignored. Data scientists collect these data sets, remove the unwanted data and then, analyse it.

Issues in data science

- **1. Identifying the data problem**
- One of the toughest **challenges of data science** is identifying the problem or the issue. Data scientists mostly start off with a huge data set that is often unstructured. They have to understand what they have to do with this data.
- **2. Finding the most appropriate data**
- As companies produce huge amounts of data every second, it is a daunting task to get your hands on the right data for analysis. This is because the correct data set will be crucial for developing the most appropriate data model. The right data having the right format will take less time to clean and analyse.

Issues in data science

- **3. Lack of skilled workforce**
- As more and more companies are becoming dependent on data science, the demand for skilled data professionals is increasing. This is one of the major **challenges of data science** at this hour. The traditional methods of working with data have changed. But, the fact is that many employees have not been able to keep up with the pace of developments.
- **4. Data cleansing**
- Data cleansing or removing unwanted data from a data set is one of the pressing challenges of data science. It is observed that companies lose almost 25% of their revenue as cleaning bad data is costly. Working on data sets consisting of many inconsistencies and unwanted information can create havoc in a data scientist’s life!