Asian Journal of Electrical Sciences
ISSN: 2249-6297 (P) Vol.9, No.1, 2020, pp.1-12

DOI: https://doi.org/10.51983/ajes-2020.9.1.2374

The Art of Data Science and Big Data Analytics: Inspecting and Transforming Data

Akella. Subhadra
Associate Professor, Department of Computer Science & Engineering, BVCITS, Amalapuram, Andhrapradesh, India.
E-mail- asubhadra2012@gmail.com
(Received 10 December 2019; Revised 2 January 2020; Accepted 23 January 2020; Available online 29 January 2020)

Abstract - **Data Science is associated with new discoveries, the discovery of value from the data. It is a practice of deriving insights and developing business strategies through transformation of data in to useful information. It has been evaluated as a scientific field and research evolution in disciplines like statistics, computing science , intelligence science , and practical transformation in the domains like science, engineering, public sector, business and lifestyle. The field encompasses the larger areas of artificial intelligence, data analytics, machine learning, pattern recognition, natural language understanding, and big data manipulation.**

It also tackles related new scientific challenges, ranging from data capture, creation, storage, retrieval, sharing, analysis, optimization, and visualization, to integrative analysis across heterogeneous and interdependent complex resources for better decision-making, collaboration, and, ultimately, value creation. In this paper we entitled epicycles of analysis, formal modeling, from data analysis to data science, data analytics -A keystone of data science, The Big data is not a single technology but an amalgamation of old and new technologies that assistance companies gain actionable awareness. The big data is vital because it manages, store and manipulates large amount of data at the desirable speed and time. In particular, big data addresses detached requirements, in other words the amalgamate of multiple un-associated datasets, processing of large amounts of amorphous data and harvesting of unseen information in a time-sensitive generation.

As businesses struggle to stay up with changing market requirements, some companies are finding creative ways to use Big Data to their growing business needs and increasingly complex problems. As organizations evolve their processes and see the opportunities that Big Data can provide, they struggle to beyond traditional Business Intelligence activities, like using data to populate reports and dashboards, and move toward Data Science- driven projects that plan to answer more open-ended and sophisticated questions.

Although some organizations are fortunate to have data scientists, most are not, because there is a growing talent gap that makes finding and hiring data scientists in a timely manner is difficult. This paper, aimed to demonstrate a close view about Data science, big data, including big data concepts like data storage, data processing, and data analysis of these technological developments, we also provide brief description about big data analytics and its characteristics , data structures, data analytics life cycle, emphasizes critical points on these issues.
Keywords: **Data Science Big Data, Data Analytics, Epicycles, Business Intelligence (BI).**

I. INTRODUCTION

Data analysis is tough, and a part of the matter is that few people can explain the way to roll in the way. It's not that

there are no people doing data analysis on a daily basis, It's that the people who are really good at it have yet to en lighten us about the thought process that goes on in their heads. Unfortunately, the method of knowledge analysis isn't one that we've been ready to write down effectively. But in our opinion, none of these really addresses the core problems involved in conducting real-world data. Analyze a data analysis may appear to follow a linear, one-step-after-the other process. Consequently, prescriptive decision taking strategies, business rules, actions, and recommendations are disseminated to decision makers for the purpose of taking of data. Data scientist refers to those people whose roles very much centre on data. Descriptive analytics refers to the type of data analytics that typically uses statistics to describe the data used to gain information, or for other useful purposes. Predictive analytics refers to the sort of knowledge analytics that creates predictions about unknown future events and discloses the explanations behind them, typically by advanced analytics. Prescriptive analytics refers to the type of data analytics that optimizes indications and recommends actions for smart decision-making. Explicit analytics focuses on descriptive analytics typically by reporting, descriptive analysis, alerting, and forecasting. Implicit analytics focuses on deep analytics, typically by predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery. Deep analytics refers to data analytics which will acquire an in-depth understanding of why and the way things have happened, are happening, or will happen, which can't be addressed by descriptive analytics.

The art of data science has attracted increasing interest from a wide range of domains and disciplines. Some examples are that data science is the new generation of statistics, is a consolidation of several interdisciplinary fields, or is a new body of knowledge. Data science also has implications for providing capabilities and practices for the information profession, or for generating business strategies. Statisticians have much to say about data science, since it is they who actually created. Data science is the science of

data or data science is the study of data. Definition From the disciplinary perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology.

Accordingly, a discipline-based data science formula is given as follows: data science = statistics + informatics + computing + communication+ sociology + management | data + environment + thinking, where "|" means "conditional on[3]. Big Data is creating significant new opportunities for organizations to derive new value and make competitive advantage from their most precious asset information. For businesses, Big Data helps drive efficiency, quality, and personalized products and services, producing improved levels of customer satisfaction and profit. For scientific efforts, Big Data analytics enable new avenues of investigation with potentially richer results and deeper insights than previously available. In many cases, Big Data analytics integrate structured and unstructured data with real-time feeds and queries, opening new paths to innovation and insight Data analysis is a highly iterative and non-linear process, better reflected by a series of epicycles during which information is learned at each step, which then informs whether (and how) to refine, and redo, the step that was just performed, or whether (and how) to proceed to the next step.

Social media and genetic sequencing are among the fastest-growing sources of massive Data and samples of untraditional sources of knowledge getting used for analysis. For example, Facebook users posted 700 status updates per second worldwide, which can be leveraged to reduce latent interests or political views of users and show relevant ads. For instance, an update in which a woman changes her relationship status from "single" to "engaged" would trigger ads on bridal dresses, wedding planning, or name-changing services. Facebook can also construct social graphs to analyze which users are connected to each other as an interconnected network. Facebook released a new feature called "Graph Search," enabling users and developers to search social graphs for people with similar interests, hobbies, and shared locations [2]. Another example comes from genomics. Genetic sequencing and human genome mapping provide an in depth understanding of genetic makeup and lineage. The health care industry is looking toward these advances to help predict which illnesses a person is likely to get in his lifetime and take steps to avoid these maladies or reduce their impact through the use of personalized medicine and treatment. Such tests also highlight typical responses to different medications and pharmaceutical drugs, heightening risk awareness of specific drug treatments.

We are living in the age of big data, advanced analytics, and data science. The trend of "big data growth" has not only triggered tremendous hype and buzz, but more importantly presents enormous challenges that successively bring incredible innovation and economic opportunities. Big data has attracted intensive and growing attention, initially from giant private data-oriented enterprise and lately from major governmental organizations and academic institutions. Typical examples include large data-centric projects in Google, Facebook, and IBM. From the disciplinary

development perspective, recognition of the significant challenges, opportunities, and values of big data is fundamentally reshaping the traditional data-oriented scientific and engineering fields. It is also reshaping those non-traditional data engineering domains such as social science, business, and management.

The article is organized as follows. Section II describes related work, Section III Theory that covers Epicycles of Analysis, Formal Modelling of data, tracks the progression from data analysis to data science, Data Analytics: A key stone of Data Science, Overview of Data Analytics Life Cycle. Section IV Methodology, in this is Data Science, the main features, initiatives, activities, and status of the era of data science, characteristics, components, lifecycle; Applications of data science, what is Big data, Big data Analytics, characteristics, and platforms are summarized. In Section V the main emphasis is on Results and Discussion. Section VI covers Conclusion of this work and Future Scope, followed by References.

II. RELATED WORK

The growth and recognition of an emerging field can be effectively measured in terms of the formation width, depth, and speed of its professional communities. The data science and analytics community is growing incredibly quickly [2]. The first indicator is the emergence of dedicated publication venues in this area.

Several journals on data science have been established. These include the Journal of Data Science [JDS 2002], launched in 2002, which is devoted to applications of statistical methods at large; the electronic Data Science Journal [DSJ 2014] relaunched by CODATA in 2014; the EPJ Data Science [EPJDS 2012] launched in 2012; The International Journal of Data Science and Analytics (JDSA) [JDSA 2015] in 2015 by Springer; IEEE Transactions on Big Data [TOBD 2015] in 2015; and the Springer Series on Data Science [SSDS 2015] and the Data Analytics Book Series [DABS 2016].Other publications are in development by various regional and domain-specific publishers' and groups. Some examples are the International Journal of Data Science, Data Science and Engineering. Published on behalf of the China Computer Federation (CCF), the International Journal of Research on Data Science (IJRD), and the Journal of Finance and Data Science (JFDS).

The second indicator can be found in the creation of a data science community that is significantly enhanced by

conferences, workshops, and forums dedicated to the promotion of data science and analytics. There are also many well-established venues that either focus on specific aspects such as KDD and ICML or have adjusted their previous non data and/or analytics focus, such as the traditional AI conferences IJCAI and AAAI.

The first conference to adopt "data science" as a topic was the 1996 IFCS Conference on Data Science, Classification, and Related Methods [IFSC-96 1996], which included papers on general data analysis issues.

The IEEE International Conference on Data Science and Advanced Analytics (DSAA) [DSAA 2014] launched in 2014, was probably the first conference series dedicated to both data science and analytics research and practice. Cosponsored by ACM SIGKDD, IEEE CIS, and the American Statistics Association (ASA), it attracted wide and significant interest from statistics, industry, business, IT, and professional bodies. The IEEE Conference on Big Data is an event dedicated to broad areas of big data.

Several other domain-specific and regional initiatives have emerged, such as the three initiatives in India, that is, the Indian Conference on Data Sciences, the International Conference on Big Data Analytics, and the International Conference on Data Science and Engineering.

Several other conference series have been renamed and repositioned from their original focus on topics such as software engineering and service-based computing to connect with big data and data science, drawing mainly on key topics of interest and participants from their original areas.

Data analytics, machine learning, and big data have eclipsed the original topics of interest in many traditionally non data and/or analytics conferences, such as IJCAI, AAAI, VLDB, SIGMOD, and ICDE. Not surprisingly, some of these venues now frequently incorporate more than 50% of papers on data science matters.

The third indicator is the growth and development of professional (online) communities' and organizations established publicly or privately to promote big data, analytics and data science research, practices and education, and interdisciplinary communications.

For example:

The IEEE Big Data Initiative [IEEEBD 2014] aims to "provide a framework for collaboration throughout IEEE," and states that "Plans are under way to capture all the different perspectives via in depth discussions, and to drive to a set of results which will define the scope and the direction for the initiative."

The IEEE Task Force on Data Science and Advanced Analytics (TF-DSAA) [TFDSAA 2013] was launched in 2013 to promote relevant activities and community building, including the annual IEEE Conference on Data Science and Advanced Analytics.

The International Institute of Data & Analytics [IDA 2014] aims to bridge the gaps between academia and industry through the promotion of data and analytics research, education, and development.

The China Computer Federation Task Force on Big Data [CCF-BDTF 2013] consists of a network of representatives from academia, industry, and government, and organizes its annual big data conference with participants from industry and government.

Several groups and initiatives promote dedicated activities of analytics and data science. For instance, Datasciences.org [2005] collects relevant information about data science research, courses, funding opportunities, professional activities, and platforms for collaborations and partnership. The Data Science Community [DSC 2016b] claims to be the European Knowledge Hub for Big data and Data science.

Data Science Central [DS Central 2016] aims to be the industry's online resource for big data practitioners. The Data Science Association [DSA 2016] aims to be a "professional group offering education, professional certification, conferences and meetups" [Galetto 2016], and even offers a "Data Science Code of Professional Conduct."

Many existing consulting and servicing organizations have adjusted their scope to cover analytics, where they previously focused on other disciplinary matters. Interdisciplinary efforts have been made to promote cross-domain and cross-disciplinary activities and growth opportunities. Examples include [INFORMS 2016], Gartner, McKinsey, Deloitte, PricewaterhouseCoopers, KPMG, and Bloomberg.

III. THEORY

A.Epicycles of Analysis

A data analysis follows a linear, one-step-after-the-other process which at the end, arrives at a nicely packaged and coherent result. In general data analysis is iterative and non-linear process, containing series of epicycles in which information is learned at each step we have to learn about whether its necessary to refine, and redo, the step that was just performed, or proceed to the next step.

An epicycle may be a small circle whose center moves round the circumference of a bigger circle. In data analysis, the iterative process that is applied to all or any steps of the info analysis are often conceived of as an epicycle that is repeated for every step along the circumference of the whole data analysis process[1].

There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

These 5 activities can occur at different time scales: for instance, you would possibly undergo all 5 within the course of each day, but also affect each, for an outsized project, over the course of many months. Although there are many various sorts of activities that you simply might engage in while doing data analysis, every aspect of the whole process is consider it as an interactive process that what we call the "epicycle of knowledge analysis". These five activities may perform within the following steps:

1. Setting Expectations.
2. Collecting the data and then comparing the data to your expectations, and if the expectations don't match, this step represents that we may collect information about your question or your data. you may collect information about your question by literature search or asking experts so as to make sure that your question may be a good one.
3. Revising your expectations or fixing the information applying these three steps in an iterative manner is known as "epicycle of knowledge analysis."
4. There are two possible outcomes either your expectations match with the data or not matched with the data. If your expectations and the data match, you can move onto the next if the data don't match, there are two possible alternatives, first, your expectations were wrong and wish to be revised, or second, the check was wrong and contains a mistake . You review the check and conclude that there is an error in the data, so ask for the check to be corrected. You want to setup your expectations and your data so that matching the two up is easy.

B.Formal Modeling

Formal models allow you to identify clearly what you are trying to infer from data and what form the relationships between features of the information take [1].

The Goals of Formal Modeling:

1. One key goal of formal modeling is to develop a precise specification of the question and how the data can be used to answer that question. Parameters play an important role in many formal statistical models.
2. Another goal of formal modeling is to develop a rigorous framework with which you can challenge and test your primary results. At this point in your data analysis, you've stated and refined your question, you've explored the data visually and may be conducted some exploratory modeling. The key thing is that you likely have a pretty good sense of what the answer to your question is, but maybe have some doubts about whether your findings will hold up under intense scrutiny. Assuming you are still interested in moving forward with your results, this is where formal modeling can play an important role.

C.General Framework

We can apply the basic epicycle of analysis to the formal modeling portion of data analysis. We still want to set expectations, collect information, and refine our expectations based on the data. In this there are three phases as follows [1].

1. Setting expectations: Setting expectations comes in the form of developing a primary model that represents best sense of what provides the answer to the question. This model is chosen based on whatever information that is currently available.
2. Collecting Information: Once the primary model is set, we want to create a set of secondary models that challenge the primary model in some way.
3. Revising expectations: If our secondary models are successful in challenging our primary model and put the primary model's conclusions in some doubt, then we may need to adjust or modify the primary model to better reflect what we have learned from the secondary models.

D.Primary Model

It's often useful to start with a primary model. This model will likely be derived from any exploratory analyses that you have already conducted and will serve as the lead candidate for something that summarizes the results and matches the expectations. It's important to realize that at any given moment in a data analysis, the primary model is not necessarily the final model.

It is simply the model against which you will compare other secondary models. The process of comparing your model to other secondary models is often referred to as sensitivity analyses, because you are interested in seeing how sensitive your model is to changes, such as adding or deleting predictors or removing outliers in the data. Through the iterative process of formal modeling, you may decide that a different model is better suited as the primary model.

E.Secondary Models

Once you have decided on a primary model, you will then typically develop a series of secondary models. The purpose of these models is to test the legitimacy and robustness of your primary model and potentially generate evidence against your primary model. If the secondary models are successful in generating evidence to the primary model, then you may need to revisit the primary model and check whether its conclusions are still reasonable or not.

IV. FROM DATA ANALYSIS TO DATA SCIENCE

The development of data mining, knowledge discovery, and machine learning, together with the original data analysis and descriptive analytics from the statistical perspective, forms the general concept of "data analytics"[2].The initial data analysis focused on processing data. Data analytics is the multidisciplinary science of quantitatively and qualitatively examining data for the purpose of drawing new conclusions or insights (predictive), or for extracting and proving (fact-based) hypotheses about that information for decision-making and action. Analytics has also become more business oriented. It now extends to a variety of data and domain-specific analytical tasks, such as business analytics, risk analytics, behavior analytics, social analytics, and web analytics. Domain-specific analytics fundamentally drives the innovation and application of data science. Both domain-specific and data-specific analytics and theoretical data analytics have together formed the keystone of data science.

V. DATA ANALYTICS: A KEYSTONE OF DATA SCIENCE

In the age of analytics, what is to be analyzed, what constitutes the analytics spectrum for understanding data, and what form the paradigm shift of analytics takes are critical questions to be answered.

Analytics practices have seen a significant paradigm shift across three major stages [2]:

Stage 1: Descriptive analytics and business reporting: The major effort is on explicit analytics, which focuses on descriptive analytics and regular and ad hoc reporting. Limited effort is made on implicit analytics for hidden knowledge discovery, which is mainly achieved by using tools and built-in algorithms. Business reports generated by dashboards and automated processes are the means for carrying findings from analytics to management.

Stage 2: Predictive analytics: The major effort is on implicit analytics, which focuses on predictive modelling and business analytics.Business analytics refers to an in-depth understanding of business through deep analytics,and more effort being made to apply forecasting, data mining, and machine learning tools for business understanding and prediction. Patterns, scoring, and findings are presented to management through dashboards and analytical reports.

Stage 3: Prescriptive analytics and decision making, the major effort is on the delivery of recommended optimal actions for business decisions by discovering invisible knowledge and actionable insights from complex data, behaviour, and environment. This is achieved by developing innovative and effective customized algorithms and tools to deeply and genuinely understand domain-specific data and business.

VI. OVERVIEW OF DATA ANALYTICS LIFE CYCLE

The lifecycle draws from established methods within the realm of knowledge analytics and decision science. This criteria was developed after collecting input from data scientists and identified suitable approaches that provided solution to the problem. Here is a brief overview of the main phases of the Data Analytics Lifecycle [4][8]:

Phase 1: Discovery: In Phase 1, the team know about the business domain that consists of information like whether the organization or business unit has undergone similar projects within the past from which they can learn. The team uses the resources available to support the project in terms of individuals, technology, time, and data. Important activities during this phase include constitute the business problem as an analytics challenge which will be addressed in subsequent phases and formulating initial hypotheses to check and start learning the data.

Phase 2: Data preparation: In Phase 2 in the presence of an analytic sandbox, the team can work with data and perform analytics for the duration of the project. The team must execute extract, load, and transform (ELT) or extract, transform and cargo (ETL) to urge data into the sandbox. Data should be transformed in the ETLT process so the team can work with it and analyse it. In this phase, the team also must familiarize itself with the information thoroughly and take steps to condition the information.

Phase 3: Model planning: In this Phase 3 , the team determines the methods, techniques, and workflow that are used for the next model building phase. The team explores the information to find out about the relationships between variables and selects key variables and therefore chooses the best suitable models.

Phase 4: *Model building:* In Phase 4, mainly the team creates datasets for testing, training, and production purposes. In addition to this the team members builds and executes the proposed models that are discovered during planning phase, and identifies whether the existing tools are enough to run the models or it needs additional environment and more tools for executing models and workflows.

Phase 5: *Communicate results:* In this phase the team members along with major stakeholders, examines the results of the project and prepare reports that determines whether the project outcome is success or failure, if they follow the standards that are defined in phase 1.

Phase 6: *Operationalize:* In this phase the team releases final reports, briefings, code, and technical documents. The team prepares a user manual and deploy the project, in a production environment.

VII. METHODOLOGY

It is a skill for extracting knowledge from Data, Predict unknown from known, Find Pattern in data, Improve business outcomes using power of data.

Data Science is Multidisciplinary field that combine skills of software engineering and statistic with domain experience to support the end to end service.

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific methods, different technologies, and algorithms. It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful. Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence. Data science is all about , asking the correct questions and analyzing the raw data, Modeling the data using various complex and efficient algorithms, Visualizing the data to get a better perspective, Understanding the data to make better decisions and finding the final result.

Let us suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Fig.1 Data science.

1. *Purpose*: Data Scientists support critical business requirements by using systematic method in converting the available data in to useful information bringing the true science in to the business process.
2. *Scope*: Data Scientists perform fundamental things differently1. Mathematics, Statistics, Computational Modeling 2. Code to implement ideas through software 3.Communicate solutions, strategies and recommendations

Behavioral analysis, mitigating financial risks, targeting offers across channels, sentiment analysis, graph analysis, text analysis, revenue forecasting, Twitter analytics, Marketing mix.

VIII. CURRENT BUSINESS CHALLENGES AND INFORMATION NEEDS

A.Challenges

1. Huge amount of complex and variety of data being generated everyday from multiple sources making business requirements/decisions more composite and time consuming
2. This huge amount of data would be having enormous information, using these insights a lot of business strategies and solutions can be developed.
3. The changing economic/market/business environment has significantly pushed all the businesses towards competitive intelligence and future business plan and strategies.

B.Requirements

1. Need to collect all the required information which can be further processed, consolidated, and explored for mining the hidden insights in order to boost business performance and competitiveness.
2. Lot of insights can be hidden within large amount of multi sourced and multi structured data. We need both storage and processing platforms to better utilize data.
3. The complex data and changing business scenarios needs people, who does not simply collect and report data, but also looks it from many angles, discover hidden insights & recommends strategies.

C.Scope for improvements

1. Advanced storage platforms are available to handle complex and huge amount of data which are open source, low cost, scalable, custom built, real time & quick, automated, accurate and efficient.

2. Storage & Computational efficiency can be improved by adopting dynamic systems which provide the solution instantly based on real time & multiple data sources.

3. We can analyse data without any prior assumptions about the structure of business data. Generating hypothesis is better than testing hypothesis while dealing with recent challenge**.**

D.Data Science Components

The main components of Data Science are given below [3]:
1. Statistics: Statistics is one among the foremost important components of knowledge science. Statistics is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.

2. Domain Expertise: Domain expertise means specialized knowledge or skills of a specific area. In data science, there are various areas that we will have domain experts.

3. Data engineering: Data engineering may be a part of data science, which involves acquiring, storing, retrieving, and reworking the info. Data engineering also includes metadata (data about data).

4. Visualization: Data visualization is supposed by representing data during a visual context in order that people can easily understand the importance of knowledge. Data visualization makes it easy to access the huge amount of data in visuals.

5. Advanced computing: Advanced computing involves designing, writing, debugging, and maintaining the ASCII text file of computer programs.

6. Mathematics: Mathematics is the critical part of data science. Mathematics involves the study of quantity, structure, space, and changes. For a knowledge scientist, knowledge of excellent mathematics is important .

7. Machine learning: Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to cover the issues.

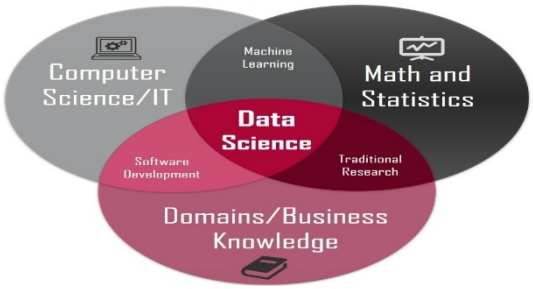

Fig.2 Data Science in Computer Science

Knowledge Discovery Process Flow through Data Science (Data Science Lifecycle)

The main phases of data science life cycle are given below [3]:
1. Discovery: The first phase is discovery, which involves asking the right questions. When you start any data science project, you would like to work out what are the essential requirements, priorities, and project budget. In this phase, we'd like to work out on all the needs of the project like the amount of individuals, technology, time, data, an end goal, then we will frame the business problem on first hypothesis level.

2. Data preparation: Data preparation is additionally referred to as Data Munging. In this phase the main tasks are Data cleaning, Data Reduction, Data integration, Data transformation, After performing all the above tasks, we can easily use this data for our further processes.

3. *Model Planning*: During this phase, we'd like to work out on the methods and techniques to determine the relation between input variables. We will apply Exploratory data analytics by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:SQL Analysis Services, R, SAS, Python

4. *Model-building***:** During this phase, the method of model building starts. We will create datasets for training and testing purpose. We will apply different techniques like association, classification, and clustering, to create the model. Some common Model tools are WEKA, SPCS, MATLAB, and SAS Enterprise Miner.

5. *Operationalize*: During this phase, we'll deliver the ultimate reports of the project, alongside briefings, code, and technical documents. This phase provides you a transparent overview of complete project performance and other components on a little scale before the complete deployment.

6. *Communicate results:* During this phase, we'll check whether we reached your goal are not, if we reach the goal, and communicate the findings and outcome with the business team.

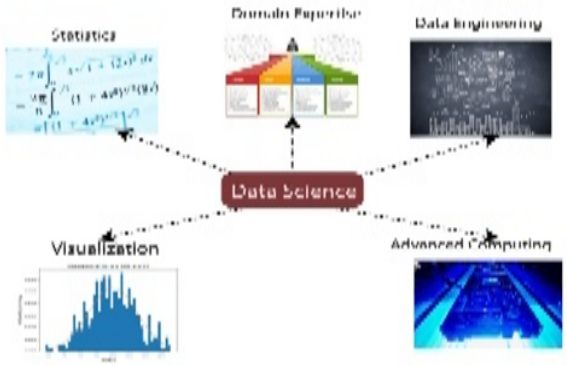

Fig.3 Data science Components

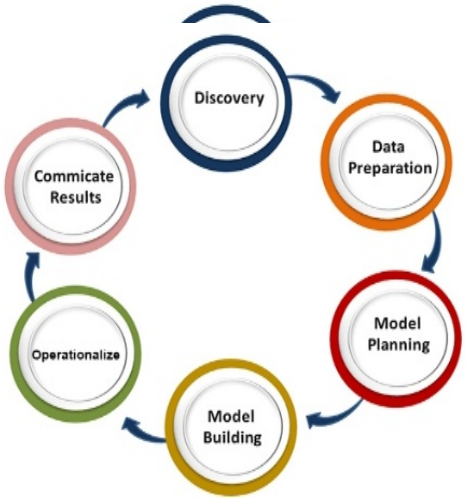

Fig.4 Data science life cycle

IX. APPLICATIONS OF DATA SCIENCE

1. Image recognition and speech recognition: Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. this type of automatic tagging suggestion feature uses image recognition algorithm, which is part of data science. When you say something using, "Ok Google", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.

2. Gaming world: In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.

3. Internet search: When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.

4. Transport: Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.

5. Healthcare: In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.

6. Recommendation systems: Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.

7. Risk detection: Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued. Now a day some of the finance companies are looking for the data scientist to avoid risk

and any type of losses and frauds with an increase in customer satisfaction.

A.Role/Position of Data Scientist

The role of data scientist starts from extraction of enormous data, perform in-depth analysis and derive valuable insights by using various statistical techniques & machine learning and ends at value creation.

B.Role of Data Scientist

1. Knowledge of disciplines like Statistics, Mathematics, Computer Science
2. Provide Strategic Business Recommendation
3. Capable of discovering and interpreting insights derived from large set of data
4. Derive value from the data
5. Analyse and develop predictive models.

C.Interconnection of Major Data Science Disciplines

Fundamentally, there are lot of similarities between conventional analytics and data science. The major factor which influences data science from other is creation of algorithms.
A. Predictive Analytics: Predicts the future probabilities and trends by using historical data. It provides clear, actionable initiatives that a business can implement
B. Big Data Analytics: Data Science is an art of bringing value and insights through discoveries in the world of big data. Data Scientists support critical business requirements by using systematic method in converting the available data in to useful information
C. Machine Learning: A part of Artificial Intelligence has been contributed to rapid development of data science.

X. BIG DATA OVERVIEW

Data is created constantly at an ever-increasing rate. Mobile phones, social media, imaging technologies to work out a diagnosis create new data, which must be stored somewhere for a few purposes. Devices and sensors automatically generate diagnostic information that must be stored and processed in real time. Keeping up huge amount of data is difficult, but substantially more challenging is analysing vast amounts of it, especially when it does not conform to traditional notations of data structure, to spot meaningful patterns and extract useful information.

Several industries have led the way in developing their ability to gather and exploit data [4]:
Credit card companies monitor every purchase their customers make and may identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
Mobile phone companies analyse subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
For companies like LinkedIn and Facebook, data itself is their primary product. The valuations of those companies are heavily derived from the information they gather and host, which contains more and more intrinsic value because the data grows.

Big Data in 2020:Big data became a big topic across nearly every area of IT.IDC defines big data technologies as a new generation of technologies and architectures ,designed to economically extracts value from very large volumes of a wide variety of data by enabling high velocity capture,discovery,and analysis[9].There are three characteristics of big data ,the data itself, the analytics of data, and the presentation of the result of the analytics.[9]

A.Big Data Characteristics [6]

Huge volume of data: instead of thousands or many rows, Big Data are often billions of rows and many columns.
Complexity of knowledge types and structures: Big Data reflects the variability of latest data sources, formats, and structures, including digital traces being left on the online and other digital repositories for subsequent analysis.
Speed of latest data creation and growth: Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.Although the quantity of massive Data tends to draw in the foremost attention, generally the variability and velocity of the info provide a more apt definition of massive Data. Big Data is described as having 3 Vs: volume, variety, and velocity. Due to its size or structure, Big Data can't be efficiently analysed using only traditional databases or methods. Big Data problems require new tools and technologies to store, manage, and realize the business benefit. Big data can be available in multiple forms, including structured and non-structured data like financial data, text files, multimedia files, and genetic mappings.

B.Data Structures

A. Structured data: Data containing an outlined data type, format, and structure that is , transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets[9].

B. Semi-structured data: Textual data files with a discernible pattern that enables parsing such as Extensible Mark-up Language [XML] ,data files that are self describing and defined by an XML schema.

C. Quasi-structured data: Textual data with erratic data formats which will be formatted with effort, tools, and time
Unstructured data: Data that has no inherent structure, which can include text documents, PDFs, images, and video.

Types of Data Repositories, from an Analyst Perspective

A. Spreadsheets and Data Marts: Spreadsheets and low volume databases for record keeping.

B. Data Ware House: Centralized data containers in a purpose-built space, Supports BI and reporting, but restricts robust analysis. Analysts must spend significant time to urge aggregated and disaggregated data extracts from multiple sources.

*C. Analytic Sandbox (workspaces)**:*** valuable data gathered from multiple sources and technologies should be strongly analysed during a non production environment.

XI. COMPARING BI WITH DATA SCIENCE

One way to evaluate the type of analysis being performed is to examine the time and the kind of analytical approaches being used. Business Intelligence (BI) tends to supply reports, dashboards, and queries on business questions for the present period or within the past. BIsystems make it easy to answer questions associated with quarter-to-date revenue, progress toward quarterly targets, and understand what proportion of a given product was sold during a prior quarter or year. These questions tend to be closed-ended and explain current or past behaviour, typically by aggregating historical data and grouping it in some way [4]. BI provides generally answers to the question related to “when” and “where” events occurred.

By comparison, Data Science uses disaggregated data during a more forward-looking, exploratory way, that specialize in analysing this and enabling informed decisions about the longer term. Rather than aggregating historical data to seem at what percentage of a given product sold within the previous quarter, a team may employ Data Science techniques like time series analysis, Time Series Analysis is to forecast future product sales and revenue more accurately than extending a simple trend line. In addition, Data Science tends to be more exploratory in nature and should use scenario optimization to affect more open-ended questions. This approach provides insight into current activity and foresight into future events, generally that specialize in questions associated with “how" and “why" events occur.

A.Big Data Platform – Apache Hadoop

Hadoop is an open-source framework to store and process huge amount of data

1. Data stored in distributed environment across clusters of computers. This is Designed to scale up from single servers to thousands of machines
2. Local computation and storage paradigm across distributed servers
3. Reliable-Data stored is replicated across different servers/nodes
4. Fault tolerant -Data is available in case of failure
5. Economical- commodity hardware's can be used to form cluster

Traditional Approach - An enterprise will have a computer to store and process huge data. Here data will be stored in an RDBMS.Google solved this problem using an algorithm called Map Reduce. This algorithm divides the main task into small pieces and assigns it to many computers that are connected over the network, and collects the results in order to form the final result dataset [10].

XII. DISCUSSIONS

Before the digital revolution came into adjust, the data at our disposal was mostly structured and relatively small in size. As a result, traditional BI tools were enough to research these small and structured datasets. However, the exponential growth of knowledge in recent years has changed the whole equation. How is it possible? Contrary to the normal datasets (that were mostly structured), the info generated today (from different sources like social media, financial transactions, and logs, multimedia files, online portals, etc.) is mostly semi-structured or unstructured. At present, more than 80% of the world’s data is unstructured. With each passing year, the info will only still increase and increase the already massive pile of knowledge . It is impossible for traditional BI tools to research such a huge volume of unstructured datasets – they demand more advanced and intelligent analytical tools for storing, processing, and analyzing data. This is where Data Science has helped make a difference. As more and more organizations are opening up to Big Data, AI, and ML, the demand for skilled Data Science professionals is ever increasing. In fact, the Harvard Business Review even hailed the job of a Data Scientist to be the Sexiest Job of the 21st century.For instance, when you connect your smartphone to smart devices and the IoT hub, you can monitor what is happening in and around your house even in your absence. Online shopping has gotten so much easier. Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define ,big data in terms of being larger than a certain number of terabytes[6].

Big Data is defined as massive amount of data which is too large and complex to be stored in traditional databases. Data has evolved over the last 5 years. Lots of data is being generated every day in every business. This data is getting used in every sector of business, like - Social Media, E-Commerce, Banking, etc. Below are some facts about Big Data for a few of the companies [11]

1. 40,000 search queries are performed on Google per second, i.e. 3.46 million searches a day
2. Every minute, users send 31.25 million messages and watch 2.77 million videos on Facebook

3. 55 billion messages and 4.5 billion photos are sent each day on WhatsApp
4. Walmart handles more than 1 million customer transactions every hour
5. By 2025, the volume of digital data will increase to 163 zettabytes

Now, the question arises, what do the businesses do with such huge volumes of data? Well, these companies collect, store and analyze this data to draw business insights.
As you see from the above statistics, it's quite evident that data will only keep increasing. All this data is of no use to us if it is not analyzed well. Big data by itself is meaningless, only when we analyze all of it, we can draw meaningful information from it and make use of it in real-time. Big data analytics features a lot of scope in various sectors. All companies have big data, and the way they analyze it to increase their revenue is known as big data analytics. Sectors like healthcare, weather outlook, government and enforcement use big data applications.

Today, data is becoming the new cash for companies. The IT industry is dynamic. IT professionals should be flexible to change according to the latest trends in the market [10].
As massive data acquisition and storage becomes increasingly affordable, a wide variety of enterprises are employing statisticians to engage in sophisticated data analysis, like emerging practice of Magnetic, Agile, Deep (MAD) data analysis as a radical departure from traditional Enterprise Data Warehouses and Business Intelligence[7].

Data analytics is a process through which data is cleaned, analyzed and shaped using tools. This data is then used to derive insights. The insights are then used for business-related decision-making purposes. As data analytics also allows enhancing business process and maximizing conversion rates, it helps the organizations in cutting unnecessary costs and reduces the value of running the corporate. As advancements within the field of knowledge analytics are being made, the method is getting automated. Machines are interpreting big chunks of knowledge in an automatic process. With the help of machines, data analysts are finding it possible to make sense of the data more quickly and easily. The newer technologies like Block chain, Internet of Things, machine learning, AI , etc. have been the foremost popular glossaries among business corridors. The most interesting thing about all the modern technology is that they are all based on data. Because of the bright future of data analytics, many professionals and students are interested in making a career in data analytics.

Data analytics is the differentiator that provides companies with a competitive edge over others. It is a fast-growing branch of study which features a bright future in India. Organizations have realized its importance and investing in data analytics tools and technologies. We can make certain that data analytics features a good future in India for years to return .The future of Big Data is bigger than anyone can even imagine; Because of advances in technology and computing, we're generating more data than ever before. Let's look at some interesting facts about data[11],

1. Less than 0.5% of all data we create is ever analysed and used.
1. 2.73% of organizations have already invested or will invest in big data by the end of 2016.
2. A 10% increase in data accessibility will end in quite $65 million additional net for the standard Fortune 1000 company.
3. Google uses about 1,000 computers to answering search query.
4. By 2020, there will be more than 50 billion smart connected devices in the world, for collecting, analysing and sharing data.
5. Last year, an estimated 1 trillion photos were taken and billions of them will be shared online.
6. By 2017, nearly 80% of photos are going to be taken on smart phones and most will become searchable data.
7. We perform 40,000 search queries every second on Google alone, becomes 1.2 trillion searches per year.
8. By 2020, about 1.7 megabytes of latest information are going to be created every second for each human on the earth .10.1 billion àpieces of content are shared via Facebook's Open Graph every day.
9. 10.70% of data is created by individuals, but enterprises are responsible for storing and managing 80% of that.

XIII. THE FUTURE OF DATA SCIENCE

With the joint efforts to be made by the entire scientific community, data science will build its systematic, scientific foundations, disciplinary structure, theoretical systems, technological families, and engineering tool sets as an independent science.

The last 50 years since the proposal of the concept "data science" has contributed to the progressive and now widespread acceptance of the need for a new science. The next 50 years of data science will extend beyond statistics to identify, discover, explore, and define specific foundational scientific problems and grand challenges. It will build a systematic family of scientific methodologies and methods and self-contained disciplinary systems. Based on the understanding of the challenges and nature of data science the development of data science may seek to [3]:

Design and develop data brain that can autonomously mimic human brain working mechanisms to recognize, understand, analyze, and learn data and environment infer and reason about knowledge and insight, and correspondingly decides actions.

Invent new data representation capabilities, including designs, structures, schemas and algorithms to form invisible data complexities and unknown characteristics in

complex data that are more visible, explicit, and more easily understood or explored.

Design to support scalable, transparent, flexible, interpretable, and personalized data manipulation and analytics in real time. And Design to storage, access, and management mechanisms, including memory, disk, and cloud based mechanisms, to enable the acquisition, storage, access, sampling, and management of richer characteristics and properties within the physical world that have been simplified and filtered by existing systems.

Create new analytical and learning capabilities, for mathematical, statistical, and analytical theories, algorithms, and models, to disclose the unknown knowledge.

Build new intelligent systems and services, including corporate and Internet based collaborative platforms and services, to support the automated or human data cooperative, collaborative, and collective exploration of invisible and unknown challenges in unknown space.

Train the next generation data scientists and data professionals who are qualified for data science problem solving, with data literacy thinking, competency, consciousness, curiosity, communication, and cognitive intelligence, to work on the preceding data science agenda.

Discover and invent data power asyetun known to current understanding and imagination, such as new data economy, mobile applications, social application and data-driven business.

XIV. CONCLUSION

Data science, big data, and advanced analytics are recognized as major eminent technologies for next-generation innovation, economy, and education. Although they are at an early stage of development, strategic discussions about the trends, major challenges, future directions, and prospects are critical for the healthy development of the field. More efforts are being made by government, industry, academia, and even private institutions on different ways to convert data for producing good decisions, and promote the research and development of knowledge science and analytics. The next generation of data science, encompassing a broad range of disciplines, science, and economy, relies heavily on the strategic planning and visionary actions that will be undertaken in prioritized data, research areas and start-ups.Without any doubt, today's questions such as "why do we need data science" will be replaced by a family of scientific theories and tools to address the visible grand challenges and significant problems facing tomorrow's big data, science, business, society, and the economy. The purpose of this paper is mainly to share an overview of the conceptualization, development, observations, and thinking about the age of data science initiatives, and this paper provides some of the key idea about characteristics, data structures used in big data and life cycle of Data analytics. Knowledge of this information will help people become active contributors to Big Data analytics projects.

XV. FUTURE SCOPE

Lots of data is being generated each day in every business sector which has given birth to Data Science & its various branches such as Machine Learning, Deep Learning, Artificial Intelligence & many more. With the help of these technologies, meaningful insights are derived from the heaps of data. So, it is not wrong to say that this makes a great sense for businesses to stay ahead of the competition. Data Analytics no matter how advanced they are, does not remove the need for human insights. On the contrary, there

is a compelling need for skilled people with the ability to understand data, think from the business point of view and come up with insights. For this reason technology professionals with Analytics skill are finding themselves in high demand as businesses look to harness the power of Big Data. A professional with the Analytical skills can master the ocean of Big Data and become a vital asset to an organization, boosting the business and their career. The job title of a 'Data Scientist' will undergo a massive transformation to include an array of diverse roles. As technology, Data Science, and AI continue to advance, Data Scientists will have to evolve to keep pace with the dynamic learning curve of Data Science. There are handful of possibilities that Data Science will bring into our world in the next few years

REFERENCES

[1] D. Roger Pengand ElizabethMatsu, *The Art of Data Science, A Guide for Anyone Who Works with Data*, Lean publishing book, 2015 - 2016 Sky rude Consulting, LLC.

[2] LONGBING, University of Technology Sydney, Australia, Data Science: A Comprehensive Overview , *ACM Computing Surveys*, Vol. 50, No. 3, Article 43, Publication date: June 2017.

[3] JavaTPoint, Data Science Tutorial for beginners, javapoint.com/data science.

[4] EMC Academic Alliance University , Data science and big data analytics ,Discovering, Analyzing, visualizing and presenting data, *EMC education services(EMC2)*

[5] T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21stCentury," *Harvard Business Review*, October 2012.

[6] J. Manyika, M. Chiu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H.Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity",McKinsey Global Institute, 2011.

[7] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, "MAD Skills: New Analysis Practices for Big Data", Watertown, and MA 2009.

[8] S. Todd, "Data Science and Big Data Curriculum" [Online].Available:http://stevetodd.typepad.com/my_weblog/data-science-and-big-datacurriculum/.

[9] D. R. John Gantz, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," *IDC*, 2013.

[10] Blog, industries' using big data, solutions of big data

[11] Quora, Future Scope of Data Science.