# Big Data Analytics and Data Science Emerging Trends and Challenges

Bikash Chandra Pattanaik [1], Snigdha Mohapatra [2], Sibani Samal [3], GYANENDRA KUMAR PALLAI[4]
[1, 2, 3] Gandhi Institute for Education & Technology, Baniatangi, Khordha, Odisha
[4]NM Institute of Engineering & Technology,Bhubaneswar,Odisha
bpattnaik@giet.edu.in, snigdhamohapatra@giet.edu.in, sibanisamal@giet.edu.in

***Abstract.*** ***Due to recent advancements in several technologies, a number of industries have seen growth in the last ten years. These innovations have transformed human existence and are boosting earnings for both individuals and businesses like Netflix, Alibaba, Flipkart, etc. Most individuals in this society today either use or are surrounded by smart things to make their lives more simple and easy. On the other hand, analytics businesses and sectors are making life easier. For instance, numerous organisations extract customer preferences for certain plays, movies, music, and other items. These suggestions are being made, and the corresponding users are receiving better services. Data Scientists carry out such tasks. As opposed to this, data science (the future of AI) is a multidisciplinary subject that combines scientific techniques, procedures, algorithms, and systems to extract information and insights from both structured (labelled) and unstructured (unlabeled) data. Additionally, Big Data Analytics is the analysis tool utilised by data scientists in data science. Many tools, like Hadoop and others, are used to analyse the vast amounts of data and forecast relevant information/make judgments. However, when we analyse, we run into issues with complexity, scalability, privacy leakage, and trust. Therefore, this article thoroughly addresses the issues and problems (arising) in this young topic (with a comparative analysis and classification).***

***Keywords- Data Science, Big Data Analytics, Challenges, Opportunities, Future Technology.***

## I. INTRODUCTION

The growing accessibility of data, storage capacity, and computational power have changed organizations across the global economy. It involves not only the popular businesses such as Google or Facebook that were "made online," but also early adopters in more conventional industries such as banking, retail and transport. [1].Around 2025, the expectation is that the Internet will exceed the scale of the brain of all those living around the planet. This firm data growth is due to developments in digital sensors, computations, communications, and processing that produced large data collections. Roger Magoulas, a scientist, coined the term Big Data (in 2005) to reflect this singularity [12]. In contrast, Big Data is a process of working like collecting or querying on ample of data. Big Data is always seen for the business process in terms of size, speed, variety, veracity and quality. Mostly it is subject to large processing systems, but most of its dimension is transferred to cloud systems where it completely satisfies the financial, technological, functional and organizational viability of big data organizations. There are several others used in today's (smart) era like data science, data analytics, big data analytics, etc. Data science continues to evolve as one of qualified professionals' most exciting and challenging career paths. Data science continues to evolve as one of qualified professionals ' most exciting and in-demand career paths. Today, productive information professionals recognize that the conventional skills of processing large amounts of data, data mining, and programming skills must be improved. To identify useful intelligence for their organizations, data scientists need to experience the full spectrum of the life cycle of data science and have a level of flexibility and awareness to optimize returns at each stage of the process. Data science is used in simple words to uncover hidden trends and knowledge from vast (unstructured) amounts of data. In other words, data science is about discovering insights from unstructured data (collected from multiple sources). For example:

- Netflix or YouTube use data mining for mining instances like pause, play or repeat about a movie/ song viewing patterns, i.e., to find user interest, and uses that to make decisions on which Netflix original/ YouTube series to produce in near future (for their viewers/ audience).
- Another example, the retail target companies describe the major customer segments in which they are focused and the particular shopping patterns in those segments that help guide marketing to different market markets (increasing profit and productivity).
- Proctor and Gamble uses time series models to perceive future demand more clearly, which helps to prepare more optimally for production levels.

Firstly, it is very necessary to understand what Data Science is and how can it add value to your business. For example, all the concepts which we see in Hollywood sci-fi movies can actually turn into reality with the help of Data Science. In general, structured data can be solved or analyzed by simple business tools but unstructured data always require

new efficient analytics methods to analysis this large amount of data. Note that in 2020, 80% of the data will be in unstructured form where several sources such as financial records, text files, multimedia formats, sensors and instruments will be collected. Further, data science helps in:

- Understand the customers' specific criteria from existing data such as past browsing history of the consumer, history of purchase, age and income.
- For predictive analytics, information analysis can be used. It is possible to collect and analyze data from vessels, aircraft, radars, satellites to create models. Not only will these models predict the weather, they will also help in predicting the likelihood of any natural calamities. It will allow us to take appropriate action in advance and save some important lives. Note that forecast and prediction are not synonym of each other, both are completely different.
- Using organisation profit via analysing user's interest, habits, etc. For example, Netflix, YouTube are analysing their viewers and providing required content to their users.

Here a big question arises "How do data scientists mine out insights"? It begins with the exploration of information. Data scientists are detectives and seek hidden trends in data to answer these problems / challenging questions. In [9], authors attempt to understand trends or features within the data. It requires a large scale of dose of innovation in analytics.


**Figure 1: Life Cycle of Data Science [9]**

Note that here tools for model planning are: R, SQL analysis services, SAS/ACESS. In summary, in social media application we can used data science for digital marketing, sentimental analysis, in marketing we use cross selling, up selling, predicting life value for customer, in travel/ tourist we use dynamic pricing, predicting fight delay, in e-healthcare we use for disease prediction, medication effectiveness, in sales we use for discount offering, demand forecasting, in automation industries we use in self driving cars, pilotless aircrafts, drones, whereas in credit and insurance we use for claims prediction, fraud and risk detection. In summary, Data science require several roles to make it successful which are data engineer, data analytics and data scientist. Data Manipulation, Data Analysis with Statistics and Machine Learning, Data Communication with Information Visualization are some popular components of data science.

Data Analytics: Data Analytics is done on raw data and made conclusions of the collected information. It is all about discovering useful information from the data to support decision-making. This process involves inspecting, cleansing, transforming and modelling data. Now, uses of Data Analytics can be discussed in several applications:

a) Social Media: It was difficult to process activity across different social media platforms before cloud drives became realistic. Cloud drives allow social media site information to be analyzed concurrently so that findings can be filtered easily.
b) Tracking Products: Long thought of as one of the kings of reliability and forethought, it's no wonder that Amazon.com uses cloud drive data analytics to monitor goods through their series warehouses and deliver items anywhere they can, irrespective of consumer proximity items. Amazon's use of cloud drives and remote storage, due to their Redshift program. Redshift offers many of the same analytical resources and processing features as Amazon for smaller organizations and serves as a data warehouse, keeping smaller companies from spending money on expensive hardware.
c) Tracking Preference: Netflix has received a lot of attention over the past decade buying DVD to select the movies on the website. One of their website's highlights is their film recommendations, which monitors users ' watching movies and suggests others they may enjoy, providing consumers with a service and promoting their product's use. The user information on cloud drives is stored remotely so that the habits of users do not switch from computer to computer. Since Netflix kept all the interests and desires of their customers in movies and TV, they were able to create a TV show that appealed objectively to a large portion of their viewers based on their established taste. The House of Cards of Netflix thus became the most profitable internet-television series ever in 2013, thanks to their data analysis and cloud-based knowledge.
d) Keeping Records: Cloud analytics enable information to be stored and processed simultaneously regardless of local database proximity. Companies throughout the United States can monitor the sales of an item from all of their divisions or franchises and modify their production and deliveries as required. If an item doesn't sell well, they don't have to wait for stock reports from the area retailers, but they can handle inventories from automatically downloaded data to cloud drives remotely. The data stored in clouds allows companies to run more efficiently and gives businesses a better understanding of the actions of their customers.

**Big Data:** The idea of Big Data means a dataset that continues to grow till it becomes difficult to manage using current frameworks and resources of database management. Data collection, processing, search, sharing, analytics, and visualization can be related to the complexity. The Big Data covers three dimensions: size, speed and variety. Some of Big Data's advantages are (currently): hospitals in the U.S., government services in Europe, retail in the U.S., and data

on manufacturing and personal location internationally. And their study confirms that in each of the mentioned domain areas, Big Data can generate value. A company using Big Data, for instance, could raise its operating margin by more than 60% [7].If US healthcare is planning to use Big Data creatively and strategically to improve efficiency, then each year the industry will generate more than $300 billion worth. Two-thirds of that would be in the form of an approximately 8 percent reduction in US healthcare spending [7].

**Data Science:** The state-of - the-art methodology of scalable machine learning from a data science perspective occurs in three stages: model building staging, model testing, and system delivery and evaluation to future data. The staging phase of data is a disc-intensive process whereas usually the stages of system delivery and software learning are computational and processes of memory intensive. Currently, most of the advancements in scalable machine learning (e.g. Madlib[4], Apache Mahout[5], etc.) are happening in the massively parallel environment of data base storage. While this is advancement, it is not possible to implement all algorithms using the database set-theoretic algebra. Linear algebra-based algorithms requiring an inversion of the matrix, decomposition of the original value, optimization of the iterative objective function, etc. are very difficult to implement in databases on a scale. While on distributed storage systems some of the machine learning algorithms can be applied, most of the algorithms are best implemented as in-memory operations. Clearly, hardware and software choice optimize these algorithm groups. Just leading metrics will make these choices. In one of our graph computing investigations, we found that the graph theoretical algorithms performed magnitude orders more rapidly in shared memory architectures than in shared storage systems [6]. Sadly, in various scalable computer architectures, we do not have such tests for machine learning algorithms to make data science decisions for learning purposes. The accessibility and affordability of the market today drive the selection of hardware configurations to deploy machine learning solutions.

Given the need for efficiency, we are unable to predict a data analysis algorithm's output as we have been successful with traditional High-Performance Computing (HPC) codes. For the different classes of machine learning algorithms, we do not understand the trade-offs of using usable resources to optimize hardware. This lack of predictability can be attributed to the fact that data analysis algorithms output is a feature of design, software and algorithmic workflow. Reasoning this problem from the perspective of the law of Amdahl, we cannot guarantee predicted linear / sub-linear scalability with HPC on architectures of Big Data.HPC problems are forward issues that are better understood in terms of parallelism and scalability, whereas Big Data issues are reverse usability issues. Often studied and less appreciated is the scalability of data-intensive computing algorithms.

Note that when we want to measure Data accuracy at Scale, we need to work with Big Data and need to determine results with different features. As discussed in [10], accuracy is always in paradox in case of mining information or prediction. Hence, the further remaining part of this paper is organized as: Section 2 will discuss tries attempts related to data science and data analytics in the past decade. Section 3 will discuss motivation section, i.e., reason working towards data science and data analytics. Section 4 discusses importance of Data Science and Big Data Analytics in today's (smart) era. Further, Section 5 discusses essential part of this work, i.e., a taxonomy which has comparison with several terms like Data Science, data engineer and Big Data analytics, etc. Further, several identified towards data science and Big Data Analytics are discussed in section 6. Section 7 discusses several challenges faced towards data science in data analytics in current era. Section 8 discusses several opportunities for Data Science and Big Data analytics in near future. In last, section 9 concludes this work in brief with some future scope. Note that the term 'smart objects' will be used as smart devices or internet of things or internet connected things interchangeably (throughout this work).

## II. RELATED WORK

Earlier, in the days of time and motion studies in the 19$^{th}$century, analytics have been used in industry. "Analytics" is not a new term for contextual discourse when we look back at the time of world war in the 1940s where more productive methods and ideas were required to optimize production with a smaller number of resources. Earlier it was named quantitative methods used in business in the late nineteenth century by Frederick Winslow Taylor [4]. After that, as computers were used in Decision Support Systems (DSS) in the late 1960s, it gets more coverage. Throughout the 1970s, very few information were collected using manuals from the domain experts. The impetus behind this wave of analytics was to achieve the results at the best with limited resources, and these types of decision support are called Operations Research and Management Science. In the 1980s, the way companies collected their business-related data was expressed in a slight change. Such systems help improve the collection, processing and relationship between organizational data fields while significantly reducing information duplication. All data can be easily accessed from anywhere and at any time through Enterprise Resource Planning (ERP). It collects the information from every corner of the world and transforms it into a schema.

Instead, in the 1990s, there was a need for a more flexible reporting system which contributed to executive information systems. Such systems were designed as scorecards and graphic dashboards.

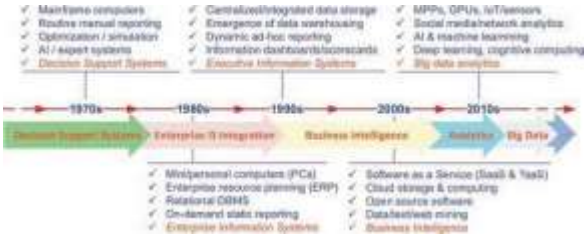

**Figure 2: Evolution of Analytics and Its Terminologies [8]**

Earlier in the 2000s, the Business Intelligence (BI) systems were named for these DW-driven decision support systems. As a requirement for a globalized competitive marketplace,

decision-makers needed the most up-to-date information in a very digestible way to address business issues and take advantage of market opportunities in a timely manner. Due to the large and feature-rich data collected in DW, emerging technological trends such as data mining and text mining have become common to "mine" corporate data to "discover" new and useful knowledge nuggets to enhance business processes and practices. The word "Big Data" was coined to reflect these issues that were brought to us by the new data sources. Several developments have been made in both software / algorithms to overcome the "Big Data" challenges. A new buzz word "deep learning" arrived with big data and improved computing capabilities. Generally speaking, deep learning will make artificial neural networks (ANN) look in a better way.

But now things are changes and to predict earlier is very hard, "What the next decade will bring to us, what are the new terms that will be used to call analytics." Today, the fact is that the rise in its popularity is very recent, but analytics is not fresh. In this competitive market, it's a good opportunity to increase the revenue and reduced the cost by building better goods, detecting fraud before it occurs, and increasing customer engagement by targeting and customizing with the help of data analytics i.e., all with analytics and data energy. Note that figure 2 shows evolution of data analytics, year by year, techniques by techniques.

Hence in this section, we have discussed several tried attempts by several data engineer or scientist in the previous decade with respect to data analytics and evolution of analytics terminologies (since 1950s to 2019). Now, next section will discuss about motivation behind this work or facts or reason which make our interest to work in this area.

## III. MOTIVATION

In the past decade, data mining was popular in mining (refining) information. But as we (people) moves in 21st century, they started uses many smart devices to do their daily routine task. Also, industries used such smart devices in automation, manufacturing for increasing profit of their product. But apart from profit or other task, these devices are highly useful in some essential applications like e-healthcare, defense, etc. These devices in healthcare system generated huge data, which require to be analyzed by skilled and trained people and result need to be provided to doctor for best possible cure/ solution for a disease. Saving human lives and providing convenient and better life to their citizen is always a goal of every nation. For that, several terms like data science is come into picture, analyzing large data using machine learning techniques (as a popular techniques). In general, Data Science is more about Predictive Causal Analytics and Machine Learning. Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning. Data science solve this problem efficiently, i.e., refine or extract uncovered/ hidden pattern/ information from this large data. Hence, due to high necessity of data science in each possible application make us to write an article on respective topic/ area (also its components like data engineer, data scientist, etc., and relation with data analytics).Hence, this section discusses about motivation behind this work with some real word' example. Analytic is being done by data scientist and data science group member. Uncovered patterns are determined by using efficient tools and algorithms (on this large collection of unstructured data).Note that the terms "analyst" and "data scientist" is not exactly synonymous, but also not mutually exclusive. Now next will discuss about importance of data science and big data analytics in today' sera and in next decade (with several examples).

## IV. IMPORTANCE OF DATA SCIENCE AND BIG DATA ANALYTICS

By using analytics and deep learning to make better decisions and boost recruiting, Data Science effectively adds value to all business models. It is also used to crunch the previous data and to forecast possible situations and threats in order to avoid them. In fact, evaluating this information can really help to set up a workflow. Some applications of Data Science are defined as follows:

i. Internet search: Search engines works so fast on queries that they will provide results in fraction of seconds and it will happen with the help of data science
ii. Digital Advertisements: Data science techniques are used throughout the digital marketing spectrum-from display banners to electronic billboards. That's the mean reason why digital ads get higher CTR than conventional ads.
iii. Recommender systems: It not only makes easy to get the relevant data in millions of products, but it also helps in adding a deal to the user experience. Several businesses are using this method to promote their products and feedback in accordance with the user's requests and data relevance. The suggestions are based on the user search history.

**Big Data:** It plays a vital role in this world because every device we used, stores some data in larger form termed as "Big Data" [9]. Big Data is data with 3 V's (Volume, Velocity and Variety) in it. Some more V's have been discussed in [9]. Big Data have its own consequences like we can use this data to predict the future (i.e., in sales department) or forecasting, or for better decision making and many more. Some applications of Big Data are:

a) Big Data for financial services: Big data is used for various financial services by banks, insurance companies, venture funds, and institutional investment banks. Massive amounts of multi-structured data residing in various fragmented structures that can be solved by big data are the common problem among all of them. Thus, big data can be used in various ways like:
   - Monitoring for consumers
   - Compliance analytics
   - Analysis for fraud
   - Operational analytics

b) Big Data in Communications: For telecommunications service providers, attracting new users, maintaining customers and expanding within the current user bases are top priorities. The

answers to these problems are the ability to combine and evaluate the volumes of customer-generated data and machine-generated data produced on a daily basis.

c) Big Data for Retail: If we know that user better and satisfy his needs according to his weblogs, transactions and social-media is the biggest game to stay for online traders.

**Big Data Analytics:**It is a process for examining the variety of data which further going to help the organizations in taking better decision. Scope of Big Data Analytics (BDA)in near future exists in:

a) Reduction of costs. Big data technologies such as Hadoop and cloud-based analytics offer substantial cost savings when it comes to processing large amounts of data.

b) Really effectively, better decision-making. By combining Hadoop's speed and in-memory analysis gives ability to evaluate new data sources, companies can quickly analyze knowledge – and make decisions based on what they have found.

c) Different goods and services. Through analytics, we can measure the needs of customer. Davenport points out that more businesses are creating new goods using big data analytics to meet the needs of consumers.

Hence, several tools (e.g., Hadoop, R, Weka, KNIME, etc.), and algorithms (e.g., Clustering, Naïve Bayes, Decision Tree, Association Rule Mining, etc.) used in producing useful decision or prediction for organization / industries. Hence, this section discusses importance of Data Science, Big data and Big Data Analytics in smart era (in detail). Now, next section will discuss about several emerging trends in data science and data analytics.

## V. EMERGING TRENDS IN DATA SCIENCE AND BIG DATA ANALYTICS (WITH AN USE CASE)

Two years ago, Gartner predicted that by 2020, Augmented Analytics (AA) will be the “dominant driver of new purchases of Business Intelligence (BI), analytics and data science and machine learning platforms and of embedded analytics” [4]. Differences among several terms like business intelligence, data science, data scientist, etc. are being discussed in section 6 (in table 1, 2 and 3, in detail). Data science is a concept that unifies statistics, data analysis and their related methods in order to understand and analyze phenomena with data. It has several sub-domains. Data Science is mainly used to render predictive analysis, prescriptive analytics, machine learning decisions and predictions. Data science uses many algorithms (in its process) such as classification, cluster analysis, data mining, machine learning and visualization. Data science is helpful in:

- Predictive causal analytics (why will happen? why will it happen?), Prescriptive analytics (what should I do? what should I do it?).
- Machine learning for making predictions and for pattern discovery

Examples of apps that integrate data material behind the scenes: the website of Amazon, the inbox of Gmail, and autonomous driving technology.

- Recommendation engines from Amazon recommend products that we can purchase, based on their algorithms.
- Gmail's spam filter is a software item, i.e. the incoming mail runs an algorithm behind the scenes and decides whether or not a message is garbage.
- Computer vision used for self-driving cars is also information item, i.e. traffic lights, other cars on the road, pedestrians, etc. can be identified by machine learning algorithms.

**Use Case: Diabetes Prevention**

What will happen if we predict about the diabetes which is going to be happened in coming years and start taking proper medication beforehand to prevent it? Moreover this, some more use cases are also present today (in several real worlds’ problem). Hence, this section discusses several emerging Trends in Data Science and Big Data analytics in detail. Now, next section will discuss comparison among terms like data science, data engineer, data analytics, business analytics, etc., in detail.

## VI. A TAXONOMY - STATE OF ART COMPARISON FOR DATA SCIENCE AND BIG DATA ANALYTICS

Over the past several years, analytics has risen rapidly in popular business lingo; the term is used broadly, but usually used to reflect quantitatively critical thinking. Technically, the "science of observation" is analytics, i.e., placed another way, the art of interpreting data for decision making. Here some differences are discussed like:

- Data Scientist: Special work in math, engineering, and business acumen capabilities. At the level of the raw dataset, data scientists are working to extract knowledge and construct data material.
- Analyst: This can mean a lot of things. Common thread is that analysts are looking at data in order to gain insight. Analysts can communicate with information at either the level of the server or the level of the summary report.

**Table 1: Big Data Vs Data Engineer Vs Data Scientist Vs Data Analytics**

| Feature | Big Data | Data Engineer | Data Scientist | Data Analytics |
|---|---|---|---|---|
| Predictions | Predict business failure | - | - | Predict future opportunities |
| Features | Cost reduction, Time saving, Better | Embrace Modern data structure | Creativity | Discover patterns |

| | decision making | | | |
|---|---|---|---|---|
| Techniques | Data visualization | Data architecture and data warehousing | Clustering and classification analysis | Uses data mining and statistical techniques |
| Data | Working with unstructured data | | Data Mining Activities | Uses historical data |
| Technologies | Technologies like Hadoop, Spark, Hive etc | Hadoop based technologies (Map reduce) | SAS/R Coding | Reporting-with data visualization software |
| Skills | Business skills | - | Statistical and Analytical Skills | Scripting & Statistical skills |

Table 2: Data Scientist Vs Big Data Professional Vs Data Analyst

| Features | Data Scientist | Big Data Professional | Data Analyst |
|---|---|---|---|
| Technologies | Statistical and Analytical Skills | Technologies like Hadoop, Spark, Hive etc | Data Warehousing |
| Data | Data Mining Activities | Working with unstructured data | Hadoop Based Analytics |
| Skills | Deep Learning principles | Familiarity with MATLAB | Scripting and Statistical skills |
| Knowledge | In depth knowledge of programming | Creativity | Reporting-with data visualization software |

Table 3: Business Intelligence Vs Data Science

| Features | Business Intelligence (BI) | Data Science |
|---|---|---|
| Data Sources | Structured (Usually SQL, often Data Warehouse) | Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text) |
| Approach | Statistics and Visualization | Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP) |
| Focus | Past and Present | Present and Future |
| Tools | Pentaho, Microsoft BI, QlikView, R | RapidMiner, BigML, Weka, R |

If we require difference between BI (Business Intelligence) and Data Science, then we can say that for small or structured data business analytics or intelligence is an optimal solution but when data is complex and unstructured, in that case, we require modern and efficient data analytics tools. Business Analytics (BA) is a new term (arising now days), completely different form data analytics, former one work entirely towards generating decision for business purpose Through efficient analytics tools we determine useful decisions and predictions (made in Data Science).

Hence, this section discusses differences several common terms used in analytics process like data engineer, data scientist, business intelligence, and business analytics, etc., in detail. Now, next section will discuss several challenges faced (identified) in data science and big analytics process.

## VII. CHALLENGES FACED IN DATA SCIENCE AND BIG DATA ANALYTICS

Big Data's main challenges are data size, volume, sophistication of computational workload and agility. Most companies struggle to deal with that data volumes. The organizations need to focus on reducing the amount of information being saved to solve this problem and leverage new storage technologies that can further boost efficiency and resource usage. The five challenge groupings are described here as:

- Insights not used in decision-making: these obstacles include corporate policy, failure to incorporate study results into decision-making processes, and lack of support for management.
- Information security, veracity, and unavailability: these problems related to the information itself, including how "dirty" it is, its availability, and privacy issues.
- Limitations of software to scale/ deploy: Problems in this section apply to methods used to gain information, deploy templates, and scaling solutions to the full dataset.
- Lack of Funds: The issues around lack of funding have an impact on what the company can purchase from external data sources, data science resources and, probably, domain expertise.
- Wrong Questions Asked: Problems are about the challenge of managing assumptions about the effects of data science initiatives and not having a clear answer question or a clear direction to go with the information available.

Although the benefits as well as the supporting reasons for analytics are clear, there are still hesitant / challenges for many companies. These are described as follows:

- Analytics Talent: A talent in analytics means people who can turned out the raw information to some meaningful information as data scientists are rare in the industry and the really good ones are very difficult to find. Several universities have initiated masters and undergraduate programs to tackle the talent gap in analytics. When analytics' popularity grows, so will the need for people with the knowledge and skills to turn "Big Data" into information and knowledge that managers and other decision-makers need to solve real-world complexities.
- Culture: As the saying goes “old habits die hard”. Changing from a traditional to a contemporary style of management is not so easy task for any organization. People don't like changing. Change means to forget what we've learned earlier in the past and learning how to do what we're doing all over again now. This means that the wisdom we have accumulated over the years (which is also defined as energy, i.e. knowledge is power) will vanish or be lost in part.
- Return on investment: The difficulty in adequately explaining the Return on Investment (ROI) is another consideration behind analytics adoptions. Since analytics projects are complex and expensive undertakings and their return is not directly and instantly linked, most executives have trouble investing in analytics, particularly on a large scale. Someone has to answer this “Will, and if so when, the value gained from analytics outweigh the investment?” It is next to impossible to convert the value of analytics into justifiable numbers. Most of the empirical quality obtained is somewhat abstract and subjective. Analytics can turn an enterprise into new and improved rates if done properly. It is necessary to bring a mixture of tangible and intangible considerations to bear in order to rationalize numerically expenditure and step

towards analytical and analytically informed management practice.

- Technology: Despite being able and adoption of technology poses other challenges for less technical enterprises. Even though it's inexpensive, setting up an analytics system often costs a significant amount of money. Despite financial means and/or a strong ROI, they may not be willing to invest in the software they need. For those, it may be less costly and easier to implement analytics as a service model (which would include both technology and infrastructure/ hardware needed to implement analytics).
- Security and Privacy: Safety is one of the major common criticisms of data and analytics. As we often hear from the media about the data bridges for sensitive information, there is no fully secured data infrastructure unless it is segregated and disconnected from all other networks (which would be contradictory to the very reason that data and analytics are available).Although the strategies to secure the data infrastructure are growing in complexity, so are the approaches and techniques used by opponents. In addition to protection, personal privacy issues also exist. However, if it is within the legal boundaries, the use of private customer data (existing or prospective) should be prevented or closely scrutinized to prevent wrong publicity and public outcry from the company.

Despite of hurdles at every step, the adoption of analytics is still inevitable for all the enterprises. The ones who succeed in doing so will be the ones fully leveraging the capabilities of analytics. Hence, this section discusses several challenges faced in data science and big data analytics process in 21st century. Now, next section will discuss many opportunities for data science and big analytics in different applications (in predicting or proving optimal decisions).

## VIII. OPPORTUNITIES FOR DATA SCIENCE AND BIG DATA ANALYTICS

As discussed in data science life cycle (discussed as figure 1), we use five stages of the data science life cycle: Capture (information collection, data recording, signal reception, data extraction); storage (data management, data cleaning, data staging, data processing, data architecture); system (data mining, clustering / classification, data modelling, data summary);Research (exploratory / confirmatory research, statistical analysis, regression, text mining, qualitative analysis); interaction (information reporting, visualization of results, business intelligence, decision making).

Towards data science area, we can provide

- Collecting right data, right time or for right application.
- Implementing analytics process with efficient tools or modern tools which can easily handle this large amount of data.
- Providing unique standard for tools and data.

Hence, this section discusses several opportunities in area of data science and big data analytics in coming future. Now next section will conclude this work in brief including several opportunities for future.

## IX. CONCLUSIONS WITH FUTURE SCOPE

Data science is necessity of today's world, i.e., necessary for industries or organization to improve their product's quality (with increasing profit). But we find that there a shortage of skilled people to make decisions (based on analysis) form this large amount of data. In near future, we can focus on providing many programs or courses at graduate level at various universities (to produce more skilled people).To make realize the value of big data, we have to grow up in a team and came up with the talent to make analytic-based decisions. This paper discussed several essential terms related to data science and data analytics in details, also provide a complete state of art comparison to avoid any conflict among several similar terms like data engineer, data scientist, data science, etc. For future with data science and data analytics, we know that cloud is connected to IoT devices for string large information (virtually). With the increase of security and reliability in the cloud, the use of data analytics in IOT also gets hiked. In near future, most of the data will be saved (stored) at clouds and this data will be access by many smart devices anytime, anywhere (for performing/ communicating smartly, refer example discussed in [11]), which invite several issues like security, privacy, trust, etc. So, overcoming such issues like security, privacy risks of cloud/ IoT devices (to make them efficient and secure against any vulnerability).

## REFERENCES

[1] T. H. Davenport & J. G. Harris, Competing on Analytics: The New Science of Winning. Harvard Business School Press, 2007.
[2] Big Data Nowby O'Reilly Media, Book. Available at: http://index-of.co.uk/Big-Data-Technologies/Big%20Data%20Now.pdf
[3] Amandeep Khurana,"Bringing Big Data Systems to the Cloud", IEEE Cloud Computing Published by the IEEE Computer Society, pp.72-75, 2014.
[4] Hellerstein, Joseph M., et al. "The MADlib analytics library: or MAD skills, the SQL." Proceedings of the VLDB Endowment 5.12 (2012): 1700-1711.
[5] Anil, Robin, Ted Dunning, and Ellen Friedman. Mahout in action. Manning, 2011.
[6] Lim, Seung-Hwan, et al. "Graph Processing Platforms at Scale: Practices and Experiences." Proceedings of the 2015 IEEE International Symposium on Performance Analysis of Systems and Software. 2015.
[7] Big data: The next frontier for innovation, competition and productivityhttp://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
[8] Coetzee, D., 1991. Illiteracy in South Africa: Some preventive policies and strategies from a developmental perspective. Development Southern Africa, 8(2), pp.215-231.

[9] Tyagi, Amit Kumar and G, Rekha, Machine Learning with Big Data (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019. Available at SSRN: https://ssrn.com/abstract=3356269

[10] Rekha, G., Tyagi, A.K. and Krishna Reddy, V., 2019. Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. International Journal of Hybrid Intelligent Systems, (Preprint), pp.1-10.

[11] Tyagi, Amit Kumar and M, Shamila, Spy in the Crowd: How User's Privacy Is Getting Affected with the Integration of Internet of Thing's Devices (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019. Available at SSRN: https://ssrn.com/abstract=3356268