# Big Data as a Driving Tool of Digital Transformation

**Harshini Priya Adusumalli**

Software Developer, Department of BigData, IBM, Manyatha Tech Park K Block, Nagwara, Bangalore, **INDIA**

## ABSTRACT

As a result of this research, it was discovered how Big Data is characterized by the five Vs: Velocity, Volume, Variety, Veracity, and Value; and how Hadoop and other tools, in conjunction with distributed computing capacity, are utilized to meet the needs of Big Data. The research defines the abilities necessary to analyze Big Data, as well as the method of Data Mining and how it generates results, and it also includes recommendations. Physicians may use data science to give the best care possible for their patients, and meteorologists can use it to anticipate the scope of local meteorological occurrences. Data science can even be used to predict natural disasters such as earthquakes and tornadoes. Capturing data is an excellent way for businesses to begin their data science journeys. They can begin evaluating the data as soon as they obtain it. Here are some examples of how people produce data and how corporations such as Netflix, Amazon, United Parcel Service (UPS), Google, and Apple exploit the data generated by their customers and workers. When a Data Science project is completed, the final output should be used to communicate new information and insights gained from the data analysis to important decision-makers.

Keywords: Digitization, Digital Transformation, Big Data, Data Science, Data Mining

**12/31/2016** **Source of Support:** Nil, **No Conflict of Interest:** Declared



## INTRODUCTION

To capitalize on the benefits of new technology, businesses must update old procedures and create new ones. This digital transformation alters how a business functions and offers value to consumers. It is a data-driven organizational and cultural shift. Big data and the competitive advantage that comes from understanding it have driven digital changes in many sectors (Pasupuleti, 2015a). The Houston Rockets NBA team used data collected by overhead cameras to assess the most effective plays, while Lufthansa studied consumer data to improve service. Organizations everywhere are changing fundamentally. Let's look at an example of how Big Data can impact a whole sector, not just one firm. In 2018, the NBA's Houston Rockets used Big Data to improve their game. The Rockets were one of four NBA clubs to implement a video tracking system. That's when they uncovered something shocking in their analysis of video tracking data (Granville, 2014). Data study

found that two-point dunks from inside the two-point zone and three-point shots from outside the three-point line give the best scoring possibilities. With this newfound knowledge, the squad increased the quantity of three-point attempts attempted. The Rockets made more three-pointers than any other team in NBA history in 2017-18, which helped them win more games than their opponents. Big Data has transformed the way teams compete in basketball. In-depth examination of how the business functions helps firms find how to enhance their processes and operations, and exploit the benefits of incorporating data science into their workflows (Pasupuleti, 2015b). Most firms recognize that digital transformation will influence their culture and how they treat data, people, and consumers. Because digital transformation affects every element of a business, it requires top-level leadership to achieve success. The assistance of the Chief Executive Officer, Chief Information Officer, and the growing job of Chief Data Officer is critical to the digital transformation process. Executives who oversee finances, employees, and daily priorities must also support them. This is a process. To succeed, everyone must support it. Fighting the difficulties that develop takes a new perspective, but Digital Transformation is the way to win today and in the future.

## DATA SCIENCE SKILLS & BIG DATA

**Goals for Data Mining:** Setting goals is the first stage in data mining. Clearly, you must determine the main questions. Beyond defining the important issues are concerns regarding the exercise's costs and benefits. You must also anticipate the accuracy and usefulness of the data mining results. In an ideal world, you could throw as much money at the problem as you wanted. However, the cost-benefit trade-off is always important in setting the data mining goals and scope. The projected accuracy of the results also affects the expenses. Data mining accuracy would cost more and vice versa (Pasupuleti, 2015c). Furthermore, the practice has diminishing rewards at a certain degree of accuracy. Thus, cost-benefit trade-offs for desired accuracy are key data mining goals.

**Picking Data:** The outcome of a data-mining effort is heavily reliant on the data used. Data are sometimes easily available for processing. Retailers, for example, frequently have extensive databases of client data. However, data mining may not be freely available. In these circumstances, you must find new data sources or design new data gathering projects, such as surveys. The kind, amount, and frequency of data gathering directly affect the cost of data mining. Data mining requires specific data to answer specific queries at acceptable prices (Tomar and Agarwal, 2013).

**Preparing Data:** Data mining requires preprocessing data. Raw data are often untidy, including errors or irrelevant data. There is also omission of data. During preprocessing, you identify unimportant data properties and remove them from further consideration (Pasupuleti, 2016). Identifying and reporting erroneous portions of the data collection is also required. Human mistake may result in accidental merger or improper processing of data between columns. Data should be checked for accuracy. Finally, you must establish a formal approach for dealing with missing data and identify its randomness or systematicity. It would be sufficient if the data were missing randomly. When data are absent in a systematic method, you must assess the outcomes. For example, a portion of people in a big data set may have declined to provide their income (Prajapati, 2013). Findings based on an individual's income would remove specifics of those who did not submit their income. As a result, the analysis is biased. Observations or variables with missing data must be eliminated from the entire analysis or sections of it.

**Changing Data:** After retaining the required data properties, the following step is to decide on the data storage format. In data mining, reducing the number of characteristics required to describe a phenomenon is key. To do so, data may need to be Principal Component Analysis (described later in the chapter) can minimize the number of characteristics without compromising information. Variables may also need to be altered to assist explain a phenomena. For example, a person's income may be reported as salary income, income from other sources (such rental properties), or government support payments. Aggregating all sources of revenue yields a representative indication of personal income. Changing variables' types is common. It may be beneficial to categorize the continuous variable for income into low, middle, and high-income individuals. A nonlinear model might represent the underlying dynamics.

**Backup Data:** The data must be stored in a format suitable for data mining. The data must be stored in a format that allows the data scientist quick read/write access. Because new variables are formed during data mining and then written back to the old database, the data storage system should be efficient. It is also vital to store data on secure servers or storage media to avoid the data mining algorithm searching for data distributed over several servers or storage media. Data storage should prioritize data security and privacy.

**Mining Data:** Data mining uses data that has been properly processed, converted, and stored. This stage covers data analysis algorithms, both parametric and non-parametric. Data visualization is a fantastic place to start. Data mining software's strong charting features assist establish a basic knowledge of the trends concealed in the data collection.

**Analyzing Mining Output:** After extracting findings from data mining, you formalize the results. Formal assessment might include evaluating the models' prediction skills on observed data to evaluate how well the algorithms reproduce data. It's called a "in-sample forecast." The results are also shared with relevant stakeholders for input, which is used to enhance the process.

The analysts employ better and enhanced algorithms to increase the quality of findings provided based on input obtained from key stakeholders.

## DATA MINING

Many terminology are used interchangeably in data science, so let's go through some of the most often used terms in this field. Data sets that are so enormous, so quickly constructed, and thus diverse as to challenge typical analysis methods such as those performed using a relational database are referred to as "big data." Large amounts of computational power in dispersed networks, as well as new tools and methodologies for data analysis, have all been developed at the same time, providing businesses with the ability to analyze massive amounts of data simultaneously (Al Mahmud, 2012). Individuals are becoming more aware of new information and ideas that are becoming available. In many cases, big data is characterized in terms of the five V's: velocity; volume; variety; truthfulness; and value. Data mining is the technique of automatically exploring and analyzing large amounts of data in order to uncover previously unseen relationships. It entails preparing data in order to make it more usable and translating it into an acceptable format for the intended use. Afterwards, insights and patterns are mined and retrieved from the data using a variety of tools and techniques ranging from simple data visualization tools to machine learning and statistical models. Machine learning is a subset of artificial intelligence that employs computer algorithms to evaluate data and make intelligent judgments based on what it has learnt without being explicitly

taught in the process. Machine learning algorithms are educated with enormous collections of data and then learn by seeing and interacting with real world instances. They do not obey algorithms that are based on rules. Machine learning is the process through which machines learn to solve problems on their own and make accurate predictions based on the data they are given. Deep learning is a specific subclass of machine learning that use multilayer neural networks to replicate the decision-making process of a person. Deep learning algorithms are capable of labeling and categorizing data as well as identifying patterns. It is what enables artificial intelligence systems to continually learn on the job and enhance the quality and accuracy of outcomes by judging whether or not judgments were accurate on the first try. Inspired by biological neural networks, artificial neural networks (sometimes known simply as neural networks) function in a manner that is substantially distinct from that of biological neural networks. In artificial intelligence, a neural network is a collection of microscopic processing units known as neurons that take in incoming data and learn to make judgments over time as they process it. Because neural networks are often layer-deep, deep learning algorithms become more efficient as data sets grow in size, as opposed to other machine learning algorithms, which may reach a plateau as data sets grow in size. Neural networks are often used in conjunction with other machine learning algorithms. Assuming that you have a general awareness of the distinctions between certain fundamental artificial intelligence ideas, there is one additional distinction that you should be aware of: the distinction between Artificial Intelligence and Data Science. It is the process and approach of extracting information and insights from massive amounts of diverse data that is known as data science. It is a multidisciplinary discipline that includes mathematics, statistical analysis, data visualization, machine learning, and other topics. It is what enables us to appropriate information, see patterns, extract meaning from massive amounts of data, and utilize that knowledge to make business-critical choices. Many artificial intelligence approaches may be applied to data science in order to get insight from it. When it comes to extracting meaning and drawing conclusions from data, it may, for example, utilize machine learning methods and even deep learning models. There is considerable overlap between artificial intelligence and data science, although neither is a subset of the other. The phrase data science refers to the complete technique of data processing, whereas artificial intelligence refers to everything that allows computers to learn how to solve problems and make intelligent judgments. The utilization of large amounts of data is possible in both artificial intelligence and data science. That is, there are extremely enormous amounts of information.

## APPLICATIONS OF MACHINE LEARNING

Everyone now deals with AI. But recommender systems are undeniably a big application. Classifications, cluster analysis, resurrecting 20-year-old marketing problems, market basket analysis, what commodities go together. That was a challenging computational challenge, but now we do it all the time using machine learning. Machine learning also includes predictive analytics. We're employing new methods to anticipate things statisticians dislike. Decision trees, Bayesian Analysis, naive Bayes, etc. The wonderful thing about them is that programs like R now require you to understand how these approaches may be utilized, not just how to use them. Anyone with a basic understanding of data science may use these strategies, but they need to understand the trade-offs between precision and recall, over sampling, and over fitting. That's definitely a few other uses of machine learning in fintech that I could discuss. One is a recommendation. So, whether you use Netflix, Facebook, or other software applications, you get

recommendations. The user has seen this program and may like to view another one. Right, or you already follow this person, therefore you should follow this other. It's kind of the same in fintech. Because you've looked at this investment concept, it would be great for you to look at this other investment idea that's comparable. Right, it's a comparable asset and firm. Or it's a comparable investing method. So we can use machine learning to make recommendations throughout fintech. Another key topic in retail banking and finance is fraud detection. Machine learning is used to identify whether a charge on a credit card is valid or not. Right, you have to learn from past transactions and construct a model, and then when the charge comes through, you have to compute everything and say, "No problem," or "hmm, not so good." Let's forward it to our fraud team for review."

## HOW DATA SCIENCE IS SAVING LIVES

Human lives are greatly impacted by using Data Science tools to interpret and analyze enormous data collections. It may help healthcare professionals manage patients better, anticipate natural disasters so people can prepare ahead of time, and much more. Using predictive analytics, data scientists in healthcare determine the best solutions for patients (Rogers, 2012). For example, gene markers, related conditions, and environmental variables are all examined. It then suggests relevant testing, trials, and treatments. Each physician has accumulated information through their education, interests, and life experiences. Predictive analytics-based data science solutions ensure that all clinicians have access to the most up-to-date illness, diagnostic, and treatment information. With such a system, all doctors have access to the same information, resulting in better patient outcomes. Consider the Boston Consulting Group and AdvaMedDx, an industry alliance of medical diagnostics businesses, who investigated the adoption of potentially lifesaving diagnostic tests for individuals with certain cancers and gene markers. Oncologist knowledge of the test and its relevance to the gene marker was shown to be the most important factor in the patient being given a test. Using data science methods can assist clinicians identify the most beneficial tests and therapies for a certain patient? There are numerous more ways to mine data, such as from electronic medical records for medical research. Schools like NorthShore University HealthSystem in suburban Chicago, a pioneer in EMR use, now teach data mining. First in America to receive the highest degree of EMR deployment for inpatient and outpatient care. So much anonymized data is now available for creative analytics research. Using data science, healthcare businesses may progress beyond simple descriptive analytics to predictive insights (Srivathsan and Arjun, 2015). Data Science technologies have been used to save lives in disaster preparedness for many years. These techniques are improving and delivering fresh data analysis, warning communities to risk faster than before. Large, high-quality data sets may be utilized to anticipate the onset of several natural disasters, potentially saving thousands of lives. Data science can anticipate earthquakes, hurricanes, tornadoes, floods, and volcanic eruptions. Recent UK study uses social media material including photographs and phrases to follow floods, storms, and other meteorological disasters. These data can be used to enhance forecasts for local weather occurrences when combined with data from scientists and weather stations. Because the value of this information is so high, schools are beginning to include it into their curricula. So the University of Chicago Graham School provides an MS in Threat and Response Management. Data science tools enable businesses to analyze massive amounts of data from many sources and display it in a way that data scientists can learn new things, perhaps saving hundreds of lives.

## How Should Companies Get Started in Data Science?

Ultimately, businesses recognize that if they can't measure it, they can't improve it. They can't cut costs until they can measure them. They can't boost earnings until they can measure them. So the first step for a corporation is to start recording data, especially expense data. And the labor and material expenses, the cost to sell one product, and the overall cost. Then you look at revenue, where does it come from? Is 80% of your income from 20% of your customers? Or is it the reverse? So, initially, start collecting data. Then you may apply algorithms and analytics to the data. The first step is to collect data. Start catching it if you aren't already. If you capture it, save it. Don't delete outdated data assuming you don't need it. Data is timeless. A century or two old data is still relevant. It affects you, your company, and your success. So save data, record it, archive it, and avoid waste. Assure consistency. Having good documentation can help someone 20 years later comprehend the data. Now. Start a firm with data preservation best practices in place. And if you're already in business, do it immediately. >> Start measuring. A decade of poor measurement is followed by a desire for data science. Data science is only as valuable as the data collected. Trash in, garbage out applies to all analyses. >> It's tough to enhance or adjust anything that isn't measured. Measuring is the first step. If organizations already have data, they should start cleaning it up. If no data exists, they must begin gathering it. >> I'm looking for a team of data scientists. First, hire workers who are interested in data science. Because engagement requires interest in your company. >> Companies should recognize that teamwork is vital. So it's not just one data scientist, but a team of data scientists with varying strengths.

## Applications of Data Science

Data science and big data are undeniably transforming corporate operations, financial analytics, and most importantly consumer interactions. It's apparent that firms can benefit greatly from data science's insights. But it's not always clear how. So here are some instances. Every day, practically everyone creates large amounts of data, often without realizing it. This digital trail discloses our online habits (Ahmed et al., 2013). If you've ever looked for or bought something on Amazon, you'll notice it starts giving suggestions. A recommendation engine is a frequent data science application. Companies like Amazon, Netflix, and Spotify employ algorithms to offer personalized suggestions based on past consumer behavior. Personal assistants like Apple's Siri employ data science to answer inquiries from consumers. Google tracks your global travels, internet purchases, and social media activity. It then analyzes that data to provide restaurant, bar, store, and other attraction suggestions based on your device's data and present location. Wearable gadgets like Fitbits, Apple watches, and Android watches contribute data about your activity, sleep, and heart rate. Knowing how customers create data, let's see how data science impacts business. The key to competition, according to McKinsey & Company, is data science. Sustaining fresh waves of innovation and productivity. In 2013, UPS revealed a new route guiding technology that would save time, money, and gasoline. Like this, data science will profoundly alter how firms compete and function. An organization's competitive edge Consider Netflix. This includes what shows individuals are viewing at what time of day, how often they pause, rewind or fast-forward through episodes, and what directors and stars they search for. Netflix knows a program will be a smash before it even starts production by assessing user preferences for directors and actors. Add this to the previous success of a show and you have a hit. Netflix, for example, knowing many of its subscribers had watched David Fincher. They also knew that Robin Wright's films

always did well, and that the British version of House of Cards did well. Netflix recognized that many Fincher fans also enjoyed Wright. All of this suggests that purchasing the series would be a wise business decision. Right. It was a smash. Netflix anticipates customer desires thanks to data science.

## CONCLUSION

The goal of analytics is to convey results to those who can utilize them to establish policy or strategy. Tables and graphs describe findings. The data scientist should next utilize the information to develop a story. Essays and reports are the end product in academics. Typical outputs range from 1,000 to 7,000 words. The end deliverable varies in consulting and business. It might be a short document of less than 1500 words including tables and graphs, or a long document of several hundred pages. The reports are used by large consulting companies like McKinsey and Deloitte to disseminate results and create expertise in certain knowledge fields.

Let's look at the Deloitte University Press' "United States Economic Forecast." This paper is a good example of a data-driven narrative delivery. The 24-page paper examines the US economy in December 2014. Contrary to common belief, the US economy and employment growth have been relatively healthy. The report is not a list of facts.

It's a well-written paper that includes Voltaire and has a clear topic. The study highlights the US economy's strengths. Among them are increasing US industrial investment and decreased oil costs, which may lead to increased consumer spending.

The Deloitte research shows market trends using time series charts. It demonstrates how the economy shrank during the Great Recession and has now recovered. The picture depicts four possible futures. A second graph illustrates changes in consumer spending. With reference to Thomas Pikkety's book on income disparity in the US. According to the Deloitte analysis, many consumers' actual salaries have stagnated but their expenditure has remained stable. Many other charts and graphs concentrate on the housing, business and government sectors. The appendix contains four tables with data from the report's four scenarios.

The "United States Economic Forecast" by Deloitte achieves its goal. The report's economic projections are based on facts and analytics. It supports the concept that the US economy is performing better than most people think. Nonetheless, the study demonstrates Deloitte's ability to analyze economic data and prescribe solutions to economic problems.

Consider presenting the findings as a deck of PowerPoint slides with eight images and four tables, without the narrative. The authors' report mentioning Piketty and Voltaire would have been lost on PowerPoint slides. It's worth reading Deloitte's report to see if the deliverable would have been as strong without the narrative.

Let us now reverse the Deloitte report. The writers must have discussed the final product before starting their analysis. They would have debated the report's main point, then sought for facts and analytics to support it. Planning and visualizing the ultimate result is critical to creating an engaging paper. Without regard for the final outcome, embarking on analytics is likely to result in a poor-quality paper where the analytics and narrative clash.

## REFERENCES

Ahmed, A. A. A., Siddique, M. N., & Masum, A. A. (2013). Online Library Adoption in Bangladesh: An Empirical Study. 2013 Fourth International Conference on e-Learning "Best Practices in Management, Design and Development of e-Courses: Standards of Excellence and Creativity", Manama, 216-219. https://doi.org/10.1109/ECONF.2013.30

Al Mahmud, T. (2012). A Survey on How Dynamically Changes Topology in Wireless Sensor Network. *ABC Journal of Advanced Research*, *1*(1), 28-34. https://doi.org/10.18034/abcjar.v1i1.3

Granville, V. (2014). Developing Analytic Talent: Becoming a Data Scientist, John Wiley and Sons, Incorporated, US.

Pasupuleti, M. B. (2015a). Data Science: The Sexiest Job in this Century. *International Journal of Reciprocal Symmetry and Physical Sciences*, *2*, 8–11. Retrieved from https://upright.pub/index.php/ijrsps/article/view/56

Pasupuleti, M. B. (2015b). Problems from the Past, Problems from the Future, and Data Science Solutions. *ABC Journal of Advanced Research*, *4*(2), 153-160. https://doi.org/10.18034/abcjar.v4i2.614

Pasupuleti, M. B. (2015c). Stimulating Statistics in the Epoch of Data-Driven Innovations and Data Science. *Asian Journal of Applied Science and Engineering*, *4*, 251–254. Retrieved from https://upright.pub/index.php/ajase/article/view/55

Pasupuleti, M. B. (2016). The Use of Big Data Analytics in Medical Applications. *Malaysian Journal of Medical and Biological Research*, *3*(2), 111-116. https://doi.org/10.18034/mjmbr.v3i2.615

Prajapati, V. (2013). Big data analytics with R and Hadoop. Packt Publishing Ltd.

Rogers, S. (2012). What is a data scientist? The Guardian. https://www.theguardian.com/news/datablog/2012/mar/02/data-scientist

Srivathsan, M., and Arjun, K. Y. (2015). Health monitoring system by prognotive computing using big data analytics. Procedia Computer Science, 50, 602-609.

Tomar, D., and Agarwal, S. (2013). A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

**--0--**