International Journal of Trend in Scientific Research and Development (IJTSRD)
UGC Approved International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

A Study on Issues, Challenges and Application in Data Science

Mukul Varshney
Computer Science and Engineering,
Sharda University, Uttar Pradesh, India

Shivani Garg
Computer Science and Engineering,
Sharda University, Uttar Pradesh, India

Jyotsna
Computer Science and Engineering,
Sharda University, Uttar Pradesh, India

Abha Kiran Rajpoot
Computer Science and Engineering,
Sharda University, Uttar Pradesh, India

ABSTRACT

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Data science is about dealing with large quality of data for the purpose of extracting meaningful and logical results/conclusions/patterns. It's a newly emerging field that encompasses a number of activities, such as data mining and data analysis. It employs techniques ranging from mathematics, statistics, and information technology, computer programming, data engineering, pattern recognition and learning, visualization, and high performance computing. This paper gives a clear idea about the different data science technologies used in Big data Analytics.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Data Science is much more than simply analysing data. There are many people who enjoy analysing data who could happily spend all day looking at histograms and averages, but for those who prefer other activities, data science offers a range of roles and requires a range of skills. Data science includes data analysis as an important component of the skill set required for many jobs in the area, but is not the only skill. In this paper the authors effort will concentrated on to explore the different issues, implementation and challenges in Data science.

Keywords: *Data science, analytics, information, data, unstructured data, preservation, data visualization, extraction*

I. INTRODUCTION

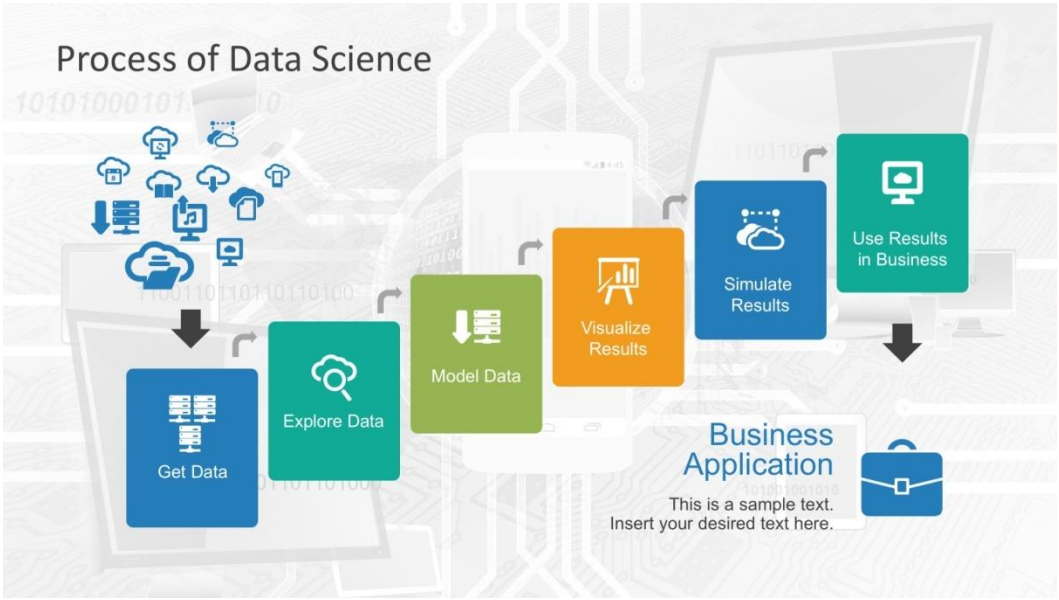
Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Although the name Data Science seems to connect most strongly with areas such as databases and computer science, many different kinds of skills including nonmathematical skills are also needed here.

Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods, but

also comprises its results. Data Science intends to analyze and understand actual phenomena with "data". In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human and social phenomena with data from a different point of view from the established or traditional theory and method. This point of view implies multidimensional, dynamic and flexible ways of thinking. Data Science consists of three phases: design for data, collection of data and analysis on data. It is important that the three phases are treated with the concept of unification based on the fundamental philosophy of science explained below. In these phases the methods which are fitted for the object and are valid, must be studied with a good perspective [4, 5].

Data science solely deals with getting insights from the data whereas analytics also deals with about what one needs to do to 'bridge the gap to the business' and 'understand the business priories'. It is the study of the methods of analyzing data, ways of storing it, and ways of presenting it. Often it is used to describe cross field studies of managing, storing, and analyzing data combining computer science, statistics, data storage, and cognition. It is a new field so there is not a consensus of exactly what is contained within it.

Data Science is a combination of mathematics, statistics, programming, the context of the problem being solved, ingenious ways of capturing data that may not be being captured right now plus the ability to look at things 'differently' and of course the significant and necessary activity of cleansing, preparing and aligning the data[7].



II. CHALLENGES IN BIG DATA ANALYSIS

1. **Heterogeneity and Incompleteness** When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work. Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient [9].

2. **Scale** Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static. First, over the last five years the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node. Furthermore, the move towards packing multiple sockets (each with 10s of cores) adds another level of complexity for intra-node parallelism. Finally, with predictions of "dark silicon", namely that power consideration will likely in the future prohibit us from using all of the hardware in the system continuously, data processing systems will likely have to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how we design, build and operate data processing components. The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goal [10, 17].

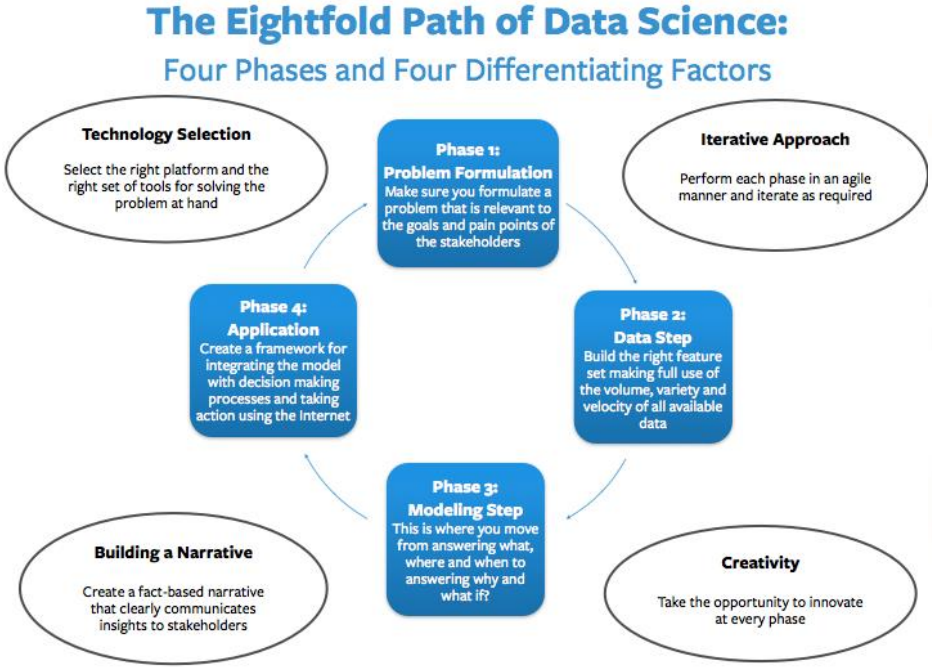
3. **Timeliness** The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge as described in Sec. 2.1, and a timeliness challenge described next. There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise

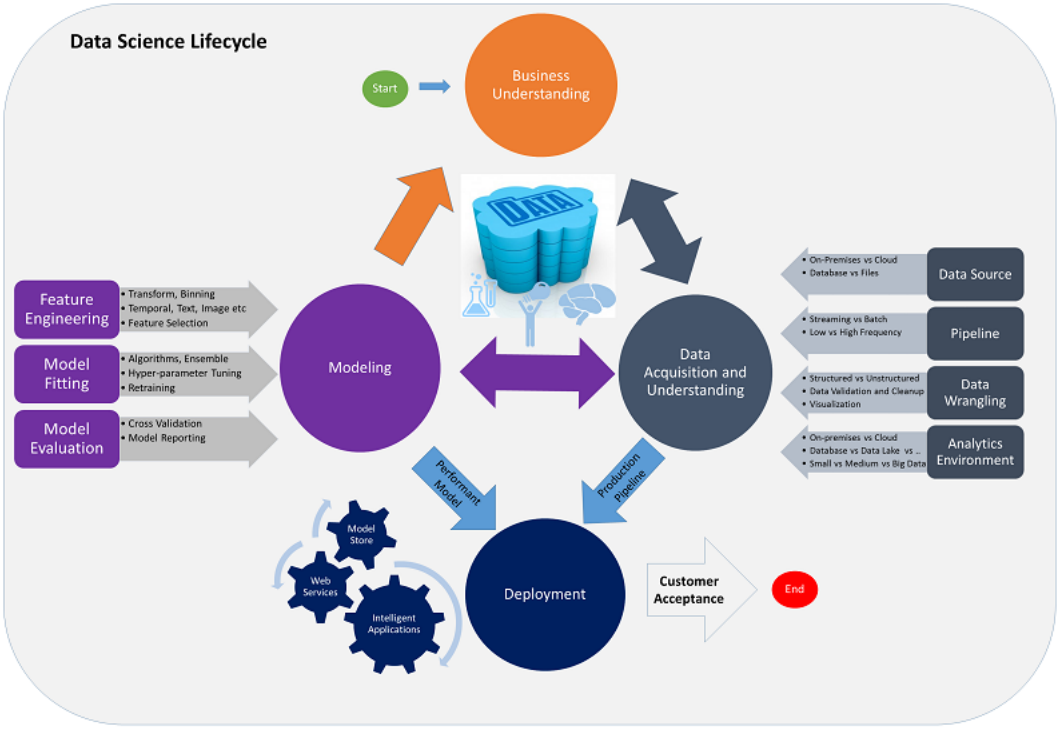
new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

4. **Privacy** The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data [16].

III. PHASES OF DATA SCIENCE

The three segments included in data science are arranging, bundling and conveying information (the ABC of information). However bundling is an integral part of data wrangling, which includes collection and sorting of data. However what isolates data science from other existing disciplines is that they additionally need to have a nonstop consciousness of What, How, Who and Why. A data science researcher needs to realize what will be the yield of the data science transform and have an unmistakable vision of this yield. A data science researcher needs to have a plainly characterized arrangement on in what manner this yield will be accomplished inside of the limitations of accessible assets and time. A data scientist needs to profoundly comprehend who the individuals are that will be included in making the yield. The steps of data science are mainly: collection and preparation of the data, alternating between running the analysis and reflection to interpret the outputs, and finally dissemination of results in the form of written reports and/or executable code. The following are the basic steps involved in data science [1, 2]





IV. TOOLS OF DATA SCIENCE

1) Python

Python is a powerful, flexible, open-source language that is easy to learn, easy to use, and has powerful libraries for data manipulation and analysis. It's simple syntax is very accessible to programming novices, and will look familiar to anyone with experience in Mat lab, C/C++, Java, or Visual Basic. For over a decade, Python has been used in scientific computing and highly quantitative domains such as finance, oil and gas, physics, and signal processing. It has been used to improve Space Shuttle mission design, process images from the Hubble Space Telescope, and was instrumental in orchestrating the physics experiments which led to the discovery of the Higgs Boson (the so-called "God particle"). Python is one of the most popular programming languages in the world, ranking higher than Perl, Ruby, and JavaScript by a wide margin. Among modern languages, its agility and the productivity of Python based solutions are legendary. The future of python depends on how many service providers allow for SDKs in python and also the extent to which python modules expand the portfolio of python apps.

2) The R Project for Statistical Computing

R is a perfect alternative to statistical packages such as SPSS, SAS, and Stata. It is compatible with Windows, Macintosh, UNIX, and Linux platforms and offers extensible, open source language and computing environment. The R environment provides with software facilities from data manipulation, calculation to graphical display.

A user can define new functions and manipulate R objects with the help of C code. As of now there are eight packages which a user can use to implement statistical techniques. In any case a wide range of modern statistics can be implemented with the help of CRAN family of Internet websites.
There are no license restrictions and anyone can offer code enhancements or provide with bug report.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- An effective data handling and storage facility.
- A suite of operators for calculations on arrays, in particular matrices.
- A large, coherent, integrated collection of intermediate tools for data analysis.
- Graphical facilities for data analysis and display either on-screen or on hardcopy

- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined
- Recursive functions and input and output facilities.

3) Hadoop

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters. Relation between Data Management and Data Analysis All that means you can scale your data up and down without having to worry about hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is not for the data beginner. To truly harness its power, you really need to know Java. It might be a commitment, but Hadoop is certainly worth the effort – since tons of other companies and technologies run off of it or integrate with it. But Hadoop Map Reduce is a batch-oriented system, and doesn't lend itself well towards interactive applications; real-time operations like stream processing; and other, more sophisticated computations [12, 16].

4) Visualization Tools

Data visualization is a modern branch of descriptive statistics. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information". Some of the tools are This software adopts a very different mental model as compared to using programming to produce data analysis. Think about the first GUI that made computers public-friendly, suddenly the product has been repositioned. "Pretty Graphs" are useless if they just look pretty and tell you nothing. But sometimes making data look pretty and digestible also makes it understood to the average person. Tableau occupies a niche to allow non-programmers and business types to do guaranteed hiccup-free ingestion of datasets, fast exploration and very quickly generate powerful plots, with interactivity, animation etc. D3: You should use D3.js because it lets you build the data visualization framework that you want. Graphic / Data Visualization frameworks make a great deal of decisions to make the framework easy to use. D3.js focuses on binding data to DOM elements. 3 stand for Data Driven Documents. We will explore D3.js for its graphing capabilities. Data wrapper: Data wrapper allows you to create charts and maps in four steps. The tool reduces the time you need to create your visualizations from hours to minutes. It's easy to use – all you need to do is to upload your data, choose a chart or a map and publish it. Data wrapper is built for customization to your needs; Layouts and visualizations can adapt based on your style guide.

5) Paxata

Paxata focuses more on data cleaning and preparation and not on machine learning or statistical modelling part. The application is easy to use and its visual guidance makes it easy for the users to bring together data, find and fix any missing or corrupt data to be resolved accordingly. The data can be shared and re-used with other teams. It is apt for people with limited programming knowledge to handle data science.

Here are the processes offered by Pixata:

- The Add Data tool obtains data from wide range of sources.
- Any gaps in the data can be identified during data exploration.
- User can cluster data in groups or make pivots on data.
- Multiple data sets can be easily combined into single AnswerSet with the help of SmartFusion technology solely offered by Paxata. With just a single it automatically finds out the best combination possible [17].

V. APPLICATIONS

Data science is a subject that arose primarily from necessity, in the context of real-world applications instead of as a research domain. over the years, it has evolved from being used in the relatively narrow field of statistics and analytics to being a universal presence in all areas of science and industry. in this section, we look at some of the principal areas of applications and research where data science is currently used and is at the forefront of innovation.

1. Business Analytics _collecting data about the past and present performance of a business can provide insight into the functioning of the business and help drive decision-making processes and build predictive models to forecast future performance. some scientists

have argued that data science is nothing more than a new word for business analytics[19], which was a meteorically rising field a few years ago, only to be replaced by the new buzzword data science. Whether or not the two fields can be considered to be mutually independent, there is no doubt that data science is in universal use in the field of business analytics.

2. Prediction _ large amounts of data collected and analyzed can be used to identify patterns in data, which can in turn be used to build predictive models. This is the basis of the field of machine learning, where knowledge is discovered using induction algorithms and on other algorithms that are said to "learn" [20]. Machine learning techniques are largely used to build predictive models in numerous fields. 3. Security _ data collected from user logs are used to detect fraud using data science. Patterns detected in user activity can be used to isolate cases of fraud and malicious insiders. Banks and other financial institutions chiefly use data mining and machine learning algorithms to prevent cases of fraud [12].

4. Computer Vision _ data from image and video analysis is used to implement computer vision, which is the science of making computers "see", using image data and learning algorithms to acquire and analyze images and take decisions accordingly. This is used in robotics, autonomous vehicles and human-computer interaction applications.

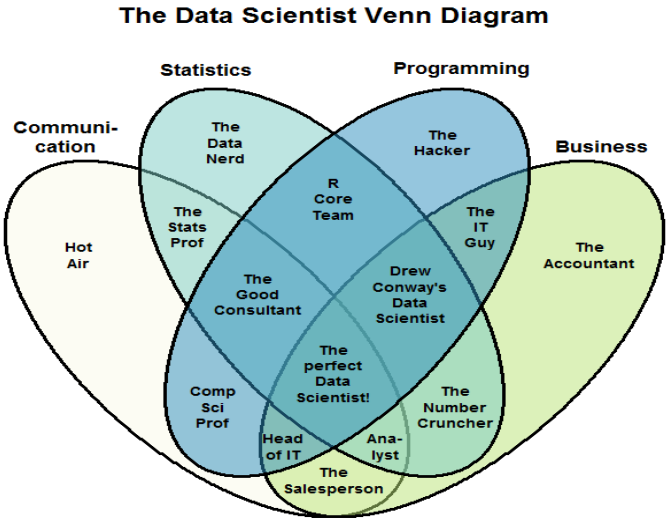
5. Natural Language Processing _ modern nlp techniques use huge amounts of textual data from corpora of documents to statistically model linguistic data, and use these models to achieve tasks like machine translation[15], parsing, natural language generation and sentiment analysis.

6. Bioinformatics: bioinformatics is a rapidly growing area where computers and data are used to understand biological data, such as genetics and genomics. These are used to better understand the basis of diseases, desirable genetic properties and other biological properties. As pointed out by michael walker _ "next-generation genomic technologies allow data scientists to drastically increase the amount of genomic data collected on large study populations. When combined with new informatics approaches that integrate many kinds of data with genomic data in disease research, we will better understand the genetic bases of drug response and disease."

7. Science and Research _ scientific experiments such as the well-known large hadron collider project generate data from millions of sensors and their data have to be analyzed to draw meaningful conclusions. Astronomical data from modern telescopes [11] and climatic data stored by the nasa center for climate simulation are other examples of data science being used where the volume of data is so large that it tends towards the new field of big data.

8. Revenue Management - real time revenue management is also very well aided by proficient data scientists. in the past, revenue management systems were hindered by a dearth of data points. In the retail industry or the gaming industry too data science is used. As jian wang defines it:"revenue management is a methodology to maximize an enterprise's total revenue by selling the right product to the right customer at the right price at the right time through the right channel. "now data scientists have the ability to tap into a constant flow of real-time pricing data and adjust their offers accordingly. It is now possible to estimate the most beneficial type of business to nurture at a given time and how much profit can be expected within a certain time span.

9. Government - data science is also used in governmental directorates to prevent waste, fraud and abuse, combat cyber attacks and safeguard sensitive information, use business intelligence to make better financial decisions, improve defense systems and protect soldiers on the ground. In recent times most governments have acknowledged the fact that data science models have great utility for a variety of missions.



VI. CONCLUSIONS

Through data science, better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

For sure the future will be crowded with people trying to applying data science in all problems, kind of overusing it. But it can be sensed that we are going to see some real amazing applications of DS for a normal user apart from online applications (recommendations, ad targeting, etc). The skills needed for visualization, for client engagement, for engineering saleable algorithms, are all quite different. If we can perform everything perfectly at peak level it'd be great. However, if demand is robust enough companies will start accepting a diversification of roles and building teams with complementary skills rather than imagining that one person will cover all bases.

REFERENCES

1) Jeff Leek (2013-12-12). "The key word in 'Data Science' is not Data, it is Science". Simply Statistics.
2) Hal Varian on how the Web challenges managers. http://www.mckinsey.com/insights/innovationn/hal_varian_on_how_the_web_challenges_managers
3) Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science.HYDROL PROCESS. 18:3637-653. http://www3.interscience.wiley.com/cgi-bin/jissue/109856902
4) Data Munging with Perl. DAVID CROSS. MANNING. Chapter 1 Page 4.
5) What is Data Science? http://www.datascientists.net/what-is-data-science
6) The Data Science Venn Diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
7) Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no. 1, 1--67. doi:10.1214/aoms/1177704711. http://projecteuclid.org/euclid.aoms/1177704711.
8) Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.
9) Peter Naur: Concise Survey of Computer Methods, 397 p. Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974
10) Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyt,"From Data Mining to Knowledge Discovery in Databases. . AI Magazine Volume 17 Number 3 (1996)
11) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics William S. Cleveland Statistics Research, Bell Labs.http://www.stat.purdue.edu/~wsc/papers/datascience.pdf
12) Eckerson, W. (2011) "BigDataAnalytics: Profiling the Use of Analytical Platforms in User Organizations," TDWI, September. Available at http://tdwi.org/login/default-login.aspx?src=7bC26074AC-998F-431BBC994C39EA400F4F %7d&qstring=tc%3dassetpg
13) "Research in Big Data and Analytics: An Overview" International Journal of Computer Applications (0975 – 8887) Volume 108 –No 14, December 2014
14) Blog post: Thoran Rodrigues in Big Data Analytics, titled "10 emerging technologies for Big Data", December 4, 2012.
15) Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012
16) T. Giri Babu Dr. G. Anjan Babu, "A Survey on Data Science Technologies & Big Data Analytics" published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 2, February 2016
17) Proyag Pal1, Triparna Mukherjee,"Challenges in Data Science: A Comprehensive Study on Application and Future Trends" published in international Journal of Advance Research in Computer Science and Management Studies , Volume 3, Issue 8, August 2015