



BIG DATA ANALYTICS

Data Scientist's Viewpoint
SHEFALI NAYAK

Preface

The book is intended to give you a Data Scientist's point of view and the thought process in the field of Big Data Analytics. At the end of this book, you should be able to define your journey on a Big Data Analytics project.

The book is designed to walk you through the various stages of a data analytics project, concepts and possible avenues when dealing with huge and overwhelming amounts of structured and unstructured data. It is crisp and concise roadmap on a Big Data project.

The book contains graphical representations and numerous examples to enable effective learning and understanding of the concepts. Also, leverages scenario examples to showcase where our understanding could falter and concepts could be applied inaccurately, to reinforce learning and fence the concept for caveats.

This is your go-to book for understanding difficult concepts in Big Data Analytics in a lucid language.

Complimentary chapters included – Short description and distinguishing factors on the various Machine Learning Models, data visualizations and storyboarding techniques; in which scenarios they are most appropriate to be leveraged to empower you to make relevant decisions related to the analytics approach.

Keep on Learning!





Contents, Disclaimer and Rights

The book cover has been designed by the author.
The examples have been created using dummy data
and the graphical and image representation in the
book has been illustrated by the author.

This book is not a piece of research and is intended to
explain the established concepts of statistics in a
simple manner.

The design and contents of the book may not be
copied, published or circulated without the consent of
the author.



Acknowledgments


I am extremely thankful and grateful to my parents, sibling and spouse for their constant encouragement and indispensable advice for the improvement of the content to shape it out in its present form.






What are we learning today?

1.	What is Big Data?	8
2.	Cloud Computing and IoT	14
3.	Big Data Analytics Roadmap	17
4.	Conceptualize	21
5.	Data Collection	29
5.1.	Primary data collection.....	30
5.2.	Secondary data collection.....	31
6.	Processing the Data	32
6.1.	Consistency	33
6.2.	Presence of Outliers.....	35
6.3.	Completeness.....	37
6.4.	Presence of Redundant Variables	39
7.	Data Preparation	42
7.1.	Data Aggregation for holistic views	43
7.2.	New Variable Creation	44



8.	Sampling Design	45
8.1.	Types of Sampling Design	48
8.2.	Random Vs Non-Random Sampling	49
9.	Exploratory Analysis	50
9.1.	Descriptive Statistics	50
9.2.	Data Visualization	57
9.2.1.	Types of Graph	57
9.2.2.	Graphs and their purpose	59
10.	Modify – The Normal Distribution	63
11.	Model Design	67
12.	Predictive Models	68
12.1.	Supervised Learning Models	69
12.1.1.	Regression Models	71
12.1.2.	Classification Models	71
12.2.	Unsupervised Learning Models	72
12.3.	Reinforcement Learning Models	74




13. Assess the Model Performance	77
14. Forward-Looking Path – Prescriptive.....	78
15. Stakeholder management.....	80
15.1. Documentation.....	80
15.2. Result Communication	82
15.3. Storyboarding.....	84
16. Measuring of Effectiveness	86
17. Reference	89
17.1. Reference 1: Images, Illustrations	89
17.2. Reference 2: Tables.....	90
About the Author and Editor.....	91



1. What is Big Data?

In less than half a century, we witnessed the landlines being replaced with the convenience of mobile technology. We now live in times where large data servers are accessible to many and supercomputers are not a myth. The landscape is constantly changing. Are we ready for this change? And how does it impact the way we respond to all that is changing around us?

Everything you just read is data!!! We are consuming this data at an exponential pace. In familiar conversation, we refer to the information we are dealing with as *Big Data* to describe an overwhelming glut of Structured, Semi-Structured and Unstructured data that it is difficult to process using the traditional database and software techniques. Simply put, Big Data is a buzz word or catchphrase used when the individual



data points can no longer be looked at in isolation to arrive at a meaningful conclusion and aggregates or derived analysis starts to matter for a holistic view of data.

One may ask why are organizations concentrating on making sense of data at this scale and *Why is Big Data so important?*

Well, it all boils down to gaining that competitive edge, distinguishing the company's product and services from that of competition, and accelerating the growth for business by knowing the requirements of the customers better.

While the advancement in technology has thrown at us a plethora of data, we observe that data has added to itself a whole new set of dimensions from the traditional garb of rows and columns. The conventional form of data has a decreased share in the overall types of data generated.

Let us look at some of the characteristics of Big Data that the researcher will require to bear in mind when dealing with research and analysis.

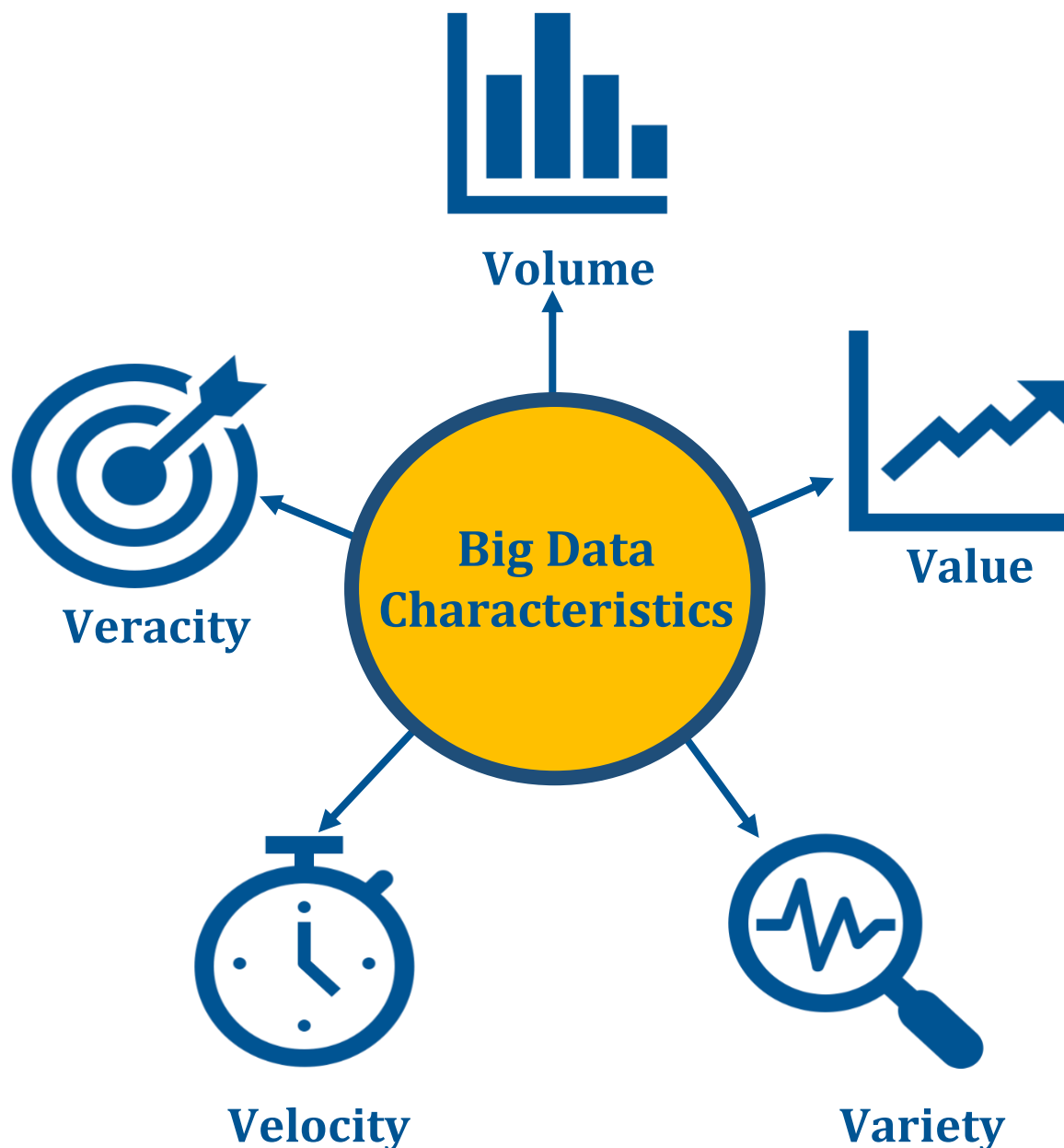


Figure 1. Pictorial Representation of Big Data Characteristics



1. Value – Insights from the data

The question the researcher requires to ask at the start of the research and analysis is Why are we processing and analyzing this data?

- i. Will the insights create a new product line, a cross-sell opportunity, or a cost-cutting measure?
- ii. Or will the data analysis lead to the discovery of a cure for a disease?

2. Volume – Amount of data

Volume refers to how much data we consume while analyzing the problem or scenario. The IoT (Internet of Things) is one of the major contributors to the exponential growth in data. The volume in Big Data requires –

- i. The storage capacity of large servers and
- ii. The processing power of supercomputers



3. Variety – Types of data


Big Data brings a lot of variety in the type of data. The data would range from being structured like the date, amount and time on a bank statement to being Unstructured like social media feeds, audio files, MRI images, web pages, weblogs.

4. Veracity – Accuracy of data

Now, Big Data brings a lot of variety in the type of data but can the data be trusted? Veracity refers to the trustworthiness of data. The data is not useful to the research and analysis if it is not accurate or reliable.

5. Velocity – Speed of the data

Velocity is the speed at which data is accessible. If it's not real-time it's usually not fast enough. The data from live audio and video streaming platforms, transactions




on credit and debit data are examples of data that requires real-time analysis.



2. Cloud Computing and IoT

Cloud Computing is defined as storing and accessing data and computing services over the internet. It has revolutionized the way we work with Big Data as we do not have to deal with the storage limitations of the personal computer and we can access the online and on-demand servers for data storage and computing. Cloud computing allows for more shared access and promotes collaboration on databases. This is a huge plus as multiple teams within the organization would require to access the data and work on the data to summarize it for various objectives. Additionally, the users do not face the limitation of sharing the same landmass as the data center and can access the data servers from remote servers.

The IoT (Internet of Things) is the application of cloud services. The Automated Teller Machine (ATM) that



helps us by dispensing the available cash in our bank account is the most common application of IoT.

The IoT (Internet of Things) is one of the major contributors to the exponential growth in the amount of data in recent years.

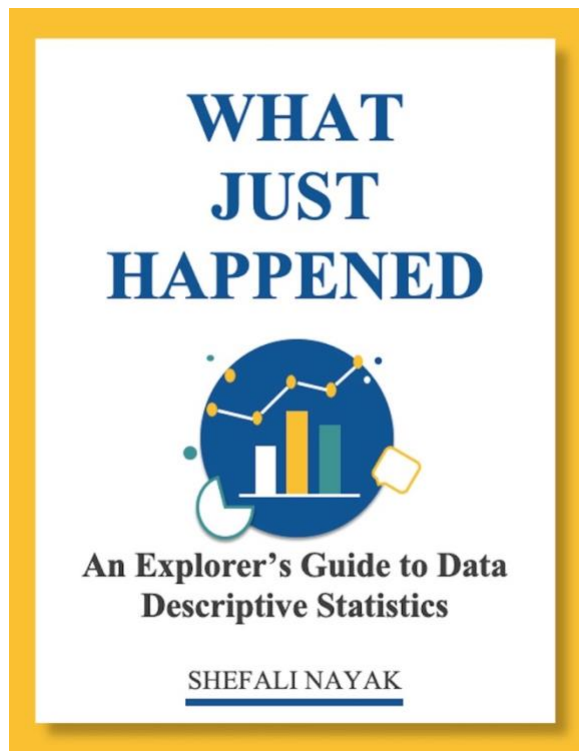
- i. Internet collects all the data consumed across the connected devices and either analyzes or sells it to third parties. The consumer starts receiving customized content and advertisements depending on the most recent searches in the internet browser.
- ii. There are also banking transaction data that are analyzed by banks to send across customized offers depending on the merchant category the customer is more likely or probable to spend on.

iii. We have wearables like the smartwatch that collect our physical statistics like the pulse rate and monitor our health and vitals over a period of time. The data is collected with the use of sensors in devices via the internet and is analyzed to arrive at data-driven insights. These insights help individuals or organizations to make more informed and data-driven decisions.



Figure 2. Pictorial Representation of Internet of Things(IoT)

Other books by the author

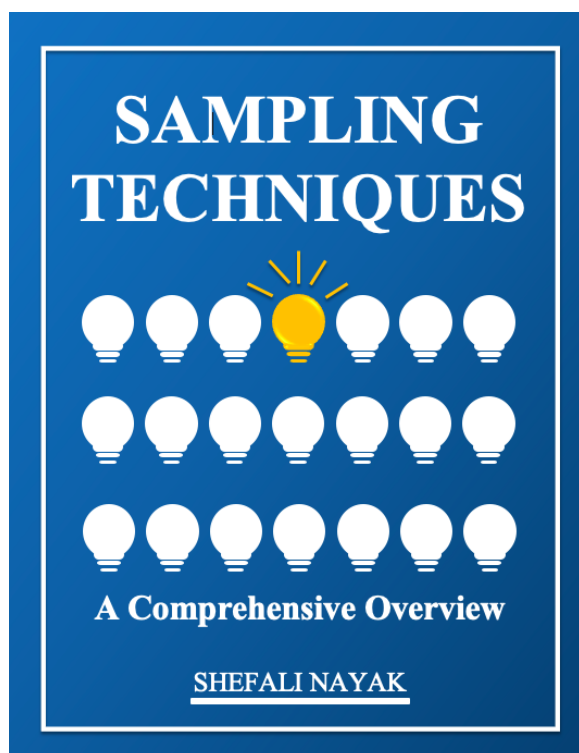


Descriptive Statistics

The book is intended to give you a comprehensive understanding of Descriptive Statistics and help you figure out "What just Happened?" through your data.

Complimentary chapters – Types of data, Normal distribution and data visualization for a holistic view of the Descriptive Statistics.

Learn more about Descriptive Statistics in lucid language.



Sampling Techniques

The book is intended to give you a comprehensive understanding of the Sampling Techniques. At the end of this book, you should be able to define the data collection process and choose the sampling technique that works best for your data.

Complimentary chapters – Types of data collection and Probability for a holistic view of the Sampling Techniques.

Follow the author



Share your review and feedback with the author about this book on social media. Let us connect, learn and collaborate on our common passion for DATA!

Instagram: [keep_on_learning_](#)