"Data, Data Analysis and Data Science, A Need Review"

Mr. Shailesh P. Thakare
Assistant Professor
Department of Information Technology
PRMIT & R, Badnera
Amravati, India
spthakare82@gmail.com

Mr. Nikhil S. Band
Assistant Professor
Department of Information Technology
PRMIT & R, Badnera
Amravati, India
nikhil.band08@gmail.com

Abstract— **A huge repository of data is generated everyday from information systems and new digital technologies now a days. Analysis of this large data requires a lot of efforts at multiple levels to extract knowledge for decision making. Recently, Data Analytics has become a hot topic in academics and industry. Data Analytics is the process of examining large amounts of data to discover hidden patterns or unknown correlations. A main part of data analytics is the need to collect, maintain and analyze large amounts of data efficiently. Therefore, data analysis is a current area of research and development. This paper aims to present basics of data, data analysis and data science which may helpful for students, fresh researchers and others in their work.**

I. INTRODUCTION

In digital world data is gathered and transmitted from various sources through high speed transmission lines which led to growth of huge amount of data everywhere. It becomes easier and cheaper to collect and store vast amounts of data now a days, all fields of human endeavor are transforming into data-intensive fields. the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications [4],[10]. To process, use and manage such huge data it requires new tools and techniques such as big data analysis. To understand these tools and techniques it is necessary to understand basic things behind it so this paper focuses on data, data analysis and data science.

II. DATA AND ITS TYPES

Representation of facts, concepts and instructions in a formalized manner suitable for communication, interpretation, or processing by humans or by some means is called as data. Data can be identified by three aspects which correspond to three ontologies, realism, nominalism and socially constructed reality and the corresponding beliefs about physical and social reality. As a result they emphasis the different possible roles of data, as:

- Record objective facts which will understood by everyone in exactly the same way.
- Record any type of concept, without guarantee to its accuracy or validity, which will be interpreted in all sorts of different ways by individuals.
- Use agreed structures and conventions for representing information, recording it and transmitting it, all in order to communicate it [1].

Data is short hand for information. Since People have a story to tell or a problem to solve they turn to data. It is used to describe things by assigning a value to them. The values are then organized, processed, and presented within a given context so that it becomes useful. There are two forms of data as: [2] [3].

Qualitative data: Representation of data either in a verbal or narrative format is qualitative data. This type of data is collected through focus groups, interviews, opened ended questionnaire items, and other less structured situations.

Quantitative data: Representation of data in numerical terms is quantitative data. Numerical terms may correspond to a specific category or label.

There are two types of quantitative data as:

- **Categorical data:** Data is placed into groups. Member of group or item cannot belong to more than one group at a time.
- **Continuous data:** It is numerical data measured on a continuous range or scale in which all values are possible with no gaps in between.

III. HOW IS DATA ANALYZED?

Data is analyzed using statistics. Essentially statistics can be classified in two ways: some help describe data and some help compare data [2].

- **Use of Statistics to Describe Data:** A ***frequency table*** presents aggregate data and gives us the number of times each particular data value occurred. A ***measure of central tendency*** is a single value that attempts to describe a set of data by identifying the central position within that set of data. The ***mean*** is the most well-known measure of central tendency that describes data. Another commonly used statistic to describe data is the standard deviation, describes the variability of the data or how closes or far away the data is to the average of the group.

- **Using Statistics to Compare Data:** Statistics allow us to compare one group with another in order to determine if the differences are **statistically significant**. When conducting tests of statistical significance, the tester determines a ***probability level*** to be used in determining significance. One of the statistics used to compare groups is the ***Chi-square***. The Chi-square is often the "go-to" statistic for categorical data, such as yes/no items or level of agreement. It is used to compare the pattern of answer options between two groups. ***Effect sizes*** are used to compare groups of data to determine how much change has occurred within a group. Effect sizes are used with continuous data.

Some common statistical terms:
Outliers: "Outliers" are values that are far from the mean; an outlier is a value that is very large or very small compared to the rest of the data set.

Small sample sizes: Small sample sizes can affect data analysis and interpretation as well.

Populations and Samples: Data is often collected to make statements or tell a story about a group or "population" of interest.

IV. METHODS FOR COLLECTING AND ANALYZING DATA

The quality and utility of monitoring, evaluation and research fundamentally relies on ability to collect and analyze quantitative and qualitative data [5].

- **For qualitative data:**

Individual interview: Is a conversation between two persons that with structure and a purpose. It is carried out to gather a knowledge or perspective on a specific topic.

Focus group discussions: A focus group discussion is an organized discussion between group of persons provide space to discuss particular topic, in a context where people are allowed to agree or disagree with each other.

Photovoice: Photovoice is a participatory method that enables people to identify, represent and enhance their community, life circumstances or engagement with a program through photography and accompanying written captions. Photovoice involves giving a group of participant's cameras, enabling them to capture, discuss and share stories they find significant.

Picture story: The picture story method enables children to communicate their perspectives on particular issues through a series of drawings they have made. The story telling can either be done in writing or verbally with a researcher. The picture story method is relatively quick and inexpensive. The picture story method provides a non-threatening way to explore children's views on a particular issue and to begin to identify what can be done to address any struggles faced by children.

Identifying participants: Qualitative research often focuses on a limited number of respondents who have been purposefully selected to participate because you believe they have in-depth knowledge of an issue you know little about it.

- **Qualitative data analysis**
  Qualitative data analysis is a process that seeks to reduce and make sense of vast amounts of information, often from different sources, so that impressions that shed light on a research question can emerge. It is a process where you take descriptive information and offer an explanation or interpretation.

- **For Quantitative data and methods**
  Quantitative data is numerical and can be collected in a number of forms. Statistical analysis is used to summarize and describe quantitative data and graphs or tables can be used to visualize present raw data.

- **Quantitative methods**
  Quantitative data can be collected using a number of different methods and from a variety of sources.

  - **Surveys and questionnaires**
  - **Biophysical measurements**
  - **Project records**
  - **Service provider or facility data**
  - **Service provider or facility assessments Data Strategies**

There are a variety of strategies for quantitative and qualitative analysis. Different strategies provide data analysts with an organized approach to working with data; they enable the

analyst to create a "logical sequence" for the use of different procedures.[3]

V. DATA STRATEGIES

There are a variety of strategies for quantitative and qualitative analysis. Different strategies provide data analysts with an organized approach to working with data; they enable the analyst to create a "logical sequence" for the use of different procedures [3].

Strategy	Involves	Reason
Visualizing the Data	Creating a visual "picture" or graphic display of the data.	way to begin the analysis process; or as an aid to the reporting/ presentation of findings.
Exploratory Analyses	Looking at data to identify or describe "what's going on"? – creating an initial starting point (baseline) for future analysis.	Like you have a choice?
Trend Analysis	Looking at data collected at different periods of time.	to identify and interpret (and, potentially, estimate) change.
Estimation	Using actual data values to predict a future value.	to combat boredom after you have mastered all the previous strategies. Also to answer PIR and Community Assessment items and tasks.

VI. DATA SCIENCE

Data Science refers to an emerging area of work concerning with the collection, preparation, analysis, visualization, management and preservation of large collections of information. Data Science includes data analysis as an important component of the skill set require for many jobs in this area, there are many challenges in involved in this work. Data Scientists plays active role in the design and implementation of four related areas Data Architecture, Data Acquisition, Data Analysis, and Data Archiving.

Data science is an interdisciplinary field requiring at least some level of expertise in three main areas. As demonstrated in the figure below (Conway, 2010), data science involves math and statistics knowledge, combined with expertise in a specific domain (the area from which the research question arises), and supplemented by hacking (command line/ text manipulation/ scripting/ programming) skills.

Fig: Data Science Venn diagram

Event data are the most important source of information. It may take place inside a machine, inside an enterprise information system, inside a hospital, inside a social network, inside a transportation system, etc. Events may be ''life events'', ''machine events'', or both. We use the term the Internet of Events (IoE) to refer to all event data available [5].

The IoE is made up of:

• **The Internet of Content (IoC):** The IoC includes traditional web pages, articles, encyclopedia,YouTube, e-books, etc.

• **The Internet of People (IoP):** The IoP includes e-mail, facebook, twitter, forums, etc.

• **The Internet of Things (IoT):** The IoT includes all things that have a unique id and a presence in an internet-like structure, etc.

• **The Internet of Locations (IoL):** Refers to all data that have a spatial dimension With the uptake of mobile devices more and more events have geospatial attributes.

Data science aims to use the different data sources to answer questions grouped into the following four categories:

• **Reporting:** What happened?

• **Diagnosis:** Why did it happen?

• **Prediction:** What will happen?

• **Recommendation:** What is the best that can happen?

Smart Phones may have following sensors and used to collect data

- GPS
- Proximity sensor
- Ambient light
- sensor
- Accelerometer
- Magnetometer
- Gyroscopic sensor
- WIFI
- Touchscreen
- Camera (front)
- Camera (back)
- Bluetooth
- Microphone

- GSM/HSDPA/LTE
- Finger-print
- Scanner

VII. CHALLENGES WHEN WORKING WITH DATA

As data nature gradually becomes more important, its study faces more and more challenges. These challenges are: [8].

Truth in Data
How do we know whether the data we have are telling the truth or giving false information?

Survival Problems in Cyberspace
Cyberspace is becoming a part of humanity's experience. How do we survive in cyberspace?

Scientific Research with Data
The discovery and exploration of phenomena and rules in data nature support the discovery of phenomena and rules in the natural world. Therefore, developing methods in data nature to explore the rules in the natural world is a important research field that will be helpful for scientific research.

Knowledge Acquisition from Data
A more important problem is how to acquire valuable knowledge from huge amount of data being generated.
Data storage, data communication, security, and computing machinery can be considered as the "hardware" in data science. Artificial intelligence (AI), statistical methods, algorithm design, and possible new mathematical methodologies can be viewed as the "software" of data science.[8]

CONCLUSION

As the data is becoming bigger and bigger, there is a need to store this data in an efficient manner. Analyzing these data is challenging for everyone now a days. In this paper we present basics of data, data analysis and data science so that it will helpful for students and new researchers to carry out their work and understand these concepts.

REFERENCES

[1]. Chapter 3: Data, Information and Meaning, https://repository.up.ac.za/bitstream/handle/2263/27367/03chapter3.pdf
[2]. Introduction To Data And Data Analysis May 2016, http://maryland.beaconhealthoptions.com/provider/forms/oms/Introduction-to-Data-and-Data-Analysis.pdf
[3]. Introduction To Data Analysis Handbook, Migrant & Seasonal Head Start, Technical Assistance Center, Academy for Educational Development, Contract with DHHS/ACF/OHS/Migrant and Seasonal Program Branch
[4]. Undergraduate Research Project in Data Science (Computer Science), Faculty Advisors: Ravi Gandham (Computer Science), Emilia Gan Computer cience) Proposed length of UGR Project: 2 quarters/student
[5]. Conway, Drew. "The Data Science Venn Diagram." March 2013.Web. Accessed 15 October 2013. SSRN: <http://drewconway.com/zia/2013/3/26/the-data-science-venn diagram>http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
[6]. An Introduction to Data Science, jeffray Stanton, Syaracuse University.
[7]. Data Scientist: The Engineer of the Future, Wil M. P. van der Aalst, Proceedings of the I-ESA Conferences 7, DOI: 10.1007/978-3-319-04948-9_2, Springer International Publishing Switzerland 2014.
[8]. Zhu, Y and Xiong, Y 2015 Towards Data Science. Data Science Journal, 14: 8, pp. 1–7, DOI:http://dx.doi.org/10.5334/dsj-2015-008.
[9]. Li M. Chen, Overview of Basic Methods for Data Science, © Springer International Publishing Switzerland 2015, L.M. Chen et al., Mathematical Problems in Data Science, DOI 10.1007/978-3-319-25127-1_2