Emerging Role of Data Scientist in Big Data

Shakti Kundu
Ph.D Scholar in CSE
DIT University, Dehradun, India

M L Garg, PhD
Professor and HOD / CSE
DIT University, Dehradun, India

ABSTRACT
In the world of Twitt, Text or Image, we have a huge amount of data. Social media, Digital Pictures, Video etc. also resulting in collection of enormous amount of data and the growth rate of data is still in running phase. In terms of figures, daily we are producing 2.5 quintal byte data in which most of the contribution comes from companies, businesses and corporate institutions. The interesting point is the companies from the world wide are using only 5% of the gathered and produced data. No doubt to handle such huge volume of data is not an easy task. so actually who do this job? The answer is the new specialist in terms of technicality i.e. Data Scientist. To gather data received from Digital Pictures, Social Media, E-mails, Live Chats, Data Warehouses, Videos, Sensors, Purchase Transaction record which are used for various business activities such as Product Design, Marketing, Sales or Distribution and on the basis of data gathering to take better decision is the responsibility of Data Scientist. The Data Scientist having multi skills such as Mathematicians, Numerologist, Analyst and Technologist work in various areas of Energy, E-commerce, Healthcare in progressive manner and they are in huge demand. The role of data scientist in the field of big data was highlighted in this paper.

Keywords
Analysis, Big Data, Data Science, Model, Process, Web

1. INTRODUCTION
To achieve a broad spectrum of preferred results, good Data Scientists are able to apply their skills and relevant knowledge. It include the ability to extract and interpret rich web data content and building rich tools that enable others to work effectively. According to some experts, the best data scientist acts as physician, besides those with backgrounds in computer science [1]. The skill-sets and competencies that Data Scientist employs vary widely. Data scientists acts as an integral part of competitive internet world, that handles a number of tasks, such as data mining and e-commerce analytics, which provides a competitive edge for business.

According to Peter Naur (1960), the term "Data Science" has existed for over thirty years and emerged as a substitute for "Computer Science". Naur published Computer Methods survey in 1974, which highlight the usage of 'Data Science' in a wide range of applications in the survey of data processing methods.

The director of HP Global Analytics says that Data Scientist in comparison of Software Engineer is wider post or responsibility. Data Scientist having skills of data specialists are also referred as Domain Experts and Manager. Having specialized skills, these Managers knew very well that how they can make use of data to take better decisions. The point that makes Data Scientist unique is their business skills. Data Scientists are having capability to progress an organization with business and IT leaders. Data Scientist normally works with the various activities going around an organization such as Data Analysis [3].

For setting future, one important question is what unique activities Data Scientist are actually doing? IBM vice-president (Big Data Products) gives answer to the raise question that imagine an Energy Company who claim 350 billion meter readings of electricity consumption for 2 Crore houses or any insurance company who can handle 70% of the valid claim. Denish Energy Company, Vestas Wind Systems is in search of Wind Turbain globally are taking help of Big Data. It's not an activity of future, rather than it has been happening presently in the various organizations and institutions world-wide. Big Data making all these activities to be happening and Data Scientists are working on that [3].

The remainder of the paper is organized as follows: In the subsequent Section 2, the Data Science Model has been discussed. In Section 3, the meaning, beginning and usage of Big Data have been highlighted. Finally, some Conclusions are given in Section 4.

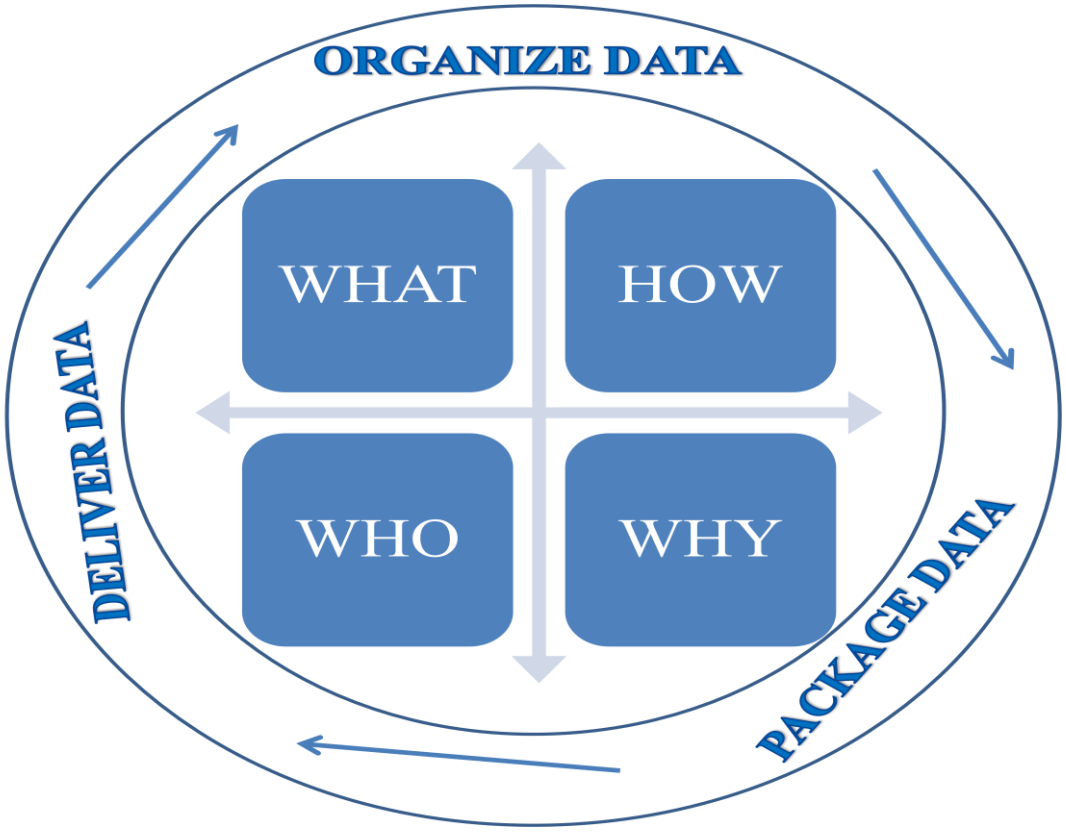

Figure 1: The Data Science Model

2. THE DATA SCIENCE MODEL

The Data Science consists of three components namely Organizing, Packaging and Delivering data. Organizing refers to planning and execution of structure of the data and physical location is planned. Packaging refers to building of prototypes, statistics performance and the creation of visualization. Delivering refers to proper packaging and obtaining the right value. The role of data science separates them from all other existing roles due to continual focus on What, Who, Why and How. A Data Scientist has focused objectives on how to achieve the relevant output from the restrained web log data. A Data Scientist knows the best categorization of people that will be involved in creating the output. Data Exploration is the practice of using visualization techniques to find unforeseen relationships between data points or sets of points from big databases. Once a relationship has been found, a similar visualization will be used to communicate that will be a reference to others. Visualization techniques can also be applied to information that is already known and it has the potential to organize large amounts of data in meaningful ways [4].

2.1 The 3 Step OPD Data Science Process

2.1.1 Organize Data

Organizing data involves gathering of relevant data in its proper format & storage and incorporating the best practices in data management.

2.1.2 Package Data

Packaging data involves in handling the raw data via logically and further representing it in new package.

2.1.3 Deliver Data

Delivering data involves by confirming that the information in the form of message had been reached and accessed by the authorized person [5].

2.2 OPD Data Science Process answers to the following:

- What is being created?
- How will it be created?
- Why is it to be created?
- Who will be involved in creating it?

All the respective questions help Data Scientist to present the data in an intelligent and effective way [5].

2.3 Origins of Data Science Model

Deriving valuable insights from web data is the general practice of Data Science. The emerging role of Data Science is capable to meet the challenges of processing huge amounts of data sets. Big Data includes structured, unstructured or semi-structured data which was produced by large enterprises. It helps in extracting new data generated from mobile social media and web. Data Science as a versatile skill-set specializes in specific domains such as computing, vision and information retrieval. Data Scientists from text retrieval and natural language processing analyze the relevant data and interpret results.

Data Science provides fruitful information from various emerging areas of computer science, such as:

- Cloud Computing
- Databases and Information Integration
- Web Information Access and Information Retrieval
- Knowledge Representation

3. BIG DATA

How a company knew about your interest and sent some information via sms. How bank knew about loss of their debt in specific area. All such cases came to know via Big Data Analysis.

Did you know that through your mobile data or computer data or Aadhar data, how one can reach to some conclusion. New accounts that have been opened in Jan Dhan Yojna does not have too much money in savings. How a company came to know that through his brand customer was satisfied or not. All of the above cases work with Big Data analytics approach.

When it was said big data analysis, then it means that it directly relates to the customer. Most of the companies prefer online selling. To find the right status, Big Data Analysis helps a lot. Say for example someone claims about particular hits but it was not confirmed till that online product was not purchased. Because according to law, until and unless online product was purchased, then only it was referred as customer. This word indicates towards incomplete or unstructured data. To conclude information from structured data, Big Data Analysis plays its role.

To understand Big Data analysis, consider case of saree manufacturer from Surat. He had recently launched his product website. His objective is to increase the sales of sarees through online platform. An arrangement was made in such a way that his website highlights at the top in various search engines (such as Google). It was accomplished through search engine optimize approach. He was very excited because every day 5000 people were visiting his website in the initial days but after a gap of 1 month his excitement goes down because among 5000 visit every day only 20% people are doing e-shopping. He was thinking what to do to improve the sales. His friend suggests him about the Big Data Analysis firm. He was confused that what kind of activity Bid Data Analysis will do. His consultant explained him that Big Data analysis will find the interest of those people who click on saree website. This analysis helps in making the online business strategy for future [8].

3.1 The meaning of Big Data

Big Data is the study of numbers which helps in reaching from analysis to conclusion. It can be in form of Petabytes (1024 Terabytes) or Exabytes (1024 Petabytes). It consists of lakhs or crores of information about the people. For example: customer contacts, social media, mobile data, and web. The numbers are although in the form of structured data but somehow incomplete and it becomes difficult to reach at the conclusion. Big Data Analytics with the help of semi-structured or unstructured helps companies and institutes in reaching some positive end. It helps in knowing new information about business activities and also came to know about some trends and some other fruitful information which was somehow not possible with numbers [8].

3.2 The beginning of Big Data

IBM in September 1956 had presented random access memory accounting machine for data processing. It was world's first disk storage product. It consists of two types. One was 305 whose capacity is 5 megabytes and weight 1 ton whereas second was 650. Telecommunication company used this machine in knowing information that whether customer was satisfied or not through the analysis.

In India, NASSCOM with partnership of Blue Ocean Market Intelligence had prepared a report stating that the market of Big Data was at present 62 arab indian rupees. In 2017, the market of Big Data will be more than double. Lots of ideas can be generated through Big Data. With support of Big Data, changes in management can be done. Big Data analysis provides some new information with any type of data and with updated information, company can make decision for its future goal or strategies. It is important for consumer point of view as it helps in increasing the number of customers [8].

3.3 Usage of Big Data

Big Data helps in making decision on business in progressive manner. In the financial sector, consumer product manufacturers and telecommunication companies are taking help to know what exactly customer wants. The people working in banks and financial market came to know about the fact that what number of loans are sanctioned and how the amount of loan will be recovered. Besides this what amount of debt in loss and what are the reasons. In the phase of supply chain and new brand, companies are taking help of Big Data analysis to know the interest of customers.

Similarly Big Data analysis helps in data processing for a specific subject related text, audio, video and new types of data in making right decision. In world, the number of consumer is in arab and they make use of different types of products. Big Data analysis helps in knowing from which product, the customer was satisfied or unsatisfied [8].

3.4 Big Data companies working in India

The Big Data analysis companies working in India are Heckyl, Sigmoid Analytics, Flutura, Indix, Fractal Analytics, Crayon Data, Germin8, Aureus Analyitics, Dataswft, C360, Metaome, Frrole, Bridgei2i, Formcept and PromptCloud.

4. CONCLUSIONS

Every day to convert Twitt's terabytes into product sentiment analysis, to filter out lakhs of call detail records in real time, to control crores of live video feeds through surveillance etc. are the activities that Data Scientist do.

A major goal of Data Science is to make it easier for others to find the relevant data with an ease. The impact of Data science technologies focus on how we access data and conduct research across different areas, such as social sciences, biological sciences, medical informatics and the humanities. Macincaze research also highlight about the fact that in an upcoming years there will be demand of One Lakh Data Scientist in India. The purpose of Data Science Model presented in this paper is to answer the present queries and through the respective answers follow a new direction for future research in the field of Big Data.

5. REFERENCES

[1] Loukides, Mike, "What is data science?", Available at: http://radar.oreilly.com/2010/06/what-is-data-science.html.

[2] IBM Expert (2014) About data scientists, Available at: http:// www-01.ibm.com/software/data / infosphere /data-scientist/ (Accessed: 12th May2015).

[3] Available at: Dainik Bhaskar Haryana Hindi Edition article dated 08jun2012.

[4] A.B.Karthick, Anand Babu, D.Maghesh Kumar and G.RajaRaja Cholan, "Visualization for STEM Subjects", Cover Story Available at: CSI Communications, Nov-2014.

[5] Available at: http://www.datascientists.net/what-is-data-

science

[6] The Journal of Data Science, Contents of Volume 1, Issue 1, Available at: http://www.jds-online.com/v1-1.

[7] Available at: http://www.stat.purdue.edu/~wsc/

[8] Available at: Dainik Bhaskar Haryana Hindi Edition article dated 03Apr2015.

[9] Plagiarism Checker | Viper Plagiarism Scanner, Available at: http://www.scanmyessay.com/

[10] Dr. Prithwis Mukerjee, "Learning Data Science – A Do-It-yourself Approach", Available at: CSI Communications, July-2014.

[11] Visualization for BigData, Available at: http:// www.datameer.com/product/datavisualization.html

[12] Data Visualization: Making Big Data Approachable and Valuable, Whitepaper, Source: IDG Research Services, August 2012.

[13] Taking the First Steps Toward Data Quality by Elizabeth Dial, Technical Solution Architect, IBM Corporation, Available at: http://www.ibm.com/developerworks/data/library/dmmag/DMMag_2010_Issue2/FeatureDataQuality/index.html