Big Data Issue & Challenge: Lack of Skill, Learn Big data through Open Courseware

Chetan Khatri
Department of Computer Science
KSKV Kachchh University, Gujarat- INDIA
chetan@kutchuni.edu.in

***Abstract—* This paper attempts to offer broad definition of Big data and it's various characteristics, highlights and differentiates Data Science and Big Data. The growth of various data formats from big giant companies and governance. This paper's primary focus is to defines Lack of Big data Skills as issue and challenge, a particular focus of this paper is to give current result of various lacking of skills in Data Science and Big data, as solution this paper offers Online Professional College degrees and MOOC(Massive Online Open Course) courses, A particular distinguishing feature of this paper is to define various MOOC Courses and resources that may help people to become Big Data expert and Data Scientist to overcome the issue of Big data, that is Lacking of various Skills.**

***Keywords—* Big Data, Big Data Challenge: Skill, Lack of Skill in Big Data, MOOC Big Data.**

I. INTRODUCTION

Big data is the data which is expensive to extract, transform, load for decision making in an Enterprise, having challenges include capture, duration, storage, search, sharing, transfer, analysis and visualization.

"Big Data is any data that is expensive to manage and hard to extract value from" - *Michael Franklin (University of Barkley)*

"Big data, have entered into world beyond just static data that collected" – *TDWI*

"Big data is overwhelming of data, this data has challenges volume of data, unstructured way of data, confidentiality" – *Adobe at TDWI*

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."
–Dan Ariely

"I'm a data janitor. That's the sexiest job of the 21st century. It's very flattering, but it's also a little baffling"
– Josh Wills, a senior director of data science at Cloudera

"Given enough data, everything is statistically significant" – *Douglas Merrill*

A. Characteristics of Big Data

Current Big data usually defined with following 5'v and C.

Volume: The amount / size of data available for processing. The characteristic notify whether data is big data or not. Before, such huge volume of data could not be stored because of storage cost and ultimately thrown away. Now this is not the case. Not even big enterprises but even small scale, mid-size businesses and even consumers can afford to store and analyze Big Data.

Variety: The different formats of data in enterprise like sensor data, flat file, xml file, documents , binary data, Relational database etc.

Velocity: The speed in which data has been generated and processed.

Veracity: The quality and accuracy of data, whether that contains missing or noisy values.

Variability: The inconstancy shown by data at times, thus hampering the process of being able to handle and manage the data effectively.

Complexity: To process the entire data is complex task as concern with Big data, it should be connected, it should be correlated for accurate decision making, complexity might face while pre-processing data.

II. DATA SCIENCE

Data Science is the field where data acquires, processes, manipulates. Data science is combination of Engineering, Mathematics, Data Warehouse, Data Mining, Database System, Machine Learning, Artificial Intelligence, Algorithm, Programming, and Statistics. The phenomenon in technology development significantly exposes the staggering growth of data, as much as growth of data goes much and much more, data science skill requires more to handle that growth and scale of new types of data from sensor, social media, website logs, click steaming etc.

In other words, data science can be broken down into four essential parts.

Mining Data: Collecting and formatting various types of data based on pattern mechanism.

Statistics: Gathered information must be analyzed.

Interpret: Representation or visualization in the form of Presentation, charts, graphs, reports.

Leverage: Studying Implication of the data, application of data, tools & technologies of data, Interaction and prediction of data.

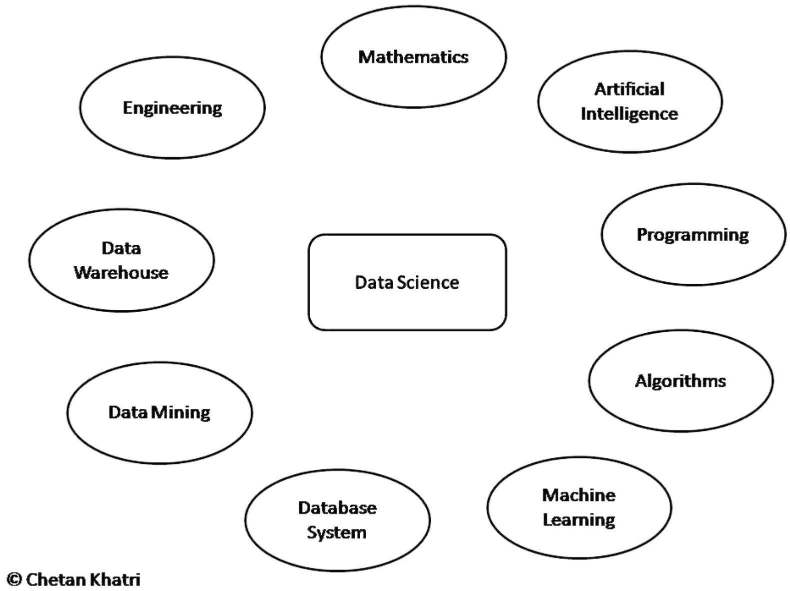

Figure 1. Define: Data Science

“I worry that the Data Scientist role is like mythical “Webmaster” of the 90s: master of all trades.”

-Aaron Kimball, CTO Wibidata.

What data science tells us: [1]

- If you are a *DBA*, you need to learn to deal with unstructured data.
- If you are a *Statistician*, you need to learn to deal with data that doesn’t fit in memory.
- If you are a *Software Engineer*, you need to learn Statistical Modelling and how to communicate with results.
- If you are a *Business Analyst*, you need to learn about Algorithms and tradeoffs at scale.

III. *DIFFERENTIATE BIG DATA AND DATA SCIENCE*

Data science is field where different areas such as Engineering, Mathematics, Data Warehouse, Data Mining, Database System, Machine Learning, Artificial Intelligence, Algorithm, Programming, and Statistics included. Where big data has modern applications and technologies to manage and process those data, Big Data includes datasets whose size and type make them impractical to process and analyse with traditional database technologies.

Data science is very impressive field in 21st century, the person who knows above skills is known as “Data Scientist”.

Data Scientist defined as “ A good Scientist understand importance of:

- Their eyes search for Information on the web.
- Algorithmic Strategizing.
- Verctorized operations.
- Have knowledge of latest tools and technologies to handle data.
- Efficient in data mining, statistics, mathematics, artificial intelligence.

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." - *HAL VARIAN, chief economist at Google.*

IV. **ISSUE: LACK OF BIG DATA SKILLS**

By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions - McKinsey Global Institute.

46% of organizations cite inadequate staffing or skills for big data analytics - TDWI Research.

More than three-fourths of 169 executives surveyed say staffing and training issues are the greatest obstacles to making the most of big data - Ventana Research.

While the majority of executives (58%) believe finding the right technology is the biggest challenge their companies face in analyzing data, the majority (56%) of IT decision-makers charged with implementing Big Data programs believe finding the right staff is a bigger challenge than finding the right technology – Avanade [2].

83% of data scientists surveyed felt that new technology would increase the demand for data scientists, and 64% believe that it will outpace the supply of available talent – EMC [3].

Today there is a shortage of trained Big Data technology experts, in addition to a shortage of analytics experts. This labor supply constraint will act as an inhibitor of adoption and use of Big Data technologies, and it will also encourage vendors to deliver Big Data technologies as cloud-based solutions – IDC [4] .

There will be a 24% increase in demand for professionals with management analysis skills over the next eight years. The need for this specialized talent is being fuelled by an increased use of business analytics by companies to better understand the explosion of data - U.S. Bureau of Labor Statistics [5] .

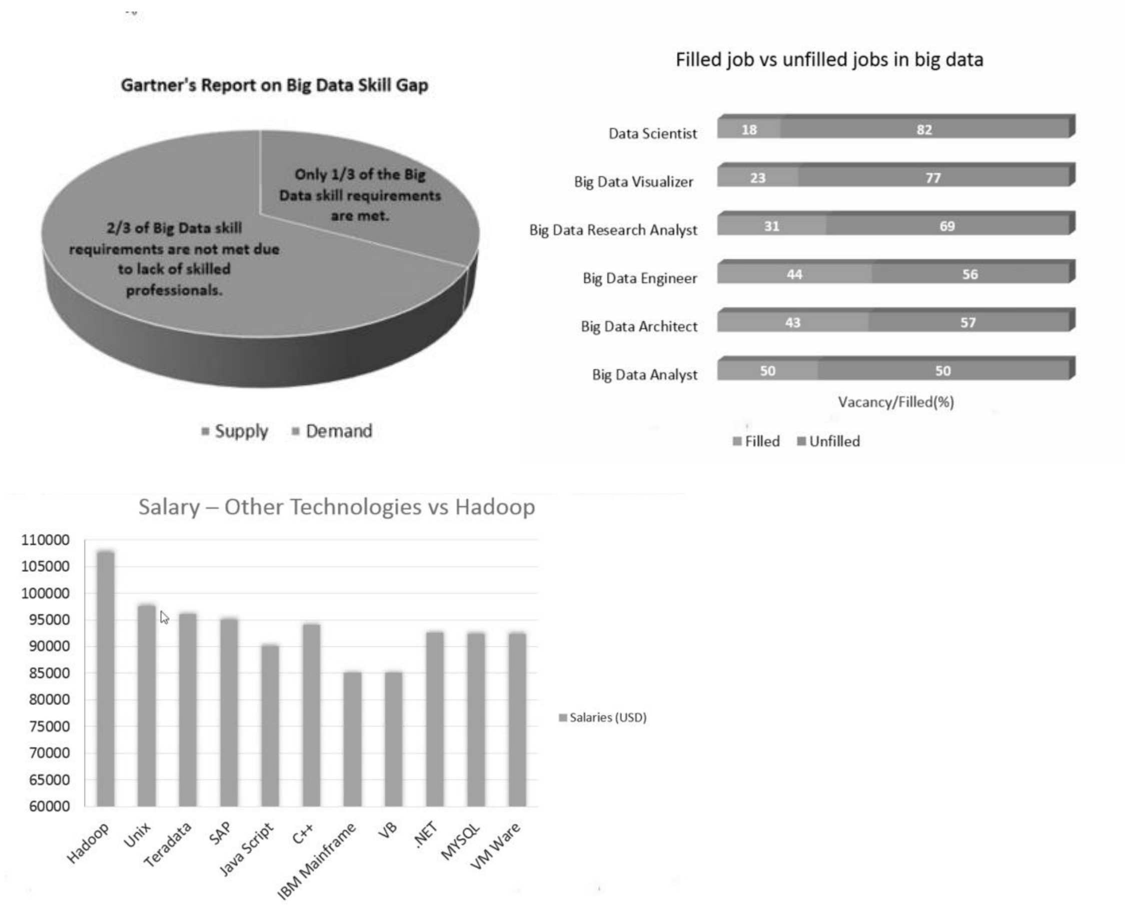

Figure 2. Increasing Demand and Lack of Big Data Skills.
Source: http://www.edureka.in/blog/increasing-demand-for-hadoop-and-nosql-skills/

V. **SOLUTION**

A. Online Big data degrees

Big data is now buzzword and rapid growth of data science already generated big demand of data scientist that diverts people to pursue data science degree.

Recently, UC Barkley started to offer Online Data science degree with $ 60,000 [6] , In the past few years, as data science has become the “sexiest” Job of the century, other top universities, like North-western and New York University, have moved into this area.

MIT also offers online Big data degree (“Tackling the Challenges of Big Data”) with $ 545 [7].

All above online degree cost a lot to people, so people not having really good financial condition cannot afford the high amount of cost to pursue big data course, and on other side Big data and Data science is sexiest job of 21st century [8] so everybody wants to become big data expert and data scientist, Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big data by 2015 [9].

So this stage must confuse people a lot that which is best path to do? Solution is Open Couseware (MOOC), researchers and students believe that MOOC (Massive Open Online Course) is golden boon for Research literacy, Open Courseware that can help people to become big data expert, Data scientist. As above high cost Online degrees from Berkeley, MIT requires computer science knowledge, MOOC also requires some sense of Computer Science knowledge.

Who offers Open Courseware programs?

- Edx.org
- Coursera.org

- Udacity.com
- BigdataUniversity.com

etc.

Above 3 are most popular portals to accomplish Open Courseware program from well knowledgeable professors at top reputed University of world such as Harvard University, Stanford University, MIT University, IIT etc.

B. Most valuable Big Data / Data Science Open Courseware's

From Udacity.com

1. Intro to Hadoop and MapReduce , How to Process Big Data by Cloudera [10].
2. Machine Learning: Supervised Learning Conversations on Analyzing Data [11].
3. Machine Learning: Unsupervised Learning Conversations on Analyzing Data [12].
4. Machine Learning: Reinforcement Learning Conversations on Analyzing Data [13].
5. Intro to Data Science Learn What It Takes to Become a Data Scientist [14].

From Coursera.org

1. Introduction to Data Science by University of Washington [15].
2. The Caltech-JPL Summer School on Big Data Analytics [16].
3. Big Data Science with the BD2K-LINCS Data Coordination and Integration Center [17].
4. Web Intelligence and Big Data by IIITD and IITD [18].
5. Statistics One by Princeton University, [19].
6. Algorithms: Design and Analysis, Part 1 by Stanford University [20].
7. Machine Learning by Stanford University [21].
8. Probabilistic Graphical Models by Stanford University [22].

From Edx.org

1. Introduction to Big Data with Apache Spark by University of Berkeley [23].
2. Introduction to Linear Models and Matrix Algebra by Harvard University [24].
3. Introduction to Probability - The Science of Uncertainty by University of MIT [25].
4. Introduction to Metrics for Smart Cities by University of SCMT [26].
5. Applications of Linear Algebra Part 1 by Davidson University [27].
6. Advanced Statistics for the Life Sciences by Harvard University [28].
7. Statistics and R for the Life Sciences by Harvard University [29].
8. Introduction to Computational Thinking and Data Science by University of MIT [30].
9. Scalable Machine Learning by University of Berkeley [31].
10. Wiretaps to Big Data: Privacy and Surveillance in the Age of Interconnection by Cornell University [32].

C. Other

1. Data Analysis Learning Path by MySlideRule [33].
2. Learn Data Science by LearnDS [34].
3. Learn R tool by datacamp [35].
4. Learn R tool by Data Science Central [36].
5. Learn R tool by Cyclismo [37].
6. Learn R tool by Code School [38].
7. Learn R tool by SwirlStats [39].

D. Books

1. *Data Integration for Dummies* a wiley brand by Brian Underdahl, Informatica.
2. *The data analytics handbook* researchers + academics by Brian Liou.
3. *The data analytics handbook* ceo's + managers by Brian Liou.
4. *The data analytics handbook* data analysts + data scientists by Brian Liou.
5. *Big data Imperatives* Apress by Soumendra Mohanty
6. *Mahout in action* for data mining with MapReduce
7. *Big Data Imperatives: Enterprise Big data warehouse, BI Implementations and analytics* by Soumendra mohanty, Madhu Jagadeesh, harsha srivatsa.

8. *Big Data: Challenges and opportunities* by Infosys Labs Briefings.
9. *Hadoop, The definitive guide* by Tom White, O'reilly.
10. *Hadoop in Action* by Chuck Lam, Manning.
11. *Planning for Big Data*, A CIO's Handbook to the changing data landscape, O'reilly radar team.

E. Youtube Channels

1. *TDWI* - TDWI is your source for in-depth education and research on all things data.
2. *EuroPython* - EuroPython isn't exactly a conference; it's a chance to hang out with friends that you haven't even met yet.
3. *EMC Academic Alliance* - EMC Academic Alliance-Technology Curriculum
4. *edureka!* - Edureka provides online training courses for BigData and Hadoop, Hadoop Admin, Cassandra, Data Science, Cloud Computing, Android Development.
5. *CS50* - CS50 is a free online class introducing students to the basics of computer science. CS50 is taught by David Malan of Harvard University.
6. *Cloudera, Inc.* - Cloudera Inc. is an American-based software company that provides Apache Hadoop -based software, support and services, and training to business customers.
7. *bipublisher* - Oracle BI Publisher is the reporting solution to author, manage, and deliver all your reports and documents easier and faster than traditional reporting tools.
8. *Hortonworks* - Hortonworks develops, distributes and supports a 100% open source distribution of Apache Hadoop for the enterprise, also training, support & services.
9. *Tech Gig* - India's Most Passionate Technology Community, Learn & stay updated on your skills, compete with fellow techies and showcase your expertise to the community.

VI. **CONCLUSIONS**

In this paper, I have proposed a various useful MOOC courses and resources of different areas of Data Science and Big Data such as Engineering, Data Warehouse, Data Mining, Database System, Algorithms, Artificial Intelligence, Programming, Mathematics, Statistics, and Machine Learning etc. I believe that entire structure can surely help to solve Lack of Skill: Big data challenge. In entire world, people can't afford high amount of money to take professional Collage Degree and other issues such as place, admission etc can also affect people to complete degrees, at this time my proposed Courses and Resources can help people to learn cutting edge tools and technologies of Big Data and Data Science, that can solve many problems including Lack of Big Data Skill. Today, that is real challenge to industry and governance.

VII. **REFERENCES**

[1] Bill hove, University of Washington, [Online]. Available. https://www.coursera.org/course/datasci

[2] [Online]. Available. Ease Big Data Hiring Pain with Cascading | CIO www.cio.com/article/.../ease-big-data-hiring-pain-with-cascading.html

[3] Intent to Plan: M.S. in Analytics and M.S. in Data Science Dakota State University and South Dakota State University. [Online]. Available. www.sdbor.edu/theboard/agenda/2013/December/CommA/III_A.pdf

[4] IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy.
[Online]. Available. http://www.businesswire.com/news/home/20120307005036/en/IDC-Releases-Worldwide-Big-Data-Technology-Services

[5] Analytics Goes to the Head of the Class: Northwestern University News
http://www.northwestern.edu/newscenter/stories/2011/12/ibm-analytics-masters.html

[6] [Online]. Available. http://datascience.berkeley.edu/

[7] [Online].Available. https://mitprofessionalx.edx.org/courses/MITProfessionalX/6.BDX/2T2014/about#overview

[8] [Online]. Available.https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[9] [Online]. Available. http://www.gartner.com/newsroom/id/2207915

[10] [Online]. Available. https://www.udacity.com/course/ud617

[11] [Online]. Available. [11] https://www.udacity.com/course/ud675

[12] [Online]. Available. [12] https://www.udacity.com/course/ud741

[13] [Online]. Available. [13] https://www.udacity.com/course/ud820

[14] [Online]. Available. [14] https://www.udacity.com/course/ud359

[15] [Online]. Available. [15] https://www.coursera.org/course/datasci

[16] [Online]. Available. [16] https://www.edx.org/course/introduction-big-data-apache-spark-uc-berkeleyx-cs100-1x#.VNYdJEeUcrI

[17] [Online]. Available. https://www.coursera.org/course/bigdataschool

[18] [Online]. Available. https://www.edx.org/course/introduction-linear-models-matrix-harvardx-ph525-2x#.VNWKuUeUdOI

[19] [Online]. Available. https://www.coursera.org/course/bd2klincs

[20] [Online]. Available. https://www.coursera.org/course/bigdata

[21] [Online]. Available. https://www.coursera.org/course/stats1

[22] [Online]. Available. https://www.coursera.org/course/algo

[23] [Online]. Available. https://www.coursera.org/course/pgm

[24] [Online]. Available. https://www.edx.org/course/introduction-probability-science-mitx-6-041x-0#.VNWK30eUdOI

[25] [Online]. Available. https://www.edx.org/course/introduction-metrics-smart-cities-ieeex-scmtx-1x#.VNWLCEeUdOI

[26] [Online]. Available. https://www.edx.org/course/applications-linear-algebra-part-1-davidsonx-d003x-1#.VNWLXUeUdOI

[27] [Online]. Available. https://www.edx.org/course/advanced-statistics-life-sciences-harvardx-ph525-3x#.VNWLh0eUdOI

[28] [Online]. Available. https://www.edx.org/course/statistics-r-life-sciences-harvardx-ph525-1x#.VNWLiUeUdOI

[29] [Online]. Available. https://www.edx.org/course/introduction-computational-thinking-data-mitx-6-00-2x-0#.VNWMF0eUdOI

[30] [Online]. Available. https://www.edx.org/course/scalable-machine-learning-uc-berkeleyx-cs190-1x

[31] [Online]. Available. https://www.edx.org/course/wiretaps-big-data-privacy-surveillance-cornellx-engri1280x#.VNWNcEeUcrI

[32] Increasing demand for 'Hadoop and NOSQL Skills'
[Online]. Available. http://www.edureka.in/blog/increasing-demand-for-hadoop-and-nosql-skills/

[33] [Online]. Available. https://www.mysliderule.com/learning-paths/data-analysis/

[34] [Online]. Available. http://learnds.com/

[35] [Online]. Available. https://www.datacamp.com/

[36] [Online]. Available. http://www.datasciencecentral.com/profiles/blogs/r-tutorial-for-beginners-a-quick-start-up-kit

[37] [Online]. Available. http://www.cyclismo.org/tutorial/R/

[38] [Online]. Available. http://tryr.codeschool.com/

[39] [Online]. Available. http://swirlstats.com/