BIG DATA- CHALLENGES AND ISSUES

# Vanshika Mudgal

*Institute of Innovation in Technology and Management, India*

vanshikasharma12312@gmail.com

***Abstract***

***The new trend to partake in 21st century would be making data driven and informed decisions. The buzz around Big Data is growing and it includes information created by machines and people, explore, pre-process, analyse and communicate results by making data-informed decisions and Walmart, US shipments making millions. Big data is at the infancy stage where all domains need to be reviewed, acquired and taken to a full adult growth life cycle and one of values we need to adopt in recent times. However, it comes with its own gridlock of issues and grand challenges mainly security, privacy, scalability, integration, inaccuracy. We then move to its solutions and future direction. In this paper we will debunk the myths around Big Data.***

***Keywords:***

***Big Data; Veracity; Valence; ETL Process; Data Profiling; Data Integration; Data Mining***

# INTRODUCTION

# Big Data is something huge and way too complex to be done by traditional systems. Big Data refers to large datasets which have so much hidden information that needs to be dealt to make the sense and value of the data from fast-moving and wide-ranging datasets. Big Data is used to analyze the information, realize hidden-difficult patterns, relationships, associations and make insightful decisions in surrounding text of decision- making. Due to the enormous data generation, multiple technologies and software have been created in view of handling additional storage capacity and real-time analysis. The term was introduced by John R. Mashey in early 1900’s. In its true character the analysis of data was done in past as well from Egyptians in 200B.C to keep the library of Alexandria or Romans trying to statistically analyse how many armies to deploy. The Big Data is used to divided in 3 phases where 1st phase wholesomely covers database management without understanding the core concept of Big Data. In early 2000’s the data was generated and big companies like Yahoo, Google and Facebook started analysing customer behaviour, clicks, search logs, etc. The understanding became clearer with help of social media, insights were gained from

# unstructured data. The third phase is burst of data generating every second and becoming unstructured. Hence, many technologies are being used to extract the value like pin in a haystack.

# To realise big data’s potential, the data should be gathered in a new way which enables it to be utilised for different purposes many times; this can be seen today in the many devices connected to the internet and the huge amount of data accesses even by individuals. By 2020, the predicted value of data is posited to double every 24 months.

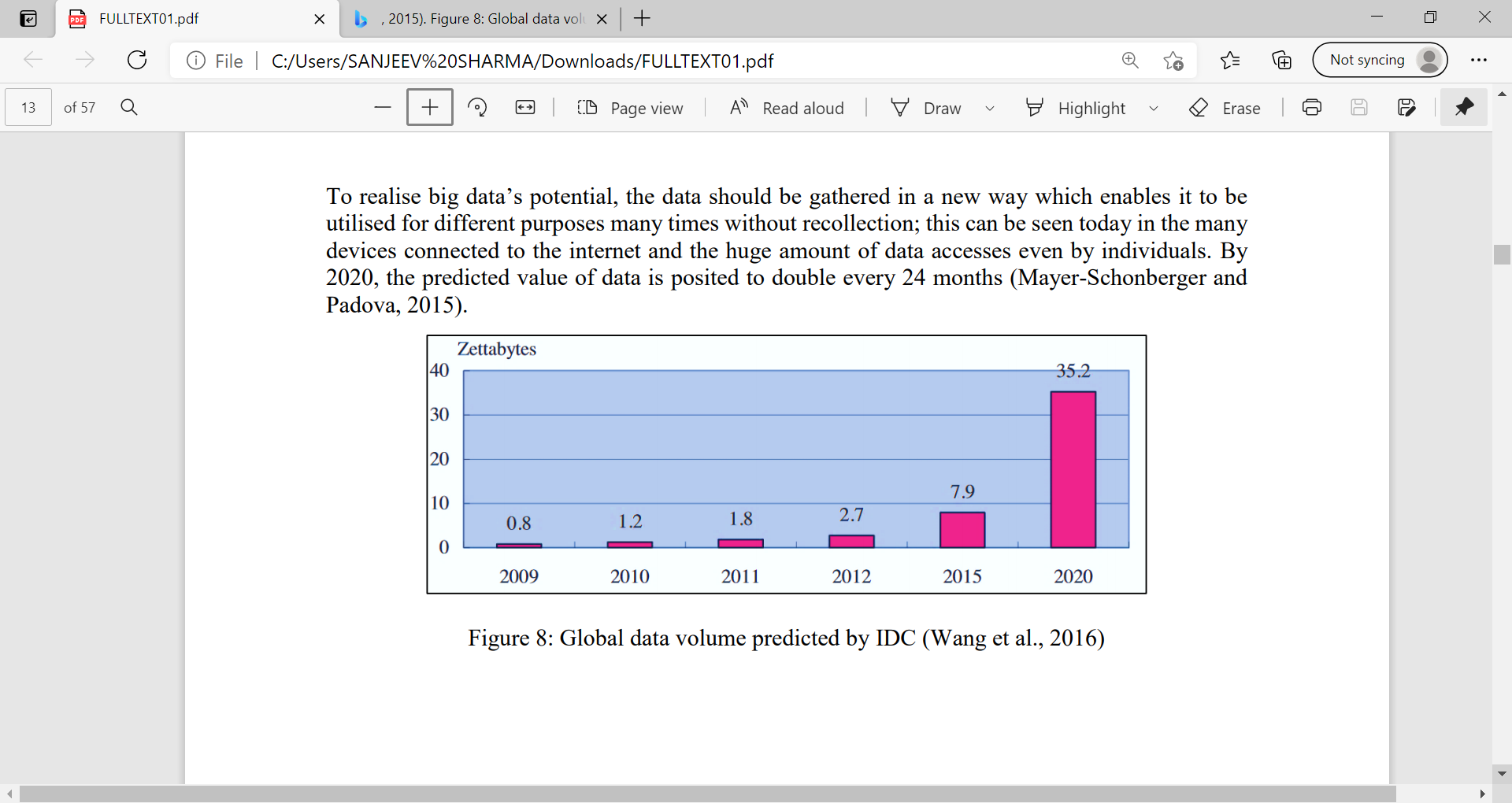


Fig 1: Global data volume predicted by IDC (Wang et al., 2016) [1]

# CHARACTERISITICS OF BIG DATA

Big Data is the tree with many characteristics where roots represent the size, flowers representing the different kinds of data, the way it grows representing speed of data fabricated. The following v’s are:

**Volume:** is the first dimension coupled with downright volume of Big Data. Data can be associated with large datasets to small inventory checkups in local stores. Every minute 204 million emails are sent, 200,000 photos are uploaded, and 1.8 million likes are generated on Facebook. On YouTube, 1.3 million videos are viewed and 72 hours of video are uploaded. The data is massive and must be handled earnestly. The data mass- produced will be in yottabytes and data is on an astronomical scale and said data is compared to Milky way Galaxy.

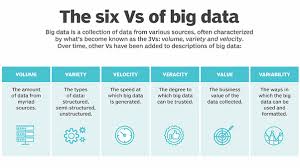
**Variety:** is the second v of data and refers to heterogeneity of data. In pre-dominantly we used to refer tables as major source of data or spreadsheets,

but now we have biodiversity of data like geographical maps, text, audio, gifs, tweets, Instagram posts, etc (Refer table 1). This is needed for real time analysis and various disparate methods have been used to analyse and make decisions regarding them. The data is amalgam of various variations like e-mails. The variety is one of the important factors for analysis and processing the reports.

|  |  |
| --- | --- |
| Old Data | Big Data |
|  |  |
| TABLES | TABLES |
| SPREADSHEETS | DATABASES |
| STOCK RECORDS | IMAGES, 3D MODELS, AUDIO |
| FINANCES | TEXT, LOCATIONS |
| PERSONNEL FILES | TWEETS, SATELLITES |

Table 1: Understanding the data

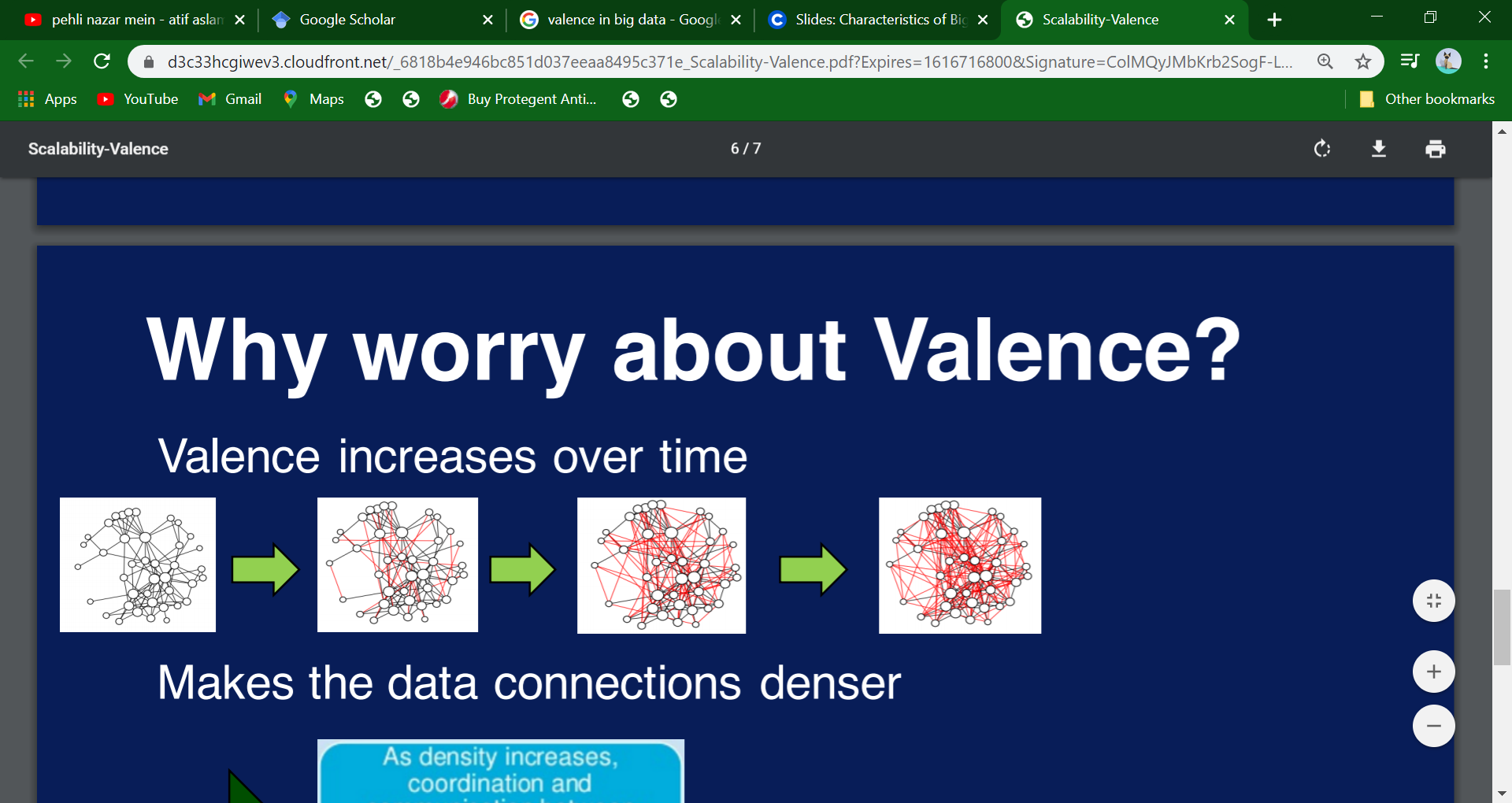
**Velocity:** is the third v of data and refers to the data incoming speed and needs to be stored and analysed. Imagine going to a new destination, you would be needing information for places to visit, weather, how to pack and what relevant must be done. So, the information must be pertinent where you want to visit your favourite café but would be seeking their schedule of 2 years back or recently updated. Thus, it makes sense to monitor the data and act on that. Batch processing was typical years earlier where data was fed into machines and processed for days.Hence, these days organizations pursue latest data are successful in market.

 Fig 2: Attributes of Big Data [5]

**Veracity:** is the fourth v of big data and refers to questioning the credibility of data acquired and finding noise, anomalies and biases. The data quality must be top notch to evaluate the decisions otherwise it’s no good. The data quality is measured upon factors like reliability, accuracy, credibility, etc. Now-a-days most data is unstructured and have no time for its quality check due to mass produce. There are fake reviews present on sites like Amazon, Myntra where product is unacceptable and yet given 5 star or great reviews. Google Flu Trends was used to analyse outbreak of influenza in 25 countries but were failed because they ignored the quality of data and did not account for uncertainties. This is akin to an art artifact having providence of everything it has gone through. Thus, the quality of data must be

checked.

**Valence:** is the fifth dimension of big data where it upholds the concept of chemistry. It means the connection in data. The higher the valency, higher the connection and leads to organizational behaviour [Fig 3]. Data can be connected directly and indirectly like two friends play the same sport and two student study at different schools but in same class respectively. The highlight is the interconnectivity increases with due time and finds possible connection in the datasets. When we go on a trip and you meet some people and after some time you plan another trip due to same likes. The data becomes denser, dynamic behaviour is present. The difficult analytical methods must be adopted and particular evens must be noted.



**ORGANISATIONAL BEHAVIOUR**

**(the denser communities grow, groups are devised, relationships are built and coalitions are formed.)**

Fig 3: Valence increases over time

**Value:** is the sixth and most valuable dimension of Big Data. The idea behind all this is to bring value to the problem at hand. The exploring,

acquiring, processing and analyzing is just done to get this from the data. The gainful insights lead to guaranteed success for various companies like Walmart, Google, Amazon, etc. The Big Data is all about making right decisions with right data. The value lies in the eyes of stakeholder and beholder and not concentrated on volume dimension or others. The value of data for big companies must be calculated through cost- risk analysis and benefits reaped and ROI must be taken in context. The enterprises can extract value from maintaining transparency, analyzing market patterns, customer behaviour, trends, etc.

These are the v’s of big data and what it encompasses.

1. **CHALLENGES AND ISSUES IN BIG DATA**

Big Data previously mentioned is a bit episode of happy moments in the series but it also contains a sad side. The flip side to big data massive prospect it comes it with challenges. According to a report from Dun and Bradstreet the enterprises just face half of problems in big data with accurate data, data quality and its analysis. The unspoken challenges are related with privacy concerns, security, integration of different types of data. (Fig 4) The current culture is not well equipped and needs to be made more dynamic.



Fig 4: Big Data Challenges [14]

The challenges are intertwined and difficult to consider in isolation, but according to King and Powell (2008), they can be split into three categories: (a) getting the picture right (i.e., summarizing the data), (b) interpreting or making sense of the data through inferences, and (c) defining and detecting anomalies. (Fernando, Catlin, 2013);

**3.1 SCALABILITY**

The extents of data mass produced these days brings one of the prospective challenges of Big data. Most of data is unstructured, transpired from documents to tweets, other sources. The stern requirements are required for storing and managing the data. The Relational Database Management System (RDBMS) makes up for traditional database but today’s data needs extensive databases or solutions to achieve their role of being reliable, available and quality enriched. Quick access, proper analysis and storage ampleness must go hand in hand to evade this challenge.

To cope up with these modern problems we require modern solutions. Techniques and massive storage device are required to be adopted. The compression is one of the oldest techniques and helps in reducing the amount of storage and bandwidth required for data sets. It also removes irrelevant or redundant data, making analysis and processing easier and faster. Second technique would be Tiering allows

the motion of data between different storage tiers, which allows an organization to ensure that the adequate data resides on the adequate storage technology. Data tiering varies in data size. Third technique would be Deduplication referring to redundant and unwanted data and helps in achieving reliable data quality. The recent storage devices Direct Attached Storage (DAS) and network storage, while network storage can be further classified into Network Attached Storage (NAS) and Storage Area Network (SAN) (MH., 2014). DAS is per se for personal use and small sized and can decrease efficiency with increased storage. The NAS is equipment based on hub of servers and more large-scale. The scalability can affect efficiency and access time and used for backups as well. These techniques can help in acknowledging and providing solutions to prospective storage challenges.

**3.2 PRIVACY AND SECURITY**

In the world where privacy is now considered as a fundamental right for every person on Earth is interesting how we ignore the terms and conditions before enjoying content of any web pages and web sites. The companies are trying to protect their data and business tactics so their contenders not get a whiff of their business plans. It is one of the budding challenges in Big Data and number one challenge which needs to be overcame. The degree of sensitivity of data depends on from accessing name, addresses to bank account numbers. Security Breaches are now common and used for wrongful purposes. The intermediary nodes as well as the existence of massive number of channels increment the chance of hackers taking benefit of susceptibility. As a consequence, organization must follow pre-eminent security practices for security and privacy perspective.

The recent security practices need to be upscaled and enriched with more security locks. The encryption, real time monitoring, identity and access control are some modern-day practices where it applies even to social media platforms messaging feature. The encryption plays an important role where the data gets encrypted into incomprehensible code which is hard for hackers to hack and gets decrypted and sent to the receiver. The authentication access given to selected set of people and login password or double step verification. Cloud Computing is one of the efficient solutions for storage and privacy purposes.

The cloud computing for private purposes is more respected but costs more and more companies due to this reason go for public services which can easily lead to leak of data with sensitive information as well. A good company will always protect its data and enforce good security practices. (Table 2).

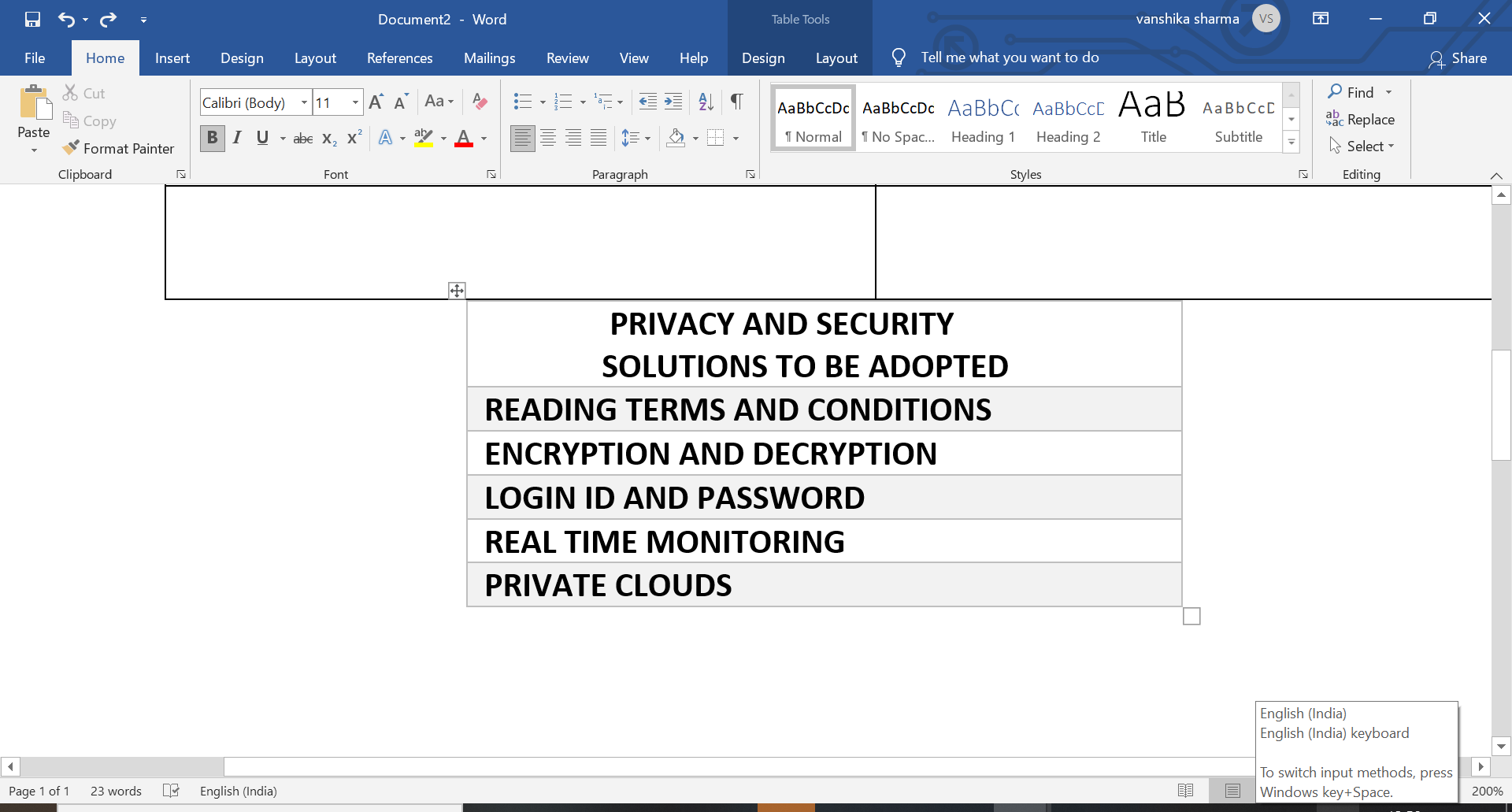


Table 2: Techniques to adopt for privacy

**3.3** **INTEGRATION OF DATA**

The daunting era of big data where data is generated by machines, by people, by databases, its integration is becoming one of the expanding challenges. The data is very heterogeneous and in typical ways need huge human intervention and attention. The data required by companies, research works, to survey need to work on different forms of data to arrive on a decision which makes it necessary to integrate them. The disparate sources and its integration go hand in hand for analytical processing. The traditional integration is done through ETL. (Fig:5) The Extract from database, Transform the extracted database into target database and Load means write into target database. It has been used with batch processing and have failed to keep up with the modern data speed.

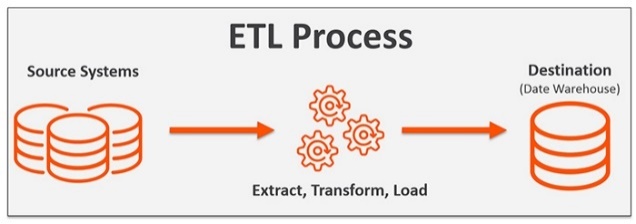


Fig 5: ETL PROCESS [21]

The elements of Hadoop software can help to integrate data in modern ways and superior to old ones. The traditional ETL process is also evolving for modern data and each software has its own way of ETL. The new techniques for big data integration are Schema Mapping, Data Fusion, Record Linkage. Schema Mapping can be considered process of two phases where global schema is created and mappings are created between global and local schema and hold attributes which have same information. Second is Record Linkage which refers to identify records that is in same logical entity across multiple data sources. In traditional ways it only worked on linking of structured data but recently the norm is to work on unstructured data and draw insights. Third is Data Fusion refers to combination of unstructured data from large datasets to structured operational database. It is a new way of integration and highly motivated by veracity. This makes us confident that results will be meaningful.

**3.4 DATA QUALITY**

High- quality data is preconditioned for big data analysis and guaranteeing the value of data. The data currently is lacking comprehensible quality standards. The data quality needs to be minded in large datasets and is essence of big data analysis. The challenge is one of the main ones which needs to be dealt in early stages and main goal must be to find the noise from data and remove it. Incompetent data quality can lead to incompetent analysis. The problems with big companies can lead to legal and judicial complications if not minded and can provide inaccurate data if its look for research purposes. The data must be conformed to users implied needs. Some organizations store data as whole, without implementing much to it which creates data pollution. Furthermore, inappropriate data can lead to risk of decision making on erroneous information. The data quality is multipart concept and wholesomely represents the real-world construct and perfect intend to use.

To cope up with this everlasting challenge we must zero on in data quality management techniques. This situation can lead to data silos which refers to data held by one particular group and is completely inaccessible and can lead to wrongful analysis within an enterprise. The techniques include Data Profiling, Data Governance, Data Matching, Data Quality Reporting, and many more are called Key Performance Indicators. (KPI’s) Data Governance is laying out data policies and standards that sets the bar and data elements which needs to be addressed. Data Profiling refers to data cleansing professionals who are appointed must have a deep understanding for data. It is often supported by businesses on many structural levels. Data Matching refers to real world alignment for data quality and is more than exact databases copying values and preventing human errors. The machine learning here plays an important role in deduplicating a database where it tells two or more records than that is the same person. Data quality reporting depends on data profiling inputs and used as KPI’s and used for measuring data uniqueness. The big data quality dimensions and metrics are Accuracy, Completeness, Timeliness, Orderliness, Uniqueness and many more. (Fig 6) The metrics are important for achieving the main purpose and achieving reliability at 100%. Today data is very digitalised and must be avoided where not needed and cross-departmental views must be holistically accepted by top level management and thus data quality can be achieved.

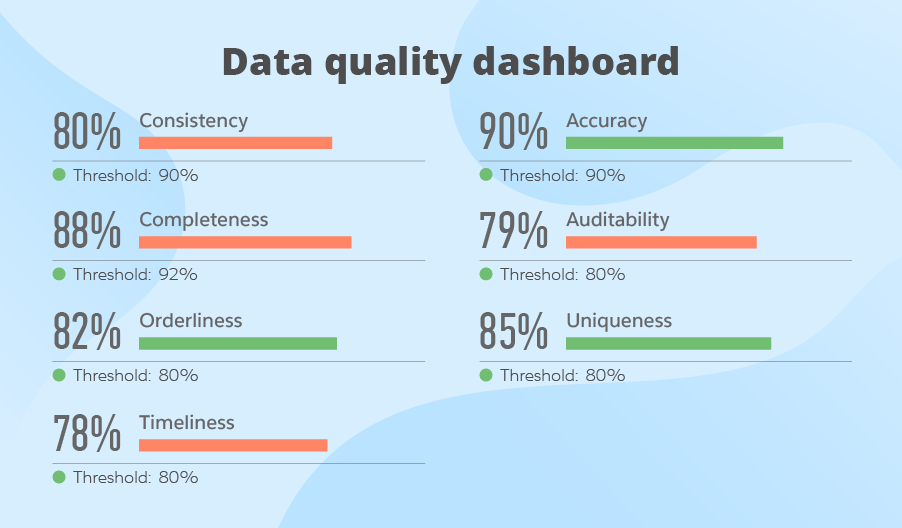


Fig 6: Data Quality Dimensions [14]

**3.5 LACK OF UNDERSTANDING OF BIG DATA**

Companies suffer from this challenge most of the time. The population is divided with being obsessed with big data era and another half being totally unenlightened. The big data is a necessity these days and must be made aware about it. Data professionals may have a clear picture what is going on, but employees may not know what actually the data is, it’s processing and sources. For instance, if employee do not recognize the significance of data storage, they might not retain the backup of sensitive data. As a result, when this data is required, it cannot be retrieved easily. The various online survey (Fig 7) shows how little IT professionals and suppliers understand this and hit the markets. The employees are not ready for the dynamic environment and laidback practices are still going on. The importance is not understandable to employees at lower and middle corporate levels. The solutions must be fast paced to enhance fast learning and a keen interest towards that topic.

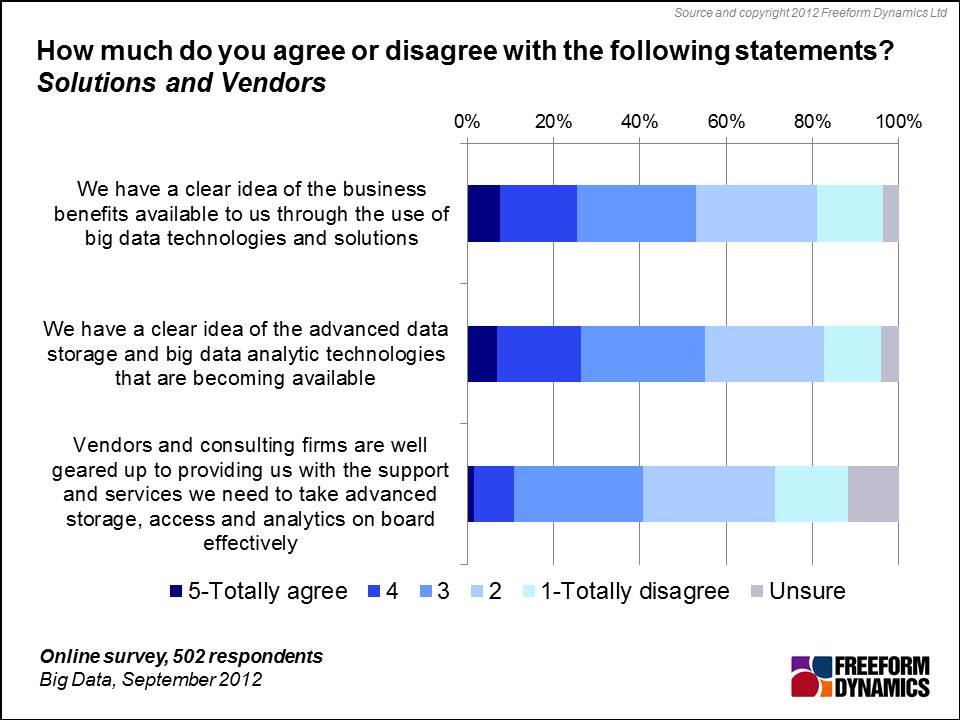


Fig 7: Online Surveys [18]

The companies and many influential organizations can help in evading this challenge. The surveys cannot be generalized for the whole population but makes us aware of one of the big data challenges.

The corporations can arrange workshops and seminars for employees to make them aware. Basic training must be given to all employees to make them understand about it. Advanced training must be given to interested and employees involved in Big Data projects. This must be treated as a habit to be inculcated in every corporate levels. The influential organizations can make dedicated social media about the topic and build a community and increase the interest of others. We also hope our research paper can help others to eradicate this challenge. These are simple basic solutions to this big challenge. (Table 3)

|  |
| --- |
| AWARENESS FOR BIG DATA |
| SEMINARS |
| WORKSHOPS |
| DEDICATED SOCIAL MEDIA |
| BUILDING COMMUNITIES |
| BASIC AND ADVANCED TRAINING |

Table 3: Techniques to make aware about big data

Hence, these are some of the challenges related to big data and proves its flip side. Any topic in this world cannot have only one side and thus a proper perspective and understanding is key existential to understand the crisis.

# DISCUSSION AND FUTURE DIRECTION

The most buzzed topic Big Data is heavily visible to all and discussed thoroughly in this paper. We started with the big data definitions, its characteristics, its applications and leading towards the flip side that is open challenges and issues faced and with modern solutions and how to tackle them and thus concluded by a discussion. The future stages of Big Data will be very advanced and bring new problems to the light. The data generating speed will quadruple in upcoming years. The data analysis will have no utility if data is not standardized. Big Data is a remarkable domain in future years with high prospects of growth and job empowerment and more data enabled decision making. The other hot topics like Cloud computing, Data Mining, Artificial Intelligence are intertangled with it and have lot to process together. The fast processing would be another future issue for its wide mass data. The data would have to be accessed fast, stored properly and retain backups, data quality practices are standardized, different types of data can be integrated in seconds. The potential powerful computers ad servers will be built while maintaining the aspect of money, privacy, reliability and others. The future research papers would be exploiting the new areas currently unknown in the Big Data era and helping mankind. The Big Data domain brandishes successfully and make major leaps in form of positive impact.

# CONCLUSION

The use of big data is healthy in many ways and for upcoming generations. The research paper covers all the topics and present a wholesome view on it. The big data applications in Indian arena have been underwhelmed and needs to use this domain fully. Big data is the Data era where it needs to be analyzed and provide successful business insights. In this research paper we enlightened issues and challenges of big data and solutions regarding it and the future scope related to this domain. The characteristics of Big Data highlights the concept around it and tells us the good data analysis is of use only when true value is found. This is one of the hot topics for research and scientific work. The people must be aware about it and hope this research paper helps them to educate about this topic.

# REFERENCES

[1] Al-Shiakli, Sarah, 2019, Big Data Analytics: A Literature Review Perspective, Pg-10

[2] A. Loyns Angela, E. Grable John E, 2018, An Introduction to Big Data, Pg-1

[3] Google, bigdataframework.com, 2019

[4] Google, courser.org/Intro to big data, 2018

[5] Google, quora.com, V’s of big data

[6] Shaman Jeffery, Kandula Sasikiran, Reappraising the utility of Google Flu Trends, 2019, Pg-1

[7] Google, bigdataframework.com, 2018

[8] Google, i-scoop.eu

[9] Google, acuvate.com, 2019

[10] Adrian Tole, Alexandru, 2013, Database Systems Journal, Pg-1

[11] Google, tiempodevelopment.com, 2020

[12] Fernando Almedia, Calistru Catilin, 2013, The main challenges and issues of big data management, Pg-16

[13]M.H Padgavankar, International Journal of Computer Science and Info Technologies, 2014, Pg-5.3

[14] Google Scholar, TWDI Predictive Analytics 2013, 2013

[15] Bertina Elisa, 2013, Big Data opportunities and Challenges, Pg-2

[16] Google, The Business Of IT Blog, 2017

[17] Google, towards data science, 2018

[18] Google, Freedom Dynamics Ltd, 2012

[19] Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI:

[20] Firmani, Donatella and Mecella, Massimo and Scannapieco, Monica and Batini, Carlo, 2016, On the meaningfulness of “big data quality”, Pg-6

[21] Google, <https://bi-survey.com/data-quality-master-data-management>

[22] Google <https://profisee.com/data-quality-what-why-how-who/>

[23]Google,[https://www.scnsoft.com/blog/guide-to-data-quality-management, 2018](https://www.scnsoft.com/blog/guide-to-data-quality-management,%202018)

[24]Google,<https://www.computerweekly.com/opinion/Lack-of-understanding-inhibits-big-data-adoption>

[25]Google, upgrad.com

[26]Google, dzone.com

[27]Google, cloudian.com

[28]Google, cazoomi.com

[29]Google, delttechnologies.com

[30]Google,techrepublic.com