

Big Data and Accuracy Challenges

Dr G. Rajitha Devi
Asst prof in Computer Science

ABSTRACT:

The term big data is defining a collection of large and composite data sets that are asperity to process using traditional data processing tools. Every day, we create billions of data all over the world. These data comes from social networking sites, scientific experiments, sensor networks, mobile conversations and various other sources. Big data divided into many dimensions: Volume, Velocity and Variety. To improve the aspect of this three dimension new dimension veracity was introduced. This paper furnish an analysis about the concept of veracity. Veracity can be inspect mainly based on three dimension objectivity, truthfulness, redibility. These dimensions are combined to form one composite index- the big data veracity index. This index is useful for evaluate efficient variations in big data quality. Companies, institutions etc., all of them use stack of data which are further used for creating reports in order to ensure cohesion regarding the services that they have to offer. The process behind the results that these entities requests represents a challenge for software developers and companies that provide IT infrastructure. The challenge is how to operate an magnificent volume of data that has to be securely delivered through the internet and reach its destination intact .This paper contributes to the big data research by dividing the existing tools to measure the suggested dimensions.

Keywords: Big Data, Sharing data, Veracity, social networking sites.

Date of Submission: 02-05-2022

Date of acceptance: 15-05-2022

I. INTRODUCTION

BIG DATA

The term “big data” is often used to report massive, complex, and real-time streaming data that require enlightened management, systematic and processing techniques to citation insights. While there is no accord on the definition and characteristics of big data, the term “big data” was initially coined to reflect the “bigness” or voluminous size of data generated as a result of using new forms of technology (e.g., social media, radio-frequency identification (RFID) tags, smart phones, and sensors). This definition was then expanded to include variety (i.e., structured or unstructured data formats) and velocity (i.e., the speed at which data are created). Over the years, others have further dimensionalized big data into veracity (i.e., messiness of data) and value (i.e., the previously unknown insights) .To have a better understanding of what Big Data means, the table below represents a comparison between traditional data and Big Data .

Table 1. Understanding Big Data *Traditional Data Big Data Documents*

Understanding Big Data

Traditional data	Big Data
Document	Photos
Finances	Audio and video
Stock records	3D Models
Personal files	Simulations
Location data	

This example provides information about the volume and the heterogeneity of Big Data. It is difficult to work with complex information on standard database systems or on personal computers. Usually it takes parallel software systems and infrastructure that can handle the process of sorting the amount of information that, for example, meteorologists need to analyze. The request for more complex information is getting higher every year. Streaming information in real-time is becoming a challenge that must be overcome by those companies that provides such services, in order to maintain their position on the market. By collecting data in a digital form, companies take their development to a new level. Analyzing digital data can speed the process of planning and also can reveal patterns that can be further used in order to improve strategies. Receiving

information in real-time about customer needs is useful for seeing market trends and forecasting. The expression “Big Data” also resides in the way that information is handled. For processing large quantities of data that is extremely complex and various there needs to be a set of tools that are able to navigate through it and sort it. The methods of sorting data differ from one type of data to another. Regarding Big Data, where the type of data is not singular, sorting is a multi-level process. Big Data can be used for predictive analytics, an element that many companies rely on when it comes to see where they are heading. For example, a telecommunication company can use data stored from length of call, average text messages sent, average bill amount to see which customers are likely to discard their services.

Data sharing, Data sharing settings

There are several settings for sharing data in your Analytics account. With the help of these settings we can customize how you share any Analytics data collection with Google (such as JavaScript tracking code, mobile SDKs, and data collected using the Measurement Protocol), so that you can optimize your data based on your own preferences. Free or Restrict These settings allow you to customize how you share data collected from websites, mobile apps, and other digital devices that only use Analytics, whether through your Analytics account or using your account Linked data, such as the number of properties and additional features set up, does not apply to your Analytics data as needed to maintain and protect the Analytics service, regardless of your data sharing settings May go All Google representatives and vendors with access to account data must agree to the terms and conditions of the Internal Access Policy Data access requires proper authentication, all access is over SSL and is logged for security review And when accessing Customer's data, representatives may only use a computer approved by Google. Change your data sharing settings

To use this feature, we must have the Editor role
(You'll need to customize your data sharing settings when you sign up for an Analytics account, but you can change the setting at any time by going back to the Admin section of the account. How to change data sharing settings)

1. Sign in to Google Analytics
- 2 Click Admin to go to the account you want to edit.
3. In the Account column, click Account Settings.
4. Change any setting and click on Save

For information about each data sharing setting and their benefits, review your account settings and additional information here so that you can enable these settings. Or decide to leave it off if all settings are turned off that your Analytics data will only be used to provide and maintain the Analytics service Google products and services .When you enable this setting, Google may access and analyze data to accurately understand users' online behavior and trends, as well as use this data to improve its products and services For example, this data can be used to improve the tools in the Google Ads system that you use to create, manage and analyze ad campaigns. Any data Google collects and uses through Google's products and services settings.

It is under the control of security and also fulfilling the objectives of GDPR.

Google is an independent controller of that data for Criteria

Collected and anonymised data may be used in the creation of publications and features. This way you'll be able to better understand what's happening in the world of your industry in case data is collected from other websites and apps for benchmarks. This data is not anonymized and cannot be used to identify your account organization or users. This data is still clearly visible from Analytics and any of your Account properties, even after you turn this setting off. You can benefit from this setting that can flow between other associated Google products, as Analytics data can be used to build better traffic and provide relevant advice. That way, we can get help with your marketing and analytics work. Sellmarking data helps you figure out where we are in your industry and how year-over-year growth in mobile traffic can help you track key market trends. Also contributes to uncovering research analyses. Technical support Analytics representatives may sometimes need to access your account to provide service and resolve technical issues When you enable this setting, support representatives can access your data to resolve technical issues. If you turn this setting off. So support representatives may not be able to resolve technical issue we benefit with this setting turned on because if you report a problem with your account, Google Support can access your account to help you troubleshoot and find a solution

Account specialist

Google sales and marketing experts trained to find ways to improve your experience with Google products. These experts will be able to see and measure the impact of the different plans and account configurations that are implemented when you first turn on the setting 360 and Standard account users can take advantage of the best marketing communications that they offer, and 360 account users can consult with their sales experts for optimization. When you turn on other settings, Google's All sales experts can also access your account, then they can make better recommendations accordingly.

Big Data Applications:

The main purpose of Bigdata applications is to search large amounts of data, business takes better decisions, Big data is used in many fields, some of these are as follows

Healthcare

Healthcare is an area in which a lot of data is generated, first use this data.

But now nine days have gone when health doctors could not use it, to detect Ebola, to correct cancer and to do much more, big data has been used and in the coming time corona virus. Same for making vaccine will be used. Big data has gained a lot in the field of medicine and researchers have lived some life through it. The defenders have seen the results. Researchers are also creating customize medicines with big data and analytics, and data analysts are using this data to develop more and more effective treatments. government (in government) - Big data analytics has proved to be very useful in the government sector. It had a vast contribution in the Barack Obama election in 2012 and also for the BJP in 2019, the government of India uses a lot of techniques to find out what voters think about the government's actions and along with it. It also uses big data to refine policies.

In social media

The data of social media is very much and any company can be successful by inspect it. With this, we can know about the nature of the user and can also know which product is running in the market nowadays and which product is going to be in demand in the coming time. Social media can provide us with important real time information and can also inform market trends, With the help of this information, companies can change their price, promotion and effort accordingly and by knowing the mentality of the consumer through the data of social media, better decisions can be taken.

In call center -

In the call center, it is used to identify the behavior patterns of the customer and staff and to know the problems that have happened, along with it is also used to capture and process the call content.

In cyber security and intelligence

In America, computer network security is being improved by analyzing large data sets. For this, the security agency collects and analyzes the data of social media and satellite.

In predicting and preventing crime

By using big data, the police department can understand the behavior of the criminal, identify crime patterns and prevent crime from happening.



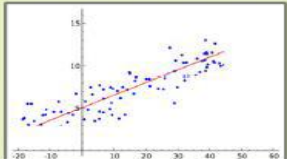
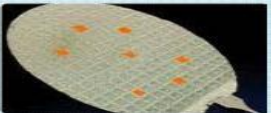





THE DEVELOPMENT OF A BIG DATA ANALYTICS CAPABILITY:

While the published research on big data is limited, there are some studies that have identified challenges associated with the success of big data projects. For instance, Kaisler and colleagues identified data storage and data transport as long-term technology issues pertaining to big data. A survey by New Vantage Group (2012) found that companies were more worried about the unstructured nature of data rather than the volume of data. Zhao et al. suggested that firms must deal with challenges pertaining to the integration of internal (e.g., transactional records) and external data (e.g., social network data). Clearly, new technology is needed to address new challenges caused by characteristics of big data; however, big data-specific technology has progressed immensely in the last few years. While we are certain that big data-specific technology will continue to progress, it is time for organizations to focus on other resources, besides technology, which are needed to build firm-specific "hard to imitate" BDA capability. For instance, Ross and colleagues assert that the majority of the big data investments fail to pay off because most companies are either not ready or do not make decisions in response to the intelligence extracted from data. McAfee and Brynjolfsson emphasize the importance of adopting data-driven decision-making culture where the senior-level executives make decisions based on data rather than on their instincts. Lack of managerial support is also cited as a critical factor affecting the success of big data initiatives. Another challenge is to recruit fresh talent and train current employees in big

data-specific skills, since working with big data requires new kinds of technical and managerial abilities, which are not commonly taught in universities.

Big data is used in organization for storing, managing, and manipulating vast amounts of disparate data at right speed at right time. To achieve the right uses the big data can be divided based on three characteristics like volume, velocity, and variety. Big data is a large or complex set of data in which cannot be managed by traditional data processing applications. The major challenges involved include analyzing, capturing, searching, sharing, storing, transferring, visualizing, queering and privacy of information. This term refers the using of predictive analysis and seldom to a particular size of data set. Big data accuracy may helps in decision making, and better decisions lead to greater efficiencies in operations, reduction in cost and it also reduces the risk. The three dimensions of big data: volume-amount of data. Variety-data in various forms .velocity-how fast data is processed. Veracity is the fourth dimension. The biases, noise and abnormality in data can be referred to as big data veracity. Compared to velocity and volume veracity is a biggest challenge. The data should be clean so that ‘dirty data’ will not accumulate in your systems.. If the data is inaccurate, is unreliable, the organization may face a big problem, especially the organization for selling information like the marketing ones. Due to the volume of information the veracity is the hardest thing to achieve with big data. The three dimensions of veracity include: objectively, Truthfulness, credibility. These dimensions may cause errors and decrease in big data quality. The Veracity issues arise due to:

1. Process Uncertainty (Processes contain randomness) Example _Uncertain travel times, Semiconductor yield
- 2.DataUncertainty (Data input is uncertain) Example _GPS uncertainty, Ambiguity, Conflicting Data, Model Uncertainty (All modeling is approximate) Example- Fitting a curve to data, forecasting a hurricane.

Process Uncertainty Processes contain "randomness"	Data Uncertainty Data input is uncertain	Model Uncertainty All modeling is approximate
 Uncertain travel times	 Intended Spelling Text Entry Actual Spelling	 Fitting a curve to data
 Semiconductor yield	 GPS Uncertainty	 Forecasting a hurricane (www.noaa.gov)
	 Testimony	
	 {Paris Airport} Ambiguity	
	 Contaminated? Rumors	
	{John Smith, Dallas} {John Smith, Kansas} Conflicting Data	

CHALLENGES:

How perfect is the sampling resolution?

- How can we manage the uncertainty, imprecision, missing values, and misstatements?
- Checks whether the data is good?
- Is the reading on time?
- Are the sampling biases understandable?
- Checks whether data is available to all?

Web has significant practical importance as online rumor and misinformation can have tremendous impacts on our society and everyday life. One of the fundamental difficulties is that data can be biased on noisy, outdated, incorrect, misleading and thus unreliable. Conflicting data from multiple sources amplifies this problem and veracity of data has to be estimated. Beyond the emerging field of computational journalism and the success of online fact-checker (e.g., Fact Checks, Claim bus) truth discovery is a long-standing and challenging problem studied by many research communities in artificial intelligence, databases, and complex systems and under various names: fact-checking, data or knowledge fusion, information trustworthiness, credibility or information corroboration for a survey and for a comparative analysis. The ultimate goal is to predict the truth label of a set of assertions claimed by multiple sources and to infer sources' reliability with no or few prior knowledge. One major line of previous work aimed at iteratively computing and updating the source's trustworthiness as a belief function in its claims, and then the belief score of each claim as a function of its sources' trust-worthiness. More complex models have then included various aspects other than

trustworthiness of source and claims belief such as the dependence between sources the correlation of claims, the notion of evolving truth.

II. CONCLUSION

In this paper we points out that big data is a collection of large and composite data set that is difficult to manage using database management tool. Management tool includes processes data like capture, storage, visualization search, sharing and analysis .In 2001, the dimension also called as 3v model were introduced .The 3vs were not enough for storing big data. So a new dimension called Veracity was introduced .Uncertainty of big data directly affect veracity Challenges are always a threat to veracity that include: How the major challenges like lack of certainty can be solved. Checks whether the data is good? How large is the coverage? How perfect is the sampling resolution? Are the readings on time? Are the sampling biases understandable? Checks whether data is available to all? So, we conclude that veracity is a new dimension that is used for conserve a balanced form of big data.

REFERENCE

- [1]. J.T. Mahoney, J.R. Pandian, The resource-based view within the conversation of strategic management, *Strateg. Manage. J.* 13 (1992) 363–380.
- [2]. R. Amit, P.J. Schoemaker, Strategic assets and organizational rent, *Strateg. Manage. J.* 14 (1993) 33–46.
- [3]. M.A. Peteraf, The cornerstones of competitive advantage: a resource-based view, *Strateg. Manage. J.* 14 (1993) 179–191.
- [4]. J.B. Barney, Looking inside for competitive advantage, *Acad. Manage. Exec.* 9 (1995) 49–61.
- [5]. D.J. Teece, G. Pisano, A. Shuen, *Dynamic Capabilities and Strategic Management*, John Wiley & Sons, 1997.
- [6]. A.S. Bharadwaj, A resource-based perspective on information technology capability and firm performance: an empirical investigation, *MIS Q.* (2000) 169–196.
- [7]. N.G. Carr, IT doesn't matter, *Educause Rev.* 38 (2003) 24–38.
- [8]. N. Melville, K. Kraemer, V. Gurbaxani, Review: information technology and organizational performance: an integrative model of IT business value, *MIS Q.* 28 (2004) 283–322.
- [9]. R.W. Palmatier, R.P. Dant, D. Grewal, A comparative longitudinal analysis of theoretical perspectives of interorganizational relationship performance, *J. Market.* 71 (2007) 172–194.
- [10]. W. Oh, A. Pinsonneault, On the assessment of the strategic value of information technologies: conceptual and analytical approaches, *MIS Q.* (2007) 239–265.
- [11]. R.M. Grant, *Contemporary Strategy Analysis and Cases: Text and Cases*, John Wiley & Sons, 2010.
- [12]. S.B. MacKenzie, P.M. Podsakoff, N.P. Podsakoff, Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques, *MIS Q.* 35 (2011) 293–334.
- [13]. J.B. Barney, D.J. Ketchen, M. Wright, The future of resource-based theory revitalization or decline? *J. Manage.* 37 (2011) 1299–1315.
- [14]. Gartner, *Gartner Survey Reveals That 64 Percent of Organizations Have Invested or Plan to Invest in Big Data in 2013*, Gartner, 2013.
- [15]. J.W. Ross, C.M. Beath, A. Quaadgras, You may not need big data after all, *Harv. Bus. Rev.* 91 (2013) 90–98.
- [16]. D.J. Teece, The foundations of enterprise performance: dynamic and ordinary capabilities in an (economic) theory of firms, *Acad. Manage. Perspect.* 28 (2014) 328–352.
- [17]. I.V. Kozlenkova, S.A. Samaha, R.W. Palmatier, Resource-based theory in marketing, *J. Acad. Market. Sci.* 42 (2014) 1–21.
- [18]. H.-C. Chae, C.E. Koh, V.R. Prybutok, Information technology capability and firm performance: contradictory findings and their possible causes, *MIS Q.* 38 (2014) 305–326.
- [19]. B. Marr, *Big Data: Using SMART Big Data. Analytics and Metrics to Make Better Decisions and Improve Performance*, Wiley Hoboken, 2015.