Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.476 – 479

RESEARCH ARTICLE

ECHO CHAMBER EFFECT IN BIG DATA

D.ARUNA KUMARI[1], N.TEJESWANI[2], G.SRAVANI[3], R.Phani Krishna[4]

Kl University, Vaddeswaram[1234]

EMAIL: aruna_D@kluniversity [1], nerellatejaswani@gmail.com [2],

gurijala.sravani2@gmail.com [3], ramadugu.999@gmail.com [4]

ABSTRACT:- Now a day's, data on the web is increasing like water in the ocean. Big data is one of the emerging area that deals with large velocity of data, is that "data sets that are too large and complex to manipulate or interrogate with standard methods or tools." Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. Data warehouses deals with multiple data sources and getting integration where as Big data covers whole WWW data that can be called as cloud data. The Big Data challenge is becoming one of the most exciting opportunities for the next years. There are many problems and applications in big data. In this paper we are going to discuss one of the problems which are called as "ECHO CHAMBER EFFECT IN BIG DATA". Echo-chamber effects occur if you design your big data model poorly. If a big data system does not consider this, it will damage the quality of the analysis. Hence we cannot extract the useful hidden knowledge. However, if the model accounts for the difference then it would improve the rigor of the prediction.

Keywords: - big data, echo chamber, volume, cloud, problems

1. INTRODUCTION

Big data is an all-encompassing terms for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. In other words big data is that which many users can be maintained and many databases can also be maintained. Big data is similar to data warehouses but data bigger, consequently requires different approaches like techniques, tools, architecture. Data bigger in the sense, if we take data of facebook or twitter data, we can see millions of posts, photos, videos per day. We can call such data as Big data.

Big data is popular term used to describe exponential growth and availability of growth, both structured and unstructured data. Both big data may be as important to business and society as the internet has become more important. It is unstructured because many texts and images are not stored in structured form and so it is in unstructured format. Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile.

Generally big data is more scalable and it is measured in pet bytes. It s a combination of various data warehouses and massive amounts of different users. Big data systems are evolution of rdbms, it is a new ways to store the data in different forms.

Big data has many different forms, different platforms, tools, and technologies. They are quite difficult to understand and use by analytics. For conversion of simple text data into multimedia data and relational data we need new data types, for that in olden days we have BI ANALYTICS (business intelligence) system.

In that BI ANALYTICS we have business data, web logs, videos, images, 3rd party applications, sensor data.

Business data is that which is action of systems run the businesses like erp system, customer relationship system etc.

3rd party applications is that like twitter, Facebook, marketing chains and large data. we use it for business should be in BI ANALYTIC form only.

Why not Big data:

Big data is not transactional in nature, and not simple or easy, not structured data ware house, not single platform, not easy or fast for analytics.

Big Data Characteristics:

Some of The key features of big data is scalability, volume, variety, velocity, variability, complexity, veracity, validity, volatility.

1. ***Volume***

Big data implies enormous volumes of data. Now that data is generated by machines, networks and human interaction on systems like social media, the volume of data to be analyzed is more massive. Yet, Inderpal states that the volume of data is not as much the problem as other V's like veracity.

2. ***Variety***

Variety refers to Variety of data , the data can contains multimedia, unstructured data, audio data…etc. Usually data is structured and unstructured. Previously we used to store data from the sources like spreadsheets and databases. Now data comes in variety form like emails, photos, videos, monitoring devices, PDFs, audio, data bases, ..,etc. This variety of unstructured data creates problems for storing and retrieving the data. Jeff Veis, VP Solutions at HP Autonomy presented how HP is helping organizations to deal with big challenges including data variety.

3. ***Velocity***

Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive , continuous and dynamic. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI, if you are able to handle the velocity. Inderpal suggest that sampling data can help deal with issues like volume and velocity. one of the big challenges is velocity, if we handle dynamic data on the web as well as on the data store, we can get more profits upon analyzing usefull and hidden and proper data.

4. ***Veracity***

Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity is like unnecessary and unknown data. We can remove

such type of data. It is like data preprocessing step in data mining. Inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems. One of the problem of Big data is volume, we can decrease it by doing data processing and null data elimination

5. Validity

Big data veracity is the issue of validity and meaning , is the data correct and accurate for the intended use. Clearly valid data is key to making the right decisions. Phil Francisco, VP of Product Management from IBM spoke about IBM's big data strategy and tools they offer to help with data veracity and validity.

6. Volatility

Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis.

Additional characteristics of big data analysis that make it different from traditional kinds of analysis are:

It can be *programmatic*, It can be *data driven*, It can use a lot of *attributes*, It can be *iterative*, It can be *quick* to get the compute cycles you need by leveraging a cloud-based Infrastructure as a Service.

If we specify the expire data for the data which is flowing on the web. We can remove the burden on the system architecture and increase the analysis speed.

PROBLEMS IN BIG DATA:

There are many problems in big data. Some of the problems in big data are:

1. Echo chamber effect
2. Google flu trends, tools in big data can be easily gamed,
3. correlations, especially subtle correlations, prone to giving scientific-sounding solutions to hopelessly imprecise questions,
4. scrubbing data , Etc

Iin this paper we are going to discuss about the "ECHO CHAMBER EFFECT" .

ECHO CHABER EFFECT:

Echo-chamber effects only occur if you design your big data model poorly. Imagine students taking a math quiz on a computer. echo-chamber effect, Big data is that which also stems from the fact that much of big data comes from the web. Whenever the source of information for a big data analysis is itself a product of big data, opportunities for vicious cycles abound. If a big data system treats these data points the same then it will damage the quality of the analysis. However, if the model accounts for the difference then it would improve the rigor of the prediction.

SOLUTION:

1. If a student answers the question incorrectly the program prompts the user with a hint or even the correct answer. The student then takes another quiz the next day with similar questions. A big data system should treat these two observations very differently. There is a much better chance the student gets the question right on day two because of the prompt from day one.
2. Consider translation programs like Google Translate, which draw on many pairs of parallel texts from different languages — for example, the same Wikipedia entry in two different languages — to discern the patterns of translation between those languages. This is a perfectly reasonable strategy, except for the fact that with some of the less common languages, many of the Wikipedia articles themselves may have been written using Google Translate. In those cases, any initial errors in Google Translate infect Wikipedia, which is fed back into Google Translate, reinforcing the error.

OTHER PROBLEMS IN BIG DATA:

1. Big data problem is that Network traffic risk analysis, geospatial classification and business forecasting. The network intrusion detection and prediction are the time sensitive applications and require highly efficient big data techniques and technologies to tackle the problem on the fly. The new technologies can help conduct big data analytics and on various applications.

2. The new technologies can help big data analytics to conduct in different applications. The techniques Hadoop distributed file system, cloud technology, hive database can be combined to address the problem like big data classification.
3. The other problem is that Management of big data this problem is that the continuity and complex parameters add extra difficulties to the big data .Hence network topology must be designed for this problem. Big data analytics can be handled efficiently with cost effective objectives.
4. The other problem is that The security mechanism in cloud technology is generally week. Hence by tampering the data at public cloud is inevitable and it's a big concern. And hence finding the robust security mechanism for the purpose of using public cloud is the biggest problem.
5. The other problem is When you combine someone's personal information with vast external data sets, you can infer new facts about that person (such as the fact that they're pregnant, or are showing early signs of Parkinson's disease, or are unconsciously drawn toward products that are colored red or purple). And when it comes to such facts, a person a) might not want the data owner to know b) might not want anyone to know c) might not even know themselves. The fact is, humans like to control what other people do and do not know about them—that's the core of what privacy is, and data mining threatens to violate that principle.
6. The another problem is A typical enterprise generally has 10x more IT employees than analysts or data scientists. The process of analysis starts with a line of business request. IT collects data from various databases and transfers it to data scientists. Large teams of data scientists are deployed who spend months (or sometimes years) querying the data. Hiring data scientists (with advanced background in statistics, computer science, and some functional expertise) to accelerate the process is difficult because people with these skills are extremely scarce. The demand and interest in data scientists is skyrocketing, as Google Trends can attest, while we are producing fewer of them. What we need is a new class of technologies that amplify the impact data scientists and allow more people to become data scientists.
7. The other is cultural issues that we must solve to unlock the latent potential of data is data silos In most enterprises, the data generated by a functional area ends up being the property of that group. This leads to two problems. First, it's difficult to get a "complete" view of the data. Consider all the silos and systems that hold data: CRM, ticketing, bug tracking, fulfillment, etc. Getting all the relevant systems to even talk to each other is a huge challenge. Second, there's significant cultural dissonance within organizations. Typically, each group controlling a data silo ends up caring more about their power and place in a department rather than the success of the organization as a whole. Organizations need to pool their data to find the answers to and get a complete view of their data.

CONCLUSION:

Big data is being used by many users and can maintain many databases it encompassing for collection of data. There are many challenges in big data and many problems in bigdata as it is not fully developed. And in our paper we are discussing about the problem "ECHO CHAMBER EFFECT IN BIG DATA" and it is mainly used for improving the quality of analysis.

REFERENCES:

1. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
2. http://www.webopedia.com/TERM/B/big_data.html
3. http://en.wikipedia.org/wiki/Big_data
4. http://www.slideshare.net/venturehire/what-is-big-data-and-its-characteristics
5. http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

