AN INFERENCE ON BIG DATA: SURVEY, CHALLENGES, REQUIREMENTS AND PROBLEMS

Blessy A V, Kavya R
18tucs027@skct.edu.in, 18tucs058@skct.edu.in
UG Students, Department of Computer Science Engineering
Sri Krishna College of Technology, Coimbatore.

ABSTRACT:
This piece of research work aims at understanding the basic of big data analysis, the risks in big data analysis and what are the different parameters in big data analysis. Big data analytics is often complex process of examining large and varied data sets, or big data, to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions. On a broad scale, data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them to help organizations make informed business decisions.

1. INTRODUCTION
Big data analytics is a form of advanced analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems. Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

1.1. WHAT’S NEW HERE?
Business executive sometimes ask us? “Isn’t ‘big data’ just another way of saying ‘analytics’?” It’s true that they’re related: The big data movement, like analytics before it, seeks to glean intelligence from data and translate that into business advantage. However, there are three key differences: There are three V’s. The first characteristics of big data, which is “VOLUME”, refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results. For example, a computer system is limited by current technology regarding the speed of processing operating. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant. To achieve higher processing speeds more computer power is needed and so, the infrastructure must be developed, but at higher costs. By trying to compress huge volumes of data and then analyses it, is a tedious process which will ultimately prove more ineffective. To compress data it takes time, almost the same amount of time to decompress it in order to analyses it so it can be displayed, by doing this, displaying the results will be highly delayed. One of the methods of mining through large amount of data
is with OLAP solutions (Online Analytical Processing).

1.2. VELOCITY

For many applications, the speed of data creation is even more important than the volume. Velocity is all about the speed that data travels from point A, which can be an end user interface or a server, to a point which can have the same characteristics as point A is described. This is a key issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.) For companies that is a challenge that most of them can’t keep up to. Usually data transfer is done at less than the capacity of the systems. Transfer rates are limited but requests are unlimited, so streaming data in real time or close to real time is a big challenge. The only solution at this point is to shrink the data that is being sent. A good example is Twitter Interaction on Twitter consists of text, which can be easily compressed of high rates. But, as in the case of ”Volume” challenge, this operation is still time-consuming and there will still be a delay in sending-receiving data. The only solution to this right now is to invest in infrastructure.

1.3. VARIETY

Third characteristics of Big data it represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, data sent from browsers, simulations etc. The challenge is how to sort all this data so it can be “readable” by all users that access it and does not create ambiguous results. The mechanics of sorting has two keys: variables at the beginning that transmit data and the system that receives it and interpret it so that can be later displayed. The issue of these two key aspects is that they might not be compatible regarding the content of the data transferred between them. For example, a browser can send data that consists of user’s location, favorite search terms and so on. Meanwhile, the big data systems receive all this information unsorted, so it’s difficult for it to understand whether this user is from London or from Orange. To avoid this mess created in big data solutions, all systems that send data should be standardized so that can send data in a logical array that, afterwards it can be easily analyzed and displayed in a proper manner.

1.4. VALUE

The fourth “v” is “VALUE ”and is all about the quality of data that is stored and the future use of it. Large quantity of data is being stored from mobile phone call records to TCP/IP logs. The question is if all together can have any commercial value. There is no point in storing large amount of that if it can’t be properly managed and the outcome can’t offer insights for a good development.

1.5. VERACITY

The fifth characteristics of big data and came from the idea that the possible consistency of data is good enough for big data. For example, if A is sending an email to B, B will have the exact content that A send it, if else, the email service will not be reliable and people will not use it. In big data, if there is a loss regarding the data stored from one geo-location, is not an issue, because there are hundreds more than can
cover the information. Current technology that is the software technology try to overcome the challenge that “V’s” raises. One of these is Apache Hadoop, which is an open source software that its main goal is to handle large amounts of data in a reasonable time. What Hadoop does is dividing data across multiple systems infrastructure in order to be processed. Also, Hadoop creates a map of the content that is scattered so it can be easily found and accessed. Because it’s accessible does not make it ethical or safe. True in all settings. In conflict/crisis settings, the privacy challenge may soon become a security risk.

2.2. ANALYTICAL CHALLENGE
1. ARROGANCE/NAIVETY
Believing that the truth is in the (big) data, that the (big enough) data ‘speak for themselves’, that all it takes is to beep digging and mining to unveil the truth, that more is always better etc. Apophenia—seeing patterns where none exists. If you torture the data long enough, it will eventually start talking. S&P index and butter production in Bangladesh? (Leinweber, 2007).

2.3. “MAKING SENSE OF THE DATA”
Not having the right computing capacities. Unstructured data. Sentiment analysis. Translation. Sample bias. No understanding of context.

2. THE DIFFICULTIES AND REQUIREMENTS OF BIG DATA

2.1. DATA CHALLENGES
How do we access data and protect their producers?

I. AVAILABILITY
Depends on technological penetration and use. But mobile phone use is increasing fast and supports/will support Internet access.

II. RELIABILITY:
Attempts at playing the system(s), plus suppressing signals? Especially prevalent in conflict settings.

III. ACCESS
Not all data produced in easily accessible, storable (e.g. Twitter). This is compounded in poor countries. How to access data from a mobile phone carrier?

IV. PRIVACY

3. OPERATIONAL/SYSTEMIC CHALLENGES:
- Data privacy? Intellectual property?
- Changing systems, policies, frameworks, decision making processes.

3.1. RISKS IN BIG DATA
Risk 1: Security and privacy of individuals and communities may be jeopardized—the dark side of Big Data. Conflict zones!
Risk 2: Creating new digital divides, between the Small and the Big world, as well as within them, between communities. Who has access, who has the
knowledge? Those who can answer the questions are often those who ask the questions.  

RISK 3: RESPOND FOOLISHLY. i.e. in a way that may exacerbate tensions, poverty.

4. **TOP 5 BIG DATA PROBLEMS**

4.1. **FINDING THE SIGNAL IN THE NOISE**

It’s difficult to get insights out of a huge lump of data. Intellect soft Big Data Scientist and author of the book “Social Network Analysis for Startups”, Maksim Tsvetovat said that in order to use Big Data “There has to be a discernible signal in the noise that you can detect, and sometimes there just isn’t one. Once we’ve done our intelligence on the data, sometimes we have to come back and say we just didn’t measure this right or measured the wrong variables because there’s nothing we can detect here.” He went on to say that in its raw form, Big Data looks like a hairball and scientific approach to the data is necessary. “You approach it carefully and behave like a scientist which means if you fail at your hypothesis, you come up with a few other hypothesizes, and maybe one of them turns out to be correct.”

4.2. **DATA SILOS**

Data silos are basically Big Data’s kryptonite. What they do is store all of that wonderful data you’ve captured in separate, disparate units, that have nothing to do with one another and therefore no insights can be gathered from this data because it simply isn’t integrated on the back end. Data silos are the reason you have to number crunch to produce a monthly sales report. They’re the reason that C-level decisions are made at a snail’s pace. They’re the reason your sales and marketing teams simply don’t get along. They’re the reason that your customers are looking elsewhere to take their business because they don’t feel their needs are being met and a smaller, more nimble company, is offering something better. The way to eliminate data silos? Integrate your data.

4.3. **INACCURATE DATA**

Not only are data silos ineffective on an operational level, they are also fertile breeding ground for the biggest data problem: inaccurate data. According to a recent report from Experian Data Quality, 75% of businesses believe their customer contact information is incorrect. If you’ve got a database full of inaccurate customer data, you might as well have no data at all. The best way to combat inaccurate data? Eliminating data silos by integrating your data.

4.4. **TECHNOLOGY MOVES TOO FAST**

Larger corporations are more prey to data silos, for such reasons as they prefer to keep their databases on-premises, and because decision making about new technologies is often slow. One example cited in the CapGemini report is that stalwarts like telecoms and utilities “…are noticing high levels of disruption from new competitors moving in from other sectors. This issue was mentioned by over 35% of respondents in each of these industries, compared with an overall average of under 25%.” In essence, traditional players are slower to move on technological advances and are finding themselves faced with serious competition from smaller companies because of this. Big Data is also fast data. Paul Maritz, Pivotal Chief Executive Officer of the EMC Federation, wrote in a recent CapGemini Report that, “If you can obtain all the relevant data, analyze it quickly, surface actionable insights, and drive them back into operational systems, then you can affect events as they’re still unfolding. The ability to catch people or things “in the act”, and affect the outcome, can be extraordinarily important, valuable and disruptive.” The ability to make snap decisions and quickly move on Big Data insights is the advantage SMEs have over large corporations.

4.5. **LACK OF SKILLED WORKERS**

CapGemini’s report found that 37% of companies have trouble finding skilled data analysts to make use of their data. The best bet is to form one common data analyst team for the company, either through re-skilling your current workers or recruiting new workers specialized in big data. You need to find employees that not only understand data from a scientific perspective, but who also understand the business and its customers, and how their data findings apply directly to them.
5. CONCLUSION

Building a viable solution for large and complex data is a challenge that companies in this field are continuously learning and implementing new ways to handle it. One of the biggest problems regarding Big Data is the infrastructure’s high costs. Hardware equipment is very expensive for most of the companies, even if Cloud solutions are available. Each Big data system requires massive processing power and stable and complex network configurations that are made by specialists. Besides hardware infrastructure, software solutions tend to have high costs if the beneficiary doesn’t opt for open source software. Even if they chose open source, to configure there is still needed specialists with the required skills to do that. The downside of open source is that maintenance and support is not provided as is the case of paid software. So, all that is necessary to maintain a Big Data solution working correctly needs, in most cases, an outside maintenance team. Software solutions are limited by hardware capabilities. Hardware can only be as fast as current technologies can offer. A software solution just sends the tasks in order to be processed. The software architecture tries to compensate for the lack of processing speed by sorting and ordering the requests, so that processes can be optimized in order to achieve best Performance. To sort through data, so that valuable information will be extracted for further use, requires human analysis skills. A computer program can only do what is programmed to do, it cannot see grey areas and cannot learn or adapt to new types of information unless is programmed to handle it. Therefore, human capabilities are used to sort data with a set of tools which speed up the process. All this will only increase the time that results will be displayed and so, the analysis of the results, in order to evaluate current position or forecast, will decrease the beneficiary’s time for taking measures or plan properly.

REFERENCES


