Topic Name: - Three V's of Big Data Analysis Using Mathematical Module

Vrushali Desale [1] ,Damini Deore [2]
Department of Computer Engineering
[1]Assistant Professor,D Y Patil College of Engineering , Ambi, Talegaon
[2]D Y Patil College of Engineering , Ambi, Talegaon

Abstract-*The hurdles of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals in required amounts and software that possess ability to process data at a high velocity. The term, Big Data 'has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. Big Data is still a novel concept, and in the following literature we intend to elaborate it in a palpable fashion. It commences with the concept of the subject in itself along with its properties and the two general approaches of dealing with it. The comprehensive study further goes on to elucidate the applications of Big Data in all diverse aspects of economy and being. The utilization of Big Data Analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers has been discussed in depth. Aside this, the incorporation of Big Data in order to improve population health, for the betterment of finance, telecom industry, food industry and for fraud detection and sentiment analysis have been delineated. The challenges that are hindering the growth of Big Data Analytics are accounted for in depth in the paper. This topic has been segregated into two arenas- one being the practical challenges faces whilst the other being the theoretical challenges.*

Keywords-Big Data, Volume, Velocity, Variety and Traditional Data.

I. INTRODUCTION

There is no hard and fast rule about exactly what size a database needs to be in order for the data inside of it to be considered "big." Instead, what typically defines big data is the need for new techniques and tools in order to be able to process it. In order to use big data, you need programs which span multiple physical and/or virtual machines working together in concert in order to process all of the data in a reasonable span of time.

Getting programs on multiple machines to work together in an efficient way, so that each program knows which components of the data to process, and then being able to put the results from all of the machines together to make sense of a large pool of data takes special programming techniques. Since it is typically much faster for programs to access data stored locally instead of over a network, the distribution of data across a cluster and how those machines are networked together are also important considerations which must be made when thinking about big data problems

Datasets considered for Big Data

The uses of big data are almost as varied as they are large. Prominent examples you're probably already familiar with including social media network analyzing their members' data to learn more about them and connect them with content and advertising relevant to their interests, or search engines looking at the relationship between queries and results to give better answers to users' questions.

But the potential uses go much further! Two of the largest sources of data in large quantities are transactional data, including everything from stock prices to bank data to individual merchants' purchase histories; and sensor data, much of it coming from what is commonly referred to as the Internet of Things (IoT). This sensor data might be anything from measurements taken from robots on the manufacturing line of an automaker, to location data on a cell phone network, to instantaneous electrical usage in homes and businesses, to passenger boarding information taken on a transit system.

How to analyses Big Data ?

Perhaps the most influential and established tool for analyzing big data is known as Apache Hadoop. Apache Hadoop is a framework for storing and processing data in a large scale, and it is completely open source. Hadoop can run on commodity hardware, making it easy to use with an existing data center, or even to conduct analysis in the cloud. Hadoop is broken into four main parts:

The Hadoop Distributed File System (HDFS), which is a distributed file system designed for very high aggregate bandwidth;

- YARN, a platform for managing Hadoop's resources and scheduling programs which will run on the Hadoop infrastructure;
- MapReduce, as described above, a model for doing big data processing;
- And a common set of libraries for other modules to use.

II. THREE V'S OF BIG DATA

3Vs (volume, variety and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing. According to the 3Vs model, the challenges of big data management result from the expansion of all three properties, rather than just the volume alone -- the sheer amount of data to be managed.

That, of course, begs the question: what is big data? The answer, like most in tech, depends on your perspective. Here's a good way to think of it. Big data is data that's too big for traditional data management to handle. Big, of course, is also subjective. That's why we'll describe it according to three vectors: volume, velocity, and variety -- the three Vs

Volume

Volume is the V most associated with big data because, well, volume can be big. What we're talking about here is quantities of data that reach almost incomprehensible proportions. Facebook, for example, stores photographs. That statement doesn't begin to boggle the mind until you start to realize that Facebook has more users than China has people. Each of those users has stored a whole lot of photographs. Facebook is storing roughly 250 billion images.

Can you imagine? Seriously. Go ahead. Try to wrap your head around 250 billion images. So, in the world of big data, when we start talking about volume, we're talking about insanely large amounts of data. As we move forward, we're going to have more and more huge collections. For example, as we add connected sensors to pretty much everything, all that telemetry data will add up. Or, consider our new world of connected apps. Everyone is carrying a smartphone. Let's look at a simple example, a to-do list app. More and more vendors are managing app data in the cloud, so users can access their to-do lists across devices. Since many apps use a freemium model, where a free version is used as a loss-leader for a premium version, Saabs-based app vendors tend to have a lot of data to store.

Taoist, for example (the to-do manager I use) has roughly 10 million active installs, according to Android Play. That's not counting all the installs on the Web and is. Each of those users has lists of items -- and all that data needs to be stored. Taoist is certainly not Facebook scale, but they still store vastly more data than almost any application did even a decade ago.

Then, of course, there are all the internal enterprise collections of data, ranging from energy industry to healthcare to national security. All of these industries are generating and capturing vast amounts of data.

That's the volume vector.

Velocity

Remember our Facebook example? 250 billion images may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day. A day. So that 250 billion number from last year will seem like a drop in the bucket in a few months.

Velocity is the measure of how fast the data is coming in. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

Here's another example. Let's say you're running a presidential campaign and you want to know how the folks "out there" are feeling about your candidate right now. How would you do it? One way would be to license some Twitter data from Grip (recently acquired by Twitter) to grab a constant stream of tweets, and subject them to sentiment analysis.

That feed of Twitter data is often called "the firehouse" because so much data (in the form of tweets) is being produced, it feels like being at the business end of a firehouse.

Here's another velocity example: packet analysis for cyber security. The Internet sends a vast amount of information across the world every second. For an enterprise IT team, a portion of that flood has to travel through firewalls into a corporate network.

Unfortunately, due to the rise in cyber-attacks, cybercrime, and cyber espionage, sinister payloads can be hidden in that flow of data passing through the firewall. To prevent compromise, that flow of data has to be investigated and analyzed for anomalies, patterns of behavior that are red

flags. This is getting harder as more and more data is protected using encryption. At the very same time, bad guys are hiding their malware payloads inside encrypted packets.

Or take sensor data. The more the Internet of Things takes off, the more connected sensors will be out in the world, transmitting tiny bits of data at a near constant rate. As the number of units increase, so does the flow.

That flow of data is the velocity vector.

Variety

You may have noticed that I've talked about photographs, sensor data, tweets, encrypted packets, and so on. Each of these are very different from each other. This data isn't the old rows and columns and database joins of our forefathers. It's very different from application to application, and much of it is unstructured. That means it doesn't easily fit into fields on a spreadsheet or a database application.

Take, for example, email messages. A legal discovery process might require sifting through thousands to millions of email messages in a collection. Not one of those messages is going to be exactly like another. Each one will consist of a sender's email address, a destination, plus a time stamp. Each message will have human-written text and possibly attachments.

Photos and videos and audio recordings and email messages and documents and books and presentations and tweets and ECG strips are all data, but they're generally unstructured, and incredibly varied.

All that data diversity makes up the variety vector of big data.

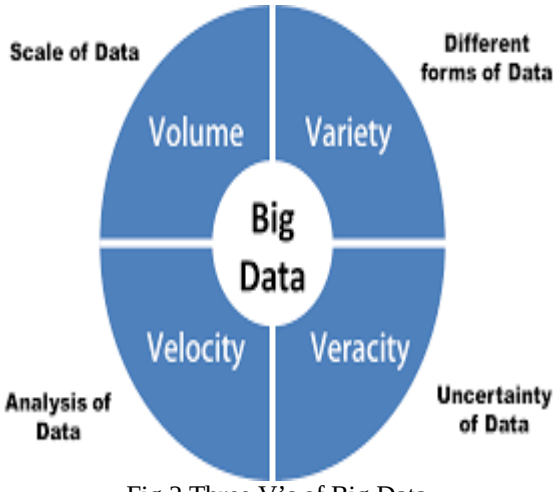

Fig 2 Three V's of Big Data

III. THREE V'S ANALYSIS

The predictions from the IDC Future Scope for Big Data and Analytics are:

1. Visual data discovery tools will be growing 2.5 times faster than rest of the Business Intelligence (BI) market. By 2018, investing in this enabler of end-user self-service will become a requirement for all enterprises.

2. Over the next five years spending on cloud-based Big Data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions. Hybrid on/off premise deployments will become a requirement.

3. Shortage of skilled staff will persist. In the U.S. alone there will be 181,000 deep analytics roles in 2018 and five times that many positions requiring related skills in data management and interpretation.

4. By 2017 unified data platform architecture will become the foundation of BDA strategy. The unification will occur across information management, analysis, and search technology.

IV. MATHEMATICAL ANALYSIS

The big data, such as the web usage data of Internet, real time traffic information, rapidly changes over time. The analytical algorithms needed to process these data quickly. The dynamic problems, sometimes termed as non-stationary environments, or uncertain environments, dynamically change over time. Swarm intelligence has been widely applied to solve stationary and dynamical optimization problems. Swarm intelligence often has to solve optimization problems in the presence of a wide range of uncertainties. Generally, uncertainties in optimized problems can be divided into the following categories. The fitness function or the processed data is noisy. The design variables and/or the environmental parameters may change after optimization, and the quality of the obtained optimal solution should be robust against environmental changes or deviations from the optimal point. The fitness function is approximated [9], such as surrogate-based fitness evaluations, which means that the fitness function suffers from approximation errors. The optimum in the problem space may change over time. The algorithm should be able to track the optimum continuously. The target of optimization may change over time. The demand of optimization may adjust to the dynamical environment, for example, there should be a balance between the computing efficiency and the computational cost for different computing loads. In all these cases, additional measures must be taken so

that swarm intelligence algorithms are still able to solve satisfactorily dynamic problems.

V.CONCLUSION

This literature survey discusses Big Data from its infancy until its current state. It elaborates on the concepts of big data followed by the applications and the challenges faced by it. Finally we have discussed the future opportunities that could be harnessed in this field. Big Data is an evolving field, where much of the research is yet to be done. Big data at present, is handled by the software named Hadoop. However, the proliferating amounts of data is making Hadoop insufficient. To harness the potential of Big Data completely in the future, extensive research needs to be carried out and revolutionary technologies need to be developed. Summarizing, Peter Sondergaard, Senior Vice President of Gartner Research famously stated, ―Information is the oil of the 21st century and analytics is the combustion engine.

REFERENCES

[1] Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013 [2] http://lsst.org/lsst/google
[2] http://en.wikipedia.org/wiki/Parkinson‘s_law
[3] http://www.economist.com/node/15557443
[4] http://www.youtube.com/t/press_statistics/?hl=en
[5] http://www.forbes.com/sites/bernardmarr/2015/04/21/how-bigdata-is-changing-healthcare/
[6] http://www.forbes.com/sites/bryanpearson/2015/04/10/exercise-inservice-fitbit-omni-channel-begs-for-omni-prescience/
[7] http://www.engadget.com/2015/04/10/jawbone-up3-shipping-april- 20th/
[8] http://www.samsung.com/uk/consumer/mobiledevices/wearables/gear/SM-R3500ZKABTU
[9] http://healthdataalliance.com/
[10] http://www.ibm.com/software/data/bigdata/industry-healthcare.html
[11] http://www.firstpost.com/business/big-data-booster-shothealthcare-industry-needs-2160271.html
[12] Chester Curme, Tobias Preis, Eugene Stanley, Helen Susannah Moat, ―Quantifying the semantics of search behavior before stock market moves‖; CrossMark, December 2013
[13] http://www.wsj.com/articles/how-computers-trawl-a-sea-of-datafor-stock-picks-1427941801
[14] Nitish Sinha, ―Using Big Data in Finance: Example of sentimentextraction from news articles‖; FEDS notes, March 2014
[15] Baker, Malcolm and Jeffrey Wurgler, 2007. "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, vol. 21(2), pages 129-152.