Big Data and its Impact on Society and Industry and the Growing Need for Big Data

Alonzo L. Foster[1], Jr., Dewayne R. Brown[2]

North Carolina Agricultural and Technical State University, School of Technology,

Department of Computer Systems Technology[1,2]

Abstract: This paper explores big data and its impact on society and industry. Data collected from many streams such as social data, machine data, and transactional data are collected at rates that reaching overwhelming numbers. Many question its importance and relevance to society, but when evaluated and analyzed, we can see how it has made advancements in technology and in social development that affects each of us on a daily basis. The more data that is collected, the more accurate information we have to meet the needs of an ever changing society. Technological advancements have changed how this data is collected and stored and this is seen in many industries and companies and their use of big data to determine their own relevance in society and how they economic and social development. SAS (Statistical Analysis System) is one of the leading providers of big data and is the standard for data management. Oracle, IBM, and Teradata are some of the leading providers of terabyte scale data to companies that require data management. The following will provide some insight on how big data is generated, analyzed and used to meet the needs of our ever changing society.

Keywords: big data impacts, data types, datification, big data uses, big data defined.

INTRODUCTION

Big data has become a new buzz words in the world of technology, and many think it is something new that has just surface in the past few years, but it has been around for a few centuries. Since the first US census in August of 1790, big data has been a part of our lives. When you consider how science takes a test group of data to make a hypothesis, big data uses larger groups of data to get a more accurate analyses, which leads to a better hypothesis. Only one example of how big data is changing our lives. Structured and unstructured data, exponential growth, and accessibility, all have caused the increase of demand and need for ways to organize, analyze, and store data in society. We will look at how data has changed society in the many facets of industry. (Hendricks, 2015)

TYPES OF DATA

It's easy to look at structured data, for example: a spreadsheet, that due to its high degree of configuration and ability to integrate it into a database so that it can be retrieved by simple searches, through algorithms and search engines. Unlike unstructured data, for example: email, which comes with many challenges as it relates to organization and retrieval. Because unstructured data has no pre-defined data model, and can often be very word heavy, it makes it impossible to humanly organize such data in a manner that it can be retrieved. Hence the reason why it was often omitted when compiling usable data to analyze. With new technologies, there is no need to omit data because it is unstructured. When we think of a spreadsheet, we automatically think of a system that allows us to organize our thoughts. When we think of an email, we basically think of abstract thoughts sent to another user with no organization. It may have a subject, but how many of us add other information, sometimes wordy thoughts, and our own personal emotions into an email which makes it next to impossible to organize that data into a database. Though structured or unstructured, data is the key component to analyzation and informed decisions. There had to be a way to use all the data. (Marr, 2015)

Aside from structured and unstructured data, data can also be classified as internal and external. These definitions are somewhat self-explanatory. Internal data is data that comes from within the company and already is accessible or can be accessible when needed, external data is data that comes from outside resources and takes more effort to attain. Examples of Internal data are as follows:

- Customer feedback
- Sales data
- Employee or customer survey data
- CCTV video data
- Transactional data
- Customer record data
- Stock control data
- HR data (Marr, 2015)

Most of this data is considered to be structured data, but there are some that may be considered unstructured, so there is no definite rule to how data is categorized. Just with anything, most people want things that are harder to achieve which remains true for external data. External data can either be private or public. Public meaning mostly free or paying a third party to obtain the data or private data, which requires one to obtain by paying the a third party or obtaining from another business. Examples of external data is as follows:

- Weather data
- Government data – such as census data
- Twitter data

- Social media profile data
- Google trends or Google Maps (Marr, 2015)

When looking for data, it is important to know the hierarchy for searching data:

1. Internal structured data - the easiest to find and the least inexpensive to obtain.
2. Internal semi-structured
3. Internal unstructured
4. External structured
5. External unstructured (Marr, 2015)

It is always best to start at the top when seeking for answers, instead of starting with external unstructured data because many times the answers we seek are right within our reach. There are seven main ways of collecting data that are listed and described below:

1. Created data is data that is created and captured manually, such as employee and customer surveys, focus groups. This type of data is basically structured and can be external or internal.
2. Provoked data when you invite people to give their opinion or review or rate your product or service. As with created data, this data is mostly structured and can be either internal or external.
3. Transaction data is generated every time a customer buys something. Whether online or in store. Transaction data is usually structured internal data.
4. Compiled data usually comes from other companies such as credit reporting companies which is compiled from many different sources outside of the company. Compiled data is usually structured external data.
5. Experimental data is usually a combination of created and transactional data. This is just like an experiment in school that requires you to design an experiment and use a focus group (created) and then observe the data that is created (transaction). Experimental data is mainly structured and can be internal or external.
6. Captured data such as google searches or GPS data that is generated through your phone. This data is one that is causing the most concern with consumers because it is basically gathered without consumer permission. Captured data is mostly unstructured and can be internal or external.
7. User-generated data such as Facebook post, Tweets, and videos posted on YouTube and even comments made in blogs or articles. User-generated data is usually unstructured and can be internal or external. (Marr, 2015)

Whether structured or unstructured, internal or external, it is important to understand that neither is more important than the other, data is based on need and what you are trying to achieve.

In late years, other forms of data have surfaced and the term datification has come about, which means taking something that was basically invisible and turning it into data. Most of our activity as humans and computer based activity leaves a data trace and this is called data, and it can be obtained and analyzed as well. Though we have always had the ability to record conversations, storage was always an issue, so with recent technologies, even conversations have become usable data. Data can come from our conversations, activates, photos and videos, sensors and of course the internet. The internet produces activity data. When we visit sites, there is logged data of our usage, logs of the time we spend on the internet and all this is logged by the internet service provider and is usable data that can be used and sometimes sold. Wearable devices collect data such as our physical activity, heartrate, and sleep patterns. This is all data that can be used by health care professionals. Conversation data such as SMS messages, audio recordings of telephone calls, and emails is another form of data. Photo and video data, collected from digital cameras and cell phones are useable and storable data. Sensor data such as the data that is obtained from GPS sensor, accelerometer sensors, gyroscope, proximity sensor, ambient sensor, and NFC (Near Field Communications) sensor. All have recently become more useable data than in years past.

REPORTING DATA

All this talk about types of data and collecting data, it is of no importance if it is not usable and can provide insight that people can understand. Here is where the end user comes in and has the ability customize reports from all the data that is created or obtained. Reports can be generated in several forms or visual aids. They can be in charts, tables, spreadsheets, and some can come in the form of numerical data only, and others may be detailed summaries with some tables and charts to help with understanding and clarification. Charts and graphs such as line graphs, bar graphs, pie charts, and scatter charts are the most popular and most requested due to the ability to illustrate the relationship between the data. When data is received in a spreadsheet, there are still things that may not be clear just by looking at numbers and dollars, but turn that spreadsheet into a pie chart and it becomes clearer and understood much faster. New data visualization is what companies are using to report data. This is software that allows the user to connect to a data source and organize the data into meaningful reports. This allows companies to use automation as way of gathering data instead of manually taking raw data and trying to create reports. This method is dated and very slow and doesn't fit in to the ever changing society and getting things done faster. Most of the software that is used for reporting is open source and is mainly user-friendly, which doesn't require an expert to run, allowing the company to have many users that extract collected data at any given time or automatically generate reports on a daily, weekly or monthly basis.

BIG DATA DEFINED

Doug Laney, who is an industry analyst, verbalized the definition in 2001 which we use today as the three Vs of big data: volume, velocity, and variety. Factors that make volume relevant is the increase of transaction based data that has been stored over the years and the influx of unstructured data such as social media that has emerged. The decrease in storage cost has made it easier to store this data, but making sense of the data and finding the relevance of the huge volumes of data. One of the largest

demands on technology is speed (velocity), so with the increase in how fast data is created, there must be ways to manage data velocity. Data used to be stored in numerical format which took time, but now data is stored in many formats, such as spreadsheets, financial transactions, audio, video, and email. SAS (Statistical Analysis System) added two additional components to the definition of big data known as variability and complexity. With inconsistent peaks of when data gathered, variability has a major part in how data is received. Once data is received, it has be organized which brings about another complex undertaking to transform and link data across various systems. Having the ability to manage and merge all these different varieties of data, the increasing velocity of data, and mass volumes of data on a daily basis is now being addressed by what we know today as big data. (SAS, 2015)

When we think about big data, most often we think, "how does this data relate to me, and why should I care?" Well, the answer lies in the following: big data is used to determine cost reduction of many of the products and services consumers use on a daily basis; we always want things right away and big data is used to determine reduction of time for something simple as a back transaction; looking at how we want things to be better, big data is used in new product development and optimization of old ones; sometimes humans err in business decisions that affect us every day, so big data aides in making smarter business decisions. Big data is used to enhance the quality of many products by finding root of failures and issues, which will save time and money. Traffic data is used to optimize routes for delivery of goods and services to consumers. SKUs and barcodes are a major part of our lives and they help determine prices and maximize profit and aid in inventory management. Couponing, which is a huge industry in itself uses big data to determine products and services that have the greatest demand and offer discounts based on that data. When in a particular area, big data is used to send personalized information to mobile devices to make life easier for all consumers. Big data manages the products and services purchased by consumers and allow merchants to offer rewards to their valued customers. Big data is working behind the scenes to protect us from fraudulent behavior by using clickstream analysis. So we see that big data affects many aspects of our lives from minor purchases to protecting us from those who want to destroy our reputations through identity fraud.

When considering the many streams of data that create what we know as big data, those streams have been narrowed down to the following varieties that include transactional data, machine data, and social data. Retailers, and B2B (Business to Business) companies, produce masses of transactional data daily. Considering their product codes, customer and payment information, and distribution and manufacturing data, and much more. There is no wonder why transactional data is such a large stream. Scanners, tracking devices, and other monitoring devices are what generate machine data. This data is probably the fast known form of data since there has been

machine data tracked at a rate of 40 terabytes of data per second. Facebook (2.7 billion Likes daily), Twitter (230 million Tweets daily), YouTube (60 hours of video uploaded daily), and many other social media sites produce usable data that provides customer behavior insight, that can be used in CRM (customer relationship management) data that can be analyzed to determined business trends and relevance. (Evans, Ciobo, Miller, Wall, 2015)

The business of storing data is a growing as well as the demand for big data. Just as there is exponential growth in the way data is growing, the business is growing exponentially as well. It is estimated that the big data market will reach $41.5 billion by 2018. This rate is 6 times the growth rate of the entire information technology market. By 2020 that growth rate is expected to reach double digit growth with only two additional years of expansion.

Figure 1 - Industries that use Big Data to transform Business Models

Big data is taking business to new heights and even forcing companies to create new business models. The above chart outlines what big data is doing in different markets and how it is improving performance. Growth of these magnitudes is having a huge effect on the economic development worldwide, all while reducing cost and increasing profits for those who choose to incorporate big data into their business practices.

Big data has taken storage of data to a new level. From the table top storage devices that were used decades ago by many companies, they now have transitioned to data centers where multiple storage units are housed.

Figure 2 - Data Center Storage Cabinets

These data centers house cabinets for storage drives that measure in petabytes instead of terabytes. The above image shows what a typical data center floor may look like. Multiple cabinets housed in a secure location. These data centers require little manpower to manage because there is little maintenance that is required at this level. There are times when drives fail and have to be changed out or replaced, but they are built to last with little to no

maintenance. The work is usually done offsite at remote locations away from the data centers. Security for these data centers are usually very tight and are usually located in unmarked buildings, unlike the glamourous buildings that may house the functioning departments of the business. This is because threats to companies happen and with the amount of data that is store in these data centers, a threat to this amount of data could be a huge detriment to the company. In some cases there are mirrored data centers in other locations as a form of backup just in case one is disabled, the business will still be able to function. They also require back up power supply to protect against power failure as seen in the architecture in the image below. (Evans, Ciobo, Miller, Wall, 2015)
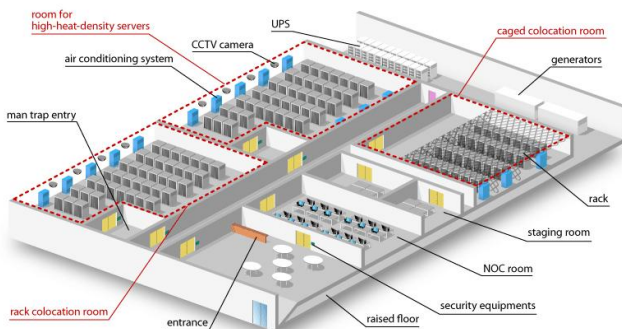

Figure 3 - Data Center Facility Architecture

These storage devices produce a high amount of heat, so a well cooled data center is extremely necessary, as you see in the architecture above, there are multiple air conditioning systems in each room, which also requires backup of power. Though it may appear that the cost of operation of data centers would be extremely high, the benefit of big data is still worth the cost and allows for growth and expansion.

Smaller companies are using scalable software such as Hadoop which parallelizes larger data sets across less expensive hardware units which allows them to have the petabyte environment without the cost of having the huge data centers. Making what may be seen as too costly for smaller business, more accessible and attainable. Hadoop is a free, Java based programming framework that allows for the large data sets to be processed on a distributed computational environment. (Nassirian, Abbott, Euwings, deVera. 2015)
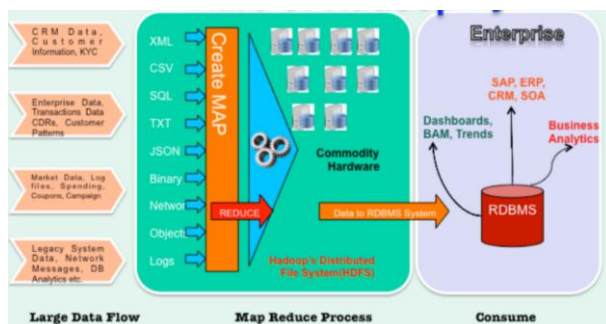

Figure 4 - Basic Hadoop Architecture

Hadoop breaks down big data into smaller blocks, as seen in the figure above, and stores them onto clusters of commodity hardware. Hadoop brings the benefits of computing power, flexibility, fault tolerance, low cost, and scalability to the small companies and allows them to experience all the benefits of big data without very little overhead for storage facilities and high cost of hardware. Computing power is increased by adding more nodes. Not requiring the traditional relational database, the flexibility of not preprocessing data before storage, means you can store data now and use it later. There is no limit to what type of data is stored. The fault tolerance protects against hardware failure. Where there is any failure of a node, the jobs are sent to other nodes to make sure the distributed computing remains active. It also allows for backup of data. Because this is an open source framework, the cost is minimal because it uses commodity hardware. Hadoop is an architecture that brings many benefits to its users and which allows them to have all the benefits of big data on a much smaller scale than those who use data centers. (Nassirian, Abbott, Euwings, deVera. 2015)

BIG DATA COMPANIES

Big data companies worldwide are constantly developing new and innovative ways to gather, analyze and distribute data. Develop more user-friendly methods that are faster and more effective to meet the growing demand for data, analyzation, and distribution. SAS, Teradata, IBM, and Oracle are the leaders in big data hardware and software. They are constantly developing software that can analyze data faster and more efficient with less human interaction.

Though this may decrease jobs for humans, it will better optimize the data that is being collected which will aid in creating other jobs. They are developing more streamline software making it more accessible and economical for smaller companies to utilize big data which will increase the possibility of small business growth. Development of more efficient hardware to decrease energy consumption and increase fault tolerance on equipment all while finding ways to increase volume, velocity, and variety. With the exponential growth of big data, the demand for more effective and efficient equipment and software is huge demand as well.

hiQ is one of the companies that focuses on gathering public data or the general internal data, or some combination of both of them. Using more machine learning technologies that decrease time and lessens the work that humans have to perform which optimizes data. They are also prideful in creating user-friendly data that allows for easier generation of data reports. (Oaks, 2015)

SumAll is a New York based company that specializes on social marketing and focusing on social data. This includes the big companies like Facebook, Ebay, Twitter and other popular companies that generate social data that is usable by multiple companies.

Splunk was originally started to be a log-analysis company that specialized on in visual analytics, but has since partnered with Tableau to aid in reporting tools as well. They also specialize in transactional data as well as customer experience to allow companies to see real time trends. (Oaks, 2015)

Duetto specializes in hotel optimization and search options to aid individuals in searching for hotels. Creating revenue and the ability for consumers to get in on the money making opportunities. They also make it easy for hotels to personalize prices based on demand in certain areas and increased demand in the vicinity of these hotels. Take into consideration other business that may be in the area and other incentives that may be in the area to make their hotel more appealing. (Oaks, 2015)

Ayasdi specializes in taking big data into the health fields. Using big data to find cures for cancer and other disease. By using cancer research and other research to develop software and make it available to health care professionals so they can effective and efficiently develop methods to help fight and find cures for cancer and a host of other diseases. (Oaks, 2015)

Box focuses on virtual storage and is finding ways to optimized data storage. They are looking at storage on large and small scales to cut down on the need for physical storage, which will optimize storage facilities and make data more accessible from anywhere. (Oaks, 2015)

These companies are all working diligently to meet the demand of big data growth and development. Some of them have been around for years, and others are new start-ups but area taking on huge challenges and making strides to bring big data to everyone and make it more economical for all businesses to be able to benefit and tailor big data to their company needs.

EFFECTS ON SOCIETY

Big data is transforming businesses all around the world by allowing companies to better understand and target customers, optimize and improve business processes, improve the health of consumers, maximize security and reduce fraudulent activity, increase business and individual performance, and improve the individual cities and infrastructure. By expanding the parameters of how data is obtained, measured and analyzed, companies are now better equipped to understand their customers and see the behaviors and preferences, allowing them to make adjustments and offer what is in demand based on a larger number of consumers instead of small test groups that don't represent the whole. (Marr, 2015)

Big data allows companies to see trends and patterns which allow them to be prepared for demand on a much larger scale than before. It allows companies to target certain groups of people in certain areas which will eliminate the guessing game that a business in a certain area will be successful. If they know ahead what the demand is in a certain area, they can meet the demand with more confidence in knowing what consumers are looking for. (Marr, 2015)

Big data optimizes and improves business processes by allowing retailers to use predictive models for stock based on social media buzz. Using data gathered from GPS allows companies to optimize delivery routes which helps consumers get the products much faster. Customer reviews on delivered items helps companies adjust packaging

needs to ensure safe delivery of products which reduces waste from damaged goods. Geographic positioning and radio frequency identification are used to track delivery trucks which allows consumers to almost pinpoint where their goods are and when they will arrive, freeing up time to do other things instead of spending all day waiting on packages to be delivered. Big data is also used in house to optimize business processes and monitor energy consumption on site. Data shows how often a facility is used and at what times it is used and can determine when it is important to adjust comfort settings, when to turn lights on and off. Big data can optimize computer and data warehouse performance. Big data is used to prevent failure of equipment by looking at operation logs to see minor changes in equipment and sending alerts when there is abnormal activity which reduces downtime of loss of productivity. (Marr, 2015)

Wearable sensors whether worn for personal preference or doctor ordered, is making consumers more aware of their health and allowing them to do more preventive care than reactive care which is saving lives and bringing awareness that once was never important to consumers. Social media buzz is forcing companies to provide healthy options in food, exercise options for employees, and even bringing awareness to the foods that have been made available to consumers for years. Big data allows health care professionals to pull from huge amounts of data instead of small test groups to see trends in medical advances and treatment options. Taking the guesswork out of medical practice. (Marr, 2015)

Big data is bringing change in security of homes and business and by using crime reports and GPS information to locate high crime areas, allowing law enforcement agencies to place officers in those areas to decrease crime. Identify theft is on the rise, but there are already companies that are seeing ways to limit this type of activity by using trends and data in certain areas to create safer and more useful methods of detecting fraud. Improving card readers to lessen the attacks and provide more security to big box retailers and internet retailers which protect the consumer and the sellers. (Marr, 2015)

Big data is aiding in driving people and business performance, by using educational data to determine where to recruit for certain types of employees based on skill, quota demand, and best educational institutions. Improving customer satisfaction by using customer feedback and ratings to determine where to make adjustments in how they do business. (Marr, 2015)

Big data is used to optimize city roads and energy use. Monitors have collected traffic data in order to optimize traffic flow in cities and highways. Knowing when there needs to be more development based on past data and future predictive models based on industry growth, population growth, and social impact. Also, looking at data collected from environmental reports that allow companies to make adjustments in disposal and safer ways of manufacturing in an effort to protect the planet. All of which affect cities and infrastructure. (Marr, 2015)

These are just a few of the ways that big data is affecting our society and how it is changing. Big data is growing at rapid rates and with the ever changing society there are bound to be new forms of data created that will also impact society. Big data has many useful and powerful methods that can and will continue to improve the quality of life for humanity.

CONCLUSION AND FUTURE STUDY

Big data has transformed the data is captured, analyzed and distributed. There is constant technological advances that will help big data reach its exponential growth in the years to come. Data has evolved from basic text data to be in many forms and with the changes that take place in society data evolving to be able to include more recordable and retrievable data. There are many companies that have played a role in the development of hardware and software to aid in the development of big data. Big data a revolutionized how business operate and how they grow. It is challenging them to develop new business models to stay relevant in the marketplace. Larger companies are realizing that they must incorporate big data into their data models to remain competitive and to keep up with the ever changing society. Big data is not leaving out the small business, there have been software and hardware developed that will allow small businesses to remain competitive. Software like Hadoop, makes it possible to do the same thing that larger companies are doing on a smaller scale. Big data has taken the market by storm and is growing exponentially as well. Predictions say that it will reach $41.5 billion by 2018 which is 6 times faster rate than that of the entire information technology market. Proving that big data is here to stay. Future study should focus in the rapid growth and need for big data versus the rate at which software and hardware developments are being made. At the rate of growth of the needs, it doesn't appear that development of software and hardware will meet the demand.

REFERENCES

1. Oaks, Jessica. *Smart Data Collective*. N.p., 14 July 2015. Web. 17 Nov. 2015. <http://www.smartdatacollective.com/>.
2. Marr, Bernard. *BIG DATA USING SMART BIG DATA, ANALYTICS AND METRICS TO MAKE BETTER DECISIONS AND IMPROVE PERFORMANCE Marr, Bernard (2015-01-09). Big Data: Using SMART Big Data, Analytics and Metrics To Make Better*. West Sussex: John Wiley and Sons Ltd, 2015. Print.
3. Doran, Brandon. *Bright Planet*. N.p., 28 June 2012. Web. 19 Sept. 2015. <http://www.brightplanet.com/2012/06/structured-vs-unstructured-data/>.
4. *SAS*. N.p., 8 Oct. 2014. Web. 29 Oct. 2015. <http://www.sas.com/>.
5. Nassirian, Babak, William Abbott, David Euwings, and Dwight deVera. *Business Intelligence Blog*. N.p., 12 Mar. 2012. Web. 28 Sept. 2015. <http://www.arcplan.com/>.
6. Hendricks, Drew. *Inc.*. N.p., 22 Jan. 2015. Web. 16 Oct. 2015. <http://www.inc.com/>.
7. Evans, Hugo, Marco Ciobo, Jason Miller, and Dan Wall. *ATKearney*. N.p., 8 May 2012. Web. 15 Oct. 2015. <www.atkearney.com>.
8. (Evans, Ciobo, Miller, Wall, 2015)