Data, from Small Solutions to Big Problems

Before the first digital computing machine was invented the concept of a Bit was created. This led to advances in manufacturing and help facilitate the industrial revolution. As this idea evolved onto a digital platform, the bits were grouped together into Bytes. These grouping eventually became large enough to need new terminology to describe them. After time, this electronic information became so large that traditional processing techniques were inadequate. In modern computing, a program will run until completion or if the program takes over a certain amount of time. Big Data will almost always go over the allotted time because of the very large data sets being processed. This requires adaptation of traditional methods for such a large scale.

The first known bit was not electronic, but rather a hole punched through a stiff piece of paper that was fed into a mechanical loom to produce specific patterns by Basile Bouchon in 1732 [1]. This technique was modified a couple times and later applied to informatics and then numeric counting by Charles Babbage. Babbage’s machine is considered to be the first computer because of its ability to process and store data for computing purposes [2]. The encoding of text occurred for the first time with the invention of Morse Code in 1844. This is the first time bits were paired. In 1928 logarithmic bits were introduced. This allowed complex data to be abstracted into almost any number of bit. In the next several decades, problems started to arise as the digital evolution took off. The people in charge of these creations implemented standards and control formats, but some of these standards were not adhered to. Small anomalies started to arise in the behavior as the size and complexities of the machines grew. Fast forward to the 1990’s the creation of the internet. Now more than ever the evolution is in full force and more humans have the ability to modify digital processes and also become more reliant on them, resulting in Big Data. Large pieces of software are now keeping accounts of large fractions of the world population. This poses more problems as the amount of data grows there is a lower ratio of man to machine. . If a person wishes to try to contribute as minimally as possible to the growth of Big Data the most important thing to do is to be aware of what is called a “digital footprint”. Every time you sign up for a new email or create any sort of account online, there are records held in a server database somewhere. The information that an individual places on a website or any other form of electronic communication is known as a digital footprint.

More specifically, Big Data can be classified into three different areas known as the three V’s. The three V’s are volume, velocity, and variety. Examples of Variety are Black Box Data, Social Media records, and Infrastructure Data. Black Box Data consists of flight records such as aircraft components, voice recordings, GPS signals, trajectory, and more. Social Media Data, such as Facebook or Twitter, keep track of information on connections, views, and large amounts of text. The rest include Stock Exchange Data, Power Grid Data, Transportation Data, and Search Engine Data [3]. These information banks are so vastly large that traditional data processing techniques have once more become obsolete. The data contained in these records can be one of three different types Structured, Semi structured, and Unstructured. Structural data, such as SQL, has relationships that tie them together making for better access to records that may be disconnected. Semi structured data has some relational data only for basic organization. Finally, unstructured data has no relationships at all and is used in raw data processing. Big data is important in providing detailed analysis on large scale data sets in real time. This gives companies better decision making facts to optimize performance of the company

With Big Data comes big problems. In some aspects Big Data security concerns its self with most of the same topics as traditional security. There are however, some additional item that need to be taken into account. SQL injection occurs when a person manipulates a data query to bypass security protocols. In a server the data contained is separated into tiers or sections. As these grow it is difficult to fit all of the necessary information in to the tier designated. This leads to something called “auto tiering” of data controllers. In auto tiered data of storage mediums, the controllers for this data will lose track of which tier has what data. Input validation is needed to maintain correctness of records. Real time analysis of security threats can be very difficult to detect as the intruder could acquire a vulnerable connection disconnect from this Big Data center before being noticed. On a traditional computing system these concerns are irrelevant because they can be processed in minutes. If the same methods were used on a Big Data system, the result would take months if not years to receive. One additional problems is access. With such large data sets it is very difficult to find one single item in the million, billions or more records contained in the set. This is not very easily solved but more easily managed with the differing big data types mentioned earlier. The issue of concurrent access is another major problem. That is to say when two or more entities modify the same records simultaneously, which one should be kept. Again no real solution exists but at the moment records are checked for differences and merged together, another technique is to notify and reload the data to the user that attempted to save last [4].

This leads us to our next topic of security. According to DigitalGuardian.com, the largest data breech in history occurred around March of 2012 [5]. A company called Experian had acquired a smaller company called Court Ventures which processed public records for analytics. In the process acquiring this company, there were certain detail that were not taken into consideration. If the proper precaution been taken by fully analyzing the company’s new assets, this entire event would never have taken place. The smaller company “Ventures” held a contract with another company called “U.S. Info Search” which allowed Ventures to search Info Search’s records of sensitive data for court proceedings. After Experian acquired Ventures there was information sold to a number of third parties, one of which included selling access to a company called “Vietnamese fraudster service”, in this deal Experian had inadvertently given a foreign company rights to look up American names, address, social security numbers, and more. The Secret Service notified Experian of what had transpired, but at that point it was too late. There is no way to know for a fact, but it’s estimated that more than 200 million records were “stolen” over the course of ten months.

Education and careful analytics are the key to preventing future disasters. More so than that, the use of proper technologies and techniques will reduce the run time resulting in faster analysis and more accurate results. When technology is handed from on person to the next, even without intention, a careful coder can become responsible for a big mess. Not fully understanding what that person has been given, it is easy to achieve an undesired effect. If this continues who knows what the future holds for the symbiotic relationship between man and machine.