Big Data: A Technology Review

Damon A. Runion, Ph.D.
Research Scholar, Universidad Central de Nicaragua
San Antonio, Texas, USA

ABSTRACT

This technology review paper will examine the rapidly evolving information technology field of Big Data. Detailed coverage of the concept will presented from historical, implementation, and technological perspectives.

Keywords

Big Data, Advanced Analytics, Business Intelligence, Relational Database Management Systems (RDBMS), Management Information Systems, Information Technology.

1. INTRODUCTION

This review provides a solid foundation for understanding the complexity of big data and how it occupies a critical place in the information technology ecosystem. The section starts off examining the rise of big data and its role in modern information technology. Next the review focuses on the qualities of big data with an examination of what “makes” big data. Then the review focuses on the evolution of big data from prior methods and technologies related to data management and analysis. Finally the review defines many of the key technologies and operational methods used in the realm of big data.

2. THE RISE OF BIG DATA

Since 1990, roughly aligned with the advent of the modern Internet, the volume of data all around the world has grown tremendously. On a daily basis, businesses capture trillions of bytes of data about their operations, suppliers, customers, and transactions. In addition, there are millions of sensors in countless locations from phones to cars to appliances capturing immediate, real-time values [1]. Big data – effectively the accumulation of this data – has grown exponentially [2]. A report by IDC in 2011 stated that the overall amount of data in the entire world was 1.8 zettabytes {approximately 2 billion terabytes} [3]. To put such a significant volume of data into perspective consider the following entities that are equivalent 1.8 zettabytes:

- The stored results of every person on the planet having 215 million Magnetic resonance imaging (MRI) scans per day [3].

- Two hundred billion High Definition (HD) movies (approximately 2 hours long) {It would take 47 million years for someone to watch 200 billion movies if they watched movies 24 hours a day, non-stop} [3].

- The storage capacity of about 57 billion Apple iPads with 32 gigabytes of storage. Fifty-seven billion iPads in terms of physical space could do the following:

  - Form a wall that would be 61 feet high and extend from Alaska to Florida

  - Form the basis of a new Great Wall of China that would be just as long as the original but two times taller

  - Form a wall around all of South America that would be 20 feet high

  - Completely cover up to 86% of Mexico City

  - Form a new Mt. Fuji that would be 25 times taller than the real Mt. Fuji [3]

From 2006 to 2011, global data volume grew a striking nine times. The expectation is that at a biennial pace data volumes will continue increase by at least two hundred percent [3].

There is no segment of the economy, organization, or any user of technology that has not been impacted by the wave of increased digital information. Consumers of products and services stand to reap benefits from the application of big data [1]. There are over 30 million networked sensors operational in the retail, transportation, utility, and other sectors. The growth of these networked sensors is proceeding at an astounding 30 percent rate annually [4].

Martin Hilbert and Priscila López studied storage and computing capabilities on a global scale from 1986 to 2007 [5]. Their work showed a significant growth of 23 percent per year over that timeframe, but more interestingly, they found computing capability grew at a faster rate (58% per annum). Their study also examined the effects of increased digitization or the shift from using analog methods of recording data. Calculations based on study data revealed that digital data grew from constituting 25 percent of formatted data in 2000 to an astonishing 94 percent share in 2007.

This massive growth of data volumes has outstripped previously used technologies such as very large database systems (VLDBs) and brought about the utilization of a new term - big data. Big data is not only an extension of traditional database storage in size but also in structure. Most of what is referred to as big data is unstructured data with a potential greater value if analyzed in real time. Such large data sets offer new approaches to answering questions and gaining significant insight that can only come from the comparison and contrast between many occurrences of an event. With new opportunities related to big data, new risks also emerge particularly in relation to the management of the data [1].

Businesses, industry, and government have all begun to appreciate the benefits of exploiting big data. In addition, many public entities have released details of their plans to cultivate big data and develop relevant applications [6]. Big data has become a frequent topic for research and news media coverage. Popular journals like Nature have published extensive special editions on the subject [7]. Such widely dispersed public recognition of big data clearly signifies that we are in the "Age of Big Data" [1].

It makes sense that one the largest sectors of data growth fueling the big data age is the Internet sector. Facebook, the popular social media site, records over ten petabytes of new data every month [1, 2]. Google receives and responds to over 20 petabytes of requests per day [1, 8]. The volume of data has indeed exploded on a global scale. Facebook postings occur globally around the clock due to the ease of use of mobile devices. More powerful mobile devices also make it easier create and upload video content [9]. Quickly appearing data presents one of the main challenges related to big data; what is the best way to collect and manage such widely distributed sets of data [1]?

Recent technological innovations such as the Internet of Things (IoT) and cloud computing have also contributed significantly to the growth of the data. IoT refers to the widespread use of sensors in many devices and environments capturing details of activity at particular points in time. IoT is also crossing over to the domain of household use with the prospect of having household appliances, such as refrigerators becoming self-monitoring, alerting consumers to spoilage or low levels of food items. Cloud computing environments are the ideal place to store the feeds from IoT but with the expected data volumes, information systems used to store that data will quickly be outstripped. Little consideration has been made of the ability to extract value from these new sets of data [1, 4].

3. QUALITIES OF BIG DATA

Big data is a simple term yet the simplicity of the words that comprise the term does not reflect the complexity and depth of its meaning. On its face, big data implies large or burgeoning data sets. While the data sets are big, there are other aspects that define the meaning of big data. At its most basic level, big data refers to sets of data which cannot be managed, accessed, or processed within the "typical" time and resource constraints of traditional software and systems used in the IT industry [1, 10]. This means the data is either too big or not compliant with storage in a relational database management system (RDBMS).

Big data also has different meanings in various contexts. Exploratory scientists look at large volumes of data differently than data analysts or even systems engineers. Data can be the key to a medical breakthrough or provide insight to potential defects in a manufacturing process. To establish a baseline perspective of big data, it will help to look at the technological aspects and factors related to big data [1].

From a technical perspective, a few key conceptual points underscore big data. Big data is recognized by what are known as the three "V's" – Volume, Variety, and Velocity. Volume relates to the size of the data sets. Variety refers to the mixture of types of data elements, structures, and file types such as spreadsheets, free-text, PDFs, video files, etc. Velocity refers to the both the speed at which new data is being generated and the speed at which it can be consumed to satisfy analytical needs. Because of the three "V's", big data does not fit comfortably within established RDBMS frameworks. Structured Query Language (SQL) is used to execute database transactions and requires a standardized method of storage and arrangement of data in order to function properly [1]. There are also physical boundaries to database processing, and when thresholds of data exceed a defined level, query performance begins to degrade as the database kernel either exceeds memory space or processor boundaries [11]. An agreed upon range of data volume that corresponds to big data is data that is between the hundreds of terabytes to several petabytes in size [10]. The research firm Gartner, a trusted independent advisor on technology, offered a clear definition of big data: " 'Big data' is high -volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making" [12].

However, some do not entirely agree with this definition of big data. IDC defined big data as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis” [3]. IDC's definition goes beyond the Gartner definition by including recognition of the value of data. This recognition highlights the fact that capturing data for the sake of capturing it makes very little sense in most circumstances. Big data must be utilized to unearth hidden patterns or relationships within data sets that are created rapidly, contain varying types of content, and are massive in scale. As explained by Jay Parikh, Vice President of Infrastructure Engineering at Facebook, “If you aren't taking advantage of the data you're collecting, then you just have a pile of data, you don't have big data” [1, 13].

The US National Institute of Standards and Technology (NIST) provides the following definition of big data: “Big data is the term used to describe the deluge of data in our networked, digitized, sensor-laden, information driven world. The availability of vast data resources carries the potential to answer questions previously out of reach” [14].

This definition places more emphasis on the technology components of big data rather than business or operational value. As envisioned by NIST, big data will only be able to provide meaningful value if new technologies or methods can be created to capture, manage, and process these large data sets [1].

4. DEVELOPMENT OF BIG DATA

The general idea of big data has its roots in the “database machine” movement of the early 1980's [15]. The database machine was to be a purpose built hardware and software technology unit designed to store and allow analysis of data. The database machine was conceptualized because the limitations of individual mainframe computers became evident to many system administrators [16]. Mainframes are based on the concept of time-sharing or shared resources, with data volumes rising many organizations soon realized that a small pool of users was quickly using more than their share of system resources. This realization led to the concept of a “share nothing” or a standalone parallel database framework [17]. A share nothing architecture provides each database environment with dedicated storage, memory, and processing capabilities. Oracle and Teradata emerged as the commercial leaders in the share nothing database space [1].

Despite the increased size and processing capabilities offered by novel approaches to database management, new issues quickly appeared. The advent of the World Wide Web and the associated query indexing and caching of results led to a rapid growth of databases dedicated to storing data related to those tasks. Google created the Google File System (GFS) and MapReduce software components to address the content storage and access issues related to query technology [18]. Aside from the growth of data as a result of Internet searches, there was also tremendous growth arising from other areas. There was a rise in user generated content, storage of information from networking and telecommunication devices, and widespread transactional data causing traditional data management technologies and frameworks to burst at the seams. In 2011 EMC in conjunction with IDC released a detailed study entitled Extracting Values from Chaos [3]. This study was the first written document that utilized the term “big data.” Both business and higher education recognized the new concept of big data and realized the profound impact it would have on computing and also society [1].

In early to mid-2010, virtually all leading software and Internet companies commenced significant efforts to create powerful big data offerings. Most notably Oracle created the Big Data Appliance, a hardware/software engineered system running an Open Source version of Hadoop [19]. Microsoft, Amazon, EMC, and Google have also spearheaded both internal and commercial offerings to claim a stake in the big data marketplace. In addition, a large number of startup firms sprang up to develop solutions in the big data space. Academic researchers have focused on machine learning and artificial intelligence which highlight much of the big data value proposition [20, 1].

Big data also garnered the focus of many governments all around the world. The United States commenced a focused effort by setting aside US$200 million to seek innovative and highly useful big data solutions. In early 2012, President Obama's technical advisors announced the “Big Data Research and Development Initiative” [21]. Later in 2012, the United Nations published a report outlining the ways in which big data can be utilized to provide greater benefits for citizens [1, 22].

4.1 Technologies Used with Big Data

Big data processing, storage, analysis, and retention rely on a widely dispersed group of technologies and methods which span several disciplines, including computer science, statistics, mathematics, and management. Any organization with intentions to derive value from big data must recognize the fact that an interdisciplinary approach will likely yield optimal results. Much of the processes and frameworks developed for traditional data management with appropriate updates and modifications can be applied to large data. However, new categories and techniques for data management and analysis have been created from principles of advanced computer engineering, mathematics, and statistics. Academic institutions have also spearheaded efforts to create computing paradigms to meet the needs of business and governments.

The realm of technologies related to big data is constantly evolving from both a physical hardware perspective as well as a software perspective. The following section will attempt to detail much of the notable technologies and practices in the realm of big data.

4.2 Software and Systems Used to Support Big Data

Big data is a constantly evolving field. There are frequent enhancements to existing technology and creation of new methods and tools to assist in the management, storage, and analysis of large data sets. The following list of technologies is not comprehensive nor is it exclusively dedicated to big data [10].

Relational database: A relational database is data storage mechanism that orients data storage into a series of tables. A collection of tables is organized into a schema and tables have defined relationships through key columns. Every table contains a data field called a primary key that uniquely identifies the rows in the table. These primary keys also serve as foreign keys in other tables where there is a logical relationship between the two tables. Tables can also have indexes defined as part of their structure. Indexes are used by the query engine to quickly locate records to satisfy a request. Relational database engines along with their disk storage and memory management components comprise a relational database management system (RDBMS). RDBMSs are accessed primarily through Structured Query Language (SQL) or vendor proprietary extensions to SQL such as PL/SQL (Procedural Language/SQL) for use in an Oracle database environment [10].

Structured Query Language (SQL): SQL is a standalone programming language created for the storage, editing, and deletion of data in RDBMSs. SQL was developed using principles of relational algebra and tuple relational calculus. Edgar Codd drafted a conceptual paper, A Relational Model of Data for Large Shared Data Banks in 1970; this work is recognized as the first documentation of what would become SQL. In 1979, Relational Software, Inc., released a version of SQL named Oracle (to which the company later changed its name). This was the first version of SQL available to the public [23]. SQL contains three subset frameworks, data definition language (DDL), data manipulation language (DML), and data control language (DCL). DDL is the set of commands used to create, alter, and delete tables or views. DML are commands used to alter the column level elements of a table for singular or multi-row operations. DCL is a set of security commands allowing users the power to grant other users rights to modify tables and other objects [10].

Business intelligence: Business intelligence (BI) is a group of technologies, architectures, and design methodologies where raw data is transformed or utilized directly to create additional business value through data analysis. BI is typically used to create reports and dashboards comprised of tabular and graphical displays of data that can be manipulated in an interactive manner with list, checkbox, and other selection mechanisms. BI is often utilized to support business processes such as human resource management, financial operations, and supply chain and order management. Many firms and organizations use BI to help refine their operations and gain insight into internal and external operations in order to achieve advantages over their competitors [10].

Non-relational / NoSQL database: A Non-relational or NoSQL (“Not only SQL”) database is a data storage framework using a categorization method not based on tables related by common or shared columns as is used in a relational database. Common types of data structures can be document storage (as in the case of JavaScript Object Notation (JSON) documents), graphs, or key-value pairs. The desire to use a NoSQL approach stems from a few factors ranging from higher availability to a faster capability to scale a growing data set with additional hardware [10].

MapReduce: MapReduce is method of programming used on clustered data sets by distributing an algorithm to subsets of the data for execution or further lower level distribution. MapReduce derives its name from its combined processing model of executing Map and Reduce procedures. A Map procedure executes a filtering routine on a data set or file (such as ordering a list of transactions by products). A Reduce procedure then executes an aggregating mathematical procedure (such as counting the total number of products by each type). While MapReduce is essentially a very simple programming concept, the true value is gained through the parallelization and distributed nature of its operation, thus enabling massive data sets to be processed. The name MapReduce itself was exclusively associated with the Google owned technology for which it was developed; however numerous clones have been developed using alternate programming languages so that the term is now a generic computing process. [8, 10].

Hadoop: Hadoop is an open-source data storage framework based on a distributed system paradigm. Hadoop is managed by the Apache Software Foundation. Hadoop consists of the following core modules:

- Hadoop Common: libraries and inter-module utilities

- Hadoop Distributed File System (HDFS): the controller for data management across commodity machines

- Hadoop YARN: the resource management component applying compute resources to the managed clusters

- Hadoop MapReduce: the programming interface used to introspect and analyze large data sets [10]

Cassandra: Cassandra, a data management framework, is based on the NoSQL architecture and utilizes commodity hardware to create large data stores. Cassandra is open source software managed by the Apache Software Foundation. Cassandra uses a row store methodology for storing data where each table is organized by a primary key with an initial value of a partition key. Table contents are clustered based on the key value, but additional indexing options exist for the core columns of the table. The University of Toronto conducted research which shows that Cassandra is the fastest of all NoSQL databases available for use by the public [24, 10].

Bigtable: Bigtable is proprietary data storage framework built by Google and designed to run on the Google File System. It is operates as a multi-dimensional sorted map with data elements being stored across hundreds to thousands of machines. This framework can scale into the petabyte range. Data elements in the framework are indexed using a set of three elements, a timestamp, a column key, and a row key. This indexing strategy is very similar to the methods used in relational databases [25, 10].

Cloud computing: Cloud computing is the framework and operations of delivering computing capabilities (software, servers, and programming environments) as a service via an open network, the Internet, or over a private network. The main concept of cloud computing is that computing shall be available via an on-demand, as-needed basis like electrical power. It is through the clustering of servers and shared access to an instance that cloud computing provides greater value over traditional on-premise computing models [10].

Data warehouse: A data warehouse is a purpose-built database used for reporting and analysis. A data warehouse pulls data from transactional systems of record, as well as supporting systems, to allow for a unified view of an organization's business. Occasionally a data warehouse is referred to as an Enterprise Data Warehouse (EDW). Data warehouses are often used to create reports for senior management focusing on the overall health of lines of the business. Data warehouses also excel at linking data together that corresponds to multiple facets of a transaction, such as inventory on hand, shipment delays, product defects, sales effectiveness, and customer satisfaction. Data is loaded into a data warehouse using data movement logic called an Extract, Transform, and Load or ETL routine. Frequently an operational data store (ODS) is put in place before the denormalization of data going into the data warehouse to allow for more transaction-oriented reporting [10].

Data mart: A data mart is business unit or departmentally focused database used for analysis of transactional or other data via business intelligence tools. Frequently a data mart is a subset of a data warehouse, but occasionally data marts are deployed in a standalone manner with the sponsoring department responsible not only for the data in the system also the hardware and software [10].

Excel: Excel is a central program in the Microsoft Office Suite. Excel is primarily a spreadsheet application, but has grown over the years through its ease of use to be a primary data analysis tool. According to a 2012 survey of 798 statistics and analytics professionals by KDNuggets [26], 29% reported using Excel as their primary analysis tool; this made it the number 2 tool of choice. Excel can also be extended using the Microsoft supplied Analysis ToolPak to allow it to perform a range of statistical tests, including Analysis of Variance, Correlation, F-Test, and Regression [1].

Extract, Transform, and Load (ETL): ETL refers to a computing process that is part of database administration responsibilities. ETLs perform the following tasks: 1) Extract data from a system of record or another system where transactions are stored. 2) Transform data to resolve errors or join elements together that are commonly used in conjunction with analysis. 3) Load data into a target database schema such as a data mart, an operational data store, or an Enterprise Data Warehouse [10].

Metadata: Metadata fundamentally is data about data. While sounding ambiguous, metadata consists of the descriptive elements that categorize a set of data or data elements for another purpose. Metadata exists in two main forms. One is structural metadata. Structural metadata defines the types and format of data such as the database column type or parameters of a field in a table. The other form is descriptive metadata. Descriptive metadata offers categorization or other types of qualities about the actual data content of a cell. An example of descriptive metadata element would be the classification of certain telephone records as being from a specific geographic location [10].

R: R is an open source programming language used to perform statistical analysis on a wide variety of data sets. R has become the analysis tool of choice for most statisticians and data miners due its simple interpreter framework and low cost of ownership. R supports almost all statistical techniques. According to a 2012 survey of 798 statistics and analytics professionals by KDNuggets [26], 31% reported using R as their primary analysis tool; this made it the number 1 tool of choice. Given the dominance of R in this space, database vendors like Oracle and Teradata recently added R support to their product offerings [1].

Semi-structured data: Semi-structured data is a form of data that does not readily comply with the requirements of relational databases in terms of size, format, and volatility. Semi-structured data does however have characteristics that allow categorization and placement into hierarchies. It does however have internal metadata qualities that allow for rapid record identification. Two examples of semi-structured data are XML documents and JSON documents [10].

Structured data: Structured data is a form of data stored in a fixed format. This allows the data elements to be accessed in a declarative manner and allows the creation of indexes or pointers to rapidly find particular data values. Relational databases and spreadsheets represent the two most common forms of structured data [10].

Unstructured data: Unstructured data is data without a formal structure or assigned index. Unstructured data must be analyzed in its entirety to be categorized. Some examples of unstructured data include email messages, comment fields on forms, or untagged video and audio [10].

4.3 Procedures, Tests, and Methods for Big Data Analysis

Most of the procedures, tests, and methods used to perform analytics on big data rely on computer science and statistics. The list below is not exclusively oriented toward big data and many of the entries are proven standard statistical techniques [10].

Statistics: Statistics is a scientific branch of inquiry related to the collection, organization, analysis, and presentation of data. Statistics is closely related to mathematics although it is regarded to be a separate field in its own right. Statistics is concerned with determining the relationships between factors known as variables and how strongly relationships between variables impact each other. Statistical significance is considered to be the case when one factor drives the outcome of another factor [27, 10].

A/B Testing: A/B Testing represents a statistical testing procedure where a control group is measured against one or more test groups to evaluate what interactions will affect a variable of interest. A/B Testing has been used quite heavily in e-Commerce and is usually focused on what types of web presentation techniques influence consumer decision. Analysts have looked at layouts, image placement, colors, and fonts as influencing parameters [10, 28]. In the context of big data, A/B Testing can be used to rapidly test a large set of data by quickly creating groups split along a small set of values. Using other statistical methods, analysts can ensure their decisions are statistically significant and in turn validate a large data set.

Association rule learning: ARL is a statistical testing method where groups are analyzed using a variety of algorithms to create expected outcomes or rules. ARL is considered to be a data mining procedure [29]. One of the most common association rules is market basket analysis, which is a routine typically used by retailers to determine if there is any association between the items customers purchase together. One goal of ARL is to identify patterns to generate promotional efforts based on product affinity [10]. In the context of big data, ARL can help identify clusters or groups within larger populations for segmentation and further analysis.

Classification: Classification is a set of statistical procedures utilizing a small set of data, called a training set, to group large data sets along desired categories. A statistician will create a training set using specific parameters. Then the training set is applied to a larger data set using a rules-driven engine. Classification is most commonly used with data sets related to customer decisions and behavior. Standard tests assess churn rate, buying patterns, frequency of service requests, or rates of product returns. Classification in the context of automation is referred to as a supervised learning process since the operator provides a training set and the process does not examine the data set on its own to identify categories [30, 10].

Cluster analysis: Cluster analysis is used to classify a large and diverse set of elements into smaller similarly organized units. Cluster analysis is similar to classification, but it does not utilize a training set and data elements are repeatedly sorted and examined by processes until natural divisions begin to appear. Cluster analysis is most commonly used for customer segmentation when the volume of customers is quite high [31, 10].

Crowdsourcing: Crowdsourcing is a method where interested individuals are contacted, typically via social media, to participate in or contribute to a project or research effort. One example of crowdsourcing is an effort by Lego Corporation that allows customers to submit new ideas for toys. Once an idea receives 10,000 votes it moves to an internal corporate review process [32]. In the context of big data, crowdsourcing can result in vast amounts of rapidly collected data which in turn creates storage and processing demands [10].

Data fusion: Data fusion is a process where data from multiple sources is combined and analyzed to detect patterns or uncover hidden qualities. One example is the process of taking data from a social media platform such as Twitter and correlating a spike in tweets with a rise in sales of a particular item [10]. Data fusion is an especially well-suited practice with respect to big data because it can help provide value through rapid insight into heterogeneous data.

Data mining: Data mining is a subset of computer science in which large data sets are analyzed for patterns using techniques such as machine learning, statistics, and artificial intelligence. The end state of a data mining process is a reduced set of actionable information in a consumption ready format. A frequently discussed application of data mining is credit card fraud detection. This type of data mining occurs through the deep analysis of spending habits of a customer, including factors such as location and amounts spent [10].

Ensemble learning: Ensemble learning is the process of joining together several predictive models to further extract data from a complex data set. The individual predictive models are derived from machine learning or statistical models. Ensemble learning is regarded as being a supervised learning activity [10, 33].

Machine learning: Machine learning, a subset of artificial intelligence, focuses on the creation and analysis of programs and systems that can learn and provide output for decision-making. An example of machine learning would be the process of training a system to determine if email messages are spam messages or genuine email traffic. Once tests have confirmed reliability, a routine can be implemented to automatically process messages and filter messages that are spam into a separate email folder [30, 10].

Natural language processing (NLP): NLP is a subset of artificial intelligence that draws heavily on the field of linguistics. NLP examines the ability of computers to understand and interact with humans in a verbal form. One goal of NLP is to minimize the degree of manual interaction a human must have with a computer, limiting it to a series of spoken commands. Another goal of NLP is to allow computers to converse with humans in a natural manner. Voice prompting in the case of fielding initial calls to a customer service line is an example of NLP [34, 10].

Artificial neural networks: ANNs are software-based models based on the structure and activity processes of living beings. ANNs are developed to mimic the processes of the brain and seek to validate advanced machine learning and recognition of visual phenomena such as pattern recognition. Depending on the complexity of the ANN and the desired tasks, an ANN can represent a supervised or unsupervised learning framework. An example of an application built using an ANN would be a program designed to detect fraud in Medicare claims [35, 10].

Social network analysis: Social network analysis corresponds to the application of electronic network theory to establish a pattern of relationships between two or more participants in social networks. Through social network analysis, individual people are considered to be nodes in a web of relationships connected by ties to the environmental framework. Nodes are evaluated by their thickness to depict values such as counts of relationships or length of time associated with those relationships [10, 36].

Predictive modeling: Predictive modeling is where a structured set of rules are applied to a data set to create a model for repeated testing. Predictive modeling strives to determine a particular outcome or evaluate how likely an outcome might appear in future situations. Examples of predictive modeling include random forests, Naive Bayes, and majority classifier [37, 10].

Regression: Regression is a form of statistical testing used to evaluate the relationships among variables. Typically regression is used to measure the effects that independent variables have on one or more dependent variables. This analysis allows a researcher to reach a level of statistical confidence that an incident is caused by another factor. Regression is frequently used in manufacturing processes to identify input parameters that could be contributing to defects in outputted products [10].

Sentiment analysis: Sentiment analysis is a data analysis framework where text analysis and natural language processing are used determine subjective or emotional meaning in a data set. Sentiment analysis is used widely with data sets pulled from social media platforms. In this context, a series of tweets or postings are analyzed to determine the overall feeling regarding a product or statement. For instance, a cell phone vendor may use sentiment analysis to gauge customer reaction to plans it recently announced for a phone with a new advanced video recording feature [10].

Spatial analysis: Spatial analysis is the application of statistical techniques to data that has been encoded with geographical qualities so that relationships can be measured in space and time. Geographic information systems (GISs) store and manage geographic properties of homes, retail establishments, and/or public places. Spatial analysis utilizes data from a GIS along with other data sets such as product sales and applies techniques like regression to determine the relationships between elements such as purchasing habits and residence location. For instance, individuals who live close to a swimming pool might be more likely to purchase sun block or swimming accessories at the start of the swim season [10].

Supervised learning: Supervised learning, a machine learning process, is where a program or procedure is taught how to rank or measure a set of data through the use of a training set created by a human. Supervised learning is frequently used for refined classification of items where a number of subjective factors are used to determine the desired groupings [38, 10].

Simulation: Simulation is a method of modeling the outcome of a series of interrelated events with a wide degree of permutations or options. Simulations are typically conducted a multitude of times with slight modifications made to variables with each run. One commonly used set of simulations are Monte Carlo simulations. Monte Carlo simulations are a set of algorithms using repeated random sampling following differing assumptions to arrive at a consistently occurring phenomenon. The output of simulations is usually provided in a histogram which allows for rapid visual identification of an outlier or leading candidate [10].

Time series analysis: Time series analysis, a data analysis procedure, is based on analyzing a series of entities over a set period of time in order to detect trends or causes behind certain occurrences. Time series analysis can also be used in a forecasting manner where a repeated series of concurrent events can be used derive a statistical significance of recurrence. Time series analysis is frequently used in sales management environments to measure and predict the flow of a sales cycle. It has also been used in the life sciences arena to predict and determine the progression of both illnesses and treatments [10].

Unsupervised learning: Unsupervised learning, a type of machine learning, is when an algorithm is used to analyze a data set. The algorithm develops groupings and structure on its own rather than through the use of a training set inputted by a human. One example of unsupervised learning is cluster analysis [10].

Visualization: Visualization is a key presentation option when dealing with big data. Visualization is the process of using charts, images, diagrams, or other animated displays to convey details about a set or subset of data. Visualizations help overcome human limitations with respect to analyzing large data sets. Visualizations can be used to identify outliers or show the relative intensity of one data group to another [10]. Some commonly used examples of visualizations are as follows:

- Tag cloud: A tag cloud, as shown in the image below, is a list of terms in a data set with the most commonly occurring terms displayed in larger font [10].

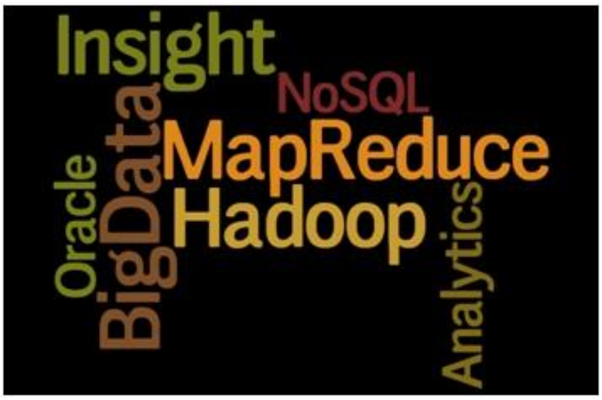

Figure 1: Example of a Tag Cloud

- Scatter plot: A scatter plot is diagram where data is displayed as a grouping of elements aligned with a measure on each axis. Occasionally color or another indicator will be used to show intensity thus becoming a third axis. Scatter plots are very useful with large data sets because they clearly highlight outliers or extremely different cases [10].

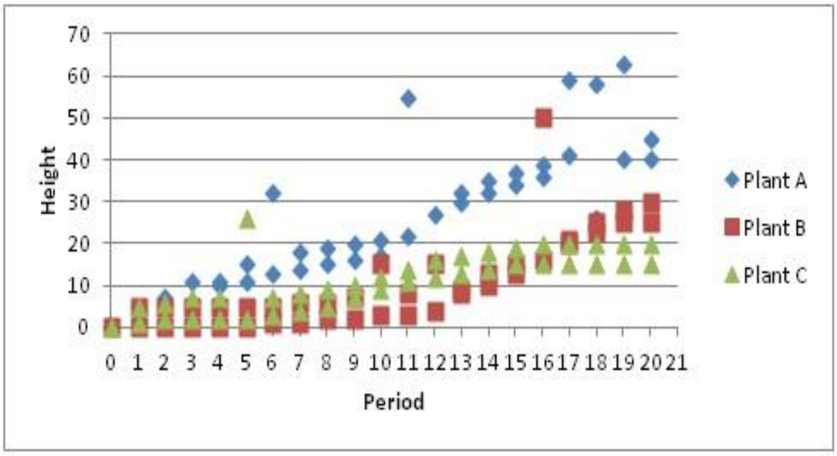

Figure 2: Example of a Scatter plot

5. CONCLUSION

Big Data is a rapidly evolving information technology field with solid roots in established computer science and information system practices. As organizations and governments continue to recognize the advantages to be gained through the use of Big Data and Advanced Analytics the expectation is that further developments and enhancements will occur.

6. REFERENCES

[1] Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19(2, 171-209.

[2] Cukier, K. (2010). Data, data everywhere: A special report on managing information. Economist Newspaper.

[3] Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1-12.

[4] Chui, M., Löffler, M., & Roberts, R. (2010). The internet of things. McKinsey Quarterly, 1-9.

[5] Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science, 60-65.

[6] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

[7] Nature. (2008). Specials: Big Data. Retrieved from Nature: http://www.nature.com/news/specials/bigdata/index.html

[8] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[9] Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt.

[10] Manyika, J., Chui, M. B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. New York: McKinsey Global Institute.

[11] McLellan, C. (2013). Big data: An overview. Retrieved from ZDNet: http://www.zdnet.com/big-data-an-overview_p2-7000020785

[12] Gartner. (2013). IT Glossary - Big Data. Retrieved from Gartner: http://gtnr.it/1rP0ANn

[13] Constine, J. (2012). How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day. Retrieved from TechCrunch: http://bit.ly/1nKWiHM

[14] National Institute of Standards and Technology. (2014). NIST Big Data Working Group. Retrieved from National Institute of Standards and Technology:: http://bigdatawg.nist.gov/home.php

[15] Su, S., Chang, H., Copeland, G., Fisher, P., Lowenthal, E., & Schuster, S. (1980). Database machines and some issues on DBMS standards. Proceedings of the May 19-22, 1980, national computer conference (AFIPS '80) (pp. 191-208). New York: ACM.

[16] Goda, K. (2009). Database Machine. In Springer, Encyclopedia of Database Systems (pp. 714-714). New York: Springer.

[17] DeWitt, D., & Gray, J. (1992). Parallel database systems: the future of high performance database systems. Communications of the ACM, 35(6), 85-98.

[18] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[19] Garlasu, D., Sandulescu, V., Halcu, I., Neculoiu, G., Grigoriu, O., Marinescu, M., & Marinescu, V. (2013). A big data implementation based on grid computing. Roedunet International Conference (RoEduNet) (pp. 1-4). Sinaia, Romania: Institute of Electrical and Electronics Engineers.

[20] Bryant, R., Katz, R. H., & Lazowska, E. D. (2008). Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. Washington, DC: Computing Research Association.

[21] Lazar, N. (2012). The Big Picture: Big Data Hits the Big Time. Chance, 47-49.

[22] United Nations. (2013). Big Data for Development: A primer. New York: United Nations.

[23] Chamberlin, D. D. (2012). Early history of SQL. Annals of the History of Computing, IEEE, 78-82.

[24] Hewitt, E. (2010). Cassandra: the definitive guide. Sebastopol, CA: O'Reilly Media, Inc.

[25] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M. .., & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS), 26(2, 4).

[26] KDnuggets. (2012). What analytics data mining, big data software you used in the past 12 months for a real project? Retrieved from KDnuggets: http://bit.ly/1s7epZB

[27] California State University - Long Beach. (n.d.). PPA 696 RESEARCH METHODS TESTS FOR SIGNIFICANCE. Retrieved from California State University - Long Beach: https://www.csulb.edu/~msaintg/ppa696/696stsig.htm

[28] Martin, B., Hanington, B., & Hanington, B. M. (2012). Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions. Beverly, MA: Rockport Publishers.

[29] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207-216.

[30] Alpaydin, E. (2004). Introduction to machine learning. Cambridge, MA: MIT Press.

[31] Bailey, K. D. (1994). Typologies and taxonomies: an introduction to classification techniques. Chicago: Sage.

[32] Schoultz, M. (2014). Lego Innovation: An example of Crowdsourcing Design. Retrieved from Digital Spark Marketing: http://bit.ly/1rM1zhc

[33] Oza, N. C., & Russell, S. (2000). Online ensemble learning. AAAI-00 Proceedings, 1109.

[34] Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

[35] Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks. Cancer, 91(S8), 1615-1635.

[36] Marcus, S., Moy, M., & Coffman, T. (2007). Social network analysis. In D. Cook, & L. Holder, Mining Graph Data (pp. 443-467). Hoboken, NJ: Wiley.

[37] Dickey, D. A. (2012). Introduction to Predictive Modeling with Examples. Proceedings of 2012 SAS Global Forum (p. 337). Orlando, FL: SAS.

[38] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.