***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

Abstract

This paper revolves around the research behind “Big Data” and addresses the main problem with assuring the quality of incoming data. What I hope to accomplish at the end of this research is to develop an algorithm that automatically schedules data cleansing while minimizing the amount of idle time a server has to undergo for this process.

Background

The term “Big Data” is a very a critical in the eyes of most companies nowadays. It is the biggest game changer in data analytics where it’s considered an asset for every organization to possess to gain that competitive advantage. The collection and analysis of data is valuable to every organization because it gives them the opportunity to depict a trend in consumer investments and therefore align what the company offers to what the customer’s needs. Even though “Big Data” emits positive results with higher returns, it still has its pitfalls. Since there are several examples, the main focus of this paper involves the difficultness of assuring high data quality during the process of collecting and analyzing large amount of datasets. Clean data can be defined as removing any erroneous data entries in which data duplication proves to be the common area of struggle for scientists. This has been the biggest struggle for every company, in fact it was estimated that it cost “… U.S. businesses more than $600 billion a year” as stated by The Data Warehousing Institute [1] due to the infinite number of data quality problems. There has been many platforms that manage “Big Data” like Apache Hadoop, Data Warehouse, MapReduce or MongoDB but even with the greatest frameworks available, there are many ways data can become erroneous due to the fact that it is always changing simultaneously.

Many companies, especially ones that involve a large number of consumers or shareholders, have a difficult time depicting trends due to the unorganized information that is being received and stored in the servers. To counteract against this problem, many in the scientific community have been developing algorithms and or programs to detect these anomalies. Even though they have been proven to increase the level of efficiency and effectiveness in data normalization, there are considerations to consider which will be discussed later on in this paper.

The Problem Statement

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

Existing algorithms have been developed by many data scientists that concentrates on the idea of performing data cleansing tasks. These solutions had partially answered the main problem but doesn’t make it as efficient as it should be. The algorithm’s scope only covers for manual data cleansing and as a result, it increases the server idle time which goes against company policy in terms of “reliability”. This has been a continuous struggle for many to defeat because of the factors that needs to be considered when making this the most cost-efficient way of getting this process integrated.

Main Problem to Address: Can we assure high data quality is being met when collecting and analyzing information while maintaining consistency on a daily basis?

Sub Problems

1. ***Energy Consumption***
   To consider the toll on energy consumption for a single host server when processing various amounts of un-normalized datasets when running this algorithm.

2. ***Expenditure Scope***
   To consider the expenses (human factor + money factor) on a single host server when processing various amounts of un-normalized datasets when running this algorithm.

3. ***Amount of server offline time***
   To consider the amount of time a single host server has to go “offline” when processing various amounts of un-normalized datasets while running this algorithm.

Hypotheses

1. ***Energy Consumption***
   The amount of energy a single host server requires will decrease as the algorithm divides huge datasets into smaller increments to process.

2. ***Expenditure Scope***
   The scope (human factor + money factor) will increase as more highly trained data scientists will be needed in order to maintain the performance of each server.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

3. ***Amount of server offline time***
   The amount of offline time a server faces will decrease when running this algorithm during the experiment.

Delimitations

a. This proposal will not include the use of Apache Hadoop as it requires an extensive amount of knowledge and experience that I do not possess.

b. The development of this algorithm will be derived and integrated from credited Data Scientists that base their research on existing algorithms.

c. The extent of this proposal is to only evaluate and formulate an algorithm, but not deploying in an environment best suitable for this situation.

d. This proposal will not use any physical hardware (Data Servers, sufficient processing power, sufficient memory, and sufficient bandwidth) that represents a data center as I do not have the funding nor the resources available to me.

e. This proposal requires the formal request to have multiple data scientists work in a cross functional team. Given the constraints with time and limited knowledge, this will only be conducted with the main researcher of this paper.

f. This proposal will be limited to data that is only captured at real-time to endure the realistic view of how this algorithm will be adapted in a data center environment.

Terms and Definitions

1. ***Clean Data*** – Datasets that doesn’t contain any of the following: duplicate data entries, incorrect information per entry, improper data value format, proper relationships between the main identifier and its child values that represents an entity, and null/missing data bits that are required.

2. ***Data Normalization*** – A fully formed database that contains multiple tables and values that all correlate to one another which is also known as a “relational database”. Each

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

record or row represents a value of a specific entity (can be a consumer) which can tie to other table values that an individual can also possess.

3. ***Erroneous Data*** – Data entries in a dataset that are the wrong data type (ex: A text value is placed in a cell that is formatted to only contain numeric values).

4. ***Data Servers or Database Servers*** – A physical machine that is made up of both hardware and software, used to run a functional database. Mostly found in data centers that every organization possess where they store any information that is valuable to the company.

5. ***Offline Servers*** – In order for a database to be backed up or altered, data servers must be in a state of “shut down” or “offline” in order to protect the integrity of that data.

6. ***Apache Hadoop, Data Warehouse, Map Reduce and MongoDB*** – Open-source software platforms that is integrated with specific frameworks to extract, transform and load (ETL) data into database servers. These platforms are widely used by data scientists everywhere due to the effectiveness of data manipulation, but is very complex to use and or understand.

7. ***Apache Spark*** – A similar open source software platform just like Apache Hadoop, but the key difference is the fact that it is much simpler to learn and use than the others available on the market. It can also process, analyze, and manipulate data faster than Hadoop, but cannot process as much data as Hadoop.

8. ***VMware*** – Is a virtualization software that allows a host to run multiple virtual machines that can act like a server.

9. ***Reliability (Corporate Policy)*** – As a client-tell company, it is important to maintain the server uptime (a server being online for 99.9999999% of the time) since business transactions are always being communicated on a daily basis. Manipulating data means shutting the server off, but has to be fast enough that it doesn’t interfere any business flows.

Assumptions

1. Will have a fully equipped server with proper specifications to run algorithm and process large amount of datasets.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

2. Apache Hadoop integrated with Map Reduce will not be used due to the complexity of the platforms.
3. Apache Spark should be ready for install and the tools integrated in the software is up to date.
4. If need be in terms of expanding physical hardware, material is being provided for.
5. Datasets given are taken via real-time and hence un-normalized
6. The time it takes to process the datasets should be a fraction less than the algorithms developed by scientists mentioned in other literatures.

Importance of Study

Ever since Big Data became a valuable asset to obtain, organizations gained numerous advantages in building better relationships with their consumers. Even though it’s considered a gold mine, there are many challenges that data scientists face on a daily basis. Since datasets come in large volumes due to the continuous flow of business transactions, the main struggle is keeping the data clean and consistent for business leaders to use when generating monthly numbers. The algorithms formulated by researchers in the scientific community definitely tackles this specific problem. However, it still cannot overcome the process of automating this form of data cleansing to reduce the time a server must go offline (The negative outcomes of an offline server was discussed earlier in the “Terms and Definition” section). My proposal stresses on the idea of scheduling automated data cleansing where it breaks the datasets into small increments to reduce the amount of power and resources needed. The main benefit of this technique is that it reduces the amount of data to clean which reduces the amount of energy consumption and offline server time. These reductions sums up to less expenses being allocated in order to maintain the integrity of the data that is constantly changing and the integrity of the organization’s clients.

Literature Review

All of these articles I reviewed, revolves around my research question: What could we do to assure data quality is always met when collecting and analyzing information? Proceeding forward, the next session of the literature review will expand on the methodology and conclusions stated by credited researchers of whom wants answers to that big question.

The first aspect to my research question is the amount of software frameworks available to data scientists that make cleaning data efficient and effective like Hadoop, MapReduce, or MongoDB. With an infinite amount of data to handle, using commercial platforms is a common practice done by data scientists to organize, analyze, and manage databases. Researchers

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

suggest companies invest more time and effort with using Hadoop, which is a widely used tool that works side-by-side with data warehousing (cyber warehouse that stores relational databases) and Map Reduce (a program that extracts data and normalizes them). In respects to a group of researchers, algorithms have been deployed using Hadoop crossed with MapReduce where sample datasets were processed through a single node machine and results shows that it was able to analyze huge amount of clusters (unstructured data – sample size contain 10,000 data fields) within an efficient amount of time [2] [6]. To compliment that experiment, Barna and Divesh made a great point of how difficult it is to maintain clean data if it wasn’t for licensed frameworks (Hadoop, MapReduce, or MongoDB) and the best way to understand the situation is by viewing it from a perspective of the world-wide web. It was stated that 14% of a single webpage is not well-formed. Webpages extend to the millions due to a variety of ecommerce sites and if it isn’t well-formulated, data collection may not be as consistent as one would hope [8]. The point made here was the fact that since raw data is never skewered to perfection and always changing, analytical tools must be used to see the pattern of erroneous information and perform bulk changes. Most researchers articulate on the use of Hadoop for specific advantages that includes: exceptional handling of raw, unstructured and complex data, and the flexibility to program in other languages. On the opposite end of the Hadoop spectrum, it’s extremely difficult for others to learn the full use of the framework, and the fact that there are not enough data scientists out there that carry the expertise to operate it [3] [7]. Due to my knowledge and skill based limitations, I will not be using Hadoop in carrying out my research as it requires expertise in the field of Data Science of which I do not possess.

Another aspect to my research question discussed my numerous scientists and researchers is the ideology of expanding the project scope in terms of resources. Miryung Kim and 3 other researchers conducted an experiment where they interviewed 16 data scientists from eight different Microsoft organizations and discovered that even working with the brightest minds is still not enough. The fact is that there are few too little data scientists out there which is the effects of “Data Science” not being a popular field of study [5]. Since data mining is bigger than ever now that everything revolves around the growth in technology, it gets more complex to maintain its quality while limiting on the usage of time and energy. The interview conducted was definitely credible since all members of the survey belong in the same community and raise similar demands with bringing more scientists into their teams. Another point made by Archana and Yanpei expanded on increasing the number of scientists in a team that possess different expertise and at the same time, forming a centralized team out of it to increase the quality of work rather than spreading it out in different departments [3]. Even though frameworks are provided for the researchers, it’s still a complex platform that not most engineers and scientists possess which can slow down the process of maintaining such large amounts of datasets that makes up databases. A group of researchers (Ken, Xiaohua, and Kui) stood out to me where it was mentioned about outsourcing to another third party service. I somewhat agree with in a way where the best thing is to outsource to third party cloud servers that can remotely integrate, analyze, or deal with data anomalies. This method would save companies from using

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

internal resources like the man power and the tools necessary for data cleaning [10]. However in today’s practices, data confidentiality is an important value to keep in mind and most organizations would like to protect their consumers in any way possible (which means not disclosing consumer information to any party not acknowledged during the agreement). Increasing the amount of data scientists in the industry is a simple answer to my research question where it was concluded from others that students at high schools and universities should be involved in the world of “Data Science”. I agree with that this is something great to consider, but poses a lot of challenges since not a lot of institutions teach “Data Science” and the fact that it requires a lot of expertise makes it more difficult for people to pick up on as mentioned by Mark Gibbs in his article explaining the main challenges of Data Analytics [4].

The last aspect which cross paths with the base of my research involves formulating algorithms to reduce the workload in the maintenance of data quality. Ikbal, Hadeel, Mohamed, Rachida and Chafik proposed an algorithm called the “BDQ Evaluation Algorithm (Big Data Quality)” which extracts raw data, transforms it through sets of metric conditions that needs to be met depending on the desired outcome, and then is queried into filtered datasets [9]. They conducted an experiment where they used the BDQ Evaluation Algorithm along with the proper MapReduce Framework, and analyzed datasets of the 6441 participants in a Sleep Heart Health Study (SHHS). This resulted in the finding of data anomalies and was then filtered out to create a legitimate normalized dataset. Instead of other solutions involving expansion of data scientists or the in-depth use of a platform, this method of creating algorithms that does the job of analyzing data deems as a more effective and efficient way of maintaining the quality of data. How this correlates to the other solutions is the fact that it integrates the use of the software framework, MapReduce, and teams up data scientists together that carry a variety of expertise and experience. The only constraint this methodology can come across is the amount of time and effort to develop and test these algorithms which means developing cross functional teams of data scientists with a high level expertise. As mentioned before, there are not enough data scientists out there to do this job which makes it difficult to proceed with future development of this method.

After reviewing a lot of the literature presented by credited researchers, I found that they lacked the idea of “automating” data analytics. My research will look at the different factors of data mining, capitalize on having a scheduled based automated program that audits and clean data to simplify the analytical process performed by Data Scientists. The end goal of this proposal is to develop or build off on existing algorithms that auto-detect data anomalies and corrects them without having someone manually normalize the datasets which can reduce the project scope for each department in any organization. The challenge I foresee coming is my lack of experience, and expertise in data analytics. To counteract against this unbalance, I will derive my research from credited researchers or data scientists on existing algorithms and evaluate them for potential future expansion.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

The Research Methodology

The Sample

For the purpose of this specific research, the sample being considered would be classified as probability sampling using cluster sampling. Datasets would be collected from different organizations across the province where the collection of data per company varies on their consumer needs. Each sample is made up of entirely different sets of data including the actual values and the way they are formatted (Example: A name of an individual is a value and is formatted as a text). This randomness would allow my algorithm to process various entries that can make it more prone to sporadic datasheets. Each collected dataset should contain a minimum of 5000 entries and a maximum of 10 000 entries to simulate real-time data flows that occurs from continuous business transactions in an organization. The population consists of organizations who depends on business transactions to make forecasts of consumer trends. Since the number of companies in BC can range from a few hundred to a few thousand and given the limited resources, the sample can scaled down to 10 different organizations.

The Instrument and How It Is Used

The main researcher does not have the main resources available that are widely used for testing all of which includes the following: a single database server, Apache Hadoop, and an immense amount of CPU and memory. In this scenario, the next best tool available would be implementing the algorithm using VMware with an Intel i7 processor, 16GB of RAM and 2TB of hard drive space. This will be used to host the framework “Apache Spark” (Refer to “Terms and Definition”), which is not used for commercial purposes that has the intention of handling extremely large amounts of data. After the algorithm has been formulated and analyzed, it would be deployed in the framework that would extract the datasets given from each organization in the sample, and then manipulated to formally cleanse the data. The other benefit of using Apache Spark rather than Hadoop are the built-in tools that Spark provides to make this a much easier process when using the platform, which benefits the researcher who has minimal experience with data analytic software.

Data Collection and How It Is Applied

The data required for this study involves random datasets from the 10 different participating organizations in BC. Each dataset should be un-normalized and full of erroneous entries to simulate how data centers receive information on a daily basis. The main focus of the results includes the following: CPU usage (GHz - Gigahertz) per time increment, energy consumption in (W - Watts) per time increment, memory usage in (% - percentage used) per time increment, and the time it takes (seconds per cell in a datasheet) to extract the data, analyze and manipulate it. The time it takes to process these datasheets is extremely important to the research as it gives a definite answer to whether or not running scheduled automated data cleansing makes a bigger difference than the current way of analyzing this information.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

The datasheets will be processed using Apache Spark as mentioned earlier and will produce a variety of numbers that will be sorted in a graphical representation. This would allow fellow readers to visually compare and conclude if this algorithm proves more beneficial than others. Of course, the datasheets collected from each organization would have fallen under the agreed terms between the board and the head researcher as the information involves a lot of legal actions that protect an individual’s privacy. The licensed agreement hasn’t been written yet due to time constraints, but it must be approved and consented by both parties before any research begins.

Data Being Applied to Support each Sub-Problem

The section mentioned earlier explains what is being collected to run the experiment and how it is being processed. We will now review the true meaning of the data results and how it correlates to each sub problem.

1. ***Energy Consumption***
   To consider the toll on energy consumption for a single host server when processing various amounts of un-normalized datasets when running this algorithm.

   a) *Data Collection and How it Will Be Treated*
      The data required for this specific sub problem will focus on the total energy consumption (measured in Watts) per time increment. When the datasets becomes larger in size and are more unorganized, it takes more energy and power to process everything in one take. This is can be seen as a cause-and effect relationship, or more specifically a positive correlation. The sum of the energy consumption will be factored from the start of this experiment till the normalized dataset is outputted in another file. After the results are collected and visually represented, the mean will be calculated to get the average use of energy per data entry in a given dataset. To support this sub-problem, the correlation coefficient between the size of the dataset and the amount of power it takes to process each dataset will be calculated. This should show the strength of the relationship between the two. The results can conclude that if datasets become larger in size and are more unorganized, it would require more processing power and time to process them in order to produce accurate results. Increasing these factors means an increase in the consumption of energy, which can correlate to sub-problem 2. Adding to this experiment, the researcher will perform a hypothesis test to compare between his newly developed algorithm and the existing algorithm. This will demonstrate whether or not this new and improved method is the best fit in today’s given environment. This testing will result in either accepting or rejecting the hypothesis in regards to energy consumption.

2. ***Expenditure Scope***

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

To consider the expenses (human factor + money factor) on a single host server when processing various amounts of un-normalized datasets when running this algorithm.

a) *Data Collection and How it Will Be Treated*

The data required for this specific sub problem will focus on the total CPU usage (GHz - Gigahertz) per time increment, and the total memory usage in (% - percentage used) per time increment. When the datasets becomes larger in size and are more unorganized, it takes more CPU and memory to process everything in one take. This is can be seen as a cause-and effect relationship, or more specifically a positive correlation. The sum of the CPU and memory usage will be factored from the start of this experiment till the normalized dataset is outputted in another file. After the results are collected and visually represented, the mean will be calculated to get the average CPU and memory usage per data entry in a given dataset. To support this sub-problem, the correlation coefficient between the size of the dataset and the amount of CPU and RAM it takes to process each dataset will be calculated. This should show the strength of the relationship between the two. The results can conclude that if the usage of CPU and RAM increases, more expenses will have to be considered just to keep the quality of performance consistent for daily use. Adding to this experiment, the researcher will perform a hypothesis test to compare between his newly developed algorithm and the existing algorithm. A test will demonstrate whether or not this new and improved method is the best fit in today’s given environment. This process will result in either accepting or rejecting the hypothesis in regards to expenses. This is crucial as it provides accurate data that will be generated into formal business reports for the organization’s leaders.

3) Amount of server offline time

To consider the amount of time a single host server has to go “offline” when processing various amounts of un-normalized datasets while running this algorithm.

a) *Data Collection and How it Will Be Treated*

The data required for this specific sub problem will focus on the total time it takes (in seconds) to extract, analyze and manipulate the data. When the datasets becomes larger in size and are more unorganized, it takes more time to produce accurate results. This is can be seen as a cause-and effect relationship, or more specifically a positive correlation. The sum of the completion time will be factored from the start of this experiment till the normalized dataset is outputted in another file. After the results are collected and visually represented, the mean will be calculated to get the average time in seconds per data entry in a given dataset. To support this sub-problem, the correlation coefficient between the size of the dataset and the amount of time it takes to process each dataset will be calculated. This

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

should show the strength of the relationship between the two. The results can conclude that the more time it takes to process each set of data, the more time the server has to be offline in order to protect data integrity. Having more offline time is not ethical as it goes against a company’s policy of providing support and services 99.9999999% of the time. Adding to this experiment, the researcher will perform a hypothesis test to compare between his newly developed algorithm and the existing algorithm. A test will demonstrate whether or not this new and improved method is the best fit in today’s given environment. This process will result in either accepting or rejecting the hypothesis in regards to the amount of time a server stays offline for.

The Qualifications of the Researcher

The research will be conducted by Khang Tran. He is a 3rd year computer science student with some experiences in data analytics. He has never worked with “Big Data” since the data he is limited to during his are only small random datasets given by the professor himself. The challenge being foreseen is the lack of experience, and expertise in data analytics. To counteract against this unbalance, he will derive his research from credited researchers or data scientists on existing algorithms and evaluate them for potential future expansion.

The Proposal Outline

The Proposal Outline

This study’s main objective is to develop an algorithm that automatically schedules data cleansing while minimizing the amount of idle time a server has to undergo for this process. The process of going through this experiment will present an analysis and conclusion of a cause-and-effect relationship between the size of the data and the major factors as discussed earlier in the literature. The next section will discuss the formal steps on how to carry this study out.

The Steps Taken to Accomplish the Study

Step 1: Study and learn the existing algorithm in place from past literatures or interviewing data scientists in charge of the research.

Step 2: After gaining knowledge and insight on existing processes, the researcher must familiarize himself with the tools available that includes getting to know the following: VMware, Apache Spark, and MySql.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

Step 3: Start to develop the new algorithm while building the host server that will be the platform processing the datasets.

Step 4: Once the algorithm is formulated, run it through use cases or test cases to stress on the ability for it to perform the way it’s supposed to without any errors.

Step 5: If there are any bugs with the proposed algorithm, the researcher must debug the program and make sure all areas of the process is stressed to the limits to assure reliability.

Step 6: Prepare and finalize the formal written agreement that allows both parties the acknowledgement of how the data will be used in the experiment, in which includes the consent and protection of consumer privacy policies.

Step 7: Once the contract has been signed and both parties come to an understanding, the researcher must obtain a dataset that contains the values needed for the experiment.

Step 8: Run both the pre-existing and improved algorithm on the given datasets using the platform built earlier. If there are any errors or bugs, the researcher must address them and troubleshoot immediately.

Step 9: Record the numerical results and convert them to a graphical representation to display a meaningful way of interpreting the difference between the two algorithms.

Step10: Write the final report stating how the experiment was conducted, the captured results, and what this means in the industry.

The Timeline of the Proposed Study

This research will be conducted over the course of 24 weeks. The breakdown of the timeline can be represented below.

Week #	1 - 4	5 - 8	9 - 15	16 -17	18	19	20	21 - 22	23	24
Step	1	2	3	4	5	6	7	8	9	10

The Budget for the Proposed Study

Since the research is conducted the Khang Tran alone, most of the instruments to conduct this experiment is free except for any hardware upgrades (CPU, and RAM). The OS and platform used to run this experiment are all open-sourced, whereas the physical hardware upgrades is at Khang’s personal expenses.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

Cited Work

[1] Abdullah, Noraini, et al. "Data Quality in Big Data: A Review." (2015). Document.

***** Remember to check for grammar: Use of “Which” needs to be corrected (Need commas or get rid of it)

[2] Awadallah, Dr. Amr and Dan Graham. "Hadoop and the Data Warehouse: When to Use Which." (2012). White Paper.

[3] Ganapathi, Archana and Yanpei Chen. "Data Quality: Experiences and Lessons from Operationalizing Big Data." (2016).

[4] Gibbs, Mark. "Network World." *Not Enough Data Scientists? Use AI Instead* (2014). Article.

[5] Kim, Miryung, et al. *The Emerging Role of Data Scientists on Software Development Teams* (2016). Article.

[6] Lathiya, Piyush and Dr. Rinkle Rani. "Improved CURE Clustering for Big Data using Hadoop and Mapreduce." (2016). Article .

[7] McGuire, Tim, James Manyika and Michael Chui. "Why Big Data is the new competitive advantage." (2012).

[8] Saha, Barna and Divesh Srivastava. "Data Quality: The other Face of Big Data." (2014).

[9] Taleb, Ikbal, et al. "Big Data Quality: A Quality Dimensions Evaluation." (2016). Article.

[10] Yang, Kan, Xiaohua Jia and Kui Ren. *A Secure and Verifiable Access Control Scheme for Big Data Storage in Clouds* (2014). Article.