DECISION SCIENCES INSTITUTE

Usage of Hadoop and Microsoft Cloud in Analytics

Ceyhun Ozgur
Valparaiso University
Email: ceyhun.ozgur@valpo.edu

Jeffrey Coto
Valparaiso University
Email: Jeffrey.Coto@valpo.edu

Elyse "Bennie" Myer-Tyson
Valparaiso University
Email: bennie.myertyson@valpo.edu

David Booth
Kent State University
Email: dvdbooth8@gmail.com

ABSTRACT

This paper examines analytics and how changes have given rise to several new programs. This paper also provides discussion on these programs and how the programs approach large analytics and allay the concerns of ever-growing data files. After examining the basics, (Hadoop and Cloud services) further exploration develops by looking at complementary software and programs that facilitate Hadoop in the Cloud (such as Apache's Drill or Amazon's S3). Lastly, the paper examines the benefits and drawbacks of these solutions overall in order to give a clear picture of the program landscape. In conclusion, the benefits outperform the drawbacks.

KEYWORDS: Hadoop, Cloud, Big Data, Analytics, Apache's Drill, Amazon's S3

INTRODUCTION

As the field of data analytics continues to grow, there is an increasing amount of data incoming from a single user (McGuire, Manyika & Chui, 2012). This means that greater processing power is necessary to run the standard analyses as well as more complex relational analyses in order to understand how these data interact with each other. The fundamental problem of this expanding model is that there is not enough power on a single computer to match these demands. Therefore, one must resort to using networks of computers set up into nodes in order to run these procedures and interpret the results effectively. This is best executed through using cloud software, as it allows for one to access data and software remotely. The largest concern, however, is security. How does one keep data and business proceedings secure when accessing from multiple points and across multiple nodes? The solution to this problem comes from the advanced security systems being put into place by several software companies in order to ensure that these data are protected at all times. The best possible solution is to use Hadoop in the cloud, as it is one of the most common analytics packages and is highly adaptable to remote usage.

BACKGROUND

WHY USE HADOOP?

Hadoop is an open source program package, meaning that the code of the program is available to all for modifications with little to no restrictions (Ozgur, Myer-Tyson, Coto, Bolerjack, & Shen, 2017). Thus, anyone can use the program as a platform upon which to build personal operations. In addition to anyone being able to access Hadoop, its open source nature allows for high degrees of personal modifications to best suit specific datasets. The two most important modules of Hadoop for use in the cloud are the Distributed File-System, MapReduce, (Dhyani & Barthwal, 2014). These modules are necessary for a basic understanding of the program and smooth application thereof. (Ozgur et al., 2017).

DISTRIBUTED FILE-SYSTEM

The Distributed File System (DFS) is what makes storing and accessing data from multiple locations, as well as linking said locations, work smoothly. Normally, the individual node would determine a file system, as they are typically part of a computer's OS. Hadoop, however, uses its own file system instead of the one on the node, which allows access to data from a computer using any supported OS (Ozgur et al., 2017). This is crucial for cloud support as the interaction between nodes is a key feature of using the cloud. Additionally, the DFS forgoes the problem of file incompatibility across networks as it uses its own standard file.

MAPREDUCE

MapReduce provides the basic data examination tools for Hadoop. It is named after the key functions it performs: reading the data from the database, putting it into a format that can be analyzed (called a map) and carrying out mathematical operations, such as sum or mean, within the dataset (reduce) (Ozgur et al., 2017).

LITERATURE REVIEW

THE USAGE OF HADOOP

Hadoop's lack of rigidity in terms of both software and hardware usages means that companies are able to modify or create systems as necessary, saving money in the long run by allowing choice of vendor. Furthermore, the choice to use cloud interaction between the private company computers and personal computers reduces overhead cost and allows for work away from the desk. It has become the most widely used analytics software for nodes that are not specially created for large data processing, which makes it ideal for use in the cloud with large data files. Hadoop was created to fill the need to analyze large data files and currently can handle petabytes of data with ease. This also makes it popular among businesses, as it is practical as well as time and cost efficient (Dhyani & Barthwal, 2014). Most large online presences use Hadoop, as anyone is able to download and modify the software (Dhyani & Barthwal, 2014). These modifications made by either individuals or businesses are often shared with Hadoop's development team allowing the product to be improved upon given real-world scenarios and concerns (Dhyani & Barthwal, 2014). This further bolsters the appeal of Hadoop in the cloud as the development team is kept well informed about the challenges and preferences of those working with Hadoop on a daily basis. This collaboration in development is critical for several

reasons. First the maintenance of open-source software and second for paving the way through innovative changes in big data analytics.

Even in its raw state and without the complicated data analysis tools, Hadoop itself can be incredibly complex and challenging to work with (Dhyani & Barthwal, 2014). This is why Apache, and other firms such as Cloudera, developed several commercial versions of the software (Dhyani & Barthwal, 2014). These software packages make the entire process of utilizing Hadoop simpler, from installation to usage and troubleshooting. These commercial packages also include training and support to streamline the usage process (Dhyani & Barthwal, 2014). As the software becomes updated for use in the cloud, these processes will likely be updated to reflect the change in user preference and the shift in software usage. This necessarily means that moving to Hadoop in the cloud, rather than simply on a node or network, will be a simple, quick transition for businesses.

Companies are free to expand their usage of Hadoop when strategic expansion is needed; this includes adding new software, reimagining old software, and even upgrading to cloud connectivity. The support garnered from the analytics community as well as from those who use Hadoop for their own personal needs, has led to the software being accessible for everyone. This again bolsters the efficacy and ease of use of Hadoop in the cloud, as the professional as well as personal use issues are being heard throughout the community (Ozgur et al., 2017).

WHY USE CLOUD COMPUTING?

Microsoft Cloud

Microsoft Cloud, supported by Dell, is versatile and low-cost. This software can communicate with a number of other devices, cloud structures, and management systems, which gives it a leading edge when deciding on how to run large-scale analytics programs (Dell, 2017 March 24). Microsoft's Cloud software runs 80% faster than the major competitors in the field (Dell, 2017 March 24). This speed, combined with a lower starting operation cost, customizable payment options, and high flexibility in applications, creates perfect software for those who need to stretch their resources. Regardless of what materials one has at one's disposal, the IT support behind the Microsoft Cloud program is designed to support any issues that may arise (Dell, 2017 March 24). This makes personal and professional cloud usage practical, as one is able to receive support independent of location.

The cloud is designed for efficiency wherever possible; as such, the backbone of the program is built upon few key components (Dell, 2017 March 25a). The ability to automate the provisioning and deployment processes of an application within the cloud greatly reduces the entire time load when calculating large analytics results. This makes the cloud ideal for personal use as well as professional, as one is able to quickly achieve the desired results. The cloud can be adapted from private to personal, and vice versa, and integrates across the IT network for both Dell and other businesses (Dell, 2017 March 25a). This solution utilizes both Dell's resources and those already available, creates a stable mix of public access and protected data and programs. This mixture is ideal for businesses who are concerned with security but who also want to expand their capabilities beyond what can be done in the office. The International Data Corporation (IDC) claims that revenue growth could be increased by as much as 100% when utilizing the cloud strategy (Dell, 2017 March 25a).
Dell creates a highly secure environment in which to utilize its cloud software, thus reducing the risk and fear of a breach. The security solution provided is layered, which allows total control

over whether or not an application, dataset, program, or other resource is private- or public-access (Dell, 2017 March 25b). This careful monitoring and detailed security structure means that there is a low chance of an outsider being able to access industry secrets or upcoming project results. Furthermore, having such fine control over security will also reduce the risk that a former employee is still able to get into the cloud once they have been let go, a very real concern for today's businesses. Regardless of availability level, Dell takes necessary precautions to protect sensitive information without hindering the development of new ideas. Virtualization of the data center, including physical servers, storage, and networking hardware and software, leads to a more efficient analytics environment (Dell, 2017 March 25c). The environment is set up to maximally use assets, lower costs associated with large-scale analytics, reduce the need for managerial oversight, and quicken the delivery of IT solutions to any problems that may arise (Dell, 2017 March 25c). Regardless of virtualization level, Dell has support to identify goals, meet requirements, and create a plan to aid in automating quotidian tasks as well as unifying data centers for easy management (Dell, 2017 March 25c). This also makes cloud computing more practical and appealing given that the level of automation and interaction can be varied depending on security concerns and personnel skill, which makes personal and professional computing smoother and safer.

OpenShift

The OpenShift program provided through NTT DATA, the company which has recently acquired Dell (and therefore Microsoft Cloud), facilitates the ease of automation in cloud services in order to give users more time to develop or work on projects (NTT DATA, 2017a). OpenShift reduces the time spent on creating applications to run data, thus leading to more time to work on software coding and innovative deployment. Additionally, a reduction in applications needed facilitates easier personal use of the nodes. It also has a wide variety of programming languages from which to choose, thus reducing potential frustration by supporting a wide variety of skill sets (NTT DATA, 2017a) and furthering personal capabilities.

The platform itself is open source with standardized components to ensure applications are usable across computers as well as eliminate the possibility of locked files (NTT DATA 2017a), which meshes well with Hadoop's flexibility across different operating systems and reduces conflict between personal and professional cloud usage. The program is simple to use and allows for near-instant startup by giving users specific database cartridge support and scaling automatically to the development tools in use with a highly interactive interface (NTT DATA 2017a). Data cartridge support means that the extensions of servers, like those found in Oracle, can be preserved without requiring additional hardware or software (Oracle Corporation, 2015). These cartridges also facilitate personal use of the cloud and access of Hadoop by allowing minimal installations to take place and freeing up both hard drive space and processing power to run the analyses necessary.

Virtual workspaces, including desktops and applications, accelerate the outcomes of analytics (NTT DATA, 2017b). Managed Virtual Workspace Services, a subscription-based, remote-management service that is easy to use, allows for quick resource allocation as well as instant IT access if problems arise (NTT DATA 2017b). This instant access can be highly appealing to the personal cloud user as there is little worry about unsolvable problems or project-stalling program failures. Usage of this software includes access to the NTT DATA Global Network Operations and Delivery Centers, which have experts to aid in the facilitation of virtual

workspaces and support them regardless of whether or not these services are being used at the personal or professional level. Workspace-as-a-Service, also part of NTT's package, can support more than 500 virtual desktops at once, to dramatically increase the productivity of a workspace (NTT DATA 2017b). Employees can run the latest versions of their applications on any device, whether it belongs to them or the company, without worrying about capital costs.

FACILITATING HADOOP IN THE CLOUD

Apache drill

Apache's Drill displays a numerous amount of applications and upgrades such as the ability to perform non-Search Query Language (SQL) databases and file systems (The Apache Software Foundation, 2017). Drill is the only columnar query engine that supports complex data (The Apache Software Foundation, 2017). This columnar format means that data is stored the same way it would be in Excel; the key difference is that in querying the data, the unused columns are immediately skipped over rather than being processed. This familiar data storage format makes reading the data easier for personnel at any skill level and can continue to facilitate the usage of Hadoop in the personal cloud. This saves the user time and decreases the strain on the network or node. It features an in-memory shredded columnar representation for complex data, which allows Drill to achieve columnar speed with the flexibility of an internal JavaScript Object Notation (JSON) document model (The Apache Software Foundation, 2017). "Shredded" columnar representation means that the data is free of the coding required to run the program and is instead put into a relational format, most likely a table (Delconte, 2013).This again makes the data more readable and approachable to personnel as they are quickly able to visualize which data are being queried and potentially included in a project.

This application allows searches to occur between the non-SQL databases as well as discriminate information from local files (The Apache Software Foundation, 2017). The toggle ability outperforms others since Drill can take full advantage of the efficiency with the search. Automatization of the Drill platform optimizes the processing by utilizing the data store's internal capabilities and provides intuitive extensions to the SQL so that on can easily query complex data (The Apache Software Foundation, 2017). This ease of use allows the employees to continue running programs or working with the software outside of the workplace environment without needless complications or frustrations. Another consumer benefit of Drill is that data is stored on the same nodes, giving support to locally stored data (The Apache Software Foundation, 2017). This is a huge benefit because it requires a small amount of IT interaction. The small IT interaction allows the user to search raw data on-site eliminating the transforming or processing of data (The Apache Software Foundation, 2017). This advanced compilation and re-compilation techniques increase productivity of the processing (The Apache Software Foundation, 2017). The featured JSON data model within Drill enables queries on nested data resulting in rapidly evolving structures commonly applied to non-relational data stores and modern applications (The Apache Software Foundation, 2017).

Because Drill supports both standard SQL and non-SQL, end users such as analysts are able to function within the standard business intelligence (BI)/analytic tools such as Tableau, Qlik, MicroStrategy, Spotfire, SAS and Excel (The Apache Software Foundation, 2017). This platform provides freedom for the end users to interact with non-relational data stores by leveraging Drill's Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers (The Apache Software Foundation, 2017). Accessing Drill's virtual datasets create realms for

multifaceted, non-relational data to be mapped into BI-friendly structures (The Apache Software Foundation, 2017). The Apache's Drill multifaceted capability allows the program to run on Mac, Windows and Linux (The Apache Software Foundation, 2017). This exceptional performance is best visualized when it is functioning in a cluster of servers dedicated to the running and usage of the Drill program. Such high performance is achievable because Apache's Drill internal foundation is built with symmetrical architecture and is complimented with simple installation (The Apache Software Foundation, 2017).

Query engines (QE) have been producing rational and non-rational data sets for decades but Drill is the first QE that encompasses user needs with both application flexibility and end product speed (The Apache Software Foundation, 2017). The design of Drill includes the aforementioned columnar engine which can support complex and large data files. Drill also compiles based on the data file and recompiles at execution times, thus making it adaptable to personal needs. The program has specialized memory management to reduce the footprint of it on memory usage in the computer as well as to eliminate garbage files or needless collections of files. The locality-aware execution of Drill reduces network traffic when Drill is located on the same node as the data, and the cost-based optimizer further enables quick processing and manipulation of the datastore (The Apache Software Foundation, 2017).

WHY HADOOP IN THE CLOUD?

Quoble's Hadoop service

With its unlimited scale and on-demand access to compute and storage capacity, cloud computing is the perfect match for big data processing. Qubole's Hadoop as a Service offering has several advantages over on-premise solutions (Quoble, Inc., 2017).

Quoble's Hadoop service has clusters in the cloud which are scalable depending on processing requirements (Quoble Inc., 2017). This makes the Quoble Hadoop solution highly flexible and practical for businesses which are beginning their expansion into the cloud. This elastic feature means nodes are automatically added to or removed from clusters based on data size to improve performance (Quoble Inc., 2017).

Hadoop's integration outperforms when in its native ecosystem, which includes MapReduce, Hive, Pig, Oozie, Sqoop, Spark, and Presto (Quoble Inc., 2017). Systems such as the Quoble Data Service have the ability to manage Hadoop-based clusters, which can eliminate the need for duplicate installations on multiple nodes. This means that users have more time to manage the nodes and develop scaled clusters (Quoble Inc., 2017).

An added feature to using Hadoop in the Cloud, as provided by Quoble, is the minimal operational and upfront capital costs. There are zero on-site hardware or IT support demands (Quoble Inc., 2017). Further cost containment occurs when utilizing the spot instant pricing compared to on-demand instances (Quoble Inc., 2017).

Cloudera

Cloudera demonstrates a superior elastic platform, allowing infrastructures to be leveraged. This allows for the independent computation and storage of files and decreases the rigidity of movement. This autonomy allows for the growth and reduction of clusters dynamically while still strengthening the ability to perform clone net-new clusters. These clones create copies of existing clusters and upload them to the net, which drives the ability to create ad-hoc, transient workloads (Mokhtar, Ghat, Robinson & Mukil, 2016).

With Cloudera's current structure, flexibility enhances the businesses' ability to minimize cloud lock-in (Gutow, 2016). According to Gutow (2016), Cloudera wishes to ensure that companies maintain the fastest, easiest, and most secured Hadoop usage. The ease of use is achieved by running Cloudera workloads on public clouds, including in multi-cloud environments.

Enterprise, or professional, grade Cloudera maintenance and performance helps manage big data by reducing risk on large platforms (Gutow, 2016). This extenuates comprehensive manageability, product availability, enforced security, and production workloads. All of these benefits are now governance required for all deployed Hadoop distributions (Gutow, 2016), meaning there is no unsupported version or service incompatibilities.

HADOOP AS A SERVICE IN THE PUBLIC CLOUD

The primary function of today's cloud storage is unstructured data. This makes sense, given the large data files produced and analyzed by cloud software; these data sets do not lend themselves to high organization or structure (Kumar, 2013). Public cloud is basically constructed for the use of multi-tenancy, isolating access, data and individualized security (Kumar, 2013). These multi-tenancy capabilities mean that users are able to share the input or output from a single instance of the program rather than each needing to run the software, which cuts down on processing issues. Processing issues are now resolved with the development of the Hadoop distributions as seen with IMB BigInsights, Cloudera, Cloudera Distribution Hadoop (CDH), Hortonworks, and MapReduce (Kumar, 2013). These cloud storage technologies can now be run on a variety of public clouds as seen in MS Azure, IBM, AWS (Amazon Web Service), SmartCloud, and Rackspace (Kumar, 2013). These public clouds offer infrastructure as a service (IAAS). The advantage of running in a public cloud is the ability of the infrastructure to be shared among the independent customers. The disadvantage is the limited virtual machine (VM) control (Kumar, 2013). With limited rack awareness, which is a middle-level sorting of the different nodes wherein 30-40 of them are stored together to minimize communication time, the required availability and performance of the cluster may be impacted by users running the VM but the enterprises can pay for these Hadoop clusters on demand to reduce the likelihood of communication interference (Kumar, 2013). As a private cloud user, there is an option for creating private networking solutions using virtual local area networks (VLANs). Even though the VLAN is an option, the Hadoop cluster performance really should be utilizing a separate isolated network since there is a high level of network traffic between the nodes (Kumar, 2013). This is true in most cases with only exception being AWS' Elastic MapReduce (EMR) (Kumar, 2013). In the case of the AWS EMR, the user must install and configure the Hadoop cluster on the cloud (Kumar, 2013).

Amazon's EMR provides a complete cluster cycle solution that allows an easy and fast way to run MapReduce jobs with the benefit of not installing a Hadoop cluster on the Cloud (Amazon Web Services, 2017). This means that the data comes back once a cycle has completed rather than waiting for all of the data to come back at once. The advantage point is that an

organization can internally develop Hadoop programming expertise to run MapReduce jobs in the workloads (Amazon Web Services, 2017). This in return can allow easy selection and us of the open source projects, spending more time on increasing the value of your data (Amazon Web Services, 2017).

Hadoop on S3

Data storage in Hadoop is accomplished within the Amazon S3 platform instead of the Hadoop Distributed File System (HDFS) (Amazon Web Services, 2017). The advantages of the Amazon S3 storage are the capabilities of bucket versioning and elasticity (Amazon Web Services, 2017). Bucket versioning provides fail-safe measures to ensure that projects are not lost and that different steps of the analytic process are kept away from each other so it is easy to separate parts of projects. Lastly, the Amazon S3 system has personal data loss protection schemes (Amazon Web Services, 2017). The disadvantage is that the S3 performs slower than HDFS (Amazon Web Services, 2017).

HADOOP IN PRIVATE CLOUD

The private cloud deployment establishes the same consideration for Hadoop (Amazon Web Services, 2017). The private cloud allows greater control over personalized infrastructure enabling provisional bare-metal servers, where the VM is installed directly to the hardware rather than operating through the OS of the system, or the ability to create separate isolated network clusters (Amazon Web Services, 2017). Other ingenuity of the private cloud is the Platform as a Service (PaaS) layer. The PaaS layer provides pre-built patterns for deploying Hadoop clusters easily (e.g. IBM offern patterns for deploying InfoSphere BigInsights on their SmartCloud Enterprise) (Amazon Web Services, 2017). In addition, there is an option of deploying a “Cloud in a Box” like the IBM PureData System (Amazon Web Services, 2017). This offers Hadoop ready within the private data center. The primary purpose for private cloud deployment involves data security and access control over data as well as improved visibility and control of the Hadoop infrastructure (Amazon Web Services, 2017).

CONCLUSION

BENEFITS OF HADOOP IN THE CLOUD

Hadoop is the best solution to a number of problems that exist in the analytics field today. The program is highly adaptable and is very powerful in terms of the number of cases it can run at once. Additionally, it reduces overhead costs by being open source software and working across all operating systems, thus reducing the need for specialized equipment.

Lack of rigidity and reduced overhead cost means the company using Hadoop is more able to focus on the projects at hand. They are therefore more capable of focusing on results rather than spending time on coding or in maintaining the nodes that run Hadoop. Cloud interaction furthers the efficiency by reducing the nodes needed further and by allowing for personal node usage as well as private nodes.

Hadoop and the cloud both have highly responsive IT. This means that there is little down time waiting for the developers to patch a problem in the software, or for maintenance of the Cloud

structure. This can further reduce overhead costs by working around node failures or any corrupted program files.

The cloud is also one of the most rapid and capable programs available to those who are working in analytics. The Cloud can run up to 80% faster than competing services, making it vital for any business looking to maintain a competitive edge. This software is capable of being accessed anywhere with these high speeds, thus increasing the output of personal nodes in the network.

Cloud lends itself to the process of automation and can therefore reduce time needed to input the data or to calculate large analytics results. Personal machine are therefore made more capable of running large data files, an issue that has plagued the data analytics community for some time. This is compounded by the Cloud requiring fewer applications to run, having greater programming language options, and reduced need for on-site IT. All of these things lead to software that is highly versatile and will not require maintenance by or training for company employees.

Drill service can combine Hadoop and the Cloud in a familiar, easy-to-read format that mimics Excel while being able to process only the needed columns. This cuts down on the time necessary to analyze data as the program selects only the relevant columns to analyze. Data are also more approachable in this format as they are easier to visualize and read, which can increase understanding of the projects even in the layperson.

Drill's ease of use allows employees to continue working on the software outside of the workplace environment, which makes personal nodes even more attractive. The portability and flexibility of these networks make them highly adaptable to each business's and employee's needs, thus solving problems of frustration or miscommunication. The minimized interactions between SQL and non-SQL are also a large burden-reliever, as one need not worry about switching between languages. In addition, there is a great amount of local file support, which can bolster node speeds or allow for work to continue should there be an internet outage (and thus a lack of access to the Cloud). This is in addition to Drill already cutting down on time used by compiling and recompiling the data files at execution times, thus furthering potential productivity on tight deadlines.

Quoble's service of Hadoop in the cloud is scalable depending on processing requirements. This means that there are few, if any, extraneous nodes put into the cluster. Cloudera is able to further what this can do by duplicating nodes as backups in the event of a node failure. Amazon's EMR is also capable of this, meaning that one never needs to worry about whether or not files can be recovered.

According to Gutow (2016), Cloudera wishes to ensure that companies maintain the fastest, easiest, and most secured Hadoop usage. The Hadoop distributions currently in Cloudera are being transferred into a system where there are no unsupported versions. This means that there will be few, if any, service or version incompatibility errors which reduce headaches and costs for businesses. In addition, users are able through Hadoop in the Cloud to share input or output from a single instance of the program, which means that not everyone needs to be running the software. This cuts down on time spent in redundant actions and cost to ensure that everyone is able to run the program in a timely manner (which may not be feasible for all personal nodes).

By using the public cloud, the infrastructure can be shared among consumers and employees alike, which allows for transparency and trust as well as for solid project backbones.

Amazon's EMR service allows one to run Hadoop without any software installation on a machine. This can make a personal node even more appealing, as there is little chance of receiving corrupted files or of having someone steal industry secrets. Rather, one is able to restrict access into and out of the cloud and therefore maintain a tight control over who has access to ongoing projects and data. The EMR service also allows for a business to create their own Hadoop programming or to utilize others' work in order to reduce the workload on a single node.

EMR comes with bucket versioning, which gives a failsafe against losing files. Each step of the analytic process is stored independently, which means that a failure of one node or bucket will not result in a loss of all the progress up to or following that point. amazon's S3 system furthers this security by offering personal data loss protection.

The advantage point is that an organization can internally develop Hadoop programming expertise to run MapReduce jobs in the workloads (Amazon Web Services, 2017).

DRAWBACKS OF HADOOP IN THE CLOUD

Open-source means few secrets or industry-standard techniques, as anyone is able to use this software or create programs within it. Therefore, a business may lose competitive edge by working with this new software through either a similar technique to other firms or by having to deal with an increased wait time caused by several firms and personal users all requiring IT support for these services. There are also potential security issues, as open source software tends to be the most easily compromised given the large audience of users. This raises the question: how does one keep data and business proceedings secure when accessing from multiple points and across multiple nodes? This is especially a concern in a day and age where software hosted by a separate business or not installed directly on company hard drives may cause breaches of security at worst or improper installations (and therefore more downtime) at best.

Additionally, any automatic process may be both a benefit and a drawback. Automating programs requires less human input, but that may cause several problems. A program failure may go unnoticed without being monitored, or the wrong file may be analyzed, which costs rather than saves time. Additionally, a program may execute only partially or incorrectly if there are errors which are not caught or which can be ignored to allow the program to continue. The biggest example of this is in SAS, an industry favorite. An incomplete or incorrectly read data file may still be analyzed, but with only a fifth or even a tenth of the data set one wishes to examine, which could lead to incorrect conclusions or magnified effects.

Automation reduces human input and may be more error-prone.

Another fear of automation is how Quoble's software determines the nodes to be used in a network. A backup node may be deleted because it is not strictly necessary, which could potentially result in data loss for the company. And, if most of the software is automated, which requires little human input, there may not be a recent version of the data set with its manipulations and outputs saved. Worse, if this lack of human input means that no one has

seen the data post-analysis, then one may have to begin again at square one, which further increases time spent on projects and defeats the purpose of using these programs. Even if one is able to recover the files, this may come at the cost of a node or program being down for hours or days while attempting to fix the problem via IT.

While outside IT support is useful in that it reduces the immediate cost to a business, there are several concerns that go along with using an outside service. The largest is security: the more people brought in on a project, the less likely the business is to be able to find who leaked information in the event of secrets getting out. This could mean that businesses would be compromised in every way, as they have no way to know if the person leaking information is even one of their employees. Further, though less severe, there is also a concern about the time it takes to explain specific setups to an outside IT supporter. While they are trained to help with their specific program, a business's use of such customizable software may not be something one or several technicians have seen before. This could increase the time it takes to resolve the issue, rather than decrease it, as one would have to either explain the system or be transferred until reaching someone who can help with that specific configuration.

The glaring flaw in any attempt at using Cloud storage is fairly obvious: how is one to get around Internet outages? They happen to every business for a variety of reasons (accidents, construction, or natural disasters are only a few of the mundane challenges to maintaining communications, especially online). These down times may strip a company of any ability to work on a project regardless of what software they use if storage and applications are all off-site or online.

Further, the assurance of a well-working machine may come with a high price tag. It is possible to run a smooth operation with the low-cost options, but there are several more expensive premium packages which guarantee better safety and confidentiality when working with businesses and their data analytics. This is not bad in theory, but in practice it can minimize or even override the costs saved by initially choosing these programs. Thus, it is in the best interest of a business to consider what the premium package of their preferred software looks like before attempting to make a shift to save capital and labor costs. This also means determining whether or not to run a dedicated server; a VLAN can operate Hadoop clusters, but they have optimal performance when working on their own server. Therefore, the time as well as material cost could increase when attempting to get the most out of Hadoop, which could put a business in a precarious position of paying more than the software is worth to run if they require a sudden spike in usage or bandwidth needs.

AWS EMR seems to fix most of the problems mentioned above, but it is not without its issues. The program requires use of the cloud, which could increase data concerns as well as concerns about whether or not the programs truly are secure if they are in cloud storage. How can a business ensure that no outsider is getting into their security once they rely on the cloud? This concern is compounded by the fact that storage in EMR Hadoop installations is accomplished by S3 rather than HDFS, which was noted to eliminate many concerns and work quite quickly. Therefore, there could be a reduction in possible communication if S3 does not support multiple OSes in the nodes or if there are queries going back and forth between SQL and no-SQL formats. S3 is also slower than HDFS which could erode some of the theoretically saved time when choosing this software.

The disadvantage is that the S3 performs slower than HDFS (Amazon Web Services, 2017).

WHY THE BENEFITS OUTWEIGH THE DRAWBACKS

While there are some significant concerns which should not be ignored, the overall analysis of drawbacks and benefits comes out with a net positive. There are several services which a business could mix and match in order to get the unique service it needs. Furthermore, using multiple layers of difference services could increase security and keep a business competitive by still being a unique process. This still keeps the time and cost benefit analysis at the forefront, as every business is able to weigh its own risk assessment and act accordingly.

Security, being the biggest concern, also has a variety of options to alleviate potential problems. The usage of several software services at once means that security can be placed at the forefront in the network and node configurations. Utilizing Dell's cloud as well as company resources and any other programs can create a stable mix of public and private access of protected data in the cloud. This could be facilitated through pass keys only given to employees in order to access certain features of the cloud storage, which eliminates or at least greatly reduces the potential of leaked information or having projects hacked. This also maintains the personal and private node balance which allows for greater flexibility in business operations when working with large data files.

Lastly, the mixed service solution fixes the problems of downtime or internet outages; there could be nodes dedicated to hard drive installations of these programs, which would allow for continued access to and development of programs in the event of a loss of service. Additionally, these mixed solutions create multiple places to back up data and have file versions, which again reduces the fear of data loss. So, not only do mixed solutions solve security problems, but they can also greatly reduce or eliminate the concern over program access interruptions.

REFERENCES

A. Gutow (2016, March 29). Decreasing operating cost in the cloud environments. [Web log comment]. Retrieved from http://vision.cloudera.com/decreasing-operating-costs-in-cloud-environments/

Amazon Web Services (2017). Use your favorite source application. Retrieved from https://aws.amazon.com/emr/

Cloudera, Inc. (2017, March 10). Hadoop in the cloud. Retrieved from: https://www.cloudera.com/products/cloud.html

Delconte, S. (2012 October 23). Manipulating XML data in SQL server. [Web log comment]. Retrieved from: https://www.simple-talk.com/sql/database-administration/manipulating-xml-data-in-sql-server/

Dell (2017 March 24). Cloud computing. Retrieved from: http://www.dell.com/en-us/work/learn/dell-cloud-computing

Dell (2017 March 25a). Build your cloud. Retrieved from: http://www.dell.com/en-us/work/learn/dell-cloud-computing-build-cloud

Dell (2017 March 25b). Cloud security. Retrieved from: http://www.dell.com/en-us/work/learn/dell-cloud-computing-security

Dell (2017 March 25c). Data center virtualization. Retrieved from: http://www.dell.com/en-us/work/learn/dc-virt

Dhyani, B. & Barthwal, A. (2014). Big data analytics using Hadoop. *International Journal of Computer Applications 108*(12). 0975-8887

McGuire, T., Manyika, J., & Chui, M. (2012, July/August). Why big data is the new completive advantage. *Ivy Business Journal. Retrived from http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/*

Mokhtar, M., Ghat, D., Robinson, H. & Mukil, S. (2016, September 22). Intigration, elasticity, agility and cost performance benefits on AWS. [Web log comment]. Retrieved from http://blog.cloudera.com/blog/2016/09/apache-impala-incubating-vs-amazon-redshift-s3-integration-elasticity-agility-and-cost-performance-benefits-on-aws/

NTT DATA (2017a). Hosted private cloud. Retrieved from: https://us.nttdata.com/en/services/cloud-services/hosted-private-cloud

NTT DATA (2017b). Virtual workspace services. Retrieved from: https://us.nttdata.com/en/Services/Cloud-Services/Virtual-Workspace-Services

Oracle Corporation (2015, October 12). What is a data cartridge? Retrieved from: https://docs.oracle.com/cd/B10501_01/appdev.920/a96595/dci01wht.htm

Ozgur, C., Myer-Tyson, E., Coto, J., Bolerjack, K., & Shen, Y., (2017). A comparative study of network modeling using a relational database (e.g. Oracle, MySQL, SQL) vs. Neo4j. *MWDSI Proceedings 2017.*

Quoble Inc. (2017 February 20). Apache Hadoop as a service. Retrieved from: https://www.qubole.com/hadoop-as-a-service/

The Apache Software Foundation (2017 February 22). Apache drill: Schema free SQL query engine for Hadoop, NoSQL, and Cloud Storage. Retrieved from: http://drill.apache.org

V. Kumar (2013, May 28). Running hadoop in the cloud. [Web log comment]. Retrieved from http://www.ibmbigdatahub.com/blog/running-hadoop-cloud

Zujie,R. Jian, W., Weisong, S., Xianghua, X., & Min, Z. (2014). Workload analysis, implications, and optimization on a production Hadoop Cluster: A case study on Taobao." *IEEE Transactions on Services Computing 7*(2), 307-321.