World Journal of Computer Application and Technology 2(3): 73-81, 2014 http://www.hrpub.org
DOI: 10.13189/wjcat.2014.020303

Environmental Conditions' Big Data Management and Cloud Computing Analytics for Sustainable Agriculture

Duncan Waga[1,*], Kefa Rabah[2]

[1]Jaramogi Oginga Odinga University of Science and Technology, Kenya
[2]Kabarak University, Kenya
*Corresponding Author: wagadun@gmail.com

Abstract The World is getting swamped with data to the tune of 7ZB a year mostly emanating from the 'internet of things' devices. This data is scattered in various devices and no meaningful relationship may be derived from it, and neither can this terrific rate be managed by traditional storage or processors. It is believed that organizations that are best able to make real-time business decisions using Big Data solutions will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure. Cloud computing promises enough capacity in terms of storage and processing power to elastically handle data of such magnitude and through the use of analytics get value from it. This paper focuses on environmental conditions' data like rainfall, winds, temperature etc and the use of particular cloud computing analytical tool to get some meaningful information from it which can be utilized by farmers for strategic and successful Agriculture. Previous similar studies are discussed and recommendations given.

Keywords Cloud Computing Analytics, Internet Of Things, Big Data, Environmental Conditions, Agriculture

1. Introduction

Big Data technologies describe a new generation of technologies and architectures, designed so organizations can economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. This world of Big Data requires a shift in computing architecture so that customers can handle both the data storage requirements and the heavy server processing required to analyze large volumes of data economically Villars et al(2011). Big data refers to the ability to accumulate, structure, and interpret unstructured data. The term generally refers to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable elapsed time. And these data sets are huge. As of 2012, big data sets range from a few dozen terabytes to many petabytes of data in one single set. To put that into context, one terabyte can hold 1,000 copies of the encyclopedia Britannica while one petabyte is able to hold 500 billion pages of standard printed text. Since environmental information concerns input such as satellite images and power plants emissions, it falls by definition into the big data category. The data collected ranges from soil moisture to nitrogen levels. The use of data will allow farmers to gain a clearer picture of farming, receiving updates of the land in real time and no longer having to guess the next move. For example, the data can monitor pests, which then allows farmers to target problem areas rather than over using pest controls. This can also be seen in the use of fertilizers, which is often carried by runoff water. With constant updates the farmers can monitor what nutrients have been absorbed and which haven't. The end result is the same; farmers can give the crops the nutrients they require while reducing pollution Friedman (2013).

2. Literature Review

Various organizations and Researchers like Monsanto, known for its seed portfolio and use of biotechnology, recently bought climate corporation, a San Francisco startup, for $930 million. Climate Corporation focuses on unique crop information and insurance for farmers. Using weather and agronomic data, climate corporation tells farmers when it's best to water, spread chemicals and nutrients, and harvest. Additionally the company collects water data from the national oceanic and atmospheric administration and tracks temperature from national weather services – both sets of data are available to the farmer on a web-based platform. With the acquisition of Climate Corporation, monsanto can begin recommending some of their current products based on the data collected. Through the platform farmers can monitor crops, reduce costs of farming, but more importantly make better and informed decisions to increase crop yield. Lamonica (2013). As the local food and sustainable farming movements grow, big data is helping farmers grow produce

more food at lower costs. Libelium, a sensor-making company, practiced this new use of big data in connected vineyards in spain. Through the use of libelium's sensors, the application of fertilizers decreased by 20% while productivity went up 15%. Although other factors of production such as weather can't be changed, the sensors can match current weather patterns to previous weather and select better crops for the change of weather. Similar technology applications are in cattle management. The cows are monitored by sensors and notify farmers what cow needs milking and how long to wait in between, Andrea (2013). Cloud computing can be a powerful tool for scientists and researchers sharing massive amounts of environmental data. At the United Nations climate conference (cop 17) in Durban, South Africa this week, the European environment agency, geospatial software company esri and Microsoft showed off the "eye on earth" network. The community uses esri's cloud services and Microsoft azure to create an online site and group of services for scientists, researchers and policy makers to upload, share and analyze environmental and geospatial data. While the eye on earth network has been under development since 2008, the group launched three services for different types of environmental data at cop 17, including water watch, which Microsoft isn't the only one working on creating these types of eco big data networks. At last year's u.n. climate meeting, cop 16, google launched its own satellite and mapping service called Google earth engine, which combines an open api, a computing platform, and 25 years of satellite imagery available to researchers, scientists, organizations and government agencies. Google earth engine offers both tools and parallel processing computing power to groups to be able to use satellite imagery to analyze environmental conditions in order to make sustainability decisions. Google's earth engine, the government of Mexico created the first comprehensive, high-resolution map of Mexico's forests that incorporated 53,000 land sat images to produce a 6 gb mapping product. The Mexican government and ngos can use the map to make decisions about land use, sustainable agriculture, and species protection in combination with a growing population. Cloud computing and big data analytics will be an increasingly important way to manage a limited number of resources — energy, water, and food — as the world population explodes. There are already 7 billion people on the planet, and there will be 9 billion by 2050. Big data research is helpful in both the business world and environmental purposes too. Scientists measure and monitor various attributes of lakes, rivers, oceans, seas, wells, and other water environments to support environmental research. Important research on water conservation and sustainability depends on tracking and understanding underwater environments and knowing how they change Hurwitz (2013). Changes in these natural environments can have an enormous impact on the economic, physical, and cultural well-being of individuals and communities throughout the world. To improve their ability to predict environmental impacts, researchers at universities and environmental organizations across the globe are beginning to include the analysis of data in motion in their research. Scientific research includes the collection of large volumes of time-sensitive information about water resources and weather to help protect communities against risks and respond appropriately to disasters impacting these natural resources. Mathematical models are used to make predictions such as the severity of flooding in a particular location or the impact of an oil spill on sea life and the surrounding ecosystem. The type of data that can be used includes everything from measuring temperature, to measuring the chemicals in the water, to measuring the current flow. In addition, it is helpful to be able to compare this newly acquired data with historical information about the same bodies of water.

Many sophisticated research programs are in place to improve the understanding of how to protect natural water resources. Rivers and adjacent floodplains and wetlands, for example, need protection because they are important habitats for fish and wildlife. Many communities depend on rivers for drinking water, power generation, food, transportation, and tourism. In addition, the rivers are monitored to provide knowledge about flooding and to give communities advance warnings about floods. By adding a real-time component to these research projects, scientists hope to have a major impact on people's lives. At one research center in the United States, sensors are used to collect physical, chemical, and biological data from rivers. These sensors monitor spatial changes in temperature, pressure, salinity, turbidity, and the chemistry of water. Their goal is to create a real-time monitoring network for rivers and estuaries. Researchers expect that in the future, they will be able to predict changes in rivers the same way that weather predictions are made. Another research center based in Europe is using radio-equipped buoys containing sensors to collect data about the ocean, including measurements of wave height and action. This streaming data is combined with other environmental and weather data to provide real-time information on ocean conditions to fisherman and researchers. In both examples, sensors are used to collect large volumes of data as events are taking place. Although infrastructure platforms vary, it is typical to include a middleware layer to integrate data collected by the sensor with data in a data warehouse. These research organizations are also using external sources like mapping databases and sensors coming from other locations as well as geographical information. The data is analyzed and processed as it streams in from these different sources. One organization is building an integrated network of sensors, robotics, and mobile monitoring. It is using this information to build complicated models such as real-time, multiparameter modeling systems. The models will be used to look at the dynamic interactions within local rivers and estuary ecosystems. By incorporating real-time analysis of data into environmental research, scientists are advancing their understanding of major ecological challenges. Streaming technology opens new fields of research and takes the concept of scientific data collection and analysis in a new direction. They are looking

at data they may have collected in the past in a new way and are also able to collect new types of data sources.

Although one can learn a lot by monitoring change variables such as water temperature and water chemistry at set intervals over time, you may miss out on identifying changes or patterns. When you have the opportunity to analyze streaming data as it happens, it is possible to pick up on patterns you might have missed. Real-time data on river motion and weather is used to predict and manage river changes. Scientists are hoping to predict environmental impacts and forecast weather. They are furthering research on the impact of global warming. They are asking what can be learned from watching the movements of migrating fish. How can watching how pollutants are transported help to clean up from future environmental contamination?. If data scientists are able to take data they have already collected, they can combine it with the real-time data in a much more efficient manner. They also have the capability to do more in-depth analysis and do a better job of predicting future outcomes. Because this analysis is completed, it allows other groups needing the same information to be able to use the findings in new ways to analyze the impact of different issues. This data could be stored in a data cloud environment so that researchers across the globe can have access, add new data into the mix, and solve other environmental problems, Zik (2011).

Schadt et al (2010) focuses on Computational solutions to large-scale data management and analysis which presents cloud and heterogeneous computing as solutions for tackling large-scale and high-dimensional data sets. These technologies have been around for years, raising the question: why are they not used more often in bioinformatics? The answer is that, apart from introducing complexity, they quickly break down when a large amount of data is communicated between computing nodes. In their Review, Schadt and colleagues state that computational analysis in biology is high-dimensional, and predict that petabytes, even exabytes, of data will be soon stored and analysed. We agree with this predicted scenario and illustrate, through a simple calculation, how suitable current computational technologies really are for such large volumes of data. Currently, it takes minimally 9 hours for each of 1,000 cloud nodes to process 500 GB, at a cost of US$3,000 (500 GB to 500 TB of total data). The bottleneck in this process is the input/output (IO) hardware that links data storage to the calculation node. All nodes are idle for long periods, waiting for data to arrive from storage; shipping the data on a hard disk to the data storage would not resolve this bottleneck. We estimate that 1,000 cloud nodes each processing 1 petabyte (1 petabyte to 1 exabyte of total data) would take 2 years, and cost $6,000,000. In the calculations, 1,000 computational nodes each processing 500 GB would take 9 hours (at a rate of 15 MB/s) using large nodes at US$0.34/h. The total cost for a single analysis run would be $1,000 \times 9 \times 0.34 = \$3,060$. In reality, throughput will be lower because of competition for access to data storage caused by parallel processing. There are significant throughput instability and abnormal delay variations, even when the network is lightly utilized. In the illustrated example, 1,000 cloud nodes each processing a petabyte would take 750 days (at 15 MB/s) and cost $1,000 \times 750 \times 24 \times 0.34 = \$6,120,000$. A less expensive option would be to use heterogeneous computing, in which graphics processing units (GPUs) are used to boost speed. A similar calculation shows, however, that GPUs are idle 98% of the time when processing 500 GB of data. GPU performance rapidly degrades when large volumes of data are communicated, even with state-of-the-art disk arrays. Furthermore, GPUs are vector processors that are suitable for a subset of computational problems only. Which is the best way forward? Computer systems that provide fast access to petabytes of data will be essential. Because high-dimensional large data sets exacerbate IO issues, the future lies in developing highly parallelized IO using the shortest possible path between storage and central processing units (CPUs). Examples of this trend are Oracle Exadata2and IBM Netezza, which offer parallelized exabyte analysis by providing CPUs on the storage itself. Another trend for improving speed is the integration of photonics and electronics.To fully exploit the parallelization of computation, bioinformaticians will also have to adopt new programming languages, tools and practices, because writing correct software for concurrent processing that is efficient and scalable is difficult. The popular Rprogramming language, for example, has only limited support for writing parallelized software. However, other languages can make parallel programming easier by, for example, abstracting threads and shared memory. So, not only do cloud and heterogeneous computing suffer from severe hardware bottlenecks, they also introduce (unwanted) software complexity. It is our opinion that large multi-CPU computers are the preferred choice for handling big data. Future machines will integrate CPUs, vector processors and random access memory (RAM) with parallel high-speed interconnections to optimize raw processor performance. Our calculations show that for petabyte- to exabyte-sized high-dimensional data, bioinformatics will require unprecedented fast storage and IO to perform calculations within an acceptable time frame. Golpayegani et al(2009) paper on Cloud Computing for Satellite Data Processing on High End Compute Clusters where they focused on Hadoop and MapReduce framework of which they suggested improvement without having to changing its coding.They agreed that these tools are well suited to analysis of big data in the cloud. Baraglia et al (2010) discussed the huge deluge of data in the cloud and the fact that it can be handled by the lattars architecture albeit with shortcomings which he gave algorithmic solutions.

3. Massively Parallel Compute (Analytic Algorithms)

Parallel computing is a well-adopted technology seen in processor cores and software thread-based parallelism.

However, massively parallel processing— leveraging thousands of networked commodity servers constrained only by bandwidth—is now the emerging context for the Data Cloud. If distributed file systems, such as GFS and HDFS, and column-oriented databases are employed to store massive volumes of data, there is then a need to analyze and process this data in an intelligent fashion. In the past, writing parallel code required highly trained developers, complex job coordination, and locking services to ensure nodes did not overwrite each other. Often, each parallel system would develop unique solutions for each of these problems. These and other complexities inhibited the broad adoption of massively parallel processing, meaning that building and supporting the required hardware and software was reserved for dedicated systems. MapReduce, a framework pioneered by Google, has overcome many of these previous barriers and allows for data-intensive computing while abstracting the details of the Data Cloud away from the developer. This ability allows analysts and developers to quickly create many different parallelized analytic algorithms that leverage the capabilities of the Data Cloud. Consequently, the same MapReduce job crafted to run on a single node can as easily run on a group of 1,000 nodes, bringing extensive analytic processing capabilities to users in the enterprise. Working in tandem with the distributed file system and the multi-dimensional database, the MapReduce framework leverages a master node to divide large jobs into smaller tasks for worker nodes to process. The framework, capable of running on thousands of machines, attempts to maintain a high level of affinity between data and processing, which means the framework intelligently moves the processing close to the data to minimize bandwidth needs. Moving the compute job to the data is easier than moving large amounts of data to a central bank of processors. Moreover, the framework manages extrapolative errors, noticing when a worker in the cloud is taking a long time on one of these tasks or has failed altogether and automatically tasks another node with completing the same task. All these details and abstractions are built into the framework. Developers are able to focus on the analytic value of the jobs they create and no longer worry about the specialized complexities of massively parallel computing. An intelligence analyst is able to write 10 to 20 lines of computer code, and the MapReduce framework will convert it into a massively parallel search—working against petabytes of data across thousands of machines—without requiring the analyst to know or understand any of these technical details. Tasks such as sorting, data mining, image manipulation, social network analysis, inverted index construction, and machine learning are prime jobs for MapReduce. In another scenario, assume that terabytes of aerial imagery have been collected for intelligence purposes. Even with an algorithm available to detect tanks, planes, or missile silos, the task of finding these weapons could take days if run in a conventional manner. Processing 100 terabytes of imagery on a standard computer takes 11 days. Processing the same amount of data on 1,000 standard computers takes 15 minutes. By incorporating MapReduce, each image or part of an image becomes its own task and can be examined in a parallel manner, Farber et al (2011).

Tsuchiya 2012 examined two processing technologies of big data in the cloud and referred to works done in Fujistu laboratories. Consumption of big data can be made as easy as online media according to Foster (2009) after he studied geosciences data on the cloud.Note that two technical entities have come together. First, there's big data for massive amounts of detailed information. Second, there's advanced analytics, which is actually a collection of different tool types, including those based on predictive analytics, data mining, statistics, artificial intelligence, natural language processing, and so on. Put them together and you get big data analytics, the hottest new practice in BI today. Of course, businesspeople can learn a lot about the business and their customers from BI programs and data warehouses. But big data analytics explores granular details of business operations and customer interactions that seldom find their way into a data warehouse or standard report. Some organizations are already managing big data in their enterprise data warehouses (EDWs), while others have designed their DWs for the well-understood, auditable, and squeaky clean data that the average business report demands. The former tend to manage big data in the EDW and execute most analytic processing there, whereas the latter tend to distribute their efforts onto secondary analytic platforms. There are also hybrid approaches.

4. Vendor Products for Big Data Analytics

Since the firms that sponsored this report are all good examples of software and hardware vendors that offer tools, platforms, and services for big data analytics, let's take a brief look at the product and service portfolio of each. The sponsors form a representative sample of the vendor community, yet their offerings illustrate different approaches to big data analytics. Cloudera makes a business by distributing open source software based on Apache Hadoop. IT personnel demand a number of features and services that Hadoop lacks. To help organizations reliably use Hadoop in production, Cloudera Enterprise is specifically designed to improve the manageability of Hadoop deployments. Cloudera makes Hadoop viable for serious enterprise users by providing technical support, upgrades, administrative tools for Hadoop clusters, professional services, training, and certification. Hence, Cloudera collects and develops additional components to strengthen and extend Hadoop, while still retaining Hadoop's open-source affordability, big data scalability, and flexibility across a wide range of data types. EMC Corporation is the world's leading provider of data storage platforms and other information infrastructure solutions. In 2010, EMC acquired Greenplum and has since built it up as the EMC. Data Computing Division, which has become a leading platform for big data analytics. Greenplum

customers are some of the largest firms in the world, and they regularly deploy Greenplum products on grids or clouds to scale up to very big data. EMC Greenplum Database is known for its shared nothing massively parallel processing (MPP) architecture, high-performance parallel dataflow engine, and gNet software interconnect technology. Recently, EMC Greenplum has released Greenplum HD (an enterprise-ready Hadoop distribution), EMC Greenplum Data Computing Appliance Product Family (purpose-built for big data analytics), and Greenplum Chorus (software for collaboration over analytics). IBM has one of the largest product portfolios of any software vendor, with analytics as a significant focus in support of IBM's global campaign for Business Analytics and Optimization (BAO). Within this massive portfolio, three products stand out because of their recent contributions to enabling big data analytics. First, IBM's acquisition of Netezza in 2010 adds to the portfolio the product that invented the data warehouse appliance and defined the modern analytic database platform. Second, just announced in 2011, IBM InfoSphere BigInsights is IBM's Hadoop-based offering that combines the power of Hadoop with IBM-unique code to address enterprise requirements. Enterprise features include built-in text analytics, a spreadsheet-style data discovery and exploration tool, enterprise grade security, and administrative tools. Third, IBM InfoSphere Streams is a platform for real-time analytic processing (RTAP), which uniquely provides velocity for big streaming data analytics on structured and unstructured data. Impetus Technologies offers product engineering and technology R&D services for software product development. In the area of big data analytics, Impetus offers consulting, advisory, and professional services. Impetus' customers are large corporations that manage big data as part of operating a business, but most clients also leverage big data with analytics. Impetus helps such firms evaluate and embrace new technologies and business practices that are related to big data analytics. Impetus provides advisory consulting (to assess big data and analytic opportunities), implementation consulting (to design and develop big data analytic infrastructure and applications), and long-term support (to help clients evolve as new practices and technologies for big data analytics evolve).

Impetus Technologies provides end-to-end, vendor- and technology-agnostic advice and engineering to objectively determine what's best for the client's business goals and how to achieve the goals with new technologies and practices, Russom (2011). Radical customization, constant experimentation and novel business models will be new hallmarks of competition as companies capture and analyze huge volumes of data Brown et al (2013). As the size of data set in cloud increases rapidly, how to process large amount of data efficiently has become a critical issue. MapReduce provides a framework for large data processing and is shown to be scalable and fault-tolerant on commodity machines. However, it has higher learning curve than SQL-like language and the codes are hard to maintain and reuse. On the other hand, traditional SQL-based data processing is familiar to user but is limited in scalability. In this paper, we propose a hybrid approach to fill the gap between SQL-based and MapReduce data processing. We develop a data management system for cloud, named SQLMR. SQLMR complies SQL-like queries to a sequence of MapReduce jobs. Existing SQL-based applications are compatible seamlessly with SQLMR and users can manage Tera to PataByte scale of data with SQL-like queries instead of writing MapReduce codes. We also devise a number of optimization techniques to improve the performance of SQLMR. The experiment results demonstrate both performance and scalability advantage of SQLMR compared to MySQL and two NoSQL data processing systems, Hive and HadoopDB.

Abadi et al (2009) article discusses the limitations and opportunities of deploying data management issues on these emerging cloud computing platforms (e.g., Amazon Web Services). We speculate that large scale data analysis tasks, decision support systems, and application specific data marts are more likely to take advantage of cloud computing platforms than operational, transactional database systems (at least initially). They present a list of features that a DBMS designed for large scale data analysis tasks running on an Amazon-style offering should contain. they then discuss some currently available open source and commercial database options that can be used to perform such analysis tasks, and conclude that none of these options, as presently architected, match the requisite features. They then express the need for a new DBMS, designed specifically for cloud computing environments.

5. The Research Problem

Because of its massive data sets, today environmental sustainability meets big data. The problem is not data but interpretation. To start, environmental sustainability is a complex idea. Generally speaking, most people are not well-versed in the language concerning the idea. Overly complex and hard to understand, executives are unable to make intuitive decisions based on the information, because they cannot interpret the data. The data lacks context and therefore holds little meaning to top-level executives. Executives are lost in a sea of information, laced with confusing terminology. This lack of clarity makes business decisions even more difficult. To an outsider, resource consumption isn't measured in traditional terms, but rather in a set of foreign "currencies" without a currency converter:

- Water in kgal
- Electricity in kWh
- Heating in mBTu
- CO_2 in tons

The problem that appears to be halting the harmonious relationship between Big Data and resource consumption information is not the data. The root of the problem rests in the interpretation of the information and data sets. The numbers are already difficult to digest and unintuitive, so it doesn't help the case that there isn't a commonly shared,

universal language in the measurement of environmental sustainability. The Solution to this problem is through the development of a mathematically rigorous, yet simple and intuitive way to interpret the different data streams, companies may be able to make better business decisions. Companies currently use business intelligence and analytics to better understand and predict future trends. Through the use of Big Data, similar techniques can be applied to better understand businesses processes and environmental sustainability efforts. For instance, businesses might transform each domain-specific resource into the energy used to create that resource. This can then expressed in terms of what we call Energy Points, using the equivalent of the embodied energy of a gallon of gasoline as a unit and factoring in local parameters such as water scarcity. So instead of wondering what is the relative importance of 1,000 kgal and 1,000 kWh, businesses can simply treat each domain like Weight Watcher's treats calories – based on efficiency points. This is one way we collapse Big Data to a common metric to address resource consumption decisions.

There are other benefits that come along with using Big Data. Companies are able to share and access real-time analytics and data sets, allowing progressive companies and organizations to release data to a broader ecosystem. A relatively simple task for IT, companies can exponentially increase efficiencies and provide the new material for building tangential business models and interactions. Big Data is playing an increasingly critical role in decision making, given simple and rigorous interpretation, which is the power to help transform the sustainability from an exercise in feel good terminology, to a quantifiable approach that has a true impact on our environment. The road to get there is paved with mathematics – developing intuitive, yet mathematically rigorous ways to interpret the data says Dr. Zik.

6. Methodology

In this study a private cloud is built using Ubuntu and Eucalyptus open source software on two quad processor machines with 8GB Ram. Apache™ Flume used is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tunable reliability mechanisms for failover and recovery. Flume lets Hadoop users make the most of valuable log data. Specifically, Flume allows users to:

- Stream data from multiple sources into Hadoop for analysis
- Collect high-volume Web logs in real time
- Insulate themselves from transient spikes when the rate of incoming data exceeds the rate at which data can be written to the destination
- Guarantee data delivery
- Scale horizontally to handle additional data volume

Flume's high-level architecture is focused on delivering a streamlined codebase that is easy-to-use and easy-to-extend. The project team has designed Flume with the following components:

- Event – a singular unit of data that is transported by Flume (typically a single log entry)
- Source – the entity through which data enters into Flume. Sources either actively poll for data or passively wait for data to be delivered to them. A variety of sources allow data to be collected, such as log4j logs and syslogs.
- Sink – the entity that delivers the data to the destination. A variety of sinks allow data to be streamed to a range of destinations. One example is the HDFS sink that writes events to HDFS.
- Channel – the conduit between the Source and the Sink. Sources ingest events into the channel and the sinks drain the channel.
- Agent – any physical Java virtual machine running Flume. It is a collection of sources, sinks and channels.
- Client – produces and transmits the Event to the Source operating within the Agent

A flow in Flume starts from the Client. The Client transmits the event to a Source operating within the Agent. The Source receiving this event then delivers it to one or more Channels. These Channels are drained by one or more Sinks operating within the same Agent. Channels allow decoupling of ingestion rate from drain rate using the familiar producer-consumer model of data exchange. When spikes in client side activity cause data to be generated faster than what the provisioned capacity on the destination can handle, the channel size increases. This allows sources to continue normal operation for the duration of the spike. Flume agents can be chained together by connecting the sink of one agent to the source of another agent. This enables the creation of complex dataflow topologies. Given the rigorous demands that big data places on networks, storage and servers, it's not surprising that some customers would outsource the hassle and expense to the cloud. Although cloud providers say they welcome this new business opportunity, supporting in the cloud is forcing them to confront various, albeit manageable, architectural hurdles. The elasticity of the cloud makes it ideal for big data analytics -- the practice of rapidly crunching large volumes of unstructured data to identify patterns and improve business strategies -- according to several cloud providers. At the same time, the cloud's distributed nature can be problematic for big data analysis. "If you're running Hadoop clusters and things like this, they put a really heavy load on storage, and in most clouds, the performance of the storage isn't good enough," said Robert Jenkins, co-founder and chief technology officer of CloudSigma, a Zurich-based Infrastructure as a Service (IaaS) provider. "The big problem with clouds is making the storage perform to a level that enables this kind of computing, and this would be the biggest

reason why some people wouldn't use the cloud for big data processing." But Jenkins and other cloud providers emphasized that these challenges aren't insurmountable, and many providers already have plans to tweak their cloud architectures to improve the capacity, performance and agility of all their cloud services -- moves that they expect will also provide better support for big data in the cloud. Devising an architecture that supports big data analysis in the cloud is no more daunting than meeting the challenges of satiating the rapidly growing appetite for cloud services in general, according to Henry Fastert, chief technologist and managing partner at SHI International, a large reseller, managed service provider (MSP) and cloud provider based in Somerset, N.J.

7. Challenges of Cloud Storage and Big Data Analysis

The cloud storage challenges in big data analytics fall into two categories: capacity and performance. Scaling capacity, from a platform perspective, is something all cloud providers need to watch closely. "Data retention continues to double and triple year-over-year because customers are keeping more of it. Certainly, that impacts us because we need to provide capacity," Corvaia said. Storage performance in a highly virtualized, distributed cloud can be tricky on its own, and the demands of big data analysis only magnify the issue, several cloud providers said.SHI International's cloud strategy is built on the company's vCore model, its branding for a proprietary "finite collection of servers, storage and switching elements," replicated across SHI's cloud, Fastert said. The distributed storage architecture enables SHI to "really optimize the performance of our infrastructure because it's set up in that granular fashion," he said. "Storage is something that's also impacted by specific types of virtualization loads, and so the way in which you spread tasks across your storage … will always impact your performance," he said. "The [vCore] model allows us to spread loads based on the characteristics of those loads, and so we constantly look at the characteristics of customers' loads across our vCore infrastructure … and then we do load balancing across them from a storage-performance point of Hadoop is an open source legend built by software heroes. Yet, legends can sometimes be surrounded by myths—these myths can lead IT executives down a path with rose-colored glasses. Data and data usage is growing at an alarming rate. Just look at all the numbers from analysts—IDC predicts a 53.4% growth rate for storage this year, AT&T claims 20,000% growth of their wireless data traffic over the past 5 years, and if you take at your own communications channels, its guaranteed that the internet content, emails, app notifications, social messages, and automated reports you get every day has dramatically increased. This is why companies ranging from McKinsey to Facebook to Walmart are doing something about big data.

Just like we saw in the dot-com boom of the 90s and the web 2.0 boom of the 2000s, the big data trend will also lead companies to make some really bad assumptions and decisions. Hadoop is certainly one major area of investment for companies to use to solve big data needs. Companies like Facebook that have famously dealt well with large data volumes have publicly touted their successes with Hadoop, so its natural that companies approaching big data first look to the successes of others. A really smart MIT computer science grad once told me, “when all you have is a hammer, everything looks like a nail.” This functional fixedness is the cognitive bias to avoid with the hype surrounding Hadoop. Hadoop is a multi-dimensional solution that can be deployed and used in different way. Let’s look at some of the most common pre-conceived notions about Hadoop and big data that companies should know before committing to a Hadoop project:

8. Myths about Big Data

8.1. Big Data is Purely about Volume

Besides volume, several industry leaders have also touted variety, variability, velocity, and value. Putting all arguments about alliteration aside, the point is that data is not just growing—it is moving further towards real-time analysis, coming from structured and unstructured sources, and being used to try and make better decisions. With these considerations, analyzing a large volume of data is not the only way to achieve value. For example, storing and analyzing terabytes of data over time might not add nearly as much value as analyzing 1 gigabyte of really important, impactful information in real time. From a tool-set perspective, you might want an in-memory data grid built for real-time pricing calculations instead of a way to slice and dice historical prices into a dead horse.

8.2. Traditional SQL Doesn’T Work with Hadoop

When Facebook, Twitter, Yahoo! and others bet big on Hadoop, they also knew that HDFS and MapReduce were limited in their ability to deal with expressive queries through a language like SQL. This is how Hive, Pig, and Sqoop were ultimately hatched. Given that so much data on earth is managed through SQL, many companies and projects are offering ways to address the compatibility of Hadoop and SQL. Pivotal HD’s HAWQ is one example—a parallel SQL-compliant query engine that has shown to be 10 to 100s of times faster than other Hadoop query engines in the market today—and it was built to support petabyte data sets.

8.3. Kill the Mainframe! Hadoop is the Only the New IT Data Platform

There are many longstanding investments in the IT portfolio, and the mainframe is an example of one that

probably should evolve along with ERP, CRM, and SCM. While the mainframe isn't being buried by companies, it definitely needs a new strategy to grow new legs and expand on the value of it's existing investment. For many of our customers that run intoissues with mainframe speed, scale, or cost, there are incremental ways to evolve the big iron data platform and actually get more use out of it. For example, in-memory, big data grids like vFabric SQLFire can be embedded or use distributed caching approaches for dealing with problems like high-speed ingest from queues, speeding mainframe batch processes, or real-time analytical reporting.

8.4. Virtualized Hadoop Takes a Performance Hit

Hadoop was designed originally to run on bare metal servers, however as adoption has grown many companies want it as a data center service running in the cloud. Why do companies want to virtualize Hadoop? First, let's consider the ability to manage infrastructure elastically—we quickly realize that scaling compute resources, like virtual Hadoop nodes, help with performance when data and compute are separated—otherwise, you would take a Hadoop node down and lose the data with it or add a node and have no data with it. Major Hadoop distributions from MapR, Hortonworks, Cloudera, and Greenplum all support Project Serengeti and Hadoop Virtualization Extensions (HVE) for this reason. In addition, our research with partners has show that Hadoop works quite well on vSphere and can even perform better under certain conditions—running 2 or 4 smaller VMs per physical machine often resulted in better performance, up to 14% faster, than a native approach according to benchmarks we've done with partners.

8.5. Hadoop Only Works in Your Data Center

First of all, there are SaaS-based, cloud solutions, like Cetas, that allow you to run Hadoop, SQL, and real-time analytics in the cloud without investing the time and money it takes do build a large project inside your data center. For a public cloud runtime, Java developers can probably benefit from Spring Data for Apache Hadoop and the related examples on GitHub oronline video introduction.

8.6. Hadoop Doesn'T Make Financial Sense to Virtualize

Hadoop is typically explained as running on a bank of commodity servers—so, one might conclude that adding a virtualization layer adds extra cost but no extra value. There is a flaw in this perspective—you are not considering the fact that data and data analysis are both dynamic. To become an organization that leverages the power of Hadoop to grow, innovate, and create efficiencies, you are going to vary the sources of data, the speed of analysis, and more. Virtualized infrastructure still reduces the physical hardware footprint to bring CAPEX in line with pure commodity hardware, and OPEX is reduced through automation and higher utilization of shared infrastructure.

8.7. Hadoop Doesn'T Work on SAN or NAS

Hadoop runs on local disks, but it can also run well in a shared SAN environment for small to medium sized clusters with different. High bandwidth networks like 10GB Ethernet, FoE, and iSCSI can also support effective performance.

9. Actions to Overcome the Myths

While many of us are fans of big data, this list can help you take a step back and look objectively at the right approach to solving your big data problems. Just like some building projects need hammers and others need screwdrivers, hacksaws, or a welding torch, Hadoop is just one tool to help conquer big data problems. High velocity data may push you towards an in-memory, big data grid like GemFire or SQLFire. A need for massive, consumer-grade web scale may mean you need message-oriented middleware like RabbitMQ. Getting to market faster may mean you need to look at a full SaaS solution like Cetas, and Redis may meet your needs and find a home in your stack much easier than a full blown Hadoop environment.

10. Discussion

In the studies of Lamonica (2013) which focused on crop agriculture information, improvement of inclusion of weather conditions that would greatly enhance its performance and be generally beneficial to all forms of agriculture. Similar studies would improve by integrating any big data analytic in the market today like Hadoop.This would bring business sense from the confused form of Big data. Applications of sensors to track animals Andrea (2013) would be enhanced if such applications would be mounted on clouds so as to have a wide coverage accessibility.

Dr Zik(2009) suggested that the way to better analysis lies in mathematics and his studies found out that analytics would be improved by application of some mathematical modeling. Computational techniques break down when applied to large unstructured data, Schadt et al (2010).This is due to traditional IS limited architectures a problem eliminated by cloud computing and analytics.Elasticity and scalability is the limiting factor of big data on traditional IS. This is sorted by Meng-Ju Hsieh et al (2011) by the use of MapReduce and SQL data processing methodologies. This also gives a solution to Abadi (2009) where cloud compatible databases are required.

11. Conclusion

The Big data concept is surely the new elephant in the block. Organizations just have to embrace it to have a competitive advantage over their peers especially if they can integrate their big data into a cloud big data analytic to produce sensible relative output which an be used by

different sectors positively. The huge deluge of data provides a great opportunity of business intelligence to those who will keep up with technology. There are still lots of research opportunity to study the cloud computing analytics and database frameworks. Environmental big data would doubtlessly enrich agricultural endeavors and quality of living.

REFERENCES

[1] Cloud Computing for Satellite Data Processing on High End Compute ClustersN. Golpayegani University of Maryland, Baltimore County golpa1@umbc.edu Prof. M. Halem University of Maryland, Baltimore County halem@umbc.edu

[2] Big data, but are we ready?: Correspondence by Schadt et al. | Review by Schadt et al. Oswaldo Trelles1,5, Pjotr Prins2,5, Marc Snir3 & Ritsert C. Jansen4, Author affiliations, Oswaldo Trelles is at the Computer Architecture Department, University of Malaga, Campus de Teatinos, E-29071, Spain. Pjotr Prins and Ritsert C. Jansen are at the Groningen Bioinformatics Centre, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands.

[3] Large-scale Data Analysis on the Cloud Ranieri Baraglia1, Claudio Lucchese1, and Gianmarco De Francisci Morales1;2 1 ISTI-CNR, Pisa, Italy 2 IMT - Institute for Advanced Studies, Lucca, Italy

[4] Massive Data Analytics and the Cloud,A Revolution in Intelligence Analysis by Michael Farber et al

[5] Big data processing in cloud environments, Tsuchiya 2012

[6] Wagging the long tail of earth science:Why we need an earth science data web, and how to build it Ian Foster, Daniel S. Katz, Tanu Malik Peter Fox Computation Institute Tetherless World Constellation University of Chicago Rensselaer Polytechnic Institute

[7] Are you ready for the era of 'big data'? Brad Brown, Michael Chui, and James Manyika Data Management in the Cloud: Limitations and Opportunities Daniel J. Abadi Yale University New Haven, CT, USA dna@cs.yale.edu

[8] Computational solutions to large scale data management analysis by Schadt et al(2010) White paper Big Data: What It Is and Why You Should Care sponsored by: AMD Richard L. Villars Carl W. Olofson Matthew Eastwood, http://investeddevelopment.com/blog/2013/10/weekly-review-october-11-big-data-in-agriculture/#sthash.bjtpxjkf.dpuf, microsoft cloud to power environmental big data By katie fehrenbacher

[9] Big data analysis in the cloud: Storage, network and server challenges by Jessica Scarpati

[10] Big data analytics fourth quarter 2011 By Philip Russo

[11] SQLMR : A Scalable Database Management System for Cloud Computing Meng-Ju Hsieh ; Inst. of Inf. Sci., Acad. Sinica, Taipei, Taiwan ; Chao-Rui Chang ; Li-Yung Ho ; Jan-Jan Wu Published in: Parallel Processing (ICPP), 2011 International Conference on Date of Conference: 13-16 Sept. 2011 Page(s): 315 – 324 ISSN : 0190-3918 E-ISBN : 978-0-7695-4510-3 Print ISBN:978-1-4577-1336-1 INSPEC Accession Number:1231622