R-Wingu, the ‘Big Data’ Analytic Framework: A Solution to Intelligent Correlation of Research Output in a Private Cloud Prototype for Seamless Research Ontologies

Duncan Waga^{1,*}, Kefa Rabah², James Ogalo¹

¹Jaramogi Oginga Odinga University of Science and Technology, Kenya

²Kabarak University, Kenya

*Corresponding Author: wagadun@gmail.com

Copyright © 2014 Horizon Research Publishing All rights reserved.

Abstract ‘Internet of Things’ is choking the world with Zettabyte scale data rate of which traditional computing can neither store nor process. Outputs of various kinds of Research endeavors are held in different media over the years and cannot be made to “talk to each other” to exploit each other for beneficial relative information. The prime objective is to find a platform that can easily accommodate Cloud technology which can elastically handle the big data concept. In this study a private cloud is built using Ubuntu and Eucalyptus open source software on two quad processor machines with 8GB Ram. Using Apache Flume and Hadoop analytics, Big Data obtained from Research outputs is mounted on the R-Wingu framework with capacity to data mine unstructured, unrelated data and relate the components intelligently and may be leased as Database as a Service(DaaS). Once disparate data streams are accessible in real time, in one place and a consistent fashion, data suddenly becomes much more powerful and decisions become that much more impactful. Surveys are conducted across research communities with an aim of isolating research Gaps through various statistical tools. Sample results are discussed alongside previous studies’ outputs. Research Gaps filled are compared by poll results conducted with exiting outcomes. Establishments are urged to embrace Big Data approach for competitive advantage.

Keywords Big Data, Internet of Things, Apache Flume, Hadoop Analytics, Ubuntu Eucalyptus, Cloud Computing

1. Introduction

The Big Data concept is emergent phenomenons that has been brought about by the “internet of things”, i.e the numerous devices that are internet enabled are have the capacity to emit data in various forms. This data is so huge to the rate of 7ZB per year and traditional computing cannot

process neither can it store it. It is more of in a confused state since put together it doesn’t make any sense. Just like combining web2.0 data with weather data may not give any sensible correlations. This study will focus in various kinds of research outputs being amalgamated and analyzed together to give out useful correlations. Research could be very useful if it has a more global audience and outputs being able to be subjected to other scenario for useful different results. Most establishments boast of having safely stored huge capacities of their data yet only able to positively utilize a third of it. This is common with businesses and research communities. Cloud computing is the one of the best platforms that may handle this kind of data since it has an elastic way of resource management. R-Wingu is a platform of a private cloud computing with Hadoop and Apache Flume open source analytic tool. It is intended to harvest Big data and have it efficiently processed. In this study a private cloud is built using Ubuntu and Eucalyptus open source software on two quad processor machines with 8GB Ram of which Hadoop Assume is integrated. It will centralize research outputs and avail it to a wider global audience of which they will be able to intelligently interrogate and analyse the Big Research Data and be able to produce very puzzling but useful correlations. R-Wingu will bring together resources which at the moment is scattered and in the disparate state not very useful.

2. Literature Review

Allen et al (1979)’s study entitled Cambridge crystallographic Data center computer based search retrieval, analysis and display information which focused on data obtained from research documents and databases the latter which outgrew its specifications thereby affecting the efficacy of the framework. Big data definitely will change how research is conducted and its general view. Commenting on computational social science, Lazer et al

argue that it offers 'the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (2009, p. 722). But it is not just a matter of scale. Neither is enough to consider it in terms of proximity, or what Moretti (2007) refers to as distant or close analysis of texts. Rather, it is a profound change at the levels of epistemology and ethics. It reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality. Just as du Gay and Pryke note that 'accounting tools...do not simply aid the measurement of economic activity, they shape the reality they measure' (2002, pp. 12-13), so Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life Boyd et al(2011). Agichtein et al(2008) focused on identification of high quality content from social media related data where content was categorized under excellent, abuse and spam. They came up with various methods of isolating these categories. Content analysis involving medical information in websites, blogs and wikis was conducted by Denecke (2009) of which comparative analyses were conducted and used extraction technologies to selectively obtain medical concepts. Lee et al (2010) conducted similar experimentation but used MapReduce software framework which could easily handle the large data streams which improved the data flow by 72%. MapReduce and its popular open source implementation,

Hadoop, are moving toward ubiquitous for Big Data storage and processing. Therefore, it is essential to quantitatively evaluate and characterize the Hadoop deployment through extensive benchmarking, Huang et al(2010). In their study they present HiBench, a representative and comprehensive benchmark suite for Hadoop, which consists of a set of Hadoop programs including both synthetic micro-benchmarks and real-world applications. The Sort, WordCount and TeraSort programs contained in the Hadoop distribution are three popular micro-benchmarks

Widely used in the community, and therefore are included in HiBench. Both the Sort and WordCount programs are representative of a large subset of real-world MapReduce jobs– one transforming data from one representation to another, and another extracting a small amount of interesting data from a large data set, Huang et al(2010). In HiBench, the input data of Sort and WordCount workloads are generated using the RandomTextWriter program contained in the Hadoop distribution. The TeraSort workload sorts 10 billion 100-byte records generated by the

TeraGen program contained in the Hadoop distribution. Chen et al (2012) asserts that within the past few years, organizations in diverse industries have adopted MapReduce-based systems for large-scale data processing. Along with these new users, important new workloads have emerged which feature many small, short, and increasingly interactive jobs in addition to the large, long-running batch jobs for which MapReduce was originally designed. However, these new workloads have not yet been described in the literature. Guo's paper analyzes and designs a

monitoring system on public topic based on cloud computing and NLP technology. The system solves the internet massive data processing and computational complexity based on Hadoop platform; it realizes the analysis on the web page, extraction of public opinion and tracking technology based on NLP techniques and machine learning technology; it also can analyze the feelings on the users' comments and further determine the trend of public topic based on emotional thesaurus; finally, it provides a visual interface and the retrieval interface for users to use this system. Implementation of the system will improve the efficiency and quality of public topic Guo Li(2013). Bryant et al(2008) focused on Big-Data Technology: Sense, Collect, Store, and Analyze content. The following technologies has ensured more visibility to big data:-

Sensors: Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages.

Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet.

Data storage: Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, costs around \$100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes.

Cluster computer systems: A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing (e.g., supercomputers), where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms.

Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can rent storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own

power plants, we can foresee an era where data storage and computing become utilities that are ubiquitously available.

Data analysis algorithms: The enormous volumes of data require automated or semiautomated analysis – techniques to detect patterns, identify anomalies, and extract knowledge. Again, the "trick" is in the software algorithms - new forms of computation, combining statistical analysis, optimization, and artificial intelligence, are able to construct statistical models from large collections of data and to infer how the system should respond to new data. For example, Netflix uses machine learning in its recommendation system, predicting the interests of a customer by comparing her movie viewing history to a statistical model generated from the collective viewing habits of millions of other customers.

Reinsel et al (2011) New capture, search, discovery, and analysis tools can help organizations gain insights from their unstructured data, which accounts for more than 90% of the digital universe. These tools can create data about data automatically, much like facial recognition routines that help tag Facebook photos. Data about data, or metadata, is growing twice as fast as the digital universe as a whole, he says. Business intelligence tools increasingly are dealing with real-time data, whether it's charging auto insurance premiums based on where people drive, routing power through the intelligent grid, or changing marketing messages on the fly based on social networking responses. New storage management tools are available to cut the costs of the part of the digital universe we store, such as deduplication, auto-tiering, and virtualization, as well as to help us decide what exactly to store, as in content management solutions.

Their study concludes that an entire industry has grown up to help us follow the rules (laws, regulations, and customs) pertaining to information in the enterprise. It is now possible to get regulatory compliance systems built into storage management systems. Trelles et al(2010), exposes that today we can generate hundreds of gigabases of DNA and RNA sequencing data in a week for less than US\$5,000. The astonishing rate of data generation by these low-cost, high-throughput technologies in genomics is being matched by that of other technologies, such as real-time imaging and mass spectrometry-based flow cytometry. Success in the life sciences will depend on our ability to properly interpret the large-scale, high-dimensional data sets that are generated by these technologies, which in turn requires us to adopt advances in informatics. Here they discuss how they can master the different types of computational environments that exist - such as cloud and heterogeneous computing - to successfully tackle big data problems.

Agrawal et al(2011) worked on Scalable database management systems (DBMS) and gave alternative techniques which cause nightmares to DBMS designers in developing and deployment to internet scale applications.

3. Methodology

In this study a private cloud is built using Ubuntu and Eucalyptus open source software on two quad processor machines with 8GB Ram.

3.1. Preparing for the Installation

First, download the CD image for the Ubuntu Server remix — we're using version 9.10 — on any PC with a CD or DVD burner. Then burn the ISO image to a CD or DVD. If you want to use a DVD, make sure the computers that will be in the cloud read DVDs. If you're using Windows 7, you can open the ISO file and use the native burning utility. If you're using Windows Vista or later, you can download a third-party application like DoISO.

Before starting the installation, make sure the computers involved are setup with the peripherals they need (i.e., monitor, keyboard and mouse). Plus, make sure they're plugged into the network so they'll automatically configure their network connections.

3.2. Installing the Front-End Server

The installation of the front-end server is straightforward. To begin, simply insert the install CD, and on the boot menu select "Install Ubuntu Enterprise Cloud", and hit Enter. Configure the language and keyboard settings as needed. When prompted, configure the network settings.

When prompted for the Cloud Installation Mode, hit Enter to choose the default option, "Cluster". Then you'll have to configure the Time Zone and Partition settings. After partitioning, the installation will finally start. At the end, you'll be prompted to create a user account.

Next, you'll configure settings for proxy, automatic updates and email. Plus, you'll define a Eucalyptus Cluster name. You'll also set the IP addressing information, so users will receive dynamically assigned addresses.

3.3. Installing and Registering the Node Controller(s)

The Node installation is even easier. Again, insert the install disc, select "Install Ubuntu Enterprise Cloud" from the boot menu, and hit Enter. Configure the general settings as needed.

When prompted for the Cloud Installation Mode, the installer should automatically detect the existing cluster and preselect "Node." Just hit Enter to continue. The partitioning settings should be the last configuration needed.

4. Apache Flume

Apache™ Flume 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.

What Flume Does

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.

There are numerous technologies as compared in Figure 1 below that can be applied to the cloud as in this study.

	<i>Apache Drill</i>	<i>Apache Hive</i>	<i>BigQuery</i>	<i>CitusDB</i>	<i>Hadapt</i>	<i>HAWQ</i>	<i>Impala</i>	<i>Phoenix</i>
Owner	Community	Community	Google	CitusData	Hadapt	Greenplum	Cloudera	Salesforce
Low-latency	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Operational mode	On-premise	On-premise	Hosted, SaaS offering	On-premise	On-premise	Part of Pivotal HD appliance	On-premise	On-premise
Data shapes	Nested, tabular	Nested, tabular	Nested, tabular	Nested, tabular	Tabular	Tabular	Tabular	Tabular
Data sources	Extensible, incl. HDFS, HBase, Cassandra, MongoDB, RDBMS, etc.	HDFS, HBase	N/A	PostgreSQL, MongoDB, HDFS	HDFS/RDBMS	HDFS, HBase	HDFS, HBase	HDFS, HBase
Hadoop dependent	No	Yes	No	No	Yes	No	Yes	Yes
Schema	Optional	Required	Required	Required	Required	Required	Required	Required
License	Apache 2.0	Apache 2.0	ToS/SLA	Commercial	Commercial	Commercial	Apache 2.0/Open Source	Proprietary
Source code	Open	Open	Closed	Closed	Closed	Closed	Open	Open
Query languages	Extensible, incl. SQL 2003, MongoQL, DSL, etc.	HiveQL	SQL subset	SQL	SQL subset	SQL subset	SQL/HiveQL subset	SQL subset
Columnar storage	Yes	Possible	Yes	No	No	Yes	Yes	No

Figure 1. Numerous technologies applied to the cloud.

5. Discussion

Allen et al (1979)'s study brought a lot of light at a time very few analytics were able to cope with data of large magnitude although they ran into database capacity issues which now are addressed comfortably by the elasticity of the clouds. Agichtein et al(2008)focused mainly on high quality data from social media and managed to come up with various techniques of isolation. The techniques could not handle unstructured content which has now been sorted out through Hadoop and flume. Denecke (2009) performed his experiments on medical data in websites blogs and wikis to assist the medic in their medical diagnoses. Unfortunately traditional computation could not handle the enormous data which is currently addressed by cloud computing. Lee et al (2010) engaged the Hadoops mapReduce for an almost similar study and obtained increased speeds of analysis upto 72% of which this paper improves by embracing Cloud technology.

6. Conclusions

Generally cloud computing analytics like Hadoop mapreduce in conjunction with Apache Flume deployed in a cloud framework is a positive way of aggregating big data from many sources in an elastic environment to produce useful intelligence from the big confusion. This is a rather new concept which is open to numerous researches. Organizations would benefit immensely should the data lying idly in their archives be analyzed by such a framework. Different sectors would be advantaged in terms of inter-relations of the elements of the big data that could easily be exploited for competitive advantage.

REFERENCES

- [1] The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers and D. G. Watson
- [2] Six Provocations for Big Data, Paper to be presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011. danah boyd Microsoft Research dmb@microsoft.com Kate Crawford University of New South Wales k.crawford@unsw.edu.au
- [3] Finding High-Quality Content in Social Media Eugene Agichtein (2008)
- [4] How valuable is medical social media data? Content analysis of the medical web
- [5] Kerstin Denecke ,Wolfgang Nejdl ,L3S Research Center, University of Hannover, Germany
- [6] An internet traffic analysis method with mapReduce Youngseok Lee ; Chungnam Nat. Univ., Daejeon, South Korea ; Wonchul Kang ; Hyeonung Son
- [7] HiBench benchmark suite characteristic of the mapreduce based analysis, Huang et al(2020) Hadoop Benchmark Suite, Shengsheng Huang, Jie Huang, Yan Liu, Lan Yi and Jinquan Dai Intel Asia-Pacific Research and Development Ltd., Shanghai, P.R.China, 200241 {shengsheng.huang, jie.huang, yan.b.liu, lan.yi, jason.dai}@intel.com
- [8] Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads Yanpei Chen, Sara Alspaugh, Randy Katz University of California, Berkeley fychen2, alspaugh, randyg@eecs.berkeley.edu (2012)
- [9] Analysis and Design of Internet Monitoring System on Public Opinion Based on Cloud Computing and NLP Hui Wen, Peiguang Lin, Yaobin Hu (2012)
- [10] I D C I V I E W Extracting Value f rom Chaos June 2011 By John Gantz and David Reinsel
- [11] Sponsored by EMC Corporation Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing Alexandru Iosup, Member, IEEE, Simon Ostermann,Nezih Yigitbasi, Member, IEEE,
- [12] Radu Prodan, Member, IEEE, Thomas Fahringer, Member, IEEE, and Dick Epema, Member, IEEE (2010)
- [13] Computational solutions to large-scale data management and analysis. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Source Pacific Biosciences, Menlo Park, California 94025, USA. eschadt@pacificbiosciences.com
- [14] Big Data and Cloud Computing: Current State and Future Opportunities*Divyakant Agrawal Sudipto Das Amr El Abbadi Department of Computer Science University of California, Santa Barbara Santa Barbara, CA 93106-5110, USA {agrawal, sudipto, amr}@cs.ucsb.edu
- [15] <http://online.liebertpub.com/action/showPopup?citid=citart1&id=T1a&doi=10.1089%2Fbig.2013.0011>