Volume 3, No. 8, August 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

A NOVEL APPROACH FOR CLOUD-BASED COMPUTING USING REPLICATE DATA DETECTION

[*1] Mr. Pritaj Yadav, and [2] Mrs. Alka Gulati
[*1] Scholar, M-Tech (SE), LNCT, Bhopal
pritaj_man@rediffmail.com
[2] Associate Prof. (CSE), LNCT, Bhopal
gulatialka@rediff.com

Abstract- Cloud-based computing is an emerging practice that offers significantly more infrastructure and financial flexibility than traditional computing models. When considering cloud-based infrastructure offerings, security is a common concern. Larger enterprises may have implemented very strong security approaches that may or may not be equaled by cloud providers, but don't just assume that security is a problem. Look for the type of security functionality you would look for in an in-house solution. A documents may get mirrored to avoid delays or to provide fault tolerance. Algorithms for detecting replicate documents are critical in applications where data is obtained from multiple sources. The removal of replicate documents is necessary, not only to reduce runtime, but also to improve search accuracy. Today, search engine crawlers are retrieving billions of unique URL's, of which hundreds of millions are replicates of some form. Thus, In this paper we propose quickly identifying replicate detection to speed up indexing and searching. By efficiently presenting only unique documents, user satisfaction is likely to increase.

Keywords- unique documents, detecting replicate, replication, search engine.

INTRODUCTION

Cloud-based computing is an emerging practice that offers significantly more infrastructure and financial flexibility than traditional computing models. At the heart of cloud-based computing is utility "services" backed by a loosely coupled infrastructure that is self-healing, geographically dispersed, designed for user self-service and instantaneously scalable in response to the ebb and flow of business demands. These services are easily accessible across IP-based networks, making it very easy to take advantage of them and all infrastructure management issues are off-loaded to the cloud provider. Cloud providers today offer everything from access to raw compute or storage capacity resources to full-blown application services in areas such as payroll and customer relationship management.

Cloud computing is an emerging concept. It has many names, including: grid computing, utility computing and on-demand computing. Indeed, one of the hindrances to the development and adoption of cloud computing is the lack of understanding of what it is and isn't among both private and public sector leaders.

The term "cloud computing" has at its core a single element, computing services are delivered over the Internet, on demand, from a remote location, rather than residing on one's own desktop, laptop, mobile device or even on an organization's servers. For an organization, this would mean that, for a set or variable, usage-based fee or even possibly for free it would contract with a provider to deliver applications, computing power, and storage via the web. In a nutshell, the basic idea of cloud computing is that computing will become location and device independent meaning that it increasingly will not matter where information is housed nor where computation/processing is taking place. This enables computing tasks and information to be available anytime, anywhere from any device so long as there is access to the Internet. The cloud concept also means that, for individuals and organizations alike, computing will increasingly be viewed as an infinite, not a finite resource. This is because computing is taking on an on-demand, scalable form, as additional network bandwidth, storage, and computation capacity can be added as needed, much as people simply use and pay for more (or less) electricity as their energy needs change. For this reason, many even in the industry refer to this as the utility model of computing.

Cloud computing offers a number of benefits, including the potential for:
Rapid scalability and deployment capabilities.
Providing just-in-time computing power and infrastructure.
Decreased maintenance/upgrades.
Improved resource utilization elasticity, flexibility, efficiencies.
Improved economies of scale.
Improved collaboration capabilities.

Ability to engage in usage-based pricing, making computing a variable expense rather than a fixed capital cost with high overhead reduced information technology (IT) infrastructure needs both up-front and support costs.
Capacity for on-demand infrastructure and computational power.
Green-friendly reduced environmental footprint.
Improved disaster recovery capability.

In large data warehouses, data replication is an inevitable phenomenon as millions of data are gathered at very short intervals Data warehouse involves a process called ETL which stands for extract, transform and load. During the extraction phase, multitudes of data come to the data warehouse from several sources and the system behind the

warehouse consolidates the data so each separate system format will be read consistently by the data consumers of the warehouse. Data portals are everywhere. The tremendous growth of the Internet has spurred the existence of data portals for nearly every topic. Some of these portals are of general interest; some are highly domain specific. Independent of the focus, the vast majority of the portals obtain data, loosely called documents, from multiple sources [12]. Obtaining data from multiple input sources typically results in replication. The detection of replicate documents within a collection has recently become an area of great interest [11] and is the focus of our described effort.

Simply put, not only is a given user's performance compromised by the existence of replicates, but also the overall retrieval accuracy of the engine is put at risk. The definition of what constitutes a replicate is unclear. For instance, a replicate can be defined as the exact syntactic terms, without formatting differences. The general notion is that if a document contains roughly the same semantic content it is a replicate whether or not it is a precise syntactic match. When searching web documents, one might think that, at least, matching URL's would identify exact matches. However, many web sites use dynamic presentation wherein the content changes depending on the region or other variables.

Replication is seen as unethical when the primary intent is to deceive peers, supervisors and/or journal editors with false claims of novel data. Given the large number of papers published annually, the large diversity of journals with overlapping interests in which to publish and the uneven access to journal publication content, it is not unreasonable to assume that the discovery of such replication is rare [13]. The recent development of algorithmic methods to systematically process published literature and identify instances of replicated/plagiarized text as accurately as possible should serve as an effective deterrent to authors considering this dubious path. Unfortunately, the methods in place now have a very limited reach, and are confined to abstracts and titles only.

Replicates: where they come from? One of the main problems with the existing geospatial databases is that they are known to contain many replicate points ([7], [10], [16]). The main reason why geospatial databases contain replicates is that the databases are rarely formed completely from scratch., and instead are built by combining measurements from numerous sources. Since some measurements are represented in the data from several of the sources, we get replicate records.

Why replicates are a problem? Replicate values can corrupt the results of statistical data processing and analysis. For example, when instead of a single (actual) measurement result, we see several measurement results confirming each other, and we may get an erroneous impression that this measurement result is more reliable than it actually is. Detecting and eliminating replicates is therefore an important part of assuring and improving the quality of geospatial data, as recommended by the US Federal Standard [9].

The identification of exact replicate documents in the Reuters collection was the primary goal of Sanderson [13]. The method utilized correctly identified 320 pairs and only failing to find four, thus proving its effectiveness. In the creation of this detection method, they found a number of other replicate document types such as expanded documents, corrected documents, and template documents.

LITERATURE SURVEY

The efficient computation of the overlap between all pairs of web documents was considered by Shivakumar et al. [8]. The improvement of web crawlers, web archives the presentation of search results, among others can be aided by this information. The statistics on how common replication is on the web was reported. In addition, the statistics on the cost of computing the above information for a relatively large subset of the web about 24 million web pages which correspond to about 150 gigabytes of textual information was presented.

Many organizations archiving the World Wide Web show more importance in topics dealing with documents that remain unchanged between harvesting rounds. Some of the key problems in dealing with this have been discussed by Sigurðsson [5].Subsequently, a simple, but effective way of managing at least a part of it has been summarized which the popular web crawler Heritrix [6] employed in the form of an add-on module. They discussed the limitations and some of the work necessitating improvement in handling replicates, in conclusion.

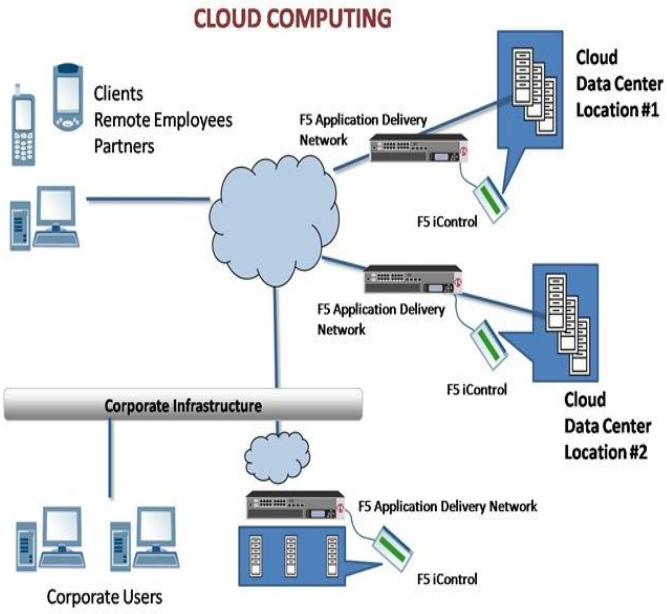

Figure 1 Cloud Computing Architecture

Theobald et al. [4] proved that SpotSigs provide both increased robustness of signatures as well as highly efficient replication compared to various state-of-the-art approaches. It was demonstrated that simple vector-length comparisons may already yield a very good partitioning condition to circumvent the otherwise quadratic runtime behavior for this family of clustering algorithms, for a reasonable range of similarity thresholds. Additionally, the SpotSigs replication algorithm runs "right out of the box" without the need for further tuning, while remaining exact and efficient, which is dissimilar to other approaches based on hashing. Provided that there is an effective means of bounding the similarity of two documents by a single property such as document or signature length, the

SpotSigs matcher can easily be generalized toward more generic similarity search in metric spaces.

Recently, the detection of replicate and near replicate web documents has gained popularity in web mining research community. This survey extents and merges a wide range of works related to detection of replicate and near replicate documents and web documents. The detection techniques for identification of replicate and near replicate documents, detection algorithms, Web based tools and other researchers of replicate and near replicate documents are reviewed in the corresponding subsections.

To improve system availability, replicating the popular data to multiple suitable locations is an advisable choice, as users can access the data from a nearby site was considered by Sun DW et al. [1]. A dynamic data replication strategy is put forward with a brief survey of replication strategy suitable for distributed computing environments. It includes:

a. Analyzing and modeling the relationship between system availability and the number of replicas.
b. Evaluating and identifying the popular data and triggering a replication operation when the popularity data passes a dynamic threshold.
c. Calculating a suitable number of copies to meet a reasonable system byte effective rate requirement and placing replicas among data nodes in a balanced way.
d. Designing the dynamic data replication algorithm in a cloud. Experimental results demonstrate the efficiency and effectiveness of the improved system brought by the proposed strategy in a cloud.

PROPOSED TECHNIQUE

The cloud enables a new set of solutions to solve perennial storage problems much more cost effectively. Data protection stands to benefit significantly from cloud-based computing options, in particular because they provide the foundation for easily accessible, affordable disaster recovery solutions. This easy access facilitates rapid implementation of off-site protection for new projects at larger enterprises, and can enable disaster recovery solutions that small and medium enterprises (SMEs) could not afford in the past. Given the increasing criticality of data, all enterprises should have a disaster recovery plan in place for at least key applications. But many do not, primarily due to cost and complexity issues. Cloud based infrastructure provides an interesting disaster recovery alternative that addresses both of these issues.

Basically Replication technology are three type:
Storage Arrays Based,
Network Based Appliances
Host Based

a) Array Based Replication:

Replication requires similar arrays at both the source and target locations, making it a poor choice in replicating data to cloud providers that likely won't have the same array you do in their cloud infrastructure.

b) Network Based Appliances:

Replication require an appliance at both the source and target locations as well, and while they are much more cost-effective to implement than array based approaches, they basically suffer from the same infrastructure issue that array based replication does: the cloud provider is unlikely to have or make available to you the same type of network appliance deployed at your site.

c) Host-Based Replication:

Replication, which basically just runs on an industry standard server, is an excellent fit, cloud providers allow you to request Windows, Linux and in some cases even other Unix servers, when you rent compute cycles from them, allowing you to replicate from servers of these same type at your location to theirs very cost-effectively.

Host Based Replication comes in two flavors. Vendors such as CA (the XO-soft product line) and Steel-Eye use block based replication approaches, while vendors such as Double Take and Never Fail use file based approaches. Both approaches can be used to replicate entire virtual machines in real time, but block based approaches offer a more comprehensive solution (due to the ability to replicate all data, not just files) when replicating physical machines to cloud-based infrastructure. Solutions that support multiple operating systems, as opposed to just Windows, can also offer more comprehensive solutions with a common management paradigm across platforms. Host based replication solutions can be configured for just a few thousand dollars, and when combined with cloud based infrastructure offer a very low cost disaster recovery solution that allows protection to be extended lower in the organization for larger enterprises and makes disaster recovery an affordable option for smaller enterprises.

This standard specifies four secure hash algorithms, SHA-1 [15], SHA-256, SHA-384, and SHA-512. All four of the algorithms are iterative, one-way hash functions that can process a message to produce a condensed representation called a *message digest*. These algorithms enable the determination of a message's integrity: any change to the message will, with a very high probability, result in a different message digest. This property is useful in the generation and verification of digital signatures and message authentication codes, and in the generation of random numbers (bits).

Each algorithm can be described in two stages: preprocessing and hash computation. Preprocessing involves padding a message, parsing the padded message into *m*-bit blocks, and setting initialization values to be used in the hash computation. The hash computation generates a *message schedule* from the padded message and uses that schedule, along with functions, constants, and word operations to iteratively generate a series of hash values. The final hash value generated by the hash computation is used to determine the message digest.

The four algorithms differ most significantly in the number of bits of security that are provided for the data being hashed – this is directly related to the message digest length. When a secure hash algorithm is used in conjunction with another algorithm, there may be requirements specified elsewhere that require the use of a secure hash algorithm with a certain number of bits of security.

Additionally, the four algorithms differ in terms of the size of the blocks and words of data that are used during hashing. Table1 presents the basic properties of secure hash algorithms.

Table: 1 Basic Properties of all four Secure hish Alogrithm

Algorithm	**Message Size** (bits)	**Block Size** (bits)	**Word Size** (bits)	**Message Digest Size** (bits)	**Security2** (bits)
SHA-1	<264	512	32	160	80
SHA-256	<264	512	32	256	128
SHA-384	<2128	1024	64	384	192
SHA-512	<2128	1024	64	512	256

PROPOSED WORK

In our proposed method ,we will rapidly compares large numbers of files for identical content by computing the hash of each file. therefore quickly identifying replicate detection to speed up indexing and searching.

Flowchart:

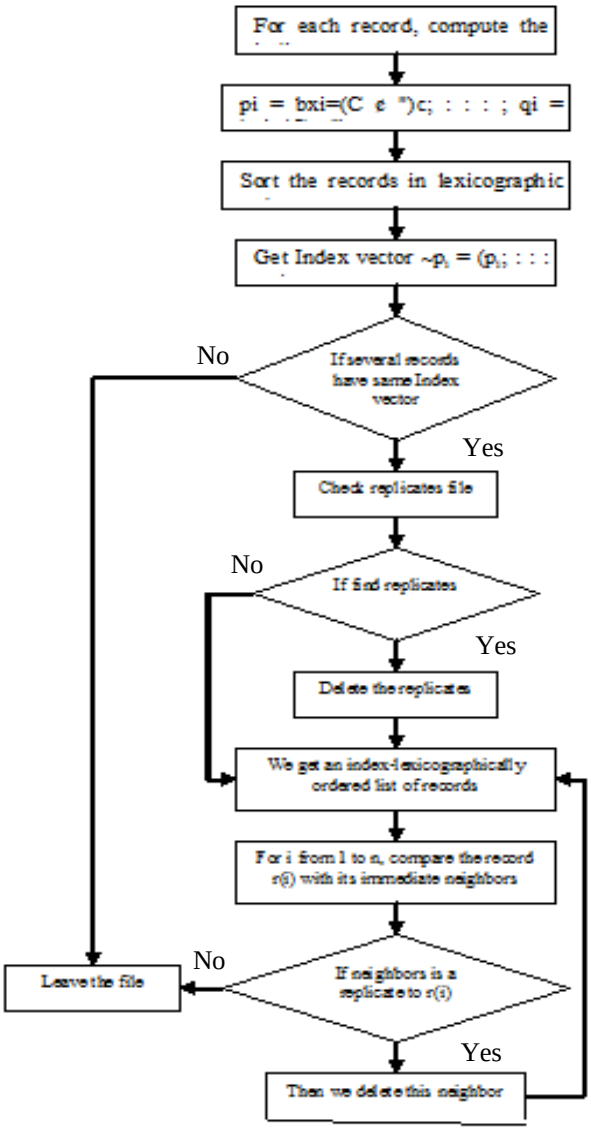

Figure: 2

EXPERIMENT AND ANALYSIS

We proposed a new Replicate Data Detection Algorithm called RDDA and its performance evaluated using multiple data collections. If you're considering what combining replication and cloud based computing can offer you, regardless of whether you're an end user or a cloud provider, look for the following features synchronous and asynchronous replication options so that the technology can be used to address both short distance and long distance requirements; understand also whether it provides real-time, scheduled, or both forms of replication. Good integration points technologies to facilitate data protection operations and server virtualization technology to lower the cost of DR operations.

A conscious approach to maintaining the write ordering established by the production application in order to maintain data integrity, it is critical that data is written to the target disk in exactly the same order that it is written to the primary disk (this is more of a concern when asynchronous replication is used).

Fault management that will automatically re-synchronize source and target devices once live network connections are re-established, look at exactly how this is done to ensure that devices can be re-synchronized with minimal bandwidth and very quickly. Integrated technologies that minimize network bandwidth requirements during normal and re-synchronization operations. Support for encrypting data both in-flight and at rest to at least a level of SHA-1 equivalence (with SHA-2 equivalence being preferred).

CONCLUSION

A new Replicate Data Detection Algorithm called RDDA are evaluated its performance using multiple data collections. The document collections used varied in size, degree of expected document replication, and document lengths. As the bar becomes ever higher for building resiliency into computing infrastructures, replication technologies will become part of the storage foundation. Cloud providers are in a good position to leverage this technology to meet existing as well as evolving customer requirements. In the near term, replication not only enables data recovery in the cloud, but server recovery in the cloud as well. Now that affordable, host-based replication approaches that can securely handle sizable data volumes through IP-based networks are available, don't overlook what the combination of replication and cloud computing have to offer regardless of whether you're an end user or a cloud provider.

Therefore, any match in even a single results in a potential replicate match indication. This results in the scattering of potential replicates across many groupings, and many false positive potential matches. This paper intends to aid upcoming researchers in the field of Replicate document detection using Cloud-based computing in web crawling to understand the available methods and help to improve perform their research in further direction.

REFERENCE

[1] Sun DW, Chang GR, Gao S *et al.*, "Modeling a dynamic data replication strategy to increase system availability in cloud computing environments". Journal of computer science and technology, 27(2), pp. 256-272, Mar. 2012.

[2] Julia Myint and Thinn Thu Naing, "Mangement of Data Replication for PC Cluster Based Cloud Storage System". International Journal on Cloud Computin Services and Architecture(IJCCSA),Vol.1, No.3, November 2011, pp. 31-41.

[3] Nathaniel Borenstein and James Blake, "Cloud Computing Standards", IEEE 2011, pp 74-78.

[4] Theobald, M., Siddharth, J., Paepcke, A., "SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections", Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, 2008, pp. 563-570.

[5] Sigurðsson, K., "Managing duplicates across sequential crawls", proceedings of the 6thInternational Web Archiving Workshop, 2006.

[6] Mohr, G., Stack, M., Ranitovic, I., Avery, D., and Kimpton, M., "An Introduction to Heritrix", 4th International Web Archiving Workshop, 2004.

[7] Scott, L., "Identification of GIS Attribute Error Using Exploratory Data Analysis", Professional Geographer 46(3), 1994, pp. 378.386.

[8] Shivakumar, N., Garcia Molina, H., "Finding near-replicas of documents on the web",Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 1590, 1999, pp. 204-212.

[9] FGDC Federal Geographic Data Committee, FGDC-STD-001-1998. "Content standard for digital geospatial metadata", Federal Geographic Data Committee, Washington,D.C., june 1998, http://www.fgdc.gov/metadata/contstan.html.

[10] Mccain, M., and William C., "Integrating Quality Assurance into the GIS Project Life Cycle", Proceedings of the 1998 ESRI Users Conference 1998. Http://www.dogcreek.com/html/documents.html

[11] Shivakumar, n. And garica-molina, h. "Finding near-replicas of documents on the web". In Proceedings of Workshop on Web Databases (WebDB'98), Valencia, Spain, March, 1998, pp. 204–212.

[12] Broder, a., glassman, s., manasse, s., and zweig, g. "Syntactic clustering of the web". In Proceedings of the Sixth International World Wide Web Conference (WWW6'97), Santa Clara, CA., April, 1997, pp. 391–404.

[13] Sanderson,m. Replicate detection in the reuters collection. Technical Report (TR-1997-5) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997.

[14] Heintze, n., Scalable document fingerprinting. In proceedings of the second usenix electronic Commerce Workshop Oakland, CA., Nov. 1996, pp.191–200.

[15] The SHA-1 algorithm specified in this document is identical to the SHA-1 algorithm specified in FIPS 180-1.

[16] Goodchild, M., and Gopal, S. (Eds.), Accuracy of Spatial Databases, Taylor & Francis, London., 1989.

