Using Hashing to Maintain Data Integrity in Cloud Computing Systems

Anthony W. Scabby and Anil L. Pereira
Department of Accounting, Computer Science and Entrepreneurship
Southwestern Oklahoma State University
Weatherford, OK 73096

Abstract
Cloud Computing technology is a recent development, and allows users to share data, software and other resources across multiple computers on the Internet. This is cost effective and can save on time, space, and licensing fees. In Cloud Computing, it is difficult to maintain data integrity because the user usually has no control over the security mechanisms that are used to protect his/her data. Data integrity could be compromised intentionally (by malware) or unintentionally (by a program bug). Such an event could be detected by a hash system. A hash is a unique string of text that is generated by applying a mathematical function to the data that is being checked for integrity. The result (called a hash value) is a string of numbers and letters which is a computational representation of the data. Data that has not been changed in any way will have the same hash value every time the same mathematical function is used. If a computer process is going to analyze data in a Cloud Computing system, then a hash value of the data could be obtained before the process starts. After the process completes, another hash value can be obtained and compared to the previous hash value. If the two values are not equal then it will show that the data has been changed in some way during the analysis, and the results of the analysis are inaccurate. At Southwestern Oklahoma State University, we are developing a model to test the effectiveness of the Hash process in Cloud Computing.

1. Introduction

Cloud computing is a fairly recent technological endeavor. As such, it is not completely understood and its potential is yet to be fully tested. In 2007, IBM joined with Google to head an initiative to do further research in the area of Cloud Computing. This was to improve understanding of what Cloud Computing was and what it could do [3]. Cloud Computing utilizes the internet to manage system resources, operating systems, and data storage. To access a Cloud, a user logs in through the Internet and gains access to the services that a Cloud provides [1]. A Cloud Computing system allows for scalability. The amount of data space, types of programs, and software tools and services can be adjusted to a required level [6].

Cloud Computing could also save money on hardware, software and data maintenance. A case study in India has shown that local governments are able to offer the services of a High Performance Computing Environment for a fraction of the cost if they use Cloud Computing services. These services can include software management systems for complaint resolution, roads and infrastructure, census, and elections. Public services allow Cloud users to be able to report traffic problems and even use an online mapping tool to specify locations where the traffic problems occur. Authorized medical professionals are able to access the Census Department to record births and deaths. Elections can be managed and scheduled online and voter lists can be maintained and easily accessed by those who are authorized to see them. Because everything is managed online on the Cloud, the local governments save money on hardware costs necessary to host the massive amount data that would be required and also on the cost of maintaining a cooling system for servers. They also save money on network managers and others with the technological expertise necessary needed to run a project with the services similar to those provided on the Cloud. A High Performance Environment could easily cost over $60,000 with hardware, software, and maintenance fees, but the same services provided on the Cloud could cost less than $10,000 in service fees [8].

Cloud Computing also has disadvantages. Shifting to a Cloud based system means giving up a certain amount of control over storage and security. This can lead to "uncertainty and fear" regarding Cloud Computing [4].

2. Data Storage and Processing in the Cloud

Maintaining Data integrity is one of the areas of concern when talking about Cloud Computing. Since you do not have direct control over your data (as it is stored off site), security is an issue. Despite the security issues, data can still be monitored and maintained in a Cloud. As shown in Figure 1, data is stored in the Cloud in several different ways, but the most common method of storage is a container called a "Blob" (or Binary Large Object) [2]. Blobs can store different types of data and can potentially be up to 1 TB in size. Blobs also allow metadata to be attached to separate files within the Blob [1].

Some Cloud Computing platforms, such as Microsoft Azure, make multiple copies of data because there is always danger of a single data copy being destroyed. If multiple data copies exist then data can still be retrieved even if several copies get destroyed [2]. Individuals and organizations pay a fee to access the Cloud's services. The fee they pay depends on the services they use and the amount of time the services are used. An organization can assign roles (such as employee, executive, guest, etc.) within the Cloud that will authorize access to certain applications or computer programs. These organizations can give their Clouds public or private access, and it is possible for two separate Clouds to interact with each other. Even with a secure storage system in place on the Cloud, it is possible that your data could become corrupted through bugs in the system, hackers, or user error. There is a computing process you can use to check if the data you have stored has been altered. This process is called "Hashing."

3. Hashing Data

A Hash Algorithm uses a mathematical function that can be applied to data. When the function is applied to data, an alphanumerical value is produced. There are many different hash functions, SHA-1, MD5, CRC etc., that each return a unique value. A set of data will produce the same hash value in most cases, as long as the same mathematical function is used and nothing in the data has been changed. The value that is produced can be thought of as a "digital fingerprint" [7]. Hashing is a good checkpoint, but it is not perfect. Though the possibility is extremely small, it is possible for two different sets of data to come up with an identical hash value. When two sets of data produce the same hash value, it is called a "Collision" [5]. Sha-1, for example, has a collision rate of 2^69, (or 5.9 x 10^20). Hash Algorithms have different Collision rates. Using multiple hashing algorithms also reduces the chance of having a collision.

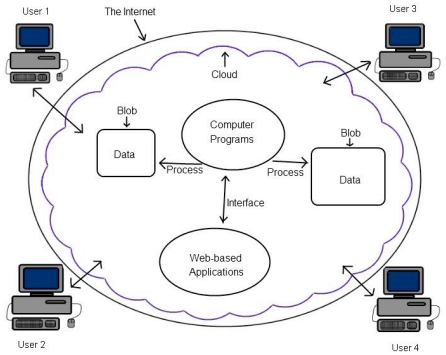

Figure 1: Computers access the Cloud's services through the Internet

4. Using Hashing in a Cloud based system

As shown in figure 2, consider the following hypothetical situation that explains how hashing data can check data integrity. In the Cloud, a person has a set of data that is needed to perform a calculation. Before the calculation is performed, a hash value is obtained from the data set. This value is recorded for future reference. The calculation is performed and then another hash value is obtained from the same data set. If the data has not been altered during the calculation then the two values should be equal. If the two values are not equal then it can be determined that something has happened to the data during the calculation and appropriate steps can be taken to address the problem.

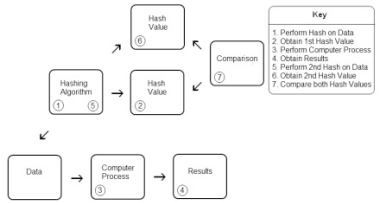

Figure 2: Hashing Data

5. Conclusion

Cloud Computing is a powerful tool, but it isn't fully understood. This misunderstanding causes hesitation and uncertainty. Using processes, such as Hashing, within the Cloud alleviates that uncertainty a little bit. Even though there is a small chance Hashing could have a Collision, it still furthers the understanding and potential of Cloud Computing. The only way to expand our understanding of Cloud Computing is to continue researching its possibilities. For future work, the performance of various hashing algorithms will be analyzed and compared in terms of their speed and effectiveness. Acceleration of the algorithms through parallelization on computing clusters and graphics processing units will be attempted.

References

[1] D. Chappell, "Introducing Microsoft Azure," Oct. 2010, http://go.microsoft.com/?linkid=9682907.
[2] D. Chappell, "Introducing the Windows Azure Platform," Oct. 2010, http://go.microsoft.com/?linkid=9682631.
[3] S. Lohr, "Google and I.B.M. Join in 'Cloud Computer' Research," *New York Times*, 8 Oct. 2007, http://www.nytimes.com/2007/10/08/technology/08cloud.html?_r=2&ex=1349496000&en=92627f0f65ea0d75&ei=5090&partner=rssuserland&emc=rss&oref=slogin.
[4] J. Martin, "Cloud computing security risks are sometimes considered greater than cloud's rewards. The industry is working to change that, and so can you," Cisco News, 7 June 2010, http://newsroom.cisco.com/dlls/2010/ts_060710b.html.
[5] T. Moses, "Exploiting weaknesses in the MD5 hash algorithm to subvert security on the web," Entrust, Jan 2009, http://www.businesssignatures.com/resources/download.cfm/23639/WP_MD5_Jan09.pdf.
[6] Microsoft, "Competitive Comparisons between Microsoft and VMware Cloud Computing Solutions," *How far will you take virtual?,* Apr 2010, http://download.microsoft.com/download/6/3/F/63F162FA-1464-4C58-ACE3-5B79B5158E7F/MSDesktopVirtCompareWhitepaperApr2010.pdf.
[7] J. Walker, M. Kounavis, S. Gueron, and G. Graunke, "Recent Contributions to Cryptographic Hash Funtions," 2009, Intel Corporation, http://www.eetimes.com/electrical-engineers/education-training/tech-papers/4201212/Recent-Contributions-to-Cryptographic-Hash-Functions.
[8] Windows Azure, "Software Services Provider Delivers Cost-Effective E-Government Solution," Nov. 2009, http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=4000005770.