ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 11, No. 1, 2016, *pp*.070-080

Modeling Cloud Storage: A Proposed Solution to Optimize Planning for and Managing Storage as a Service

Anita L. Timmons, Pavel Fomin, James Wasek
Department of Engineering Management & Systems Engineering School of Engineering and Applied Science, The George Washington University, Washington, DC
(*Received August 30, 2015, accepted October 19, 2015)*

Abstract. Cloud-computing service providers are currently viewed as the best solution to the global need for massive data systems because of their superior flexibility, scalability, and cost benefits. Cloud computing that is enabled by virtualized services is still constrained, however, by the capacities of the underlying physical systems that are combined into sharable pools of resources. The next challenge for computation systems will arise when even the cloud is not sufficient. What comes after cloud migration and adoption? In this paper, we examine how service providers can manage cloud storage resources and costs when the amount of collected data to be retained grows exponentially, to the point that it strains even virtualized resource capacities. We assess the analytical frameworks being developed to identify which storage architectures can best accommodate the specific needs of large data storage consumers. We also investigate the areas in which these fail to fully address the problem, and propose solutions. We argue that a cloud storage framework that addresses data volume, data growth trends over time, and requirements for storage management will enable service providers to manage cloud storage resources and costs in such a manner that the cloud will continue to offer the greatest benefits for the storage of massive data systems.

Keywords: Cloud storage, big data, cloud storage architecture, surface response methodology

1. Introduction

Many organizations are faced with the challenge of how to operate and manage information technology (IT) in an age of exponential data growth. As different types of data are generated, processed, and stored, data increase in terms of size (terabytes to petabytes) and velocity (the rate at which new content is added [1]. The accumulation of such large volumes of data over time is referred to as “big data,” and represents a potentially valuable source of information for organizations [2]. The use of big data is changing the ways that organizations operate and do business. Companies use big data to optimize process flow, make predictions based on statistical analysis, and a myriad of other business, scientific, and engineering analyses [2, 3]. Over time, managing data can present many challenges when developing a data infrastructure to accommodate big data [2] . Several issues, such as scaling to meet increased user requests for existing and new data, performance demands and continuous availability as the amount of data grows, workload diversity as methods for combining and analyzing data increase, and data security as data and analysis products become desirable targets for hackers, must be addressed as part of big data analysis [4].

Cloud computing, which is a solution to the issues that arise from big data [4], has emerged to provide data-intensive computing environments with a cost-effective way to manage elastic (both growth and shrinkage) requirements for data and data processing [5]. Cloud architectures specify virtualized IT services and resources that are provided over a network by a cloud service provider [6]. Cloud adoption is driven by the appeal of lower participation costs, easier scaling, and effective provisioning [7]. Cloud consumers can also take advantage of the economic impacts of a scaled-down infrastructure cost (e.g., labor, facilities, and utilities) while relying on cloud providers to provide better performance and the guaranteed service quality that satisfies requirements for security, availability, and reliability.

With the exponential growth of data collection that has evolved over time to become big data and the growing dependence on IT in almost every field of human endeavor, cloud computing storage systems are widely considered to be the most viable solution for managing and preserving data [1]. Cloud consumers require storage solutions that are sufficiently heterogeneous, scalable, and flexible to allow for storage reuse and data sharing [8]. Systems architects and engineers grapple with ways to design storage architectures that deliver the service-quality metrics essential to meet these cloud consumer requirements [9].

Because no two cloud service providers are the same, cloud consumer systems architects and engineers

require a method to model, test, and evaluate their storage needs to determine the cloud provider that will best fit their business, research, or mission needs. This has created the need for a system that can (1) predict big data patterns and trends that will affect how big data is stored and (2) identify limitations and necessary improvements in current cloud storage infrastructure. In this paper, we examine how service providers can manage cloud storage resources and costs as the amount of collected data to be retained continues to grow exponentially.

Our study's contribution is to propose a solution to data growth. The optimal cloud storage framework (CSF) will address the volume of data growth based on performance requirements for storage management. Additionally, the framework should measure and report the behavior of the cloud environment, based on the size of the data set. In this paper, we construct and run cloud simulation models to collect data to answer the following research questions:

- Based on the collected performance data, can an optimal minimal point be identified when using the CSF?
- What improvements can be made to the framework that will assist engineers in increasing cloud storage performance?

Using cloud simulation tools in the CSF to generate the performance data needed for statistical analysis can benefit service providers by allowing them to identify methods for managing cloud storage resources, costs, and big data. This framework could also assist consumers and consumer system architects and engineers to determine which cloud storage provider's services best fit their needs. The CFS is based on general systems engineering principles that will assist service providers and cloud consumers who face challenges with the migration and adoption of large data sets to the cloud environment.

1.1. Related work

Cloud storage environments are not new. Many researchers and engineers have addressed the problems and benefits of cloud storage environments, and numerous articles have discussed the challenges of data migration, quality of service, security, and unexpected hardware and software failures. Deciding whether to move to the cloud requires that businesses and organizations clearly assess their current infrastructure needs. Many articles have provided comprehensive reviews of cloud-based infrastructures and suggestions about how to manage different data types in the storage environment [5]. Kolodner et al. [5] propose a storage architecture that raises the abstraction level of storage using middleware to modify and manage data as a service. The authors discuss current offerings for cloud storage and present conceptual architectures for two scenarios. However, no quality of service or cost data are evaluated to determine how the storage solution performs for the cases presented. Another practical storage strategy is to examine the trade-offs between computation and storage usage. Yuan et al. [10] introduced a cost model that uses data sets for storage-producing algorithms to compute costs based on storage and computation. A cloud simulation tool called SwinCloud was used to test data sets with random sizes ranging from 100 GB to 1 TB. This approach—of using different data types of random sizes—takes into account the importance of size when analyzing storage functionality and performance. Unfortunately, the cost data that were used for the model and simulation came from only one service provider. Identifying trends, patterns, and limitations requires larger data sets and the comparison of several service providers' costs.

Quality of service (QoS) plays an important role in consumers' and engineers' decisions about which cloud service provider to choose [11]. Zhang et al. [11] propose an infrastructure for a large-scale cloud cluster and focus on storage components, but the size of the data set used for the experiment was relatively small (1 GB). This is a differentiated approach to providing services based on the data path and the priority of the class of user based on allocations made by a management and scheduling panel. Differentiated cloud storage resources assist with latency and throughput, but other aspects of storage resources, such as Central Processing Unit (CPU) utilization and Random Access Memory (RAM), are not addressed. Throughput is important for the user to access the data faster; however, to truly study data behavior, all resources that impact data should be evaluated and analyzed.

Other studies have provided insight into the evolution of the data center environment. Kant [12] puts into perspective why cloud consumers are adopting cloud-based storage solutions, namely, the costly challenges and issues involved in maintaining data centers [12]. With the growing use of data centers to store massive amounts of data, rapid data growth can pose significant challenges for data centers and their customers [13]. In addition, with the increase in application build-ups and Virtual Machine (VM) scale-outs, bandwidth (throughput) rates can potentially cause degradation of transfer rates, which in turn causes storage