ABSTRACT

Name of the Scholar:	**Kashish Ara Shakil**
Name of the Supervisor:	**Dr. Mansaf Alam**
Department:	**Department of Computer Science**
Title:	**An Effective Framework for Data Management in a Cloud Based System**

Cloud computing has emerged as an efficient paradigm for offering computational capabilities to its end users. The prominence and the various applications of cloud computing are prodigious and thus, it is a topic of huge significance. It offers numerous astounding features like multi-tenancy, on demand service, and pay per usage. One of the major issues in cloud computing is the management of growing volumes of data and delivery of computing services to its end users to satisfy their QoS requirements.

The data in cloud consists of high volume, variety, velocity, and veracity. Traditional systems fail to address issues raised by such data requirements. Moreover, these systems are not scalable and a large amount of cost is involved for handling such volumes of data, which can lead to a bottleneck. A system is therefore required to handle data with such characteristics. Since cloud works on the utility model of computing i.e. on a pay-per-usage model, therefore, these systems should also utilize resources efficiently leading to cost effective solutions.

This thesis presents a novel framework for data management in a cloud environment. It addresses the core data management issues arising from large data sets in the cloud. The issues addressed takes into account challenges faced due to multi-tenant architecture, resource abundance, resource heterogeneity and utility model of cloud.

We first present an overview and survey of data management techniques in cloud that have been adopted in literature. Based on this we developed a taxonomy of data management techniques in the cloud and discuss the different use cases of cloud computing along with various scheduling techniques, characteristics of data in cloud and different database models. Limitations of existing data management approaches have also been identified.

An architecture for data management in cloud is proposed to overcome limitations of the existing systems. This architecture comprises of three levels: cloud service provider level,

data center level, and client level. It identifies key elements such as scheduling, data access control, security and monitoring in the cloud. It has several features such as support for modeling cloud computing infrastructure, a user-friendly interface, and flexibility to switch between the different types of users. The proposed architecture is validated by its application in the education sector.

We demonstrate how data can be efficiently managed at the cloud service provider level. This is achieved by managing data in such a manner that resource usage is optimized. Therefore, a system is proposed, to effectively utilize resources. This system is developed based on the proposed architecture. The key elements of the proposed system include service brokers, monitor, and resource allocator (RA). An algorithm for efficient utilization of resources in the cloud is proposed, which ensures optimum allocation of resources. Its evaluation and validation have been performed by comparing it with existing algorithms. The experiments have been performed via simulation on Google cluster trace using CloudSim. Since Google cluster trace is taken as the representative of cloud environment; therefore a detailed study on it has also been performed. In order to demonstrate the generic applicability of the system, we extend the system to support other applications such as scientific workflows.

Data management at the RA element of the proposed system is performed through a proposed data management workflow, which presents how data can be efficiently stored, processed and retrieved in a scalable, fast and cost-effective manner. The system is validated by performing experiments on biometric signature samples. It uses parallel algorithms for training and is able to handle storage and processing of data sets of large volume. After the parallelized pre-processing data was reduced to 25 percent (approximately) of the original size thereby reducing the storage space. The reduction in storage space leads to faster processing and cost optimizations. The experimental results show that the proposed distributed data processing algorithm was able to achieve a speedup of 8.5x over the other existing approaches, without compromising on the efficiency (low EER). We perform a cost and benefit analysis of the proposed method and show that the system is also cost effective in comparison to other state of art methods.

Thus, the proposed framework is effective for handling data in cloud environment. In this framework resources are utilized efficiently to manage huge volume of data, this is achieved by the proposed system for optimal resource allocation. Furthermore, data is stored, processed and retrieved efficiently by these resources, which is demonstrated through results of the proposed data management workflow.