Research on Semi-Structured and Unstructured Data Storage and Management Model for Multi-Tenant

Xin Hu, Beijing Information Science & Technology University, Beijing, China
Yabin Xu, Beijing Information Science & Technology University, Beijing, China

ABSTRACT

In order to solve the problem of data isolation and storage caused by the growth of semi-structured and unstructured data in SaaS mode, the multiple-universal table data storage and management model based on XML is proposed. XML management technology is introduced into the model, and using it to improve and optimize the multiple-universal table data storage model, and the model can effectively solve the problem of storage and management of semi-structured and unstructured data. Comparative experimental results show that, the method has high storage density and access rate, can be very good to meet the customized demand for multi-tenant data.

KEYWORDS

Cloud Computing, Data isolation, Multi-Tenant, SaaS, Semi-structured data, Unstructured data, XML

1. INTRODUCTION

In recent years, the cloud computing technology has developed rapidly. As one of the three service models of cloud computing, SaaS (Software-as-a-Service) is gradually rising, it is being accepted by more and more users, and it is getting more and more widely used. SaaS is different from the traditional software of the purchase and use, but the way to use the rental software. A number of users use specific software products in a way that is shared by the Internet. SaaS service providers provide computing and storage services for different tenants by building an application instance in the cloud computing center, and provide software operation and maintenance, upgrade services, as well as data isolation and security, and provide software operation and maintenance, upgrading services and data isolation and security guarantee (Lin & Han, 2010).

With the increasing of multimedia applications and Web applications, the explosive growth of semi-structured and unstructured data has been increased. As a result, it is inevitable that the SaaS application will appear semi-structured and unstructured data. However, the existing research and application of data storage for SaaS services are based on structured data, the multi-tenant data storage scheme is based on the relational database, which is proposed for the structured data. Due to the consistency and integrity constraints of traditional relational database, it cannot solve the storage and management requirements of semi-structured and unstructured data very well. Therefore, efficient storage and fast access schemes for semi-structured data and unstructured data in multi-tenant environment are studied, the scientific and reasonable design of SaaS multi-tenant data storage and management model can not only provide an effective solution for SaaS multimedia applications, but also can be widely used to expand the scope of SaaS applications and services, its significance and importance is self-evident.

DOI: 10.4018/JITR.2019010104
2. RELATED WORK

At present, many domestic and foreign scholars have done a lot of research work on the SaaS multi-tenant data storage and management, and the representative research results are as follows.

Frederick Chong et al. (2006) as early as 2006 put forward three kinds of data sharing approaches which are the separate databases model, the shared database separate schemas model, and the shared database shared schemas model. The shared database shared schemas model can reflect that the SaaS application is based on the characteristics of sharing to achieve customization.

Waters (2006) and Aulbach et al. (2008) proposed a private table storage mode. The scheme allocates private table for each tenant and stores all the data for the tenant. This approach has good isolation and security, but if the number of tenants is too large, it will produce too much redundant data, resulting in low efficiency and waste of resources.

Greg Gianforte (2005) presented an Extension table storage mode. In this storage scheme, a large number of data were extracted, and then placed in a shared base table, while the tenant’s own custom data stored in an extension of the table, the two table are connected through the tenant key (Tenant ID) to connect. This scheme can effectively reduce data redundancy, and save more space. But due to the use of the extended table, when we need to implement the operation involving customized data, it is bound to be a multi table joint, this will lead to low efficiency.

The storage method of Universal table proposed by Craig D Weissman (2009) is a data sharing storage structure based on the metadata-driven, the core thought is to place all the tenants’ data in a large universal table for shared storage. Because we need to store data for all tenants, so the number of columns of universal table should be greater than the number of columns of the tenant with the largest number of attributes. The parts of the custom differences between tenants are set to null values, namely, the field which is not required for the tenant is set to null. The custom operation in this scheme does not require multiple tables, but because the number of table columns cannot be predicted, the space left out will lead to the waste of storage space, and be detrimental to the extendibility.

Chen et al. (2010) and Wang et al. (2011) improved the Universal table, and put forward the storage model of multiple-universal table. In the multiple-universal table data storage mode, firstly, according to the statistical information to establish a number of different columns of the table Universal, instead of universal table used to store the data of the tenant. At the same time, we can establish the corresponding metadata table to save the tenant’s customized information, which can choose a universal table to store data that the universal table meet the needs of tenants and has the minimum number of columns. Thus, the existence of the mode null value can be reduced, and the degree of the data sparseness is reduced.

Cunningham C et al. (2003) proposed the Pivot table storage mode which is a more sophisticated storage mechanism, it uses a vertical storage strategy which divides each field of each row, and gives the number of the specified row number, then places the same type of field in a table. The proposed scheme is more suitable for query, but it is difficult to reconstruct the tuple in this method, and it is not suitable for the SaaS application environment for processing transactions.

Amazon’s Dynamo (DeCandia et al., 2007) saves the business data by its tenants with Key-value storage mode, it uses a hash algorithm to segment the data in a distributed system, and then places the segmentation data on different nodes. When the reading operation is performed, it is based on the hash value of the Key to find the corresponding node, which has good usability and scalability. But the data storage content is unstructured, the storage mode stores the original form of the data value, cannot identify the data structure, and it is more complex to query data.

Based on the above research results, above multi-tenant data storage solutions are the solutions of data sharing for structured data (Aulbach et al., 2009), at present, there is no effective solution for the data sharing of semi-structured and unstructured data. However, with the increasing popularity of Internet applications and widely used, semi-structured and unstructured data quantity becomes more and more large, the proportion is bigger than before. And because the semi-structured and
Self-Organizing Networked Learning Environments
www.igi-global.com/chapter/self-organizing-networked-learning-environments/14634?camid=4v1a