Journal of Computer and System Sciences 82 (2016) 113–129

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Developing the Cloud-integrated data replication framework in decentralized online social networks

Songling Fu [a], Ligang He [b,c,*], Xiangke Liao [d], Chenlin Huang [d]

[a] College of Polytechnic, Hunan Normal University, Changsha, China
[b] Department of Computer Science, University of Warwick, Coventry, UK
[c] School of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[d] School of Computer Science, National University of Defense Technology, Changsha, China

ARTICLE INFO

Article history:
Received 30 December 2014
Received in revised form 5 May 2015
Accepted 8 June 2015
Available online 7 August 2015

Keywords:
Decentralized online social network
Cloud
Data availability
Data replication
Erasure coding

ABSTRACT

Decentralized Online Social Network (DOSN) services have been proposed to protect data privacy. In DOSN, the data published by a user and their replicas are only stored in the friend circle of the user. Although full replication can improve Data Availability (DA), pure DOSNs may not deliver sustainable DA. This paper proposes a Cloud-assisted data replication and storage scheme, called Cadros, to improve the DA in DOSN. This paper conducts quantitative analysis about the storage capacity of Cadros, and further models and predicts the level of DA that Cadros can achieve. The data in Cadros are partitioned in such a way that the overhead caused by storing the data in the Cloud is minimized while satisfying the desired DA. This paper also proposes the data placement strategies to realize the desired DA and improve other performance. Experiments have been conducted to verify the effectiveness of Cadros.

1. Introduction

In the last decade, Online Social Networks (OSNs), such as Facebook [18], Twitter and Sina Weibo [19], have gained extreme popularity with more than a billion users worldwide. OSNs allow a user to publish the data to all friends in his friend circle.

Currently, the OSN platforms are typically centralized, where the users store their data in the centralized servers deployed by the OSN service providers. The service providers can utilize and analyze these data to know the users' private information, such as interest and personal affairs, and in the worst case may sell this information to the third party. Therefore, the current Centralized Online Social Networks (COSNs) have raised the serious concerns in privacy [14,15,30].

In order to address the data privacy issue, an obvious solution is to encrypt the user data stored in the centralized server [1,2,16,29]. A typical procedure of this solution is as follows [1]. The user data are first encrypted with the secret key, and the secret key is then encrypted with the public keys of the corresponding friends. After a friend receives the encrypted data and secret key, it first decrypts the secret key with his own private key, and the user data are then decrypted with the secret key. However, the disadvantage of this encryption solution is that a user may have a large number of friends and a user may add or delete the friends over time. It is not practical to manage this many keys. Another obvious downside of this approach is that encrypting and decrypting the user data and the secret keys incur high overhead.

* Corresponding author at: School of Computer Science and Electronic Engineering, Hunan University, Changsha, China.
E-mail address: liganghe@dcs.warwick.ac.uk (L. He).

http://dx.doi.org/10.1016/j.jcss.2015.06.010
0022-0000/

Therefore, Decentralized Online Social Networks (DOSNs) have been proposed recently as a promising solution to protect data privacy [1–4]. Although the DOSN products [17] are not as popular and mature as the OSN products [18], DOSN is indeed under active research and development. In DOSNs, in order to protect the data privacy the centralized servers are bypassed and the data published by a user are stored and disseminated only among the friend circle of the user [4]. Although DOSNs can help protect the data privacy, maintaining Data Availability (DA) becomes a big challenge [11,29,32]. This is because if a friend of the user is offline, the data stored in the friend cannot be accessed by other friends.

In order to achieve good data availability in DOSN, the data replication approach has been widely used [6]. Full replication is a popular replication approach. In this approach, a certain number of copies (e.g., k copies) are created for each data item published by a user and these data replicas are stored across the user's friends in the DOSN. By doing so, if a friend is offline, the data in this offline friend can be accessed through the replicated data stored in other friends. Consequently, data availability is improved.

Although data replication helps improve DA, the following characteristics of DOSN have negative impact on its data availability.

First, the friends in DOSN are highly volatile [22,28]. Further, the studies [11,31] show that the online/offline states of individual friends show high correlation, which indicates that many friends in a friend circle may go offline in the same time duration. When this happens, there may not be enough online friends to contribute the sufficient storage to save the data (and their replicas) published by the user.

Second, in a typical DOSN, the data published by a user are distributed among the friends in his own friend circle. Some friend circles may be small (e.g., with tens of friends), which may also cause the situation during certain periods where there are not enough online friends to offer the adequate storage for the published data.

Finally, the increasingly more data are being generated on the OSNs nowadays. On the other hand, current users often use the mobile devices, such as smart phones, to access the OSN services. The storage capacity in the mobile devices is much more limited than the desktop computers used in the "old fashioned" style of accessing OSNs. Adding even more strain, a mobile device owner typically only sets a small fraction of total storage capacity in his device to be used by the OSN client app installed in the device.

There is now a dilemma. On one hand, using the centralized server to store the published data raises the data privacy concern. On the other hand, using the friend nodes as the only storage facility may raise the data availability concern although data replication helps improve data availability. In order to further improve data availability while guaranteeing data privacy, this paper proposes a hybrid data replication and storage approach which combines the DOSN with the centralized server.

Nowadays, the Cloud becomes a popular storage platform. The Cloud is very suitable to be used as the centralized server in this work, because 1) it is available all the time and 2) the storage capacity offered by the Cloud can scale up and down according to the users' demands. Thus, this work utilizes the Cloud as the centralized server and develops a Cloud-Assisted Data Replication framework in decentralized Online Social networks (Cadros).

Due to the complexity of using encryption to protect DA, Cadros employs the erasure coding technique [20] to prevent the Cloud service provider from knowing the content of the stored data. In the erasure coding technique, the original data are split into m data segments, which are then encoded into n new data segments. Any r data segments of the n encoded segments can be used to reconstruct the original data. Thus, if the number of data segments stored in the storage facility offered by a Cloud service provider is less than r, the Cloud service provider cannot reconstruct and know the original data.

The erasure coding technique can also be regarded as a data replication technique, and its redundancy degree is (n/m). Therefore, Cadros effectively employs two data replication techniques. Namely, all data replicas generated by full replication are stored in the friend circle, while less than r data segments generated by erasure coding are stored in the Cloud. Erasure coding can save storage space when n/m is less than the number of data replicas generated for each data item in full replication (k), which is typical case. The first contribution of this work is to conduct the quantitative analysis about the amount of data that Cadros can store as the result of combining the Cloud and erasure coding with DOSN.

In order to help achieve the desired DA, it is very useful to predict the user and the friends' behavior in the DOSN, and make judicious replication and storage decisions in advance with the prediction. The second contribution of this work is to analyze the probabilistic behavior of the friend circle in the DOSN and predict the values of two metrics at a future time point: i) the storage capacity that the friend circle can contribute and ii) the amount of data that the friends request to update at a future time point. Further, this work models the relation between the above two metric values and DA, and consequently predicts the level of DA that the Cloud-assisted DOSN system can achieve at a future time point.

As discussed above, erasure coding can save storage space. However, it incurs the overhead for coding and reconstructing the data. More data are stored using erasure coding, higher overhead is incurred. Ideally, the overhead should be minimized. The third contribution of this work is to develop a data partition scheme in terms of the replication techniques, i.e., decide the portion of published data that should be stored using full replication or erasure coding, so that the erasure coding overhead is minimized while satisfying the desired level of DA.

The DA prediction in the second contribution only indicates that the hybrid system has the capacity to achieve such a certain level of DA. It still depends on the underlying data placement strategy to realize the DA. The placement strategy determines how to place the newly published data replicas among the friends. Imagine if a poor placement strategy deliberately places the data replicas on those friends who are unlikely to be still online at the targeted future time point t', then the desired level of DA will not be realized at t' even if the hybrid system has such ability based on our probabilistic