ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Automatic Data Replication in Online Social Network

Supriya F. Rathod, Prof. A.V. Deorankar

M.Tech Student, Department of Computer Science and Engineering, Government College of Engineering, Amravati, India

Associate Professor, Head of Department, Department of Information Technology, Government College of Engineering, Amravati, India

ABSTRACT: Online Social Network (OSN) is a network which consists of real users that interact with each other in various ways. Nowadays the number of users in an OSN increases diversely with respect to time. Facebook is one of theworld's largest OSN; it requires centralized datacenters to provide a constant service to all its users. In this paper, we find that OSN is open to partitioning and provides a fine grained distribution all over the world and processing of data can considerably improve the performance of system without loss in its service consistency. For this purpose we introduce the concept of replica creation in the datacenters. We show that the splitting of OSN is a good scaling method for Facebook and other services of OSN. In this paper we are introducing the new OSN model, which distributes datacenters worldwide, to help decrease service latency which leads higher inter-datacenters communication load. Each datacenter of Facebook has all the data, which are updated by the master datacenter, leading to remarkable load in the new model. This model uses datacenters to store data at their geographically nearest datacenters. The regular interactions in online network between the various users can generate long service latencies.

KEYWORDS: OnlineSocial Networks, Datacenter, Data replication.

I. INTRODUCTION

With 1.13 billion daily active users as of September 2016, Facebookis the third-busiest site on the internet, according to Alexa, and has built an extensive infrastructure to support this already massive and still growing user base. The company's servers are now housed in numerous gigantic data centers around the world. Facebook has not stopped building new data centers and seeking for new data center. Each data center houses tens of thousands of computer servers, which are networked together and linked to the outside world through fiber optic cables. Every time when users share data on Facebook site, the data goes on datacenters and then that stored information and is distributed to user friends in the network. To provide a consistent view for user in the social network, Facebook stores its data on large datacenters in the U.S. and also provides full replicas in the OSN's network. The centralization of OSN datacenters in the U.S. means the wastage internet bandwidth whenever various users from outside of the U.S. request for the same data. So in this paper we introduce the concept of placing datacenters at geographically distributed various points all over the world. Users from various locations can share their data on OSN and the replica of that data is laced on its geographically nearest datacenter. Now as we know replication is the process of creating copy of documents; this document should be in the form of XML document on a local data server this decreases the load of database which automatically increases performance of model. This happens because as the time required fetching XML document data is less than that of the document in the database. As our model includes features such as security as well as availability of documents in case of any attack, the XML document made it easier. The replica created will be stored at various datacenters data replica. These datacenters are located worldwide but replicas should be placed to their geographically nearest server. We are proposing a location based load balancing concept using which user will be automatically redirected to their geographically nearest server and in case of any damage, system will automatically recover its damaged file from other datacenters as replica documents are stored in distributed datacenters.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

To design a large-scale distributed storage system, availability, scalability, user experience, and resource utilization are all under consideration. Known approaches for distributing and locating a huge amount of objects in such systems cannot meet all the requirements. In this paper, we propose a new method which uses three-level mapping to locate objects. With this method, objects can be distributed among a large number of storage devices with high scalability. Clients can access objects in parallel without consulting a central server each time, and the cost to locate an object is one step in general. Moreover, replication and weighted allocation can also be supported, both of which are needed to permit systems to efficiently grow while accommodating new technology.Use Master-Subordinate replication with read-only replicas to improve performance of queries. Locate the replicas close to the applications that access them and use simple one-way synchronization to push updates to them from a master database. Use Master-Master replication to improve the scalability of write operations. Applications can write more quickly to a local copy of the data, but there is additional complexity because two-way synchronization with other data stores is required.Include in each replica any reference data that is relatively static, and is required for queries executed against that replica to avoid the requirement to cross the network to another datacenter.

II.LITERATURE SURVEY

Previous study of A. Thomson, D. J. Abadi [2015] uses the concept of reliability protection of replicas over geographically placed datacenters or within a datacenter in the network. This work mainly focuses on providing replicas on the geographically nearest server. After this study, Guoxin Liu, HaiyingShen [2016] uses the algorithm for Automatic user data replication in datacenters. This work focus on data replica to be stored into database on local server which may increase the local database load as replica increases.Y. Zhou, T. Z. J. Fu, D. M. Chiu [2013] and M. S. Ardekani, D. B. Terry [2014], also shares the adaptive replication techniques with some works in P2P systems and in clouds, which energetically used to the number and location of data replicas. These works focus on load balancing, while *AD3* focuses on saving network load, availability of document, security.The basic idea to design data replication datacenters is based on many previous studies on online social network properties. The work in [3] studied online social network evolution patterns and user behaviour. Online social networks are categorized by the existence of different communities based on user communication, with a high rate of interaction between communities and low rate of interactions outside [2]. For very large online social network, the network communities are become untight, which leads to idea behind the result of data replication in various datacenters to create replicas based on different user communication and update rates rather than static friend communication.

III. DATA REPLICATION USING DISTRIBUTED DATACENTERS

In this section, we authenticate the profit of the new OSN model for distributed datacenters. The main aim to develop *AD*3 for OSNs is reduce the network load with low service latency. If at any instance an id is created randomly in the OSN is occurred and the user of that id is widely accessible, then we crawled that user's data in OSN. We observe all the users' IDs and the time stamps of events on their profile not including crawling event contents. All the datasets in online social network are completely protected and made private. In this section we design a model for AD3, to achieve our goal of load balancing concept in the distributed datacenters by using replica formation concept. This leads to increase document security and availability.

3.1 User Data Replication

In the existing system only inter-network communication load balancing is discussed. There is no provision is given to increase the availability and security of documents. Data Replica will be stored into database on local server which may increase the local database load as replica increases. Replica data will be stored on local server as it is, which may cause attack on replica data. Replica update will increases network congestion. To overcome all this disadvantages we proposed model of AD3 in the online social network for distributed datacenters. Replication is the process of creating copy of documents; this document should be in the form of XML document on a local data server this decreases the load of database which automatically increases performance of model.

Here we produce data replica created through different user interactions is used in the online social networks. In study it has been seen that the interactions between various users decreases with respect to their age and time. And also each user in the network has different update rates and visit rates according to their interest. This shows that each user

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

relationships do not essentially have high data visit or update rates between the friends in social networking sites. These rates depend upon the friends and over time of their updates in the sites. According to study about 90% of all friend pairs in networks have a regular interaction rate of friends visit below 0.4, with the regular interaction rate of the remaining 10% is ranges from 0.4 to 1.8[1], so the data update rate is completely depends on users activities. From above information it is clear that not all the users have same data visit rate and updates rates in the networks. Therefore users with low visit rates in communities are leads to generate replicas with low intensity. But the replicas are created for these types of users also this leads to waste of storage space in the datacenters. Thus, we are considering only those visit rates of a user's data that are communicating regularly in the network data replication. As various users have different updates and visit rates, thus we have to differenciate various users update rates for this we use equation to calculate variance by using $\sigma^2 = \sum(x-\mu)^2/(n-1)$, where x is the interaction rate of users friendship and μ is the average and n is the interaction rates number[1]. From this equation it is shown that about 10% of friend interactions in networking sites are having high variation which ranges from [0.444 to 29.66], these leads that update rates of various users can vary according to time in the network. From all this study it is clear that visit or update rates of user's data replica should be checked always with respect to time and updates.

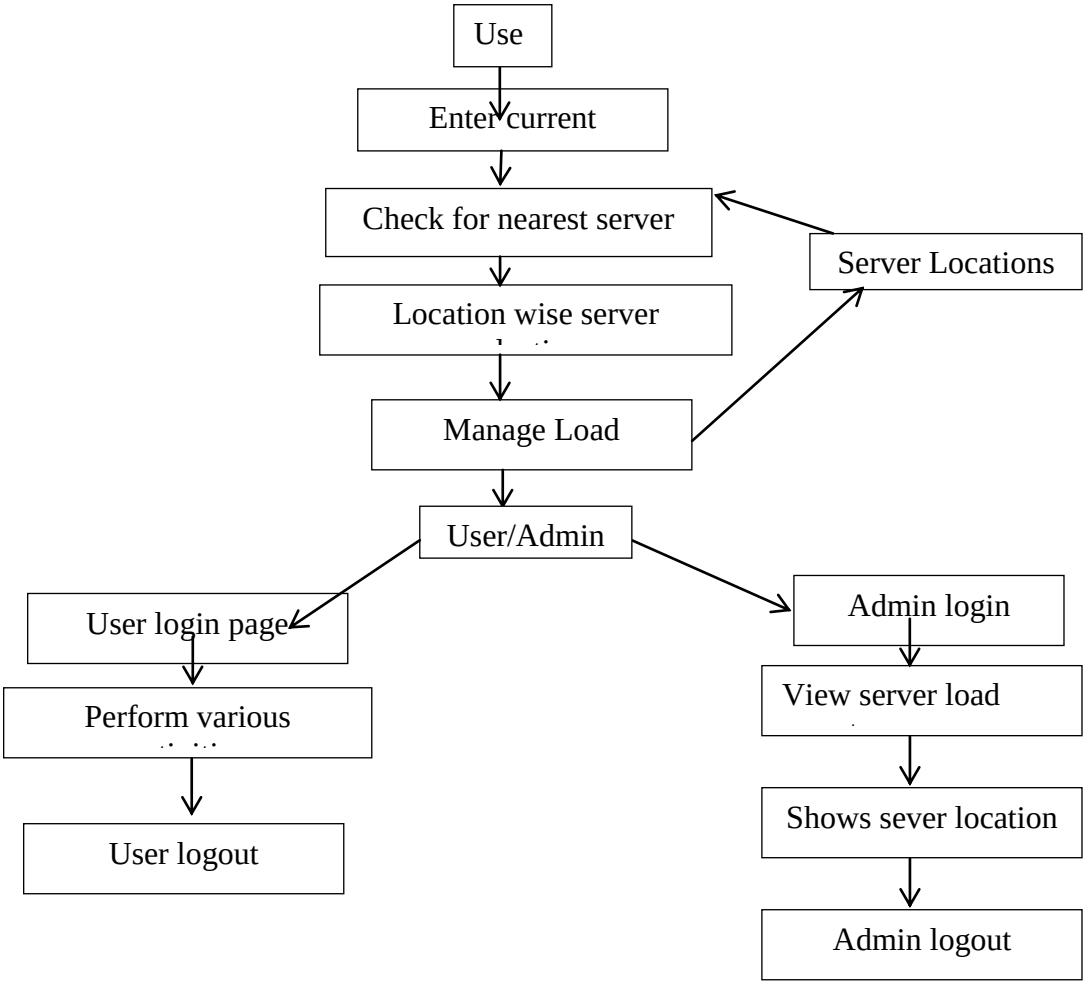

Fig. 1: Proposed Architecture of AD3

Now for performing user replica operations we need to consider news feed of various users Local-to-local (LL) and remote-to-local (RL) posts. Users in the OSNs are interconnected through their relation of friendship and are able to do

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

asynchronous communications such as posting an update in the form of text, photos, and videos. These communications creates traffic in social network and also creates some performance issue in the OSN function. To simplify the problem we represent OSN traffic patterns, we consider some built-in Facebook modes in the form of wall posts, comments, andlike tags on the pages of Facebook.

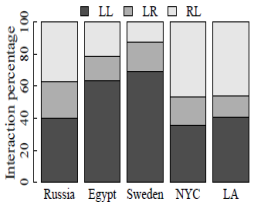

Fig 2: Wall posts byendpoint location

Every interaction on Facebook requires some connection with servers. In figure 2, while doing a wall post a user need to opens a TCP connection withFacebook site and then sends the new content the social network in return receives HTMLmarkup to present the post on wall of his/her friends. For understanding the performance of Facebook's traffic inthe OSNfirstly we need to think on the various features of the internet paths between its users and the OSN infrastructure. For that we need to discover theendpoint locations that of service interactions in altered regions. Secondly, we need to measure service latency, loss and capacity of paths betweenthese endpoints and user hosts on each datacenter. Our main goal is to measure the effect of various network characteristics on the OSN network traffic. Because we do not have direct access to the actual infrastructure of Facebook endpoints or the user hosts on the site therefore for our study we are considering measurementnodes from[2] in which are located at nearest datacenters. For characterizing user communication patterns within each datacenter, we need to address a fault of the crawl process such as the posts that are not addressed to users can reside on crawled region cannot be showed in the user news feed or his friends news feed scannedby the crawl. Now in this paper we observed that Local-to-local (LL) and remote-to-local (RL)user posts can be scanned by using news feeds oflocal users in the network. Those interactions from Local-to-remote (LR) are observed on the profiles of remote users other than local users which arenot characterized in the crawls. Various posts delivery is determined by the social graph of this interaction of user in the network. Apost from any user will be delivered to users who are friends and the addressee of that user.

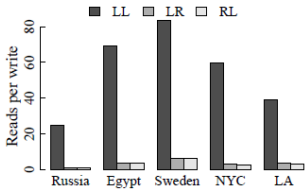

Fig 3: Delivery ratio of wallposts

As shown in the figure 3, the delivery ratio of wall posts for various users in the network. It is shown that the y-axis gives the number of users in the region that receive each post in the network and it's according to user interaction and friends of user in the network. Now according all this data LL posts will result in the largest read-to-write ratio as the sets of posted update and addressee's of user tend to havehigh overlap rate within the datacenter. LR and RL interactions of users have much lower delivery ratio than LL in the datacenter this is because friends of the local user have different sets of friends from various regions. Figure 4 shows the analysis of these regional interaction patterns of local users shows that traffic is produced and consumed in major volumes within the same region will be local user traffic. This analysis leads that locality of interest is for caching of postsoriginating in the region which is geographically close to user would reduce request delay in the communication.We believe that our study will help to improve Facebook's infrastructure to provide to local interactions andproduced replica will be stored on the nearest server in the network.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

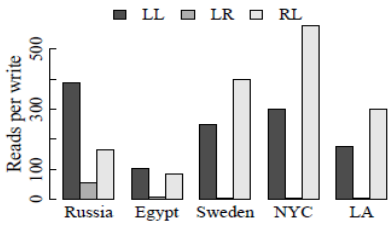

Fig 4: Delivery ratio of commentand like posts

In this waywe observed that the datacenterwhich handles the request may not be the closest one, but when requests issued from the U.S. will sometimes redirected to the datacenters in the Europe, and because of that the network shows some delay between hosts in Russia, or Egypt, while the users in Western Europe may be significant to get data in less time. We include the impact of user CDN traffic in our evaluation of the overall duration of Facebook transactions.

IV. RESULTS

Initially, we did not apply the atomized user data replication in order to see the sole effect of the selective data replication algorithm. LocMap generates a stable hit rate because an interaction between geographically distant friends always produces a miss. Due to the variation in visit rates and different interacting friends each day, the hit rate of AD3 also varies over time.In general, Software Engineering distinguishes software fault (technology) from software failure. In case of a failure, the software does not show any wrong result instead it retrieve results from another server as the replicas are present at every server. So in any case if any one of the server is failed then user will get their data from another server and they are able to get their data anywhere and at any time. A fault is a programming error that may or may not actually manifest as a failure. Figure 5 shows the comparison between the time required to fetch the data from database and XML data. This shows that time required for fetching XML data is lower than the database data this will introduce the low service latency in the network and provides idea about the minimization of network load in the system.

Prevailing data replication systems are based on creating replicas of data on datacenters but they do not provide the concept of load balancing. The fact behind this is that by providing concept of data replication user will never lost their data in social networks. On the same side we are providing network load balance. In this system user can create their profile and communicate with their friends in the network. Generally people want to become friends with people having similar interest. So, our proposed system matches the location of users by using redirection server and providing load balance on the social networks.The basic difference between the existing system and the proposed system is that concept of load balancing and security to the user documents. For providing security we store document in the form of XML documents. The reason behind the inclusion of these parameters is that we have to use this automatic data replication system for security application in the social network for user's data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

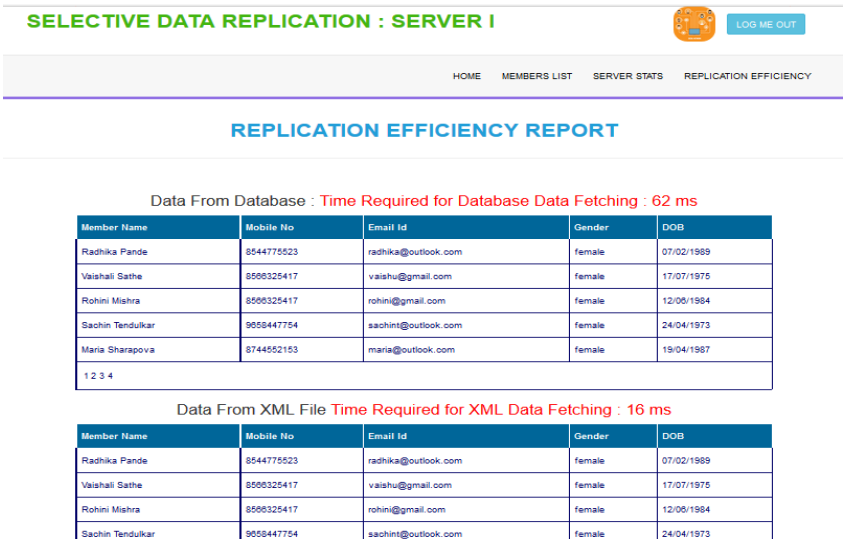

Fig. 5 Comparison between time required to Database and XML data fetching

This addition of content feature viz. security and creating replica is going to enhance the application of automatic data replication. This reduces workload in the network.Therefore AD3 architecture is proposed to overcome this entire load in the network. Data will be provided in the form of encrypted and decrypted data to increase the security of data. Replica of data is stored in the form of XML document to improve load balancing. Now, as we know replication is the process of creating copy of documents; this document should be in the form of XML document on a local data server this decreases the load of database which automatically increases performance of model. This happens because as the time required to fetching XML document data is less than that of the document in the database. As our model includes features such as security as well as availability of documents in case of any attack, the XML document made it easier. The replica created will be stored at various datacenters in the form of Data replica. These datacenters are located worldwide but replicas should be placed to their geographically nearest server. We proposed location based load balancing concept using which user will be redirected to the geographically nearest server automatically. In case of any damage, system will automatically recover the file from stored replica documents in distributed datacenters. Thus, we study how to replicate data in online social network using distributed datacenters to minimize network load. The replication of user data can helps to provide more data requests locally, leading to lower service latencies. A rarely visited replica provides a little advantage service latency reduction while at the same time it increases network load for updates, leading to higher network load.

V. CONCLUSIONS

While a new OSN model with many small, globally distributed datacenters will result in improved service latencies for users, a critical challenge in enabling such a model is reducing inter-datacenter communications (i.e., network load). Thus, we propose the Selective Data replication mechanism in Distributed Datacenters (AD3) to reduce inter datacenter communications while achieving low service latency. We verify the advantages of the new OSN model and present OSN properties from the analysis of our trace datasets to show the design rationale of AD3. Some friends may not have frequent interactions and some distant friends may have frequent interactions. In AD3, rather than relying on static friendship, each datacenter refers to the real user interactions and jointly considers the update load and saved visit load in determining replication in order to reduce inter datacenter communications. Also, since different atomized data has different update rates, each datacenter only replicates atomized data that saves inter-datacenter communications, rather than replicating a user's entire dataset. AD3 also has a locality-aware multicast update tree for consistency maintenance and a replica deactivation scheme to further reduce network load. To avoid workload congestion of datacenters in AD3, each overloaded datacenter releases its excess load to its neighboringdatacenters based on their available capacities.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Through trace-driven experiments on PlanetLab, we prove thatAD3 out performs other replication methods in reducing network load and service latency.

Until now OSN model with distributed datacenters results in improved service latencies for users but challenge is that this model are not able to reducing inter-datacenter network load. Thus, we propose the Automatic Data replication concept in Distributed Datacenters (*AD*3) to lower inter datacenter network load while achieving low service latency in OSN. This also provides documents in order to achieve security as well as availability of documents in case of any attack and manages location wise datacenter selection to enhance the security of documents stored in social network using AES algorithm. Some friends may not have regular interactions and some distant friends may have regular interactions. On basis of this AD3 provides replica activation and deactivation mechanism. In *AD*3, rather than relying on inactive friendship, each datacenter refers to the actual user interactions between friends and it jointly considers the update load and visit load in determining replication in order to achieve low inter datacenter communications. Thus by applying all this features, we summarize a concept of AD3. In all this provide document security, availability by reducing network load.

REFERENCES

[1] M. P. Wittie, V. Pejovic, L. B. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks," *ACM conext*, 2010.
[2] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The little engine(s) that could: scaling online social networks," *SIGCOMM*, 2010.
[3] M. S. Ardekani, D. B. Terry, "A Self-Configurable Geo-Replicated Cloud Storage System," in IEEE/ACM Transactions on Networking, 2014.
[4] Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. C. Lau, "Scaling Social Media Applications Into Geo-Distributed Clouds," *INFOCOM*, 2012.
[5] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the evolution of user interaction in facebook," *WOSN*, 2009.
[6] **"**Facebook." Available: http://www.facebook.com/
[7] "Socialbakers." Http://www.socialbakers.com/facebookstatistics/
[8] H. Shen and G. Liu, "A geographically-aware poll-based distributed file consistency maintenance method for P2P systems," *TPDS*, 2012.
[9] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," *ACM IMC*, 2009.
[10] M. Wittie, V. Pejovic, L. B. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks," *ACM conext*, 2010.
[11] Z. Li and H. Shen, "Social-p2p: An online social network based P2P file sharing system," *ICNP*, 2012.
[12] Guoxin Liu, HaiyingShen, *Senior Member IEEE*, Harrison Chandler, "Automatic Data Replication for Online Social Networks with Distributed Datacenters," IEEE Transactions on Parallel and Distributed Systems,2016.
[13] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: Automated data placement for geo-distributed cloud services," Usenix NSDI, 2010.
[14] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, "DONAR: Decentralized server selection for cloud services," AMC SIGCOMM, 2010.