A multilingual reference based on cloud pattern

G.Rama Rao
Department of Computer science and Engineering, Christu Jyothi Institute of Technology and Science, Jangaon

***Abstract*- With the explosive growth of data, the risk and cost of data management are significantly increasing. In order to address this problem, more and more users and enterprises transfer their data to the cloud and access the data via Internet. However, this approach often results in a large volume of redundant data in the cloud. According to an IDC report, around 75% of the data are redundant across the world. ESG indicates that over 90% of the redundant data are in backup and archiving systems. The reason behind this is that multiple users tend to store similar files in the cloud. Unfortunately, the redundant data not only consume significant IT resources and energy but also occupy expensive network bandwidth. Therefore, data reduplications is urgently required to alleviate these problems in the cloud.**

***Index Terms*- Similarity detection; Sampling; Shingle; Position-aware; Cloud.**

1. INTRODUCTION

With the explosive growth of data, the risk and cost of data management are significantly increasing. In order to address this problem, more and more users and enterprises transfer their data to the cloud and access the data via Internet. However, this approach often results in a large volume of redundant data in the cloud. According to an IDC report, around 75% of the data are redundant across the world. ESG indicates that over 90% of the redundant data are in backup and archiving systems. The reason behind this is that multiple users tend to store similar files in the cloud. Unfortunately, the redundant data not only consume significant IT resources and energy but also occupy expensive network bandwidth. Therefore, data DE duplication is urgently required to alleviate these problems in the cloud. Data duplication calculates a unique fingerprint for every data block by using hash algorithms such as MD5 and SHA-1. The calculated fingerprint is then compared against other existing fingerprints in a database that dedicates for storing the fingerprints. If the fingerprint is already in the database, the data block does not need to be stored again, a pointer to the first instance is inserted in place of the duplicated data block. By doing so, data DE duplication is able to eliminate the redundant data by storing only one data copy, thus reducing the cost of data management and network bandwidth . However, data deduplication suffers disk bottleneck in the process of fingerprint lookup. This is because massive requests going to disk drives generate a large volume of random disk accesses which significantly decrease the throughput of deduplication and increases the system latency. In the cloud, any increased latency may result in a massive loss to the enterprises . For example, according to,every 100 ms of increased latency would reduce 1% of sales for Amazon, and an extra 0.5 seconds in search page display time can cut down revenues of Google by 20%. On the contrary, any decreased latency will bring huge benefits to the enterprises. Hamilton et al.note that only a speedup of 5 seconds at Shopzill will bring about an increase of page view by 25% and taxes by 10% while a reduction of hardware by 50% and traffic from Google by 120%. Therefore, reducing the latency in the cloud environment is very important for the enterprises who store their data in the cloud.

2. RELATED WORK

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of

external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

1. A universal storage architecture for big data in cloud environment

With the rapid development of the Internet of Things and Electronic Commerce, we haveentered the era of big data. The characteristics, such as great amount and heterogeneousity, ofbig data bring the challenge to the storage and analytics. The paper presented a universalstorage architecture for big data in cloud environment. We use clustering analysis to dividethe cloud nodes into multiple clusters according to the communication cost between differentnodes. The cluster with the strongest computing power is selected to provide the universalstorage and query interface for users. Each of other clusters is responsible for storing the dataof a particular model, such as relational data, key-value data, and document data and so on.Experiments show that our architecture can store all kinds of heterogeneous big data andprovide users with unified storage and query interface for big data easily and quickly.

2. The digital universe decade-are you ready

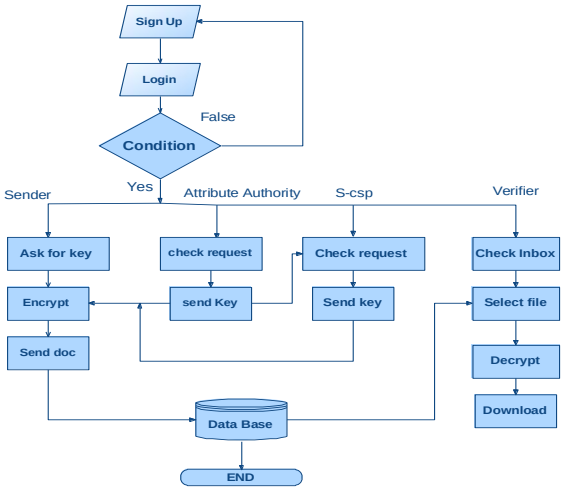
The title of that track from the 1974 B. Achman Turner Overdrive album Not Fragile aptlydescribes the state of today's Digital Universe. Between now and 2020, the amount of digitalinformation created and replicated in the world will grow to an almost inconceivable 35 trilliongigabytes as all major forms of media voice, TV, radio, print complete the journey from analogto digital**.**

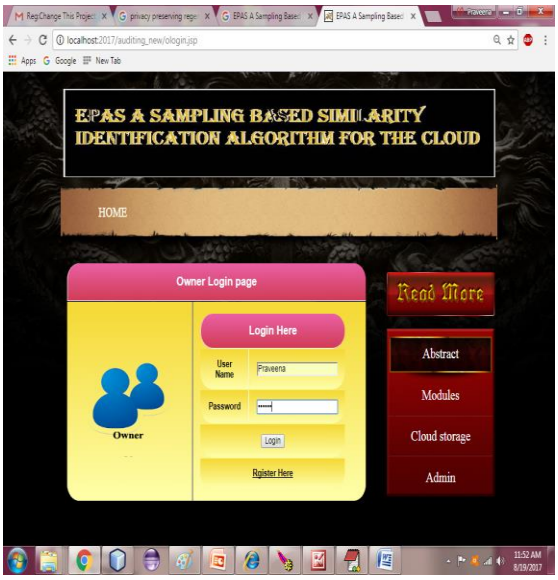
3 BACKGROUND

The explosive growth of data brings new challenges to the data storage and management in cloud environment. These data usually have to be processed in a timely fashion in the cloud. Thus, any increased latency may cause a massive loss to the enterprises. Similarity detection plays a very important role in data management. Many typical algorithms such as Shingle, Simhash, Traits and Traditional Sampling Algorithm (TSA) are extensively used. The Shingle, Simhash and Traits algorithms read entire source file to calculate the corresponding similarity characteristic value, thus requiring lots of CPU cycles and memory space and incurring tremendous disk accesses. In addition, the overhead increases with the growth of data set volume and results in a long delay. Instead of reading entire file, TSA samples some data blocks to calculate the fingerprints as similarity characteristics value. The overhead of TSA is fixed and negligible. This paper proposes an Enhanced Position-Aware Sampling algorithm (EPAS) to identify file similarity for the cloud by modulo file length. EPAS concurrently samples data blocks from the head and the tail of the modulated file to avoid the position shift incurred by the modifications. Meanwhile, an improved metric is proposed to measure the similarity between different files and make the possible detection probability close to the actual probability. Furthermore, this paper describes a query algorithm to reduce the time overhead of similarity detection.

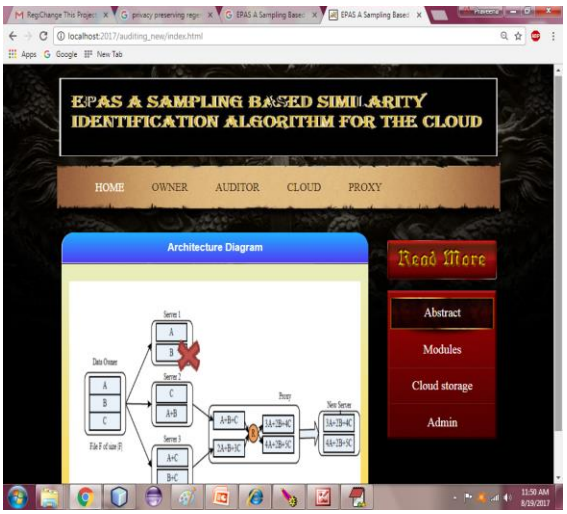
4. SYSTEM FLOW

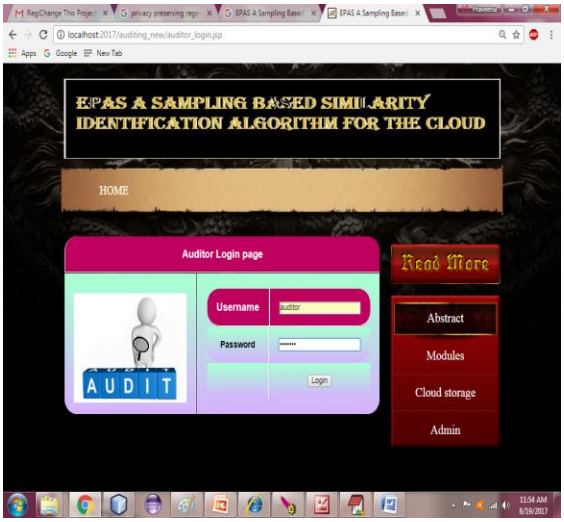
The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing Information flow and functional detail.

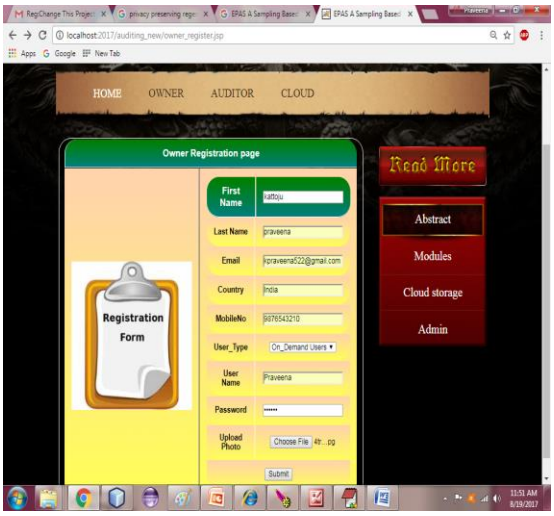


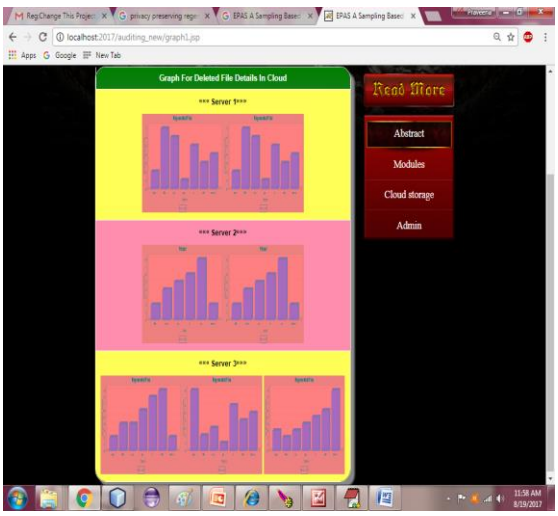


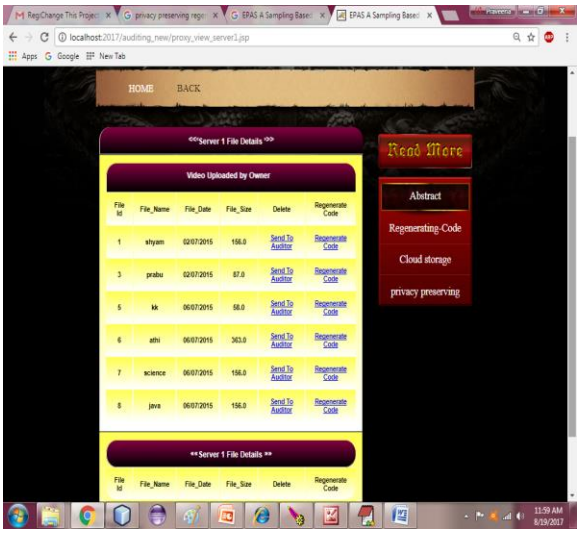
5. EXPERIMENTAL RESULTS











6. CONCLUSION

In this paper, we have studied a novel problem, cross-site cold-start product recommendation, i.e., recommending products from e-commerce websites to microblogging users without historical purchase records. Our main idea is that on the e-commerce websites, users and products can be represented in the same latent feature space through feature learning with the recurrent neural networks. Using a set of linked users across both e-commerce websites and social networking sites as a bridge, we can learn feature mapping functions using a modified\ gradient boosting trees method, which maps users' attributes extracted from social networking sites onto feature representations learned from e-commerce websites. The mapped user features can be effectively incorporated into a feature-based matrix factorization approach for cold-start product recommendation. We have constructed a large dataset from WEIBO and JINGDONG. The results show that our proposed framework is indeed effective in addressing the cross-site cold-start product recommendation problem. We believe that our study will have profound impact on both research and industry communities. Currently, only a simple neutral network architecture has been employed for user and product embeddings learning. In the future, more advanced deep learning models such as Convolutional Neural Networks13 can be explored for feature learning. We will also consider improving the current feature mapping method through ideas in transferring learning.

REFERENCES

[1] Q. Zhang, Z. Chen, A. Lv, L. Zhao, F. Liu, and J. Zou, "A universal storage architecture forbig data in cloud environment," in Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing. IEEE, 2013, pp. 476480.

[2] J. Gantz and D. Reinsel, "The digital universe decade-are you ready," IDC iView, 2010.

[3] H. Biggar, "Experiencing data de-duplication: Improving efficiency and reducing capacity requirements," The Enterprise Strategy Group, 2007.

[4] F. Guo and P. Efstathopoulos, "Building a highperformance deduplication system," in Proceedings of the 2011 USENIX conference on USENIX annual technical conference. USENIX Association, 2011, pp. 25–25.

[5] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in ACM SIGOPS Operating Systems Review, vol. 35, no. 5. ACM, 2001, pp. 174–187.

[6] B. Zhu, K. Li, and R. H. Patterson, "Avoiding the disk bottleneck in the data domain deduplication file system." in Fast, vol. 8, 2008, pp. 1–14.

[7] Y. Deng, "What is the future of disk drives, death or rebirth?" ACM Computing Surveys (CSUR), vol. 43, no. 3, p. 23, 2011.

[8] C. Wu, X. LIN, D. Yu, W. Xu, and L. Li, "End-to-end delay minimization for scientific workflows in clouds under budget constraint," IEEE Transaction on Cloud Computing (TCC), vol. 3, pp.169–181, 2014.

[9] G. Linden, "Make data useful," http://home.blarg.net/_glinden/StanfordDataMining.2006-11-29.ppt, 2006.

[10] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: listen to your customers not to the hippo," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp.959–967.

[11] J. Hamilton, "The cost of latency," Perspectives Blog, 2009.

[12] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, "Extreme binning: Scalable, parallel deduplication for chunk-based file backup," in Modeling, Analysis & Simulation of Computer and Telecommunication Systems, 2009. MASCOTS'09. IEEE International Symposium on. IEEE, 2009, pp. 1–9.

[13] W. Xia, H. Jiang, D. Feng, and Y. Hua, "Silo: a similarity-locality based near-exact deduplication scheme with low ram overhead and high throughput," in Proceedings of the 2011 USENIX conference on USENIX annual technical conference. USENIX Association, 2011,pp. 26–28.