International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 21 Issue 1 – APRIL 2016.

OPTIMAL CLUSTERING QoS HISTORY RECORD BASED SERVICE COMPOSITION FOR BIG DATA APPLICATIONS

Bavithra S[#1] and Jayamala R[*2]

[#]*M.E, Software Engineering, University College of Engineering,Tiruchirappalli ,India*

[*] *Assistant Professor, IT, University College of Engineering, Tiruchirappalli, India*

***Abstract*— Cloud stores the large amount of data and the facilitator to the emergence of the big data. Big data refers to large amount of data and cloud computing is a techniques to perform on-demand task on big data. Large data sources from the cloud are stored in a distributed database and processed through a programming model for big dataset with a parallel distributed algorithm in a cluster. Cross-cloud service composition is an approach used for processing big data applications. Cloud service composition based on QoS values history records. Hierarchial clustering and multiscale bootstrap analysis techniques are used to composite the dataset. This technique is used for clustering the dataset and provided the optimal service composition. Furthermore, it significantly reduces the time complexity of the cross-cloud service composition using the QoS history records.**

Index Terms— **big data, cross-cloud, history records, Quality of Service(QoS), service composition.**

I. INTRODUCTION

In recent years most of the business industries store the large amount of data in cloud with low cost. Cloud computing is a web administration based innovation and 'on-interest' administration whereby assets are given as shared administrations. Big data is high variety, high velocity and high volume information that utilizes distributed storage technology based on cloud computing. It processes the distributed queries across multiple datasets and returns the results. It provides economical way to support more users and new IT companies[1],[21]. Big Data Applications can benefit from cloud computing such as elasticity, environment friendly, pay-per-use.

In order to satisfy different security and privacy requirements, cloud computing have three models such as private cloud, public cloud and hybrid cloud for big data processing system. Private cloud delivery model owned by single organization and does not share resources with any other company. Private clouds[20] can offer efficient, cost-effective way to implement analysis of big data. Public clouds are access by anyone and security issues are more than the private cloud. The hybrid cloud architecture merges private and public cloud deployments. This is an attempt to achieve elasticity and security, or provide cheaper base load and burst capabilities.

Web service having functional property and non-functional property i.e Quality of Service (Qos)[1],[2]. Qos is an concept encompasses number of properties such as availability ,execution price, reliability and reputation[7] offered by an application and by the platform that hosts it. Qos is critical element for cloud suppliers, who gives the publicized quality attributes administration to the cloud clients and for cloud suppliers, who need to discover the tradeoffs between QoS levels and operational expenses. Any violation of Service Level Agreement (SLA) results in loss of administration for both cloud suppliers and cloud clients.

A. Related work

L.Zheng et al suggests the Qos-Aware service composition[1],[2],[11] based on the QoS values web service providers. Based on the QoS criteria service composition occurs in the service providers. This technique used to assess the overall quality of the service providers. Service composition is the determining factor to ensure the customer satisfaction and minimize the execution duration.

Ranking mechanism is to facilitate the process of Web service discovery, based on assessing the semantic similarity between the service parameters and request[8]. Calculating precision and recall for find the matched services based on the service level advertisement.

M.Alrifai et al suggests a worldwide optimization[2],[12] with nearby determination. Neighborhood determination depends on the utility estimation of the administration and locate the greatest estimation of the utility. Global Optimization is utilizing the aggregate and composition functions to discover the enhanced choice.

Selecting the skyline[2],[15],[16] administrations in view of the utility and composite total capacities. The skyline administrations interms of their QoS values, and a calculation that enhances the proficiency of the best in class arrangement by pruning non-skyline administrations.

L.Qi et al propose, a heuristic administration creation LOEM (Local Optimization and Enumeration Method). LOEM goes for separating the applicants of every undertaking to a little number of accessible ones by neighborhood selection[17],[18], and after that all the composite answers for locate a close to-ideal one.
Privacy maintaining information examination and knowledge publishing have gotten to be serious downside in today's present world. That is the reason particular methods of data anonymization frameworks are anticipated. To the excellent of our data, TDS methodology abuse Mapreduce[3],[19] are utilized on cloud to data anonymization and intentionally outlined a bundle of imaginative Mapreduce occupations to solidly finish the specialization calculation in an exceedingly prominently adaptable means. The anonymization is powerful to create the security on data sets and increase high adaptability.

The preeminent detriment of this anonymization is brought together methodologies probably experience the ill effects of low intensity and versatility once taking care of huge scale data sets.

Personal health record (PHR)[6] is a patient-driven of wellbeing data, that is typically outsourced to be hang on at an outsider, similar to cloud suppliers. This empowers a patient to specifically share PHR among an arrangement of clients by encoding the record under an arrangement of properties, without need to know a complete rundown of clients.

Using ABE, access policies area unit expressed supported the attributes of users or information, so as to guard the private health information hold on on a semi-trusted server, we have a tendency to adopt attribute-based secret writing (ABE) because the main secret writing primitive. This method achieved scalable and fine-grained data access control for PHRs[12]. The main drawback of this system is manual insurance climbing, frustration of missing doses, difficult for long-term medication.

B. Motivation

In past work , a history record based administration advancement strategy, named HireSome, which upgrades the creditability of administration composition[1],[9], by utilizing the history records of administrations quality connected with their executions to rate administrations' execution quality next time later on, instead of utilizing the QoS values gave by the Web administration suppliers. HireSome receives an ideal strategy to choose the best Web administration for every undertaking without producing all conceivable piece arranges, so accelerating the calculation.

The technique portrayed above spotlights on the exploration of segment administrations, which chooses the best part benefit for every assignment to produce a most qualified service arrangement. In any case, this strategy doesn't consider the Web administration structure issue globally, so the best administration arrangement acquired by HireSome possibly not the ideal one[1], i.e., an imperfect one. We propose another history-record based administration improvement strategy, which considers cloud QoS history record based administration service issue globally. So as to decrease the time complexity nature, utilizing the progressive grouping technique to group the QoS history records, then utilizing the groups to give QoS history-record based service plans. To gauge the exactness of the group utilizing multiscale bootstrap resampling.

II. PRELIMINARY KNOWLEDGE

In this segment, we quickly present some preliminary knowledge about clustering and Quality of Service (QoS) [4]criteria. cluster analysis is an intense exploratory system for finding group of comparable perceptions inside of an information set. The thought of cluster[10],[14] investigation is to utilize estimations of variables to devise a plan for gathering objects in a manner that comparable items will have a place with the same group (in some sense or another) to each other than to those in different group.

A. Quality of Service

- Execution price

  Service requester pays for the requested service operation to the service provider.

- Execution Duration

  It measures the expected time in seconds between the moment when the service request and results are received.

- Reliability

  It is a measure related to web services and the network connections between service providers and requesters.

- Availability

  Probability of that service is accessible.

- Reputation

  It is a measure of trustworthines

Table 1: Formulas of QoS Criteria

Formula	**Explanation**
$q_{Price}(s, op)$	S=service, op=operation of service
$q_{duration}(s, op) = T_{process}$ (s,op) + T_{trans}(s,op)	$T_{process}$(s,op)=sum of the processing time. T_{trans} (s,op)= past observation of processing time
$q_{availability}(s) = T_a(s)/\theta$	$T_a(s)$ = Total amount of time Θ may vary independently
$q_{reputation} = \frac{\sum_{i=1}^{n} R_i}{n}$	R_i = End User ranking n = number of times the services graded
$q_{reliability}(s) = N_c(s)/K$	$N_c(s)$ = $No\ of\ times\ services\ success$ K=total number of invocations

Table 2: Aggregation functions for computing QoS Criteria

Criteria	**Aggregation function**
Price	$q_{price}(s)=\sum_{i=1}^{N} q_{price}(s_i, op(t_i))$
Duration	$q_{duration}(p) = C\ P\ A\ (\ p, q_{duration})$
Reputation	$q_{reputation}(p)=\frac{1}{N}\sum_{i=1}^{N} q_{reputation}(s_i)$
Reliability	$q_{Reliability}(p) = \prod_{i=1}^{N} \left(e^{q_{reliability}(s_i)\cdot s_i}\right)$
Availability	$q_{Availability}(p) = \prod_{i=1}^{N} \left(e^{\prod_{i=1}^{N}\left(e^{q_{availability}(s_i)\cdot s_i}\right)}\right)$

B. *Kmeans clustering*

K-means clustering algorithm is used to representative history records without affecting privacy[1],[24]. It partitions the n objects into k cluster so that intracluster similarity is high and intercluster similarity is low.The Square error criterion is used to calculate the clustering effect

$$E = \sum_{i=1}^{k} \sum_{C \varepsilon c_i} |C - m_i|^2 \quad (1)$$

Where C = point in space representing a given object
m_i = mean of cluster c_i

Eucledian distance formula

$$\text{Dist}((x, y), (c, d)) = \sqrt{(x-c)^2 + (y-d)^2} \quad (2)$$

Where (x, y) and (c, d) are the two points in the eucledian points in datasets.

The distance between the two data points are calculated using the eucledian distance formula.

C. *Hierarchial clustering*

Hierarchic techniques begin at the leaves and consolidation the leaves hub groups together. The procedure of blending branches depends on the estimation of watched divergence coefficients between all sets of people hub in the information. The calculations are the procedures by which a progressive grouping method maps an arrangement of watched likeness or disparity coefficients to another arrangement of closeness or divergence coefficients.

Correlation Method

For information communicated as (npx) network or information outline, we expect that the information is n perceptions of p articles, which are to be clustered. The row vector relates to the perception of these objects and the column vector compares to a sample of item with size n. There are a few techniques to gauge the dissimilarities between items. For information framework , the default is

$$\mathbf{d2 = 1 - \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x})(x_{ik}-\bar{x}k)}{\left\{\left(\sum_{i=1}^{n}(x_{ij}-\bar{x})^2\right)\left(\sum_{i=1}^{n}(x_{ik}-\bar{x}k)^2\right)\right\}}} \quad \mathbf{(3)}$$

Definition 1 (Dendogram).

Dendrogram is the aftereffect of various hierarchial clustering procedure which are spoken to as the tree graph. Tree structure are built from the separation grid and to orchestrate the people into a various leveled request such that people with the most astounding comparability are set together. At that point bunches of items are connected with different gatherings, which they are most nearly look like, thus on until the greater part of the people have been set into a characterization plan. Nonetheless, bunch emerging from various separation measures and grouping systems connected to the same information will vary as per bootstrap resampling received for separation measures.

D. *Multiscale Bootstrap Resampling*

It was additionally used to compute likelihood values (p-values) for every group utilizing bootstrap resampling procedures. The two sorts of p-values that were gotten are the: around fair-minded (AU) p-values and bootstrap likelihood (BP) value. Multiscale bootstrap resampling procedure was utilized for the estimation of AU p values, which has predominance in predisposition over BP-esteem figured by the normal bootstrap resampling. The algorithms of multiscale bootstrap are:

- Generate bootstrap samples for each sample size.
- Apply hierarchical clustering to each bootstrap sample to obtain the sets of bootstrap replications of dendrograms.
- Compute bootstrap probability for each sample size.
- Using values of bootstrap probabilities, estimate the *p*-value by fitting a theoretical equation to them. The estimated *p-value* is called AU (approximately unbiased) value.

Definition 2(Bayesian Bootstrap)

Create datasets using reweighting the initial data. Assume N datapoints, assign weight provide new dataset is provides the probability resampling technique in the hierarchial clustering and provide accuracy to the clustering in the clustering tree and subtree in the hierarchial clustering.

III. SYSTEM MODEL

A. *System Architecture*

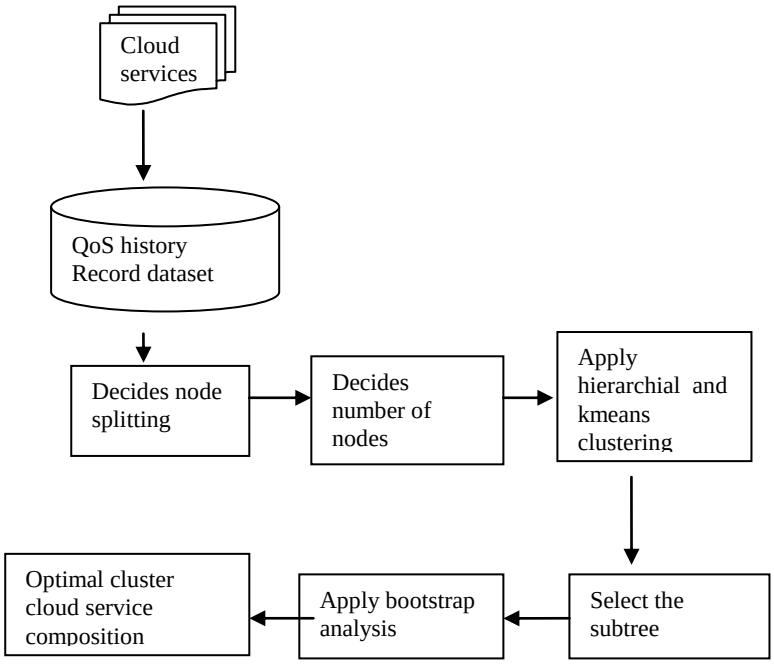

Fig. Schema of the main method

B. *Algorithm for finding optimal cluster*

Step 1. At the beginning the QoS history records of each cloud service are respectively viewed as full members of a unique huge cluster. The decision whether to split the clusters into more specific sub-clusters is taken by assessing the quality of the clusters and apply kmeans clustering.

Step 2. In the case the cluster has been evaluated worth splitting, the number of sub-clusters to generate is determined depending on cluster's properties, if the clusters selected to generate the next level of some cloud services could not be split, then the clusters will not change in the next hierarchical level.

Step 3. The Hierarchial ckustering algorithm applies a clustering algorithm having the number of sub-clusters generated in step 2 as input so as to get the soft sub-clusters.

Step 4. We use the sub-clusters' centroid QoS history records of each cloud service to generate record composition plans. And we calculate the random probability of the record composition plans.

Step 5. We accumulate the dataset of the corresponding record composition plans to the scores of each service composition plan respectively.

Step 6. Each sub-cluster has a variable to accumulate the scores of the record composition plans which the subcluster participates in. For each Web service, the sub-cluster with the highest score will be used to generate the next hierarchical level.

Step 7. The process iterates until no more clusters arecevaluated worth splitting.

Step 8. pvClust used for resampling the datasets in hierarchial clustering

It provides the Bayesian inference classification, density, uncertainity and classification. Then, it provides the optimal model for the classification

IV. CONCLUSION AND FUTURE WORK

In this, cross cloud service composition on QoS history records based on the previous one, cross cloud service composition on QoS values. Cross cloud service composition on QoS values, it calculates the QoS values each time for finding the optimized service composition is time consuming task. Using the cross cloud service composition based on history records, it searches the service present in the datasets and provided the service to the customer. Hierarchial and K-means clustering technique for clustering the history record dataset and provide the optimized service to the customer and provide accuracy for big datasets. For future work, we plan our work to investigate big data application in real cloud service system.

REFERENCES

[1] W.Dou, X.Zhang, J.Liu and J.Chen, "Hiresome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications", IEEE Trans. Parallel Distrib. Syst.,vol.26,no.2, pp.455-466, 2015.

[2] M.Shunmei, L.Zhenxing, D.Wanchun," A QoS-Aware Service Optimization Method Based on History Records and Clustering", Second International Conference on Cloud and Green Computing, 2012.

[3] X.Zhang, L.T. Yang, C.Liu and J.Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Mapreduce on cloud", IEEE Trans.Parallel Distrib.Syst., Vol.25, no.2, pp.363-373, 2014.

[4] D. Ardagna, G.Casale, M.Ciavotta, J.F Perez and W. Wang, "Quality-of-service Cloud computing: modeling techniques and their applications", Journal of Internet Services and Applications, 2014.

[5] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, ''A Privacy Leakage Upper- Bound Constraint Based Approach for Cost Effective Privacy Preserving of Intermediate Datasets in Cloud", IEEE Trans. Parallel Distrib. Syst.,vol.24,no.6, pp.1192-1202, 2013.

[6] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, ''Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption", IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 1, pp. 131-143, 2013.

[7] J. O'Sullivan, D. Edmond, and A.t. Hofstede, "What's in a Service?" Distributed and Parallel Databases, Vol.12, No.2-3, pp.117-133, 2002.

[8] H.Y. Lin and W.G. Tzeng, ''A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding,''IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 6, pp.995-1003, 2012.

[9] W. Lin, W. Dou, X. Luo, and J. Chen, ''A History Record-Based Service Optimization Method for QoS-Aware Service Composition'', IEEE ICWS, pp. 666-673, 2011.

[10] A. Klein, F. Ishikawa, and S. Honiden, ''Towards Network Aware Service Composition in the Cloud'', Int'l Conf., pp. 959-968, 2012.

[11] L. Zeng, B. Benatallah, A.H.H. Ngu, M. Dumas, J. Kalagnanam and H. Chang, ''QoS-Aware Middleware for Web Services Composition'', IEEE Trans. Softw. Eng., vol. 30, no. 5, pp. 311-327, 2010.

[12] V. Nallur and R. Bahsoon, ''A Decentralized Self-Adaptation Mechanism for Service-Based Applications in the Cloud'', IEEE Trans. Softw. Eng., vol. 39, no. 5, pp. 591-612, 2013.

[13] C. Ye, S.C. Cheung, and W.K. Chan, ''Publishing and Composition of Atomicity-Equivalent Services for B2B Collaboration'', ICSE, pp. 351-360, 2006.

[14] B. Benatallah, M. Dumas, Q.Z. Sheng, and A.H.H. Ngu, ''Declarative Composition and Peer-To-Peer Provisioning of Dynamic Web Services'', Int'l Conf. Data Eng., pp. 297-308, 2002.

[15] M. Alrifai, D. Skoutas, and T. Risse, ''Selecting Skyline Services for QoS-Based Web Service Composition'', Int'l Conf. WWW, pp. 11-20, 2010.

[16] Qi, Y., Bouguettaya, "Computing Service Skyline from Uncertain QoWS," IEEE Transactions on Services Computing , Vol.3, No. 3, pp.16-29, 2010.

[17] G. Bordogna, G. Pasi. "A quality driven Hierarchical Data Divisive Soft Clustering for information retrieval," Knowledge- Based Systems, Vol. 26, No. 1, pp. 9-19, 2012.

[18] F. Barbon, P. Traverso, M. Pistore, and M. Trainotti, ''Run-Time Monitoring of Instances and Classes of Web Service Compositions'', ICWS, pp. 63-71, 2006.

[19] S. Hossein Siadat, A. Mello Ferreira, T. Talaei Khoei and A. Reza Ghapanchi, "Performance Analysis of QoS-Based Web Service Selection Through Integer Programming", World Applied Sciences Journal, 463-472, 2013.

[20] J. Octavio, Gutierrez, Garcia, Kwang Mong Sim ,"Agent-based Cloud service composition" , Springer , vol. 10489, pp.012-0380, 2012.

[21] Ibrahim Abaker, Targio Hashem , Ibar Yagoob, "The rise of "big data" on cloud computing: Review and open research issues" , vol.31,no.7, pp.200-14, 2014.

[22] V. Narasimha Inukollu , S. Arsi and S. Rao Ravuri, "Security issues associated with Big data in cloud computing " , International Journal of Network Security & Its Applications), Vol.6, No.3, 2014.

[23] N. Ani Brown Mary and K. Jayapriya , "An Extensive Survey on QoS in Cloud computing" , International Journal of Computer Science and Information Technologies, Vol. 5 , pp. 1-5, 2014.

[24] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam, "Implementation of K-Means Clustering in Cloud Computing Environment", Research Journal of Applied Sciences, Engineering and Technology, vol. no 10, pp. 1391-1394, 2012.

S.BAVITHRA received B.E degree in Computer Science and Engineering from R.M.K College of Engineering and Technology, Chennai in 2014.She Currently pursuing her Master Degree in Software Engineering at University College of Engineering, Trichy.

Mrs.R.JAYAMALA, Asst. Professor under the Department of Information Technology at University College of Engineering (BIT Campus), Trichy. Her Research focuses on the cloud computing. She Published papers in 7 International Journals, 2 International Conference, and 4 National Conference