2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

An Intelligent Analysis and Prediction Model for On-demand Cloud Computing Systems

Xiuju Fu*, Xiaorong Li, Yongqing Zhu, Lipo Wang, Rick Siow Mong Goh

Abstract— **In this paper, an intelligent model for analyzing and predicting cloud computing resource utilization is proposed to enhance on-demand services in cloud computing systems. The model is with the capability to discover active users and mine the system storage utilization patterns. This model is also with learning capabilities to adapt the dynamics in the cloud computing platform by capturing changing patterns of system storage utilization, and it employs data mining means for computing the practical model to be used for prediction and providing inputs for intelligent management in the on-demand cloud computing system. We have evaluated the proposed analysis and prediction model in a cloud computing platform. High prediction accuracies of 95% and 86% have been achieved in 1-day ahead and 7-day ahead system utilization prediction, respectively.**

I. INTRODUCTION

Providing on-demand cloud system storage is important as it is one of solutions to realize scalable and elastic IT-enable capabilities in cloud computing systems [1][2][3][4]. For better managing on-demand cloud system storage utilization, the historical storage usage patterns should be evaluated and the discovered patterns may be further utilized for providing time-forward prediction on cloud system storage utilization. With advanced analysis and prediction on the utilization patterns of on-demand cloud system storage, qualified services can then be provided to allow users to store, distribute, retrieve and manage data according to projected demands. To realize such qualified services, intelligent model for system storage is becoming an increasingly important part of application architectures in cloud computing systems [5]. Furthermore, being able to add storage capacity quickly and deal with resource failures [6] is a key architecture theme when face to big demands from increasing users.

As on-demand storage capacity affects significantly the performance of the whole cloud computing system, which should be therefore planned in advance based on the accurate prediction on utilization. On the other hand, the realization of on-demand storage planning requires real-time monitoring and analysis of storage usage in the cloud computing system. In order to react to capacity or performance problems, real-time monitoring and analysis is needed in place. This monitoring and analysis function should be tied into key operation processes and functional areas, and needs to be fast, convenient, adaptive and accurate for providing good quality of services in an on-demand cloud computing system. All of these requests from cloud systems form the solid foundation to set up an intelligent model for analyzing and predicting cloud computing resource utilization [7][8][9][10].

Data mining approaches had been applied widely for discovering patterns from data sets in various domains including finance, manufacturing and many other industries [11][12][13]. The storage usage data of cloud computing systems is time-series representing the temporal variation of system utilization. There are many approaches dealing with time-series data [14][15][16][17][18]. In [18], data mining had been used to mine the usage patterns in cloud system to better tune the cloud storage resources and improve storage efficiency. In this study, we analyzed active users in our cloud computing system and their usage patterns are considered for predicting the system storage demand to meet the need to provide on-demand cloud storage for longer-term planning in large-scale computer systems.

In this paper, a predictive analytics model based on data mining approaches is proposed to provide intelligent analysis and prediction for cloud system storage on top of the cloud computing platform. The paper is organized as follows. In Section 2, we describe the usage data collected from the cloud computing platform. In Section 3, the predictive analytics model is introduced. In Section 4, the experimental results for predicting time-forward cloud system storage utilization are shown. In Section 5, we conclude the paper.

II. CLOUD COMPUTING SYSTEM STORAGE DATA COLLECTION

For supporting on-demand storage services for High Performance Computing (HPC) applications, research had been conducted to provide an on-demand storage provisioning solution to the A*STAR Digital Nervous System (ADNS) [19], which was designed for realizing advanced computing and distributed applications. A platform has been deployed on top of the cloud computing infrastructure and various experimental tests were conducted to evaluate the data accessing speed via such dynamic storage services in ADNS environment. We have collected storage usage from Axle cluster in A*STAR Computational Resource Centre (ACRC) [19]. Fig. 1 shows a layout of the components to facilitate the collection of storage data on Axle. The NFS server can access "/scratch" folder that is local across all computing nodes. The quota mechanism has been enabled on

Xiuju FU*(corresponding author), Xiaorong Li, and Rick S.M. Goh are with Institute of High Performance Computing, Singapore (email: fuxj@ihpc.a-star.edu.sg;xiaorong.li.2010@gmail.com; gohsm@ihpc.a-star.edu.sg)

Yongqing Zhu is with Data Storage Institute, Singapore, (e-mail: zhu_yongqing@dsi.a-star.edu.sg).

Lipo Wang is with Nanyang Technological University, Singapore (email:elpwang@ntu.edu.sg)

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

"/scratch" folder, and a simple command 'repquota' can get information for each individual user. A script was written using 'df', 'repquota' and 'top' commands to collect storage-related data. When running the script on the Dell server (Intel Xeon CPU 3.00GHz, Memory 1G), the CPU consumption is 0% and memory consumption is around 1MB only. With little overhead caused, such scripts were run on the NFS Server to get storage usage data for "/scratch" folder in Axle.

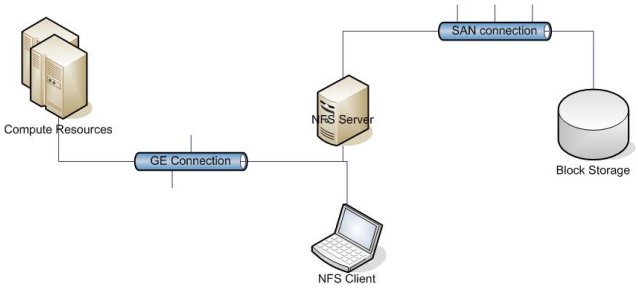

Fig. 1. Layout of the components to facilitate the collection of storage data on Axle

The AXLE computer system storage data including system level storage utilization and individual-level user storage utilization had been collected and studied. Data had been collected from 1st April of 2009 to 30th Nov 2009. The first set of data was the system level system storage utilization with four attributes including date, hour, system total storage usage, and storage utilization. The time unit to record the storage usage was half an hour. The example data had been shown in Table 1.

The second set of data is the user-level storage which includes five attributes, i.e., date, hour, user ID, user block size (KB), and the number of user files. The time unit for user-level data is also half an hour. The example user-level data were shown in Table 2.

Table 1. System-level storage usage

Date	Hour	System Total Storage Usage	Storage Utilization (%)
0801	00.00.01	4859348440	70%
0801	00.30.01	4864257352	70%
0801	01.00.01	4869259080	70%

Table 2. User-level storage usage data

Date	Hour	User ID	Used Block Size (KB)	Used Files
0801	00.00.01	3434567183	1664692	17939
0530	22.00.01	1829321659	12385752	3131
0531	21.30.01	1861565962	197106904	154187

III. The Cloud Computing Predictive Analytics Model

Instead of providing one-time storage of hard disks in computer system, it is desirable to plan on-demand storage capacity, which allows flexible services to diversified users and allows buying and managing the system hardware with lower costs [1]. Therefore, it's important to think about who will access the system actively and what regulatory requirements are involved. Especially, for longer-term and large-scale computer systems, an accurate prediction model is important for predicting the system storage demand based on the historical user usage patterns. The whole on-demand cloud computing system based on the predictive analytics model is shown in Fig. 2. It is composed of a cloud computing predictive analytics model and a cloud computing platform. The cloud computing predictive analytics model can be considered as a sub-system, which takes inputs from the other subsystem, i.e., the cloud computing platform, and outputs data storage planning and the predicted cloud computing system utilization for facilitating storage and other computing resource planning. The on-demand cloud computing system can thus meet the requirements of on-demand planning in the cloud computing applications.

A. The Cloud Computing Predictive Analytics Model

The model proposed in this study is comprised of data collection component (1), data analysis component (2), machine learning and prediction component (3), data storage planning component (4), pattern checking component (5), and cloud system utilization prediction component (6).

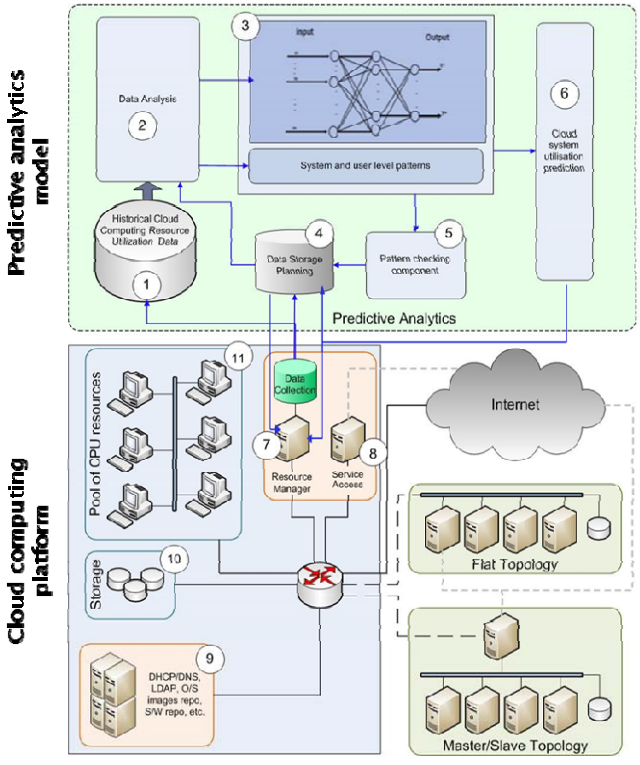

Fig. 2. The on-demand cloud computing system

The cloud computing system utilization data will be regularly added (per week or per month, etc.) in historical data collection component. Along with each update of historical

data, the new set of data is sent to the data analysis component. The analysis component employs a regression model for estimating the main trend of system utilization and detecting the active users. The regression model is shown as Eq.1:

$$U_s(t) = \sum_{i=1}^{D} \beta_i N(t-i) + \varepsilon \quad (1)$$

It is designed for estimating the main trend of system utilization. At time t, the main trend of the cloud computing system utilization U_s is estimated with only the number of users of past D days as inputs.

$N(t-i)$ is the number of users on i days ago. The predicted U_s is compared with the cloud system utilization recorded in real time. Besides the regular operation, if the difference of predicted value and actual value exceeds the tolerance value, the whole analysis and prediction system will be triggered to run immediately.

In a cloud computing system, there may exist both inactive and active users. The active users are those who actively use the computing resources and may request more storage gradually. For detecting the active users, each user's usage of the system is constantly monitored. For a user j, the standard deviation of his/her historical storage utilization S_j is calculated regularly as follows:

$$S_j = \sqrt{\frac{1}{N-1}\sum_{k=1}^{M}(x_k - \bar{x})^2} \quad (2)$$

Here M is the number of system days. x_k is the user storage usage on the k - th day. $\bar{x}$ is the average user storage usage.

A user j is identified as an active user if :

$$S_j \geq S_T$$

The default value $S_T = S(Q_3)$, i.e., S_T is set as the third quartile (Q_3) of the standard deviations of existing users. It may be tuned based on the practical need in different cloud systems.

The temporal data storage utilization of active users, system level storage utilization and the varied number of registered users are used as inputs to the machine learning and prediction component. The cloud system utilization prediction component is comprised of a linear regression model and a non-linear neural network model. Historical utilization data include system level utilization data $S_y(t)$, active user utilization data $A(t)$, and the number of users in the system $N(t)$. The utilization prediction $U_R(t)$ from the regression model is as follows:

$$U_R(t) = \sum_{i=1}^{D}\beta^S S_y(t-i) + \sum_{i=1}^{D}\beta^A A(t-i) + \sum_{i=1}^{D}\beta^N N(t-i) + \varepsilon \quad (3)$$

Past D days' data are used for prediction the utilization in the present time.

The utilization prediction from the multi-layer perceptron (MLP) neural network model is $U_N(t)$. The 2-layer MPL model is shown in Eq. 4:

$$U_N(t) = f\left(\sum_{i=1}^{K} W_i^{(2)} * \phi_i(X) + b^{(2)}\right) \quad (4)$$

Here we have $\phi_i(X) = f\left(W_i^{(1)} \cdot X + b_i^{(1)}\right)$, where $W_i^{(1)}$ is the weight vector connecting the input vector with hidden neuron *i*, and $b_i^{(1)}$ is the bias of hidden neuron *i*. and $X = (S_y, A, N)^T$.

The linear relationship between historical data and the time forward system utilization is mainly captured by the linear regression model. Moreover, the nonlinear relationship is mapped by the multiple-layer neural network model. The prediction $U(t)$ from the machine learning and prediction component is :

$$U(t) = w * U_R(t) + (1-w) * U_N(t) \quad (5)$$

where w is determined through training.

The data storage planning component receives the updates about system level and user level patterns and the utilization prediction from the data analysis component, pattern checking component and system utilization prediction component. The ratio between active users and the total number of users is calculated in this component and compared with a predefined threshold. The predicted storage value is also compared to the planned storage. When the ratio value is above the threshold or the prediction utilization value exceeds the planned storage, the pattern checking component will send warning message to the system administrator for reminding the re-evaluation of the planning in the whole on-demand cloud computing system.

B. The On-demand Cloud Computing System

The cloud computing platform provides services where users can request computational resources, e.g., network storage (10), computer nodes (11) and software tools with different network topologies via the Internet. It is a utility-based service model where large-scale of resources can serve a variety of users who have different demands. Service access point (8) provides web portal interfaces where users make authentication via LDAP server (9) and subscribe the cloud

computing resources. Resource manager (7) allocates the resources and monitors the utilization of the resources in a network log depository. Cloud system utilization prediction component (6) works with the resource manager (7) to analyze and predict resource utilization in the cloud. The historical resource utilization data are collected and stored in data collection component as the inputs to the data analysis component. Statistical and data mining methods in the data analysis component are used to extract system level, user level utilization pattern and identify active users as described in earlier section.

The active user information with the present and historical system storage utilization pattern are sent to machine learning and prediction component for predicting time-forward cloud system utilization.

The system planning component (4) receives prediction results from component (3) and subsequently make adjustment to the existing planning of the cloud computing system, i.e., the demand of users will be met promptly without delay and without sacrificing the service qualities. Besides the regular operations, the machine learning and prediction component (3) can be automatically activated to update utilization prediction when the pattern checking component (5) detects over-active users or abnormal patterns.

IV. RESULTS

A. Data Analysis for Investigating System-level and User-level Usage Patterns

The system storage usage was recorded from 1 April to 30 Nov 2009 shown in Fig.3. It is observed that the system storage usage increases from the initial usage of 4% of the total system storage in April to 99.7% of the total system storage in November. It is also observed that the system storage usage had shown several rebounding patterns, which may result from the data cleaning in the scratch folder by administrators or the data cleaning implemented by individual users.

Before planning the system storage demand in a distributed computer system, it is desirable to capture the key features which affect the variation of system storage. The key features will then be used as inputs to the data mining prediction model for predicting and planning system storage.

Firstly, the number of users is extracted in daily basis. It is observed that the number of users is increasing steadily. Fig. 4 (a) shows the linear regression in which the dependent variable is the system storage usage and the independent variable is the number of users. It shows that the number of users can capture the main trend of the system storage. The correlation coefficient is calculated between system storage usage and the number of users. We had obtained that R^2 =0.94 at p-value <0.001. Then the users are evaluated and active users are identified by considering their influence on the system storage utilization. Fig. 4 (b) shows that the usage of active users can better capture the system storage usage fluctuation. The correlation coefficient is calculated between system storage usage and the usage of active users. We had obtained that R^2 =0.994 at p-value <0.001.

Therefore, the number of users, the storage usage of active users and system level storage usage are used together to predict time ahead system storage usage and we design the data mining model using these data as inputs to our MLP neural network predictor for predicting time ahead system storage utilization.

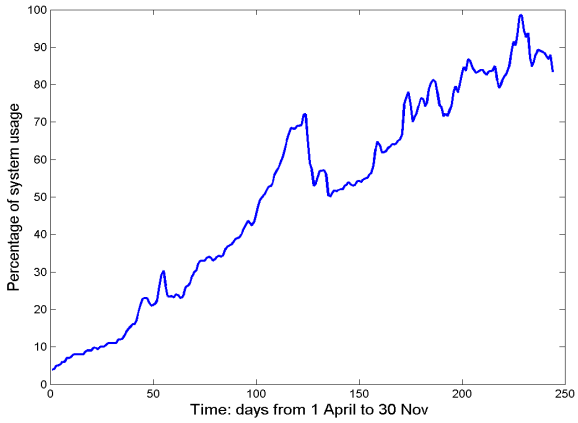

Fig. 3

B. Prediction Results for System Storage Usage

Equation (6) shows that the MLP neural network is used for predicting one-day ahead system storage $y(X,t)$.

$$U(X,t) = f(X, t - t_0) \quad (6)$$

$f(X)$ is the MLP neural network as shown in Eq.4. X is the inputs to MLP neural network.

The time-ahead prediction results of system storage usage are shown in Fig. 5 (a) and (b). The neural network predictor can adapt to predict the system usage at longer forward time, i.e., given the present and historical time computer system storage usage, the neural network model can be trained to learn longer forward time prediction patterns. The prediction results on 1 to 7 day-ahead prediction are shown in Table 3. It is observed that for longer forward time prediction, the lowest prediction accuracy is 86.43%. The prediction accuracy is with a decreasing trend from 1 to 7-day-ahead prediction.

Table 3. 1 to 7-day-ahead prediction results
(Tr set: training set; Te set: Testing set)

	Accuracy (day-ahead prediction)						
Data	1	2	3	4	5	6	7
Tr set (%)	96.9	94.37	92.31	93.63	92.97	93.0	89.76
Te set (%)	95.5	94.35	89.58	92.92	92.4	93.0	86.43

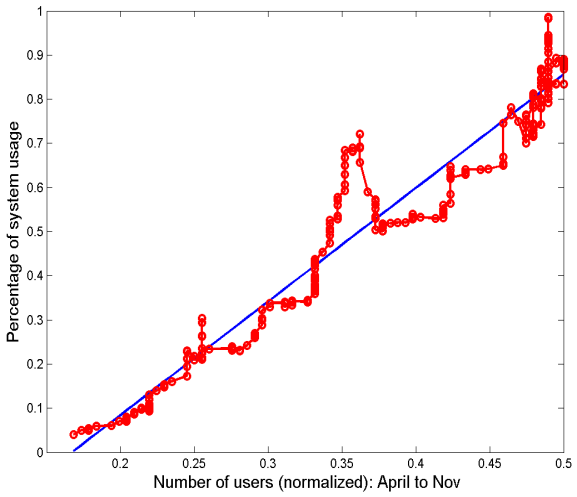

Fig. 4 (a)

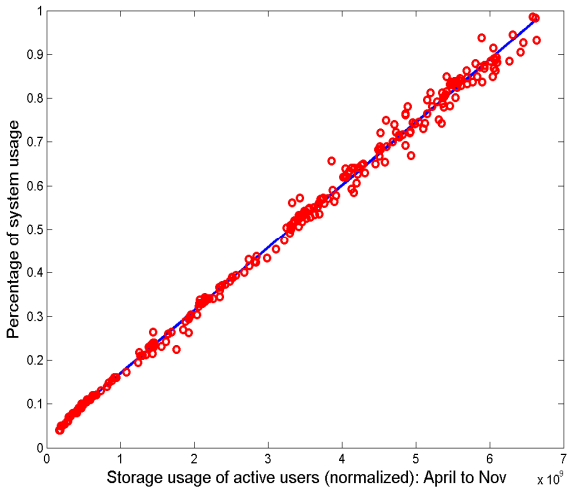

Fig. 4 (b)

Fig. 4 (a): blue line shows the regression results on the system storage usage with the number of users as the independent variable. Red line shows the actual system storage usage.

Fig. 4 (b):blue line shows the regression results on the system storage usage with the usage of active users as the independent variable and Red line shows the actual system storage usage

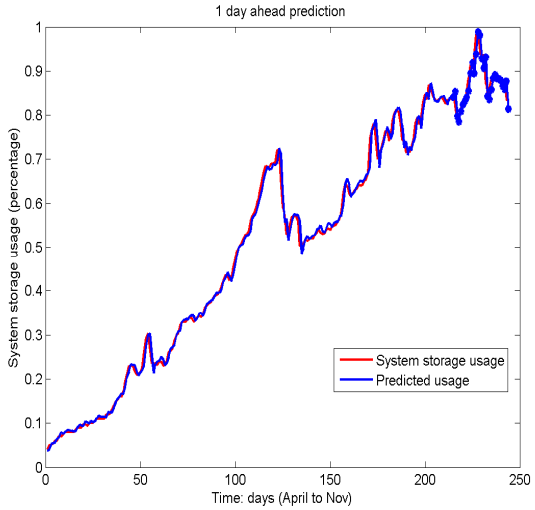

Fig. 5 (a)

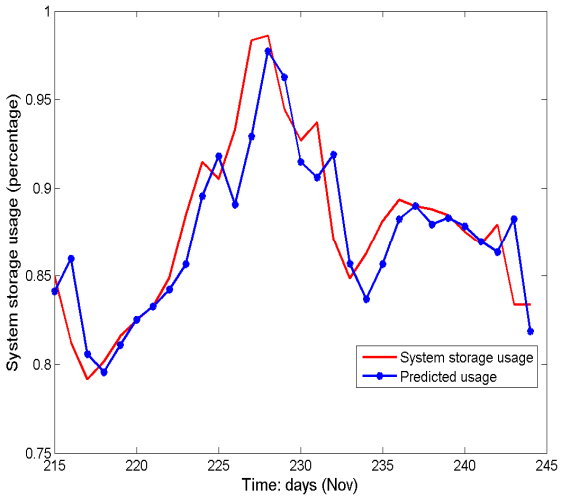

Fig. 5 (b)

Fig. 5 (a) shows the prediction over the whole period using April to Oct as training data and Nov as testing data. Fig. 5 (b) shows the 1-day ahead prediction on Nov usage using April to Oct as training data and Nov as testing data. For training set, prediction accuracy 96.94% is obtained, and high testing accuracy at 95.5% is obtained for predicting unseen Nov system storage usage.

V. CONCLUSION

In this study, a predictive analytics model is proposed to predict time-ahead system storage usage for an on-demand cloud computing system. The model is with three main functions including storage usage data preprocessing, computer system storage usage pattern extraction and the neural-network predictor. The system storage data are preprocessed to generate data ready for further data analysis. Active users and system usage patterns are then extracted to help capture the important variation patterns in computer system storage usage. Prediction is implemented for predicting system storage usage up to 7 days ahead. For 1-day-ahead prediction, high prediction at 95.55% accuracy is obtained and for 7-day-ahead prediction, 86.43% accuracy is achieved. Besides the general decreasing accuracy with longer time-ahead prediction, we also observed small fluctuation in prediction accuracy for the time forward prediction, which may be mainly caused due to the variation of user usage patterns and system behavior. This will be further study in our future research.

We highlighted that more data collection is expected for better analyzing the system utilization, user usage patterns and improving the prediction accuracy. The system features to be included for the analysis include system memory usage, CPU usage and job distributions of users, which may be important data for identifying system patterns and assisting the smart management of an on-demand cloud computing system. Furthermore, users can be categorized into different groups for the system manager to target at users with different priority and improve service quality as well.

ACKNOWLEDGMENT

Authors thank for the support of A*STAR computational resource center on data collection.

REFERENCES

[1] J. Timmermans, V. Ikonen, B. C. Stahl, and E. Bozdag, "The Ethics of Cloud Computing: A Conceptual Review," in *2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*, 2010, pp. 614–620.

[2] X. Li, H. Palit, Y. S. Foo, and T. Hung, "Building an HPC-as-a-Service Toolkit for User-Interactive HPC Services in the Cloud," in *2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA)*, 2011, pp. 369–374.

[3] "AWS | Amazon Elastic Compute Cloud (EC2) – Scalable Cloud Servers." [Online]. Available: http://aws.amazon.com/ec2/. [Accessed: 20-Jan-2014].

[4] B. C. Tak, B. Urgaonkar, and A. Sivasubramaniam, "Cloudy with a Chance of Cost Savings," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, pp. 1223–1233, 2013.

[5] R. Van den Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-Efficient Scheduling Heuristics for Deadline Constrained Workloads on Hybrid Clouds," in *2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*, 2011, pp. 320–327.

[6] M. Lu and H. Yu, "A Fault Tolerant Strategy in Hybrid Cloud Based on QPN Performance Model," in *2013 International Conference on Information Science and Applications (ICISA)*, 2013, pp. 1–7.

[7] C. Wang, J. Chen, B. B. Zhou, and A. Y. Zomaya, "Just Satisfactory Resource Provisioning for Parallel Applications in the Cloud," in *2012 IEEE Eighth World Congress on Services (SERVICES)*, 2012, pp. 285–292.

[8] S. Chaisiri, B.-S. Lee, and D. Niyato, "Robust cloud resource provisioning for cloud computing environments," in *2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*, 2010, pp. 1–8.

[9] S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimization of Resource Provisioning Cost in Cloud Computing," *IEEE Trans. Serv. Comput.*, vol. 5, no. 2, pp. 164–177, 2012.

[10] T. N. B. Duong, X. Li, and R. S. M. Goh, "A Framework for Dynamic Resource Provisioning and Adaptation in IaaS Clouds," in *2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*, 2011, pp. 312–319.

[11] S. Wang, "A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research," in *2010 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2010, vol. 1, pp. 50–53.

[12] S. Asghar and K. Iqbal, "Automated Data Mining Techniques: A Critical Literature Review," in *International Conference on Information Management and Engineering, 2009. ICIME '09*, 2009, pp. 75–79.

[13] X. Wu, "Data mining: artificial intelligence in data analysis," in *IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004. (IAT 2004). Proceedings*, 2004, p. 7–.

[14] M. Zhu and L. Wang, "Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–5.

[15] S. Gupta, and L. P. Wang. "Stock Forecasting with Feedforward Neural Networks and Gradual Data Sub-Sampling." *Australian Journal of Intelligent Information Processing Systems* 11.4 (2010).

[16] L. Wang and S. Gupta, "Neural Networks and Wavelet De-Noising for Stock Trading and Prediction," in *Time Series Analysis, Modeling and Applications*, W. Pedrycz and S.-M. Chen, Eds. Springer Berlin Heidelberg, 2013, pp. 229–247.

[17] G.Q. Dong, Fataliyev K., and L.P. Wang. "One-Step and Multi-Step Ahead Stock Prediction Using Backpropagation Neural Networks." the 9th International Conference on Information, Communications and Signal Processing (ICICS 2013). 2013.

[18] J. Deng, J. Hu, A. C. M. Liu, and J. Wu, "Research and Application of Cloud Storage," in *2010 2nd International Workshop on Intelligent Systems and Applications (ISA)*, 2010, pp. 1–5.

[19] "A*STAR Computational Resource Centre." [Online]. Available: http://www.acrc.a-star.edu.sg/. [Accessed: 20-Jan-2014].