ANOMALY DETECTION IN THE SERVICES PROVIDED BY MULTI CLOUD ARCHITECTURES: A SURVEY

Mahendra Kumar Ahirwar[1], Manish Kumar Ahirwar[2], Uday Chourasia[3]

[1]PG Scholar DoCSE, UIT-RGPV, Bhopal (Madhya Pradesh), India
[2]Asst. Prof. DoCSE, UIT-RGPV, Bhopal (Madhya Pradesh), India
[3]Asst. Prof. DoCSE, UIT-RGPV, Bhopal (Madhya Pradesh), India

Abstract

An Anomaly is abnormal activity or deviation from the normal behavior. Anomaly detection is the process of removing these abnormal or anomalous behaviors from the data or services. The services delivered to users by cloud service providers must have normal behavior. To provide services to users in the proper and normal form, anomaly detection becomes important and interested area for research work. For anomaly detection so many techniques are developed and these techniques are broadly divided into three categories: - statistical, data mining based and machine learning based anomaly detection technique. Anomaly detection techniques are used to detect and discard anomalies from the data or services. In this survey paper we provide overview of some anomaly detection techniques which are discovered recently for the tracing data. In the anomaly detection models anomalies are detected by comparing the tracing data with the actual data. On the basis of comparison deviations in the traced data or services are identified and they are considered as anomaly.

Keywords:- *Anomaly, services, tracing data, anomaly detection techniques and cloud service provider.*

----------***----------

1. INTRODUCTION

Cloud computing is an Internet-based most recent popular technology offering dynamic resources, scalable resources, on-demand, self-service and pay-per-use. Cloud computing is an active area for research and growing very fast. It provides services at low cost and low operational software and hardware expenditure's. The use of cloud computing has increased in companies rapidly because of fast access to applications and decreasing maintenance cost for cloud infrastructure.

1.1 Types of Clouds

Clouds can be classified into four categories on the basis of physical location of users. Four types of cloud are private, public, community and hybrid clouds. In the available types clouds explain benefits and limitations of each cloud types on the basis of which we can conclude that which cloud model will be suitable for us. A *private* cloud is one which is setup by single organization and installed services on its own data center. *Public* cloud services are offered by third-party cloud service providers and involve resource provisioning outside of the user's premises. The *Community* cloud can offer services to the cluster of organizations. In other words we can say that community cloud provides combinational services of a group of clouds. *Hybrid* cloud is the combination of any two or more than two types of clouds which are mentioned above means combine any two or more from private, public or community to build it.

1.2 Cloud Service Models

Cloud computing technology allows users to access information and computing resources from anywhere that a network connection is available. It provides a shared pool of services and resources including data centers (data storage space), networks (Internet), computer processing power and user applications. Web server provides services from shared pool according to 3-tier architecture. Cloud services can be divided into three types which are explained as:-

Infrastructure-as-a-Service (IaaS):- It is used to provide network for connecting users and servers and also provides virtual machines to start, stop, access and configure virtual servers and storage blocks.

Platform-as-a-Service (PaaS):- In this model a platform is provided to users which typically include operating system, programming languages, execution environments, databases, queues and web servers.

Software-as-a-Service (SaaS):- This model provides "On-demand software's" to users without installation setup and running of the applications.

1.3 Towards Multi-clouds

The terms multi-clouds or inter-clouds or cloud-of-clouds are used to refers the new concept in the field of cloud computing in which small clouds are combined to build a big cloud which provides combined services of participating clouds. Multi-clouds [7] has some benefits over single cloud such as scalability, more security, combined services provisioning but it also suffers from some drawbacks in performance diagnosis of fine-granularity, unsupervised, scalable and high efficient

multi clouds. These drawbacks can be diagnosing using CloudDiag tool which can detect and remove anomalies if available in the multi-clouds.

1.4 Anomaly Detection in Multi-clouds

There are three types of anomaly detection techniques which are available for cloud computing. These anomaly detection techniques are: - Statistical, data mining and machine learning.

1.4.1 Statistical Anomaly Detection

In this technique of anomaly detection, to identify anomalies system observes computations and generate a profile which stores a value to represent their behavior. For anomaly identification this technique used two profiles in which one is stored ideal profile and other is current profile which is updated periodically and calculates anomaly score. If anomaly score of current profile is higher than Threshold value of stored profile than it is considered as anomaly and then it can be detected.

1.4.2 Data Mining Based Anomaly Detection

Anomalies can also detected using data mining techniques like classification, clustering and association rule mining. To identify anomalies this technique added level of focus to anomaly detection. Data mining techniques used an analyst which can differentiate normal and abnormal activity within clouds by defining some boundaries for valid activities in the clouds.

1.4.3 Machine Learning Based Anomaly Detection

This approach of anomaly detection uses the concept of machine learning to identify anomaly. The ability of programs or softwares to learn and improve performance of the task or group of tasks over time is called machine learning. This technique develops a system which can improve performance of the programs on the basis of previous results. From the previous results new information is acquired and on the basis of this information even execution strategy can be changed for performance improvement if required.

Statistical anomaly detection technique is beneficial as compared to other two techniques because this technique have number of benefits over others. Firstly this technique does not require any prior knowledge domain of security risks or intrusion. Secondly this technique has capability to detect even very recent anomalies generated in the data. This also provides accurate notification for anomalies that occurs over extended time period.

The remaining parts of this paper are organized as follows:- Section 2 describes related work about various anomaly detection techniques used recently. In section 3 we conclude our paper and also provide direction for future enhancement.

2. RELATED WORK

This paper is designed for brief discussion about various anomaly detection techniques used in multi cloud architectures [11]. In these techniques for anomaly detection monitoring of tracing data is performed and then detection and diagnosis of anomalies is done. This paper is surveyed as:-

Pinpoint [5] can be used to traces the request call relationship data of service components and apply clustering data mining algorithm to classify data entries into failure or success group. From this entries classification we can identify anomalies i.e. entries available in the failure group. Pinpoint is dynamic analysis methodology framework which can determine anomaly from the traced data. Pinpoint framework is implemented on the top of J2EE platform and in this case prior knowledge of anomaly components is not required. Pinpoint is three layer framework in which first layer is communication layer that is used for tracing client requests. Second layer is failure detection which is used for internal and external system monitoring and then determines whether requests of clients are succeeded or failed. Third layer of Pinpoint framework is data clustering layer which combines tracing data of client requests and success or failure data of each request. The combined data is feed into data analysis engine which find out faulty components and their interactions.

Anomalies can be mined using feature distributions for which entropies of data are computed and stored in BigTable and use it for automatic classifying data anomalies through unsupervised learning. LERAD [8] used different approaches to tackle different types of data entries. A framework is discovered that uses both normal and anomalous data to find out characteristic features of anomalies on the basis of which anomalies can be removed.

Pip [6] model is used to compares the actual behavior of tracing data with self-defined expected data to check whether a request of user is anomalous or anomaly free. But designing of such a model is very difficult because this approach required vast amount of specific domain knowledge. Pip model is an architecture designed for distributed systems to detect structural anomalies and performance problems based of comparison between actual behavior and expected behavior of tracing data. Actual system behavior of Pip model expressed by annotation tools and system instrumentation and expected behavior by visualization and query tools. Using above tools pip allows a developer to debug and understand familiar and unfamiliar systems immediately.

Anomaly can be detected in case of categorical datasets [10]. As in each application an anomaly is abnormal data point or deviation from the normal activity. Normally anomalies are detected by creating a model for normal data and then compare current services record value to normal data model. [10] Proposed probabilistic approach for anomaly detection in which a likelihood model is builds from the training data. In this detection technique for categorical datasets Bayesian network is used which provides likelihood representation and can also detect those anomalies which are rarely appears. This approach used marginal distribution for comparison to find out

anomalies. This approach is applicable in unsupervised anomaly detection problems where unlabelled data for training is provided. In case of unlabelled data it is very difficult to find out those records which are not following normality of datasets.

Dapper [2] introduces an infrastructure for monitoring of performance of services. It stores tracing data into Bigtable [3]. This approach is unable to describe how diagnosis performed for anomalies. By monitoring tracing data we can find primary causes of performance changes between two time intervals. P-Tracer [4] can be able to identify anomalies available in the call trees and once the anomalies are detected these can be removed from the data.

In the utility clouds like Amazon EC2 online anomaly detection [11] is performed for data centers. In cloud computing size of data centers is increasing due to complexities of applications, system softwares and workload patterns. In these systems automatic anomaly detection must be operated without prior knowledge of anomalous or normal behavior of data. Chengwei Wang et al. proposed EbAT (Entropy based Anomaly Testing) model [11] which is a novel anomaly detection model. EbAT can detect anomalies by analyzing arbitrary metrics distributions instead of individual values of metrics entries. In EbAT entropy is used as a measurement parameter to get information about degree of concentration and dispersal of distributed and aggregated data metrics in the clouds. These entropy time series can be combined hierarchically and also across multiple cloud systems.

Along with the technological improvements new anomalies and advanced intrusion are appeared unexpectedly in the datasets. It is necessary to identify and protect against these anomalies. Ensemble [12] method is proposed to identify and remove such anomalies in semi-supervised mode using supervised learning algorithm. Ensemble of feature chains is improved version of Cross-Feature Analysis (CFA) and classifier chaining method. In the learning algorithms, some models underestimate the probability value and others overestimate the probability value, ensemble of feature chain model is convenient option to solve this problem. In ensemble of feature chain models, each model is learned from different chains of randomly ordered features. This method is capable to handle numeric as well as nominal featured data and it is suitable for mobile devices.

In cloud computing infrastructure, to detect anomalies data monitoring is required. Husanbir et al. [13] proposed AAD (Adaptive Anomaly Detection) framework to detect anomalies. In cloud computing due to large scale, inherent complexity and production cloud computing various runtime problems in hardware and software may be occurred. ADD [13] perform autonomic anomaly detection by monitoring cloud execution and collection of runtime performance data. The monitored data is usually unlabelled data and records about previous failures are also not available in this case ADD can detect possible failure or anomalies on the basis of monitoring data and then these anomalies are verified by cloud operators. Cloud operators confirmed these failures with its type or prove it as normal data. ADD framework has recursive learning capability to adapt itself for future anomaly detection based on verification of results.

In the cloud computing Monitoring-as-a-Service [14] framework is discovered to detect anomalies. The existing security tools of internet are not suitable for cloud computing because existing tools does not consider specific intrusions and challenges of clouds such as cross virtual machine side channel attacks and these tools are focused at only one layer of architecture. Matthias et al. proposed context based anomaly detection framework, Monitoring-as-a-Service which can tackle cloud challenges like cross-VM side channel attacks and multi-tenancy. This framework also focuses on all layers of workflow and infrastructure. In this framework machine learning and complex event processing rules are used to detect anomalies in the monitored data.

Thermal Anomaly-aware Resource Allocation (TARA) model [9] is discovered to create time varying fingerprints of the datacenters to minimize the latency of the thermal anomalies such as coldspots, hotspots and fugues and maximize the accuracy of the datacenters. TARA improves the performance of the model based anomaly detection as compared to traditional resource allocation schemes.

CloudDiag [1] can diagnosis the anomalies which are appeared in case of fine granularity, unsupervised and scalable cloud systems. CloudDiag can perform, performance diagnosis in three steps i.e. collection of data, assembling of data and diagnosis of anomalies. Diagnosis of anomalies is to find out anomalies in the tracing data and if present then remove these anomalies. Anomalies diagnosis process is done as identify anomalous categories, identify anomalous methods and locate physical node where anomalies are found. CloudDiag can also have scalability property to identify anomalies available in any other neighbor cloud of distributed system.

3. COMPARISON STUDY

In the previous section we provide various techniques and frameworks for anomaly detection. Now we can also compare merits and demerits of these techniques so that we can easily make decision regarding their applicability in different circumstances. Here we design a comparison table which provides information about anomaly detection technique, model (or framework) and its advantages and disadvantages. Table 1 shows comparison study of various anomaly detection techniques which are mentioned in the related work.

Table- 1 Comparison table of various anomaly detection techniques

S.No.	Framework/ Model	Anomaly Detection Technique	Advantages	Limitations
1	Pinpoint	Clustering data mining technique	Suitable for large and dynamic systems where it is difficult to monitor application-level knowledge of services	Performance degradation of services and issue of scalability
2	LERAD	Unsupervised learning technique	It can detect stimulated and real attacks	Unable to differentiate between true and false alarms
3	Pip	Imperative and declarative statistical tehniques	Monitoring and checking dynamic properties of programs like latency, throughput, concurrency and node failure	Require vast amount of specific domain knowledge
4	Probabilistic likelihood	Bayesian Network Technique	Can detect anomalies in case of categorical datasets	Difficult to define constraints for groups and not suitable for real valued attributes
5	P-Tracer	Supervised learning technique	Simple and easy to implement for anomaly detection	Detect only primary causes of anomalies
6	EbAT	Statistical techinique	Detect anomalies from whole metrics simultaneously instead of individual value of metrics	Neither focus on possible cases of anomalies nor evaluate scalability, cross-stack metrics and hadoop
7	Ensemble of feature chains	Supervised learning technique	Handle numeric and nominal featured data and suitable for mobile devices	It suffers from high false positive rate
8	ADD	Recursive learning technique	Tackle unlabelled data	Unable detect failure for which cloud operators are used
9	Monitoring-as-a-Service	Machine learning and event processing rules	Tackle cross-VM side channel attack and multi-tanency in clouds	Complex computation process for anomaly detection
10	TARA	Scheduling algorithms	Minimize latency of thermal anomalies and maximize accuracy of datacenters	Difficult to detect anomalies in case of high density hotspots
11	CloudDiag	Statistical technique	Diagnose anomalies in fine-grained, scalable and unsupervised cloud systems	Required specific knowledge domain for anomaly detection

4. CONCLUSIONS & FUTURE WORK

In this paper we have discussed about various anomaly detection techniques which can be used in different conditions. The above mentioned techniques can differentiate between normal and anomalous behavior of services on the basis of comparison between them. When performance of our data or service is deviate from the normal path it is considered as anomaly. Once anomaly is detected in the data it can be removed using any suitable detection technique.

In the future we can also use black box anomaly detection technique which does not require any human intervention. We can also optimize the above mentioned techniques or even combine two or more than two techniques to generate better results.

REFERENCES

[1]. Haibo Mi, Huaimin Wang, Yangfan Zhou, Michael Rung-Tsong Lyu and Hua Cai, "Toward Fine-Grained, Unsupervised, Scalable Performance Diagnosis for Production Cloud Computing Systems."IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp 1245-1254, June-2013.

[2]. B. Sigelman, L. Barroso, M. Burrows, P. Stephenson, M. Plakal, D. Beaver, S. Jaspan, and C. Shanbhag, "Dapper, a Large-Scale Distributed Systems Tracing Infrastructure," Technical Report Dapper-2010-1, Google, 2010.

[3]. F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A Distributed Storage System for Structured Data," ACM Transaction Computer Systems, vol. 26, no. 2, pp. 1-26, 2008.

[4]. H. Mi, H. Wang, Y. Zhou, M.R. Lyu, and H. Cai, "P-tracer: Path-Base Performance Profiling in Cloud Computing Systems," proc. IEEE 36th Ann. Computer Software Applications Conference (COMPSAC), pp. 509-514, 2012.
[5]. M. Chen, E. Kiciman, E. Fratkin, A. Fox, and E. Brewer, "Pinpoint: Problem Determination in Large, Dynamic Internet Services," Proc. IEEE International Conference Dependable Systems and Networks (DSN), pp. 595-604, 2002.
[6]. P. Reynolds, C. Killian, J. Wiener, J. Mogul, M. Shah, and A. Vahdat, "Pip: Detecting the Unexpected in Distributed Systems," Proc. USENIX 3rd Symposium Networked Systems Design and Implementation (NSDI), pp. 115-128, 2006.
[7]. Jens-Matthias Bohli, Nils Gruschka, Meiko Jensen, Luigi Lo Iacono and Ninja Marnau, "Security and Privacy-Enhancing Multicloud Architectures", IEEE Transactions on Dependable and Secure Computing, Vol. 10, No. 4, pp. 212-224, July/August 2013.
[8]. M. V. Mahoney and P. K. Chan. "Learning Rules for Anomaly Detection of Hostile Network Traffic", 3rd IEEE International Conference on Data Mining, pp. 601-604,2003.
[9]. Varun Chandola, Arindam Banerjee and Vipin Kumar "Model-based Thermal Anomaly Detection in Cloud Datacenters", ACM Computing Surveys,pp. 1-72, Sept.-2009.
[10]. Kaustav Das and Jeff Schneider "Detecting Anomalous Records in Categorical Datasets", 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 220-229, 2007.
[11]. Chengwei Wang, Vanish Talwar, Karsten Schwan and Parthasarathy Ranganathan "Online Detection of Utility Cloud Anomalies using Metric Distributions", IEEE Network Operations and Management Symposium (NOMS), pp. 96-103, april 2010.
[12]. Lena Tenenboim-Chekina, Lior Rokach and Brach Shapira "Ensemble of Feature Chains for Anomaly Detection", © Springer-Verlag Berlin Heidelberg, 2011.
[13]. Husanbir S. Pannu, Jianguo Liu and Song Fu "AAD: Adaptive Anomaly Detection System for Cloud Computing Infrastructures", 31st International Symposium on Reliable Distributed Systems, pp. 396-397, 2012.
[14]. Matthias Gander, Basel Katt, Michael Felderer, Adrian Tolbaru, Ruth Breu, and Alessandro Moschitti "Anomaly Detection in the Cloud: Detecting Security Incidents via Machine Learning",©Springer Verlag Berlin Heidelberg, pp. 103-116, july 2013.
[15]. Haroon Malik,Bram Adams, Ahmed E. Hassan" Automatic Detection of Performance Deviations in the Load Testing of Large Scale Systems" IEEE 35th International Conference on Software Engineering (ICSE), pp. 1012-1021, may-2013.