Compliance-Aware Provisioning of Containers on Cloud

Mehdi Bahrami* Abhishek Malvankar† Karan K. Budhraja‡ Chinmay Kundu§ Mukesh Singhal¶ Ashish Kundu‖
*Fujitsu Laboratories of America, Sunnyvale, California, USA Email: mbahrami@us.fujitsu.com
†IBM Watson Health, Yorktown Heights, NY, USA Email: asmalvan@us.ibm.com
‡University of Maryland, Baltimore County, MD, USA Email: karanb1@umbc.edu
§Researcher, India Email: ckkundu@gmail.com
‖Cloud Lab, University of California, Merced, USA Email: msinghal@ucmerced.edu
¶IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA Email: akundu@us.ibm.com

***Abstract*—Deploying applications in containers has several advantages, such as rapid development, portability across different machines, and simplified maintenance. In a cloud computing environment, container scheduling algorithms coordinate with different aspects of physical systems, such as memory allocation for tasks of different users. The scheduled containers on a host may process sensitive data. For instance, containers may process healthcare information. In that case, diverse cloud environments with different components and subsystems may lead to a potential personal health information leakage and violation of data privacy. In this paper, we introduce a novel compliance-aware analysis model for provisioning containers in the cloud, that provides a HIPAA compliance model. The proposed method dynamically analyzes different requirements of HIPAA complaint containers (HIPAA parameters) and their associated risk values. Based on the risk optimization of the compliance parameters for data security and data privacy of the containers, our proposed method determines scheduling of containers that offer the lowest risk to healthcare data and to the compliance posture of the container. The model describes the resources that are associated with high-level risks and provides real-time resource recommendation for a container scheduler to decrease the risk of HIPAA compliance violation.**

***Index Terms*—compliance; cloud computing; privacy; security; container.**

I. Introduction

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) [1], [2] is a legislative act established by the United States government to protect medical data. This protection is in the form of privacy and security for handling the data. The healthcare industry used to work with primitive technology, relying on paper medical records and used custom on-site installed infrastructure to process the data. This increased the operational cost and reduced the amount of flexibility for deploying new capabilities. To reduce cost, healthcare applications need to adapt to cloud computing environments [1]. Due to recent advances in cloud computing and datacenter design there is emergence of tools such as Kubernetes and Mesos that schedules applications across a cluster of machines. Modern scheduling applications consider various factors including physical resources in the system, and energy consumption to schedule containers to a pod or rack in a cloud computing environment. Such schedulers work well for general applications that do not have strict compliance requirements, unlike those mandated by HIPAA. Due to these requirements, it becomes important in cloud environments to determine whether the desired host is compliant for running HIPAA-sensitive applications. The problem becomes more interesting when applications are scaled across multiple geo-located cloud datacenters.

In this paper, we focus on providing a compliance-aware model for provisioning containers in the cloud where we compute a risk model that associated to HIPAA compliant parameters. This risk is computed based on the features explained later in the paper (Section IV). If the computed risk of resources is above a threshold, then we avoid scheduling on those resource on a given host. The rest of the paper is organized as follows: Section II reviews related work in this domain; Section III defines the problem statement and the motivation for this study; Section IV describes the compliance parameters including physical controls, technical controls, and administrative controls. This section also explains the proposed method for assessing the compliance parameters based on an unsupervised machine learning model; Section V describes evaluation results; and finally Section VI summarizes this study.

II. Related Work

When working in a strictly regulated environment, such as one associated with HIPAA, we need additional controls in place to satisfy compliance requirements. In one study, [3] provides several VM placement algorithms which, however, do not focus on placement of VMs based on compliance and security parameters. Although several studies have been completed about HIPAA compliant healthcare applications, to the best of our knowledge, there is not any available study that focuses on a comprehensive analysis on different HIPAA regulation parameters for cloud container scheduling.

III. Problem Definition

Consider a scenario involving data transmission across a public cloud infrastructure. The current data center designs follow a silo-ed approach [4], [5] to transfer data and schedule containers on the cloud platform. If data transmission needs

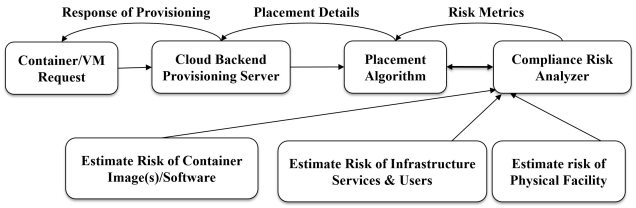


Fig. 1. Provisioning of a container based on HIPAA Compliant parameters

to occur with compliance-based data, an arbitrary rack cannot be used for provisioning to complete the data transfer. Due to compliance constraints, we need fixed compliance approved hosts which avoids elasticity in cloud for container provisioning. Consider a setting in which a large number of such HIPAA compliant applications are required to be provisioned or scheduled. In this case, a problem that surfaces is to maximize utilization of the cloud infrastructure while maintaining elasticity by removing its dependency on specific hosts for compliance specific applications. Existing containers provide some extent of data isolation but they do not have a complete deployment view of a container that follows HIPAA regulations. This calls for applying focus on devising a scheduling mechanism that is compliance-aware across data centers not constrained to a cluster, pod, rack or host in data center. The data privacy may cause a critical issue for sensitive applications, such as work in [6] and the parallel method of the data privacy [7]. Compliance-aware container placement is therefore a problem in which containers are provisioned by cloud scheduling infrastructure, only at the racks and the hosts that minimize compliance risks. For example, consider that a HIPAA compliant container has been requested to be provisioned on Amazon EC2 [8] cloud infrastructure. For privacy or compliance-sensitive data, it must be mandated that the underlying administrative management for those racks be following certain compliance requirements as well as privacy requirements.

Moreover, the parameters of compliance rules may be dynamically be changed. In order to understand these parameters (quantified as *risk*), we need to compute a risk ratio that associated to different compliance parameters including software, hardware, network, administration, and policies, which are being satisfied based on customer risk ratio request. For instance, if a customer has a sensitive information for processing a specific function or an application, then they may request the lowest threshold for data privacy and/or data security. Therefore, all required resources selected from the required resource pool can be assigned by the scheduler. The problem then becomes an optimization requirement dealing with both risk and Quality of Service (QoS) as well as target compliance parameters. This is visualized in Figure 1.

IV. Proposed Method

The proposed method is summarized in Figure 2. A quantification of risk, in the form of a *risk score*, is computed for the resources where a container can be provisioned. The resource that has lowest risk and appropriate capacity can be used to provision the container. The risk is computed by measuring the compliance rules across different components in the cloud infrastructure. A compliance-aware placement system requires all the risk parameters and an evaluation of all risk parameters. The system determines the best resource where the data transactions should be provisioned by considering risk ratio of resources. We develop an algorithm which estimates and reduces the risk that associates to data security and data privacy of the available resources based on compliance parameter across the cloud infrastructure.

Our work focuses on HIPAA compliance parameters but it is extensible to any risk parameters for compliance in general. There are the following three types of compliance parameters that cater to HIPAA compliance: *i)* physical controls, *ii)* technical controls, and *iii)* administrative controls. These parameters are obtained from HIPAA compliance requirements[1]. These parameters are evaluated and used to build the risk model of the infrastructure. The container scheduling algorithm uses the generated model to provision resources with lowest risk ratio.

In each of the subsections, the risk parameters are ordered in increasing order of risk quantification. A simplifying assumption for our work is that this may be treated in a manner similar to enumeration where only the relative difference in value matters. A more complex version of our proposed model can be easily developed by manually tuning these risk quantification values, or using an automated weighting system. Further, each risk parameter is referred to using the acronym of its category name, and the quantified risk value. For example, the first item in context of Section IV-A may be written as PC_i which has the highest risk ratio than $PC_j, PC_{j+1}, PC_{j+2}, ..., PC_{j+n}$ where $i > j$. Similarly, the second and third items (respectively) in Section IV-B and Section IV-C may be referred to as TC_2 and AC_3 respectively. Each item is a constraint for our proof-of-concept model, and can be easily modified to encompass a larger spectrum of quantification values. In addition, the model can be extended by considering more risk parameters in respect to the target compliance or specific criteria.

A. Physical Controls

Risk parameters in context of the physical facility of the architecture are summarized as follows.

Contingency operations for facilities for disasters: This parameter denotes whether the system operates and make services or data available during disaster.

Contingency operations for facilities for emergency mode operations: This parameter explains whether there is any policy implemented to access data in emergency mode.

Facility security plans implement to protect the facility and equipments from unauthorized physical access, tampering and theft: This denotes whether there are fail safe plans available in the event of a security breach.

[1]https://www.hhs.gov/hipaa/

Access control of physical facility: This parameter limits access to Protected Health Information (PHI). This parameter denotes whether the current data transfer is limited in its access to physical components.

Procedures to control and validate access to physical facilities, for control of access to software systems to users: This denotes whether a guest mode is available or used for the current transfer.

Implement policies to document physical repairs or changes to the facility: This parameter denotes the existence of any repairs to the current devices involved.

Device and media controls: This parameter provides whether the system implements controls to govern the receipt and removal of media and equipments handling PHI.

Disposal of PHI, media and electronic devices holding PHI: This parameter denotes whether the current devices are to be disposed after their use.

Record maintenance of movement of hardware and electronic media and any person responsible for the same: This parameter denotes whether there has been any recent maintenance of the devices involved.

Backup and disaster recovery policies and procedures for PHI before physical movements of media: This parameter shows whether backup and recovery processes are available when transferring the system from one architecture state to the other.

B. Technical Controls

Risk parameters specific to the technical facility for the architecture comprise of both the infrastructure and container images. The associated compliance requirements are listed as follows.

Access control policy and implemented system: This parameter denotes whether user and software authentication supported.

Encryption on stored-data: This parameter explains whether or not the data is encrypted by any means (application specific).

In-transmission-data encryption: This denotes whether or not the data is encrypted while being sent across the network, as opposed to static encryption when the data is in place.

Data integrity when in storage: This parameter shows whether the data is signed or hashed or HMACed to verify if it has been tampered with when in storage or on network.

In-transmission-data integrity: The parameter provides data integrity. However, it focuses on data integrity in network transmission.

Data backup: This parameter denotes whether the data concerned is being retained.

Disaster recovery: This parameter entails a description of the disaster recovery process and whether the devices involved are capable of such recovery.

Secure data deletion: This parameter shows how the PHI data is handled for deletion.

Logging and auditing: This parameter explains whether logs are being generated for each event that handles sensitive data or operation.

Log-in monitoring of privileged users: This parameter denotes whether privileged users are being monitored.

Log retention: This parameter denotes whether logs are being retained for a long time e.g. 320 days.

FIPS 140-2 compliant libraries: This parameter denotes whether FIPS-140-2[9] libraries are used for encryption.

Untrusted libraries in use for processing or while processing PHI: This parameter indicates trustworthiness of processing libraries.

C. Administrative Controls

Risk parameters associated with the administrative facility of the architecture are listed as follows.

User onboarding: This parameter denotes whether or not a user is currently being added to the system.

User offboarding: This parameter shows whether or not a user is currently being removed from the system.

User training: This parameter represents a custom control where a user is being trained on system administration.

D. Compliance modeling

We define a set of all compliances as follows: $C = \{c_1, c_2, ..., c_m\}$. Each compliance can be defined from different perspectives, therefore we define each parameter for each compliance as follows: $P = \{p_1, p_2, ..., p_n\}$. In this paper, we consider different rates for each $v_{ij} = c_i p_j$ where c_i denotes the compliance and p_j denotes parameter of c_i. First, we generate a static model of compliances by considering a high-dimension model of $C * P$. In order to find the decomposition of compliance risk level (R) into C and P, we use a Nonnegative Matrix Factorization (NMF) [10] by optimizing for the squared Frobenius norm which is a Euclidean norm to matrices. Other possible optimizations may be substituted in place of this norm, such as Kullback-Leibler divergence [11]. In NMF, we consider C as the weight matrix and P as feature matrix. The algorithm is based on two Singular Value Decompositions (SVD) processes for computing approximations of two parameters: *i)* the data matrix, and *ii)* positive sections of the resulting partial SVD factors (utilizing an algebraic property of unit rank matrices). The basic Nonnegative Double Singular Value Decomposition (NNDSVD) [12] [13] algorithm is better fit for sparse factorization. Its variants include NNDSVDa (in which all zeros are set equal to the mean of all elements of the data), and NNDSVDar (in which the zeros are set to random perturbations less than the mean of the data. Of these processes, NNDSVD, allows us to compute the risk model based on both compliance character (C) and its parameters (C_P).

The compliance risk model can be deigned as follows:

$$arg \min_{C,P} \frac{1}{2} \|R - CP\|_2^{Fro} = \frac{1}{2} \sum_{i,j} (R_{ij} - CP_{ij})^2$$

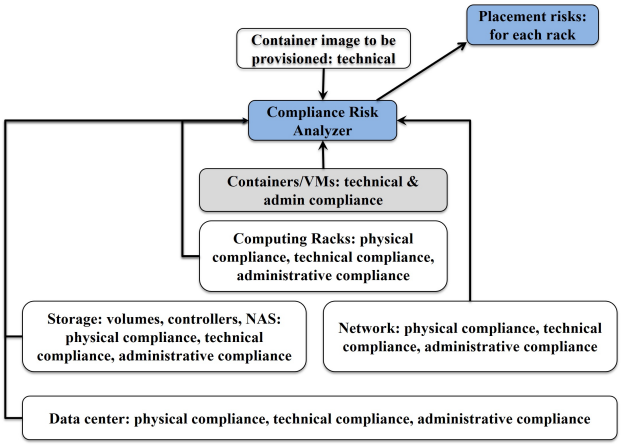


Fig. 2. Risk computation for a container

The combination of two processes with the parameter ρ, and the intensity of the regularization with the parameter α, is defined as follows:

$$\vartheta = \alpha\rho \left\|C\right\|_1 + \alpha\rho \left\|P\right\|_1 + \frac{\alpha(1-\rho)}{2} \left\|C\right\|_{Fro}^2 + \frac{\alpha(1-\rho)}{2} \left\|P\right\|_{Fro}^2 \quad (1)$$

Therefore, we compute the final combination of two processes with using the compliance risk model as follows:

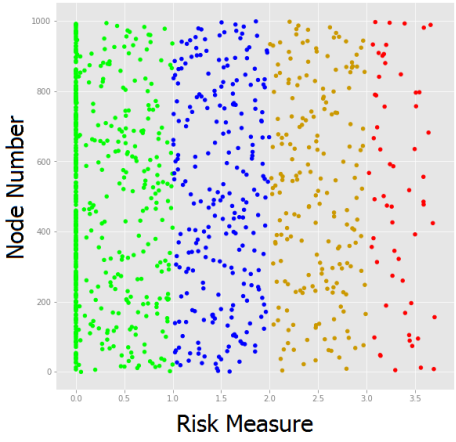
$$arg \min_{C,P} \frac{1}{2} \left\|R - CP\right\|_2^{Fro} = \frac{1}{2} \left\|R - CP\right\|_{Fro}^2 + \alpha\rho \left\|P\right\|_1 + \alpha\rho \left\|C\right\|_1 + \frac{\alpha(1-\rho)}{2} \left\|C\right\|_{Fro}^2 + \frac{\alpha(1-\rho)}{2} \left\|P\right\|_{Fro}^2 \quad (2)$$

This risk model allows container scheduling algorithm to pick low level risk ratio for healthcare applications. Since the model is generated as a part of offline computation and it can be updated based on updating the resources, the model provides the best low level risks to the scheduling algorithm. This model also enables data center administrators to monitor their resources if they have to provide low-level risk ratio resources to their healthcare providers that follow HIPPA compliance.

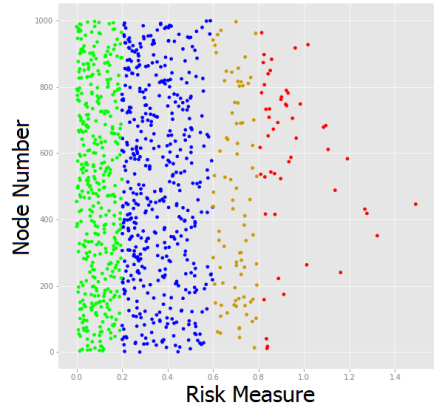
V. Evaluation

We implemented the proposed scheduling method using scikit-learn[2](a well-known machine learning library for Python) to simulate 1000 nodes in the network. We consider two components as output: first, compliance evaluation with respect to data security, and second, compliance evaluation with respect to data privacy. These two evaluations allow us

[2]http://scikit-learn.org



(a) Evaluation of data security

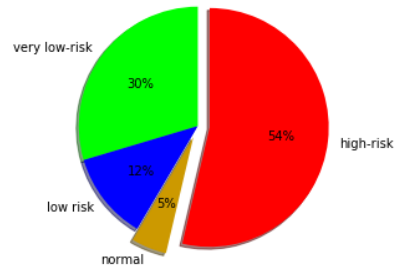


(b) Evaluation of data security

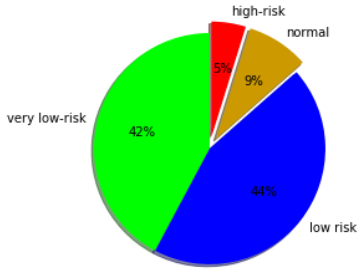
Fig. 3. Distribution of the compliance model for data security and data privacy.

to understand which resource has high-level risk ratio for data security or data privacy, based on compliance parameters. We assume that we have a normal distribution of jobs. We used Non-Negative Matrix Factorization (NMF)[14] with $2,000$ of iterations in NLS subproblem. We considered the following classification for the risk measure: $risk\ measure < 0.2$ as very low-risk, $0.2 \leq risk\ measure < 0.6$ as low-risk, $0.6 \leq risk\ measure < 0.8$ as normal, $0.8 \leq risk\ measure$ as high-risk.

We use uniform data sampling to generate samples for $1,000$ nodes. Figure 4 shows evaluation results for our simulation. In this figure, $Y - axis$ represents the number of node and $X - axis$ represents the risk measure for each node. As shown in this figure, we found that 54% and 5% of scheduling tasks have been considered as high-level risks for compliance with respect to data security and data privacy, respectively, based on HIPAA. Figure 3 shows the distribution risk ratio for all nodes for both data security (Figure 4.a) and data privacy (Figure 4.b). This report allows both provider and consumer understand the level of risk factor for both security and privacy based on HIPAA Act. A cloud provider will be able to provide a dedicated cloud service with lower risk factor



(a) Evaluation of data security



(b) Evaluation of data privacy

Fig. 4. Evaluation of the compliance model for data security and data privacy

in respect to HIPAA compliant for both data security and data privacy. For instance, the proposed method enables a cloud provider to retrieve a report from current cloud services as shown in Figure 3. Then, the cloud provider may consider a risk ratio, i.e., less than 2.0, to retrieve the list of cloud service configurations including software, container and hardware that compliant with HIPAA Act with the risk level less than 2.0. This list of cloud services may provide to a set of customers. Figure 3 also shows that the scheduling algorithm needs to remove this high-level risk resource for the pool during the scheduling process for those customers.

This evaluation also allows the network administrator to monitor compliance risk ratio for each component and it provides an action plan for improving the network component (including software and hardware). By removing the components corresponding to high-level risk ratio, the network cloud provider may claim that their network is HIPAA compliant.

VI. Conclusion

The scheduled containers on a host may process sensitive patient data such as data from a healthcare application. If a scheduling algorithm selects high-level risk resources from the pool, it may lead to personal health information leakage. Also, an unregulated environment may allow an attacker to access the container on the cloud. In this paper, we introduced a novel compliance-aware provisioning model that provides HIPAA compliant resources. The model computes risk ratio of the resources and recommend the low-level risk ratio resources to the scheduler. We performed an evaluation on our proposed method and we found in our study that 54% of cloud container resources are vulnerable to data security and 5% cloud container resources are vulnerable to data privacy when it is annotated with high-level risk based on HIPAA Act. By allowing the scheduling algorithm of container to use the HIPAA compliant model, it decreases the associated risks to both data security breach and data privacy violation in the cloud.

References

[1] C. for Disease Control, Prevention *et al.*, "Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services," *MMWR: Morbidity and mortality weekly report*, vol. 52, no. Suppl. 1, pp. 1–17, 2003.

[2] G. J. Annas, "Hipaa regulations-a new era of medical-record privacy?" *New England Journal of Medicine*, vol. 348, no. 15, pp. 1486–1490, 2003.

[3] K. Mills, J. Filliben, and C. Dabrowski, "Comparing vm-placement algorithms for on-demand clouds," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 91–98.

[4] R. Dutta, G. N. Rouskas, I. Baldine, A. Bragg, and D. Stevenson, "The silo architecture for services integration, control, and optimization for the future internet," in *Communications, 2007. ICC'07. IEEE International Conference on*. IEEE, 2007, pp. 1899–1904.

[5] K. Jang, J. Sherry, H. Ballani, and T. Moncaster, "Silo: Predictable message latency in the cloud," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 435–448, 2015.

[6] M. Bahrami and M. Singhal, "A light-weight permutation based method for data privacy in mobile cloud computing," in *2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, March 2015, pp. 189–198.

[7] M. Bahrami, D. Li, M. Singhal, and A. Kundu, "An efficient parallel implementation of a light-weight data privacy method for mobile cloud users," in *Proceedings of the 7th International Workshop on Data-Intensive Computing in the Cloud*. IEEE Press, 2016, pp. 51–58.

[8] G. Wang and T. E. Ng, "The impact of virtualization on network performance of amazon ec2 data center," in *Infocom, 2010 proceedings ieee*. IEEE, 2010, pp. 1–9.

[9] T. Caddy, "Fips 140-2," in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 468–471.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562. [Online]. Available: http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf

[11] S. Lukits, "Maximum entropy and probability kinematics constrained by conditionals," *Entropy*, vol. 17, no. 4, pp. 1690–1700, 2015. [Online]. Available: http://www.mdpi.com/1099-4300/17/4/1690

[12] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[13] A. Cichocki and P. Anh-Huy, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 92, no. 3, pp. 708–721, 2009.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.