Managing sensitive applications in the public cloud

Kirk A. Beaty*, Jenny M. Chow†, Renato L. F. Cunha‡, Koushik K. Das§, Mark Hulber*, Ashish Kundu*, Vanessa Michelini†, Elaine Palmer*
*IBM Research Division, Thomas J. Watson Research Center
†IBM Watson Group
‡IBM Research Division, Brazil Research Laboratory
§Capital One Enterprise Data Services

Abstract—Protecting the security and privacy of data is a paramount concern of enterprises in medical, educational, financial, and other highly regulated industries. While some industries have moved rapidly to take advantage of the cost savings, innovations in data analysis, and many benefits provided by cloud platforms, regulated enterprises with sensitive data have proceeded with caution. In this paper, we explore a fully public cloud-based architecture that is able to handle both service requirements and security requirements. In such a public cloud environment, the traditional notion of static perimeter-based reactive security can leave internal system components vulnerable to accidental data disclosures or malicious attacks originating from within the perimeter. Therefore, ensuring security and compliance of such a solution requires innovation and new approaches in several directions, including proactive log monitoring and analysis of virtually all parts of the cloud-based solution, full end-to-end data encryption from the client through Internet transmission to data storage and analytics in the solution, and robust firewall and network-intrusion detection systems. We discuss many of these techniques as applied to a specific real-world application known as the Watson Genomic Analytics Prototype.

1 Introduction

The cloud platform has led a paradigm shift in the way computing is delivered [1]. As per IDC and Gartner studies [2, 3], the cloud offers a platform for industry-transforming solutions. In this paper, we use the term "public cloud" to refer to a form of cloud computing in which a company relies on a third-party cloud service provider for services. This kind of cloud provides a platform for massive scale analytics of data, and leading-edge enterprises are heavily investing in cloud-based innovations and using them as a foundation for new competitive offerings.

Managing security, privacy, and compliance is an expensive (both in terms of time and cost), on-going requirement. It also requires continuous innovation and vigilance, particularly when hosting applications that handle sensitive data, such as health or education records, financial data, or even mobile-device data. Certain classes of such sensitive data are governed under regulatory compliance requirements, whereas others require enforcement of privacy policies. Data breaches, data leaks, and system compromises can cost fortunes to companies, and can damage brand reputation [4, 5]. Companies seem to face competing goals - moving to the public cloud versus maintaining compliance, fearing that systems over which they have little control will be compromised and will leak sensitive data. In fact, they can have it all – by deploying applications designed for cloud and by selecting a cloud service that provides continuous innovation in its security and compliance services.

This paper shares our experience in applying organizational practices and system security principles to protect sensitive data, which is stored and analyzed in systems in the public cloud. In particular, our systems, collectively known as Watson Genomic Analytics Prototype (WGAP), help physicians analyze genetic mutations of cancer patients, thus compliance with the Health Insurance Portability and Accountability Act (HIPAA) is a paramount concern. Our intention is to help other practitioners achieve compliance with data security and privacy regulations, as they, too, deploy sensitive applications in the public cloud.

This document is structured as follows. In the section titled "Sensitive applications in the cloud," we present the case of hosting sensitive applications in the cloud and related challenges, considering applications compliant with HIPAA. In the section titled "The use case," we describe a use case of an application we implemented, relating it to the requirements of the previous section. Finally, in the section titled "Related Work," we position our work with respect to the existing literature, and we summarize our work and findings in the section titled "Conclusions."

2 Sensitive applications in the cloud

In the following discussion we assume knowledge of a few different concepts, here defined: Systems of Record (SoR) are storage systems that are the authoritative data sources for given data elements or pieces of information (*e.g.*

the domain name server for a given domain is a SoR, for it is the authoritative domain name resolver); Systems of Engagement (SoE) are systems that focus less on discrete pieces of information, and more on peer interactions (*e.g.* social networking applications are SoE); and Systems of Insight (SoI) are the product of linking SoR and SoI for finding new relations in the meeting of historical data (SoR) and dynamic behavioral data (SoE).

A key requirement for cloud-based business transformation is the ability to seamlessly integrate data from disparate sources including the traditional SoR and born-on-the-web SoE, and deriving business insight in near real-time. A highlevel conceptual architecture of one such class of SoI solutions on a public cloud platform is shown in Figure 1 for three industry sectors: Healthcare, Retail, and Education. These solutions show a common pattern: SoR data is securely ingested, undergoes some level of transformation, and then is stored in a control repository, which is typically an object or warehouse store. Several examples of SoR data are shown in the left part of Figure 1.In a Healthcare solution, raw genome sequencing data goes through multiple levels of mapping and annotation [6] before being stored in an object store for further analysis. For a retail solution, raw data comes from multiple sources like customer profiles, transaction logs, and customer relationship management and may be stored in a data warehouse for further querying and analysis. In case of an Education solution for a K-12 school in the USA, sources of SoR data include the student information system, assessment and attendance records, curriculum and instruction, and student activity information.

On the other hand, the SoE data goes through a knowledge extraction step, which may include data mining and generation of tags, metadata and predictive models as shown in the right side of Figure 1. The resulting curated data is stored in a knowledge repository, which may either be a relational database management system (RDBMS) [7] or a database system that stores and retrieves data that is modeled in means other than tabular relations used in relational databases (i.e. NoSQL data stores [8]). Sources of SoE data may include a collection of medical articles and drug information coming from multiple sources for a healthcare solution, social data for a retail solution, and publisher content and online knowledge sources like Wikipedia for an education solution. Knowledge extraction may include pre-filtering and ranking of drugs/pathways for a healthcare solution, social influencer, sentiment analytics, and life event detection for a retail solution and determining the applicability and relevance of publisher contents by topic, grade, and level of difficulty for an education solution.

Both of the above steps for pulling information from SoR and SoE are essentially offline batch processing steps on data at rest. As shown in the middle box in Figure 1, the SoI analytics engine links and integrates data from SoR and SoE and generates business insight. Examples of business insight generation include personalized drug recommendations in a healthcare solution, or customer churn reduction in a retail solution, or student risk assessment and personalized learning in an education solution.

2.1 Challenges

As sensitive workloads and data migrate to the cloud, security and compliance of the data and the workload are becoming increasingly paramount. Traditional SoR have been implemented in non-cloud environments behind the corporate firewalls and physically isolated from the internet. SoR data often have sensitive/personal information and hence are often subject to various compliance requirements of government regulations for protecting data security and privacy. For example, in the US, there are the HIPAA [9] for healthcare, Family Educational Rights and Privacy Act (FERPA) [10,2] for education, and, globally, there is the Payment Card Industry Data Security Standard (PCI DSS) [11] black for payment standards. In such deployment scenarios, static perimeter based reactive security controls are usually deployed, with additional measures implemented as needed to meet the various compliance requirements.

With the advent of SoE, the scale, variety and velocity of data that typically needs to be analyzed has rapidly grown. Moreover, the SoE data often have less stringent security requirements, since they typically do not contain as much sensitive or personal information as the SoR data. These trends have led to many enterprises starting to embrace public cloud as the SoE deployment platform for massive scale data analytics and knowledge generation.

With the sophistication of security threats and attackers increasing, deploying a SoI solution on a public cloud requires that it meet the security and performance requirements of both SoR and SoE. SoE performance requirements can be met through design and deployment of a cloud-based elastic scale-out architecture, exploiting the capabilities of a public cloud platform (right section in Figure 1). The bigger challenge is hosting the SoR (left section of Figure 1) and the SoI near real-time analytics engine (middle section of Figure 1). One design strategy can be to have a hybrid cloud implementation, where sensitive SoR data stays in an on-premise (non-public cloud) infrastructure while the SoI analytics engine intelligently links data from SoR and SoE to deliver the required business insight. However, hybrid cloud implementations with policy-based data movement between cloud and on-premise environments can be complex and can put additional restrictions on the architecture design. Moreover, a hybrid cloud implementation of a SoI solution may prevent full exploitation of the advantages provided by public cloud computing.

Deploying regulated applications on cloud-based systems provides many advantages as stated above, but raises challenges not present when systems are hosted on the premises of the data owner. For example, vendors operating within customer infrastructure can rely on some level of preexisting perimeter security (e.g., intrusion detection, firewalls). Cloud vendors, however, must provide the network and security infrastructure. As vendor responsibilities expand, so does their level of risk. For example, in the case of electronic medical records stored in the cloud, cloud vendors, and their subcontractors are typically considered "Business Associates", and must comply with security rules [12]. Violations can incur federal civil penalties up to $50,000 per occurrence applied on a per record basis (not per incident), with a $1,500,000 annual cap. Prison terms are possible in cases of

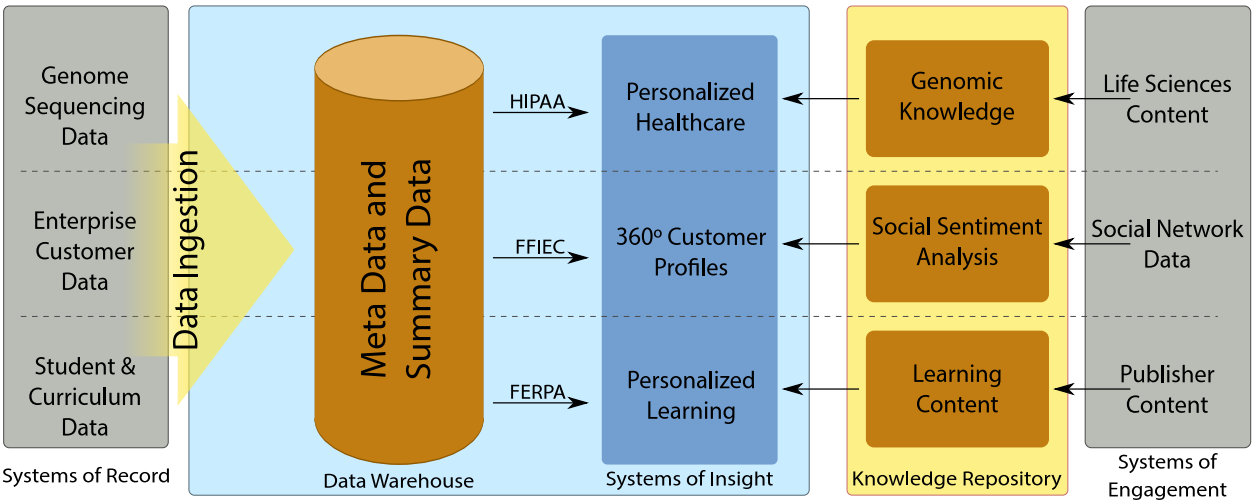

Figure 1. High-level conceptual architecture of three classes of SoI solutions on a public cloud platform. In the first row one can see how a personalized healthcare could be constructed. In the second row, it shows the same for a customer profiling application. In the third row, a personalized learning example is shown. At the extreme left, traditional SoR data are shown. At the extreme right, SoE data are displayed. These traditional systems are combined together to create SoI, while adhering to regulatory standards.

willful intent, and additional state penalties may apply [13].

2.2 HIPAA: Case for compliance

In the US, personally identifiable information (PII) that contains one or more of the eighteen identifiers (derived from Code of Federal Regulations 160.103) is considered to be protected health information (PHI). PHI is regulated by HIPAA and, more recently, by the Health Information Technology for Economic and Clinical Health (HITECH) Act and must be treated with special care. HIPAA was introduced in 1996 by the US Department of Health & Human Services (HHS) to bring the US healthcare industry into the digital age while protecting the privacy of the individual.

HIPAA has three key components: privacy rule, security rule and the transaction code set. These key components paved the way for health providers, health insurance companies and other healthcare entities to exchange necessary information for healthcare business operations with regard to treatment as well as payment. The HITECH Act of 2009 expanded HIPAA and made business associates – partners in the healthcare industry, which includes solution providers and infrastructure services – become responsible for their share of the liability in the areas of privacy and security rules. To satisfy these rules, a solution provider must ensure that PHI is protected and secure. One must have policies, procedures and HIPAA requirements spelled out clearly in the contract with clients. The team must strictly adhere to the agreements and do its part to ensure information is treated appropriately. What this means to a cloud solution is that in addition to securing the sensitive data at rest and in motion, team members in specific job roles must be educated on HIPAA and follow security policies and procedures that are put in place to ensure proper access and handling of sensitive data.

Other industries have different regulations. FERPA and PCI DSS are examples of those regulations. Another common one in the United States is the legislation and regulations from Federal Financial Institutions Examination Council (FFIEC). They all have security and deployment implications. It is imperative that the solution data compliance strategy is built upon a set of solid governance, risk and compliance principles.

2.3 Architectural considerations

Here we discuss the kind of design choices that matter to sensitive cloud applications.

2.3.1 Cloud model

An approach for designing the architecture for cloud solutions is to combine together pre-built cloud-based services based on the published application programming interfaces (APIs) made available by providers. Such an approach leads to reduced development time and costs, provided such pre-tested services fulfill functional needs. However, for sensitive applications with stringent requirements on security and compliance, the design decisions may require sub-services to be developed or updated. Consider the case of sensitive data; having other external services access such data would make the auditing of compliance difficult at best, untenable likely. Further, the acceptance of risk is much greater, as any breach of security would call for considerable contractual and operational coordination. Besides compliance, the same arguments may be made for robustness. The more the number of dependent services (sub-services) used, the harder it becomes to guarantee the uptime of the whole service (even if the dependent sub-service has its own service-level commitments). For example, if sub services go down, the main service may go offline without any simple way to recover. As an alternative to external services, cloud marketplaces (e.g. IBM Blue Mix [14]) that provide service patterns can offer sub-services to be incorporated instead of having them hosted externally and called when needed. In such a marketplace, it would be also possible to incorporate services critical for operation, and leave less important services hosted elsewhere.

Table 1
Possible multi-tenancy configurations and their security, privacy and compliance aspects for cloud deployments. For each possible end-user tenancy configuration(single-tenant, multi-tenant, multi-tenant across end-users, and hybrid tenant), its security, privacy, and compliance levels due to isolation are analyzed.

	Single-tenant	*Multi-tenant*	*Multi-tenant across end-users*	*Hybrid-tenant*
Security due to isolation	High	High	High	Optimal: Low to High
Privacy due to isolation	High	Low	Medium	Optimal: Low to High
Compliance due to isolation	Yes	No. Unless stated otherwise.	Yes. Unless stated otherwise.	Optimal: Low to High

2.3.2 Multi-tenancy

Multi-tenant deployment of the workload is an approach that has trade-offs on cost, security and compliance. By multi-tenant deployment, we mean deployment of some or all components on hardware/software shared with other cloud tenants (customers). However, in some cases, multi-tenancy may also refer to the sharing of a solution instance among multiple end-users. While the former has both security, privacy, and compliance issues, the latter has more privacy and compliance issues than security, as security of the solution stays almost identical. The single-tenant deployment model does not share any infrastructure or solution components with any other customers or end-users.

Another form of deployment is *hybrid-tenant* deployment: where the solution is deployed such that certain components are deployed in a single-tenant manner and other components are deployed on a multi-tenant manner. Table 1 summarizes the various possible tenancy configurations and the levels of security, privacy, and compliance one can achieve with them.

2.3.3 Elasticity and hardening

Due to the highly dynamic nature of cloud-based solution deployment, gone are the days of sophisticated long-range capacity planning – as well are gone the long lead times to have hardware ordered, installed, configured and operational. Apparent infinite scalability, with provisioning and installation in the order of minutes, are some of the defining characteristics of the cloud [1]. However, with the restrictions of sensitive applications and the compliance necessary, it does take more consideration than simple elastic provisioning of another server to be able to join it to an existing service cluster. Before such application servers are deployed, a stringent check of compliance is required. Delivery teams have procedures to "harden" a server for use in a production environment (cloud-based or not) that include scans of software and network components to ensure compliance to the myriad of security guidelines and best practices. Thus a new server being provisioned will be unusable to the service until these compliance scans are successfully performed. To be successful in having elastic auto-scaling for such sensitive application cloud environments, automation of the compliance adaptation process will be essential – not so easy in practice with the kind of oversight typical of delivery teams.

2.3.4 Security issues and controls

In this section, we briefly cover security components, which are commonly required in order for systems to be compliant with standards and regulations.

2.3.4.1 Data protection: Data protection is the key objective for the security of a solution. In some applications, such as healthcare, data confidentiality is a primary concern. Moreover, data integrity must be assured, so that unauthorized tampering of data is prevented, or at least, detected. Although malicious data breaches receive much public attention, inadvertent disclosures are a common source of violations [15].

Organizations must actively manage sensitive data in their control. For example, when sensitive data is no longer needed, simply deleting it is not sufficient, because data remnants can be left on the storage media, and retrieved by the next application storing data in the same location. Instead, it must be securely deleted, using specific techniques such as writing over the data with special or random bit patterns [16], for example, when disks are surrendered for reuse. Cloud object stores, by design, may spread data across multiple data centers to accomplish redundancy and to facilitate detection of corrupted data. However, cross-border restrictions on the use and migration of data limit the use of these object stores [17]. Furthermore, a well-documented procedure must be in place for backup, restoration, and disaster recovery of data.

2.3.4.2 Encryption: Sensitive data must be encrypted when at rest (at its origin and in storage) and in motion (during upload and between system components).Not only must the encryption keys be protected against disclosure and modification, but some regulations, such as those from the Payment Card Industry, require that keys be used only within hardware security modules (HSMs) [18]. HSMs protect the keys from physical tampering and unauthorized access. Keys and encryption algorithms must be strong enough to protect the data and withstand attack for as long as the data is retained, for example, six years in the case of HIPAA. The Advanced Encryption Standard with a key length of 256 bits is currently sufficient for this purpose.

Key management support and services are essential. For example, keys must be passed securely from storage to other servers, where they are used to decrypt data prior to analysis. Native database encryption, such as that provided with IBM DB2, offers built-in key management support, and can separate the keys from the data, as is required by HIPAA. Keys can optionally be stored in a key store protected by an

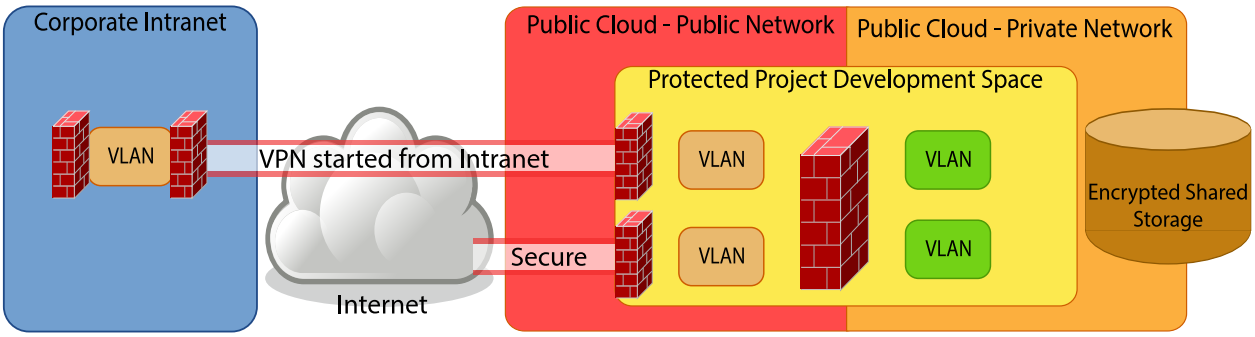

Figure 2. Network isolation for project development in the cloud.

HSM.

2.3.4.3 Network isolation: As mentioned above, cloud deployments can be very different from those in the corporate IT infrastructure. In the cloud, computing resources are accessible from the public Internet. To protect these resources, firewalls must be put in place. Figure 2 displays an overall diagram of the network isolation used in WGAP. Intellectual property must be protected when developing projects in the cloud. Networks must be fully isolated and data must be protected. Generally there will be no access to the corporate Intranet without an approved network architecture. Access from the Intranet or public Internet will be over a VPN until the project moves to production. Administrative control will remain over a VPN and public flows will be controlled and isolated. VLANs create isolated networks protected by the firewalls put in place. (VLAN: Virtual Local Area Network; TLS: Transport Layer Security; VPN: Virtual Private Network.)

To move data between the corporate environment and the cloud, one simple solution would be to create a Virtual Private Network (VPN) between these two environments. Doing so naïvely is discouraged: without sufficient governance, the corporate network would be put in danger. Without proper controls, the remote endpoint inherits the trust of the internal VPN endpoint. The remote endpoint, without corporate governance becomes an uncontrolled and unmonitored path into the corporate intranet. Therefore, until such a process is guaranteed, connections between cloud and on-premises are best restricted to one direction. This level of security has a performance impact, as the encryption present in VPNs requires more processing cycles, which must be considered when planning the size of environment required. In the cloud, access to disk is most frequently to network attached disks instead of Storage Area Network (SAN) attached disks. This network traffic follows the similar paths to other traffic, must be encrypted, and must go through firewalls. Going through these firewalls contributes to a bottleneck to the application for remote disk access. For applications which require parallel, high speed access to disk, it may become impractical from a performance perspective to go through the firewalls. From a security perspective, avoiding firewall controls is not an option. The result is the need to have highly scalable firewalls between the compute and storage infrastructure. In a traditional enterprise deployment, these issues would be solved with additional hardware switching. In the cloud, the ability to mirror these types of flow and throughput is limited by the offerings of the infrastructure provider.

Data transferring through the internet should have a second layer of encryption to ensure any termination of the Transport Layer Security (TLS) connection does not potentially expose the unencrypted contents of the data transmission. A particular example of a possible breach is a man-in-the-middle-attack (MITM). This requirement does not end at the edge of the solution. To prevent unauthorized access to sensitive data, all network data flows should be encrypted, and best practices for zone isolation must be followed in the cloud. This can be particularly challenging in cloud environments where Virtual Local Area Network (VLAN) allocation is limited [19]. It benefits cloud providers to include some level of Intrusion Prevention/Detection Systems (IPS/IDS). The audit logs and controls for those systems may not be readily available to the consumer and so additional IPS/IDS are required to satisfy security and audit requirements.

2.3.4.4 Access control: In order to allow modification or disclosure of data, systems must authenticate and grant appropriate access to different parts of the overall system to users with varying privileges. Examples of such users include application users, application administrators, emergency room personnel, database administrators, network administrators, security officers, hypervisor and operating system administrators, and hardware operators. Often, the "user" is not a human being at all, but a subsystem serving as a proxy for a human or mobile device, and accessing another part of the system. In cases where an application must enter a stored password, passwords should not be captured in audit logs. Furthermore, when a user terminates employment or no longer operates in a role, her privileges must be revoked immediately (and documented). This activity may require removing data and encryption keys from a laptop or phone.

2.3.4.5 System and application security: Following software engineering best practices, such as life cycle management, thorough documentation, and extensive testing, helps to secure an application by exposing defects that may lead to vulnerabilities. However, additional activities are recommended in the application development phase, specifically, carrying out security assessments, white box scanning source code, and performing black box penetration tests. Additionally, applications should use vetted software whenever possible, such as cryptographic libraries evaluated under Federal Information Processing Standard 140-2 and

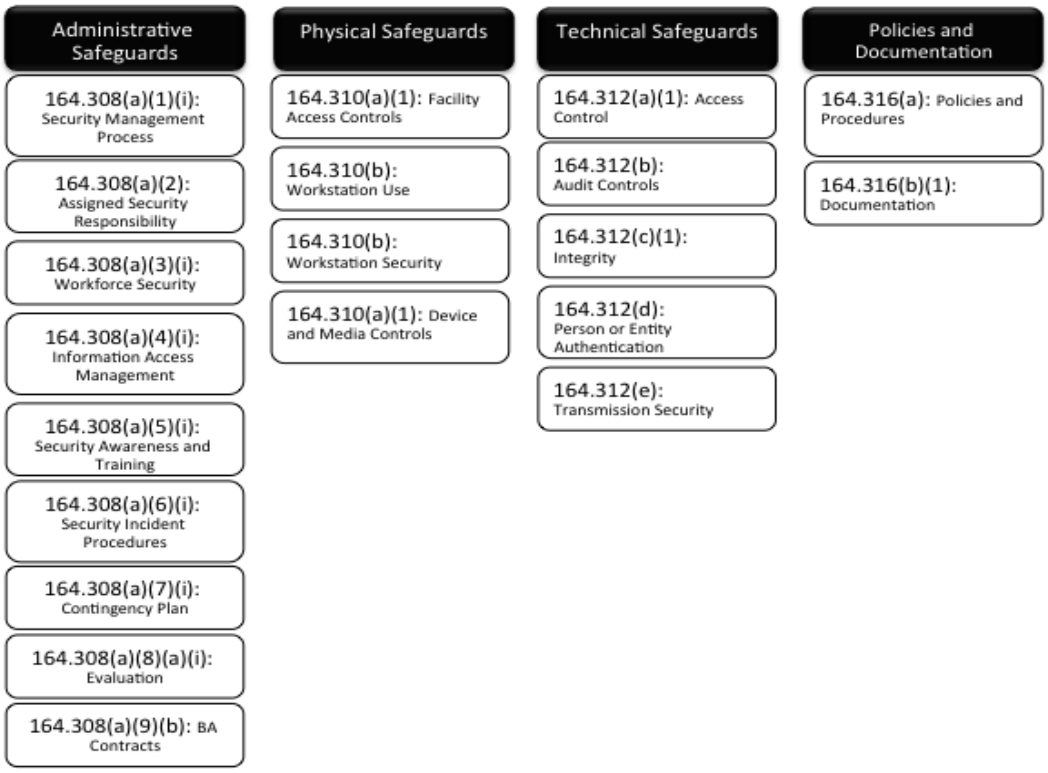

Figure 3. Health Insurance Portability and Accountability Act (HIPAA) Compliance requirements organized as four pillars: Administrative, Physical, Technical safeguards and Policies and Documentation.

operating systems evaluated under the Common Criteria. Following best practices across an entire system can overly constrain requirements and require compromises. For example, we describe one such issue related to cryptographic libraries in ***Deployment Challenges*** below.

System and infrastructure providers, too, must follow security best practices, such as deploying carefully configured firewalls, offering private access VLANs, IDS/IPS, security assessments, and penetration testing (of the entire system and infrastructure).

2.4 Organizational considerations

This section briefly covers further requirements for HIPAA compliance. In addition to the three key HIPAA components mentioned previously, HIPAA compliance has the following requirements: (1) administrative compliance, (2) physical security compliance, (3) technical compliance, and (4) policies and documentation, as summarized in Figure 3. Organizations must consider these requirements along with the architectural considerations described above. For example, organizations must manage a HIPAA-compliant solution throughout its lifecycle and throughout the organization. Furthermore, legal agreements should document and clearly delineate responsibilities, which, by necessity, are shared by customer and provider.

3 THE USE CASE

A real-world application is introduced in this section to illustrate the salient points of this article, as it has the characteristics of a cloud service with sensitive data, having security as one of its prime objectives in its design, development, and day-to-day operational delivery and management. This application is one in which the authors have been directly involved in developing and managing.

The service is known as Watson Genomic Analytics Prototype (WGAP). WGAP is a Software-as-a-Service (SaaS) offering that runs on the IBM SoftLayer public cloud [20]. WGAP helps physicians analyze genetic mutations of patients diagnosed with cancer. The input data is somatic mutations (abnormal cells compared to normal cells after DNA sequencing is performed), which a user of the service—oncologists or cancer researchers—can securely upload to the system for analysis. During the analysis, WGAP does a comprehensive search in the medical research literature (e.g. PubMed.gov) using a another service, the Watson Discovery Advisor (WDA), to identify relevant mutations, biological pathways, and drugs reported as targeting directly or indirectly those specific mutations. The result of the analysis is presented to the user in a form of reports and graphical data representations, enabling the user to drill down and review the findings and evidences provided by WGAP, as well as how the drugs target cancers on the reported pathways. The doctor combines this results from WGAP with the standard factors of age, overall health, current treatments/prescribed drugs, potential harmful side-effects, and quality of life to come up the best treatment plan.

3.1 The component architecture

Figure 4 shows a component overview of the WGAP service. Users interact with the service via the portal user interface.

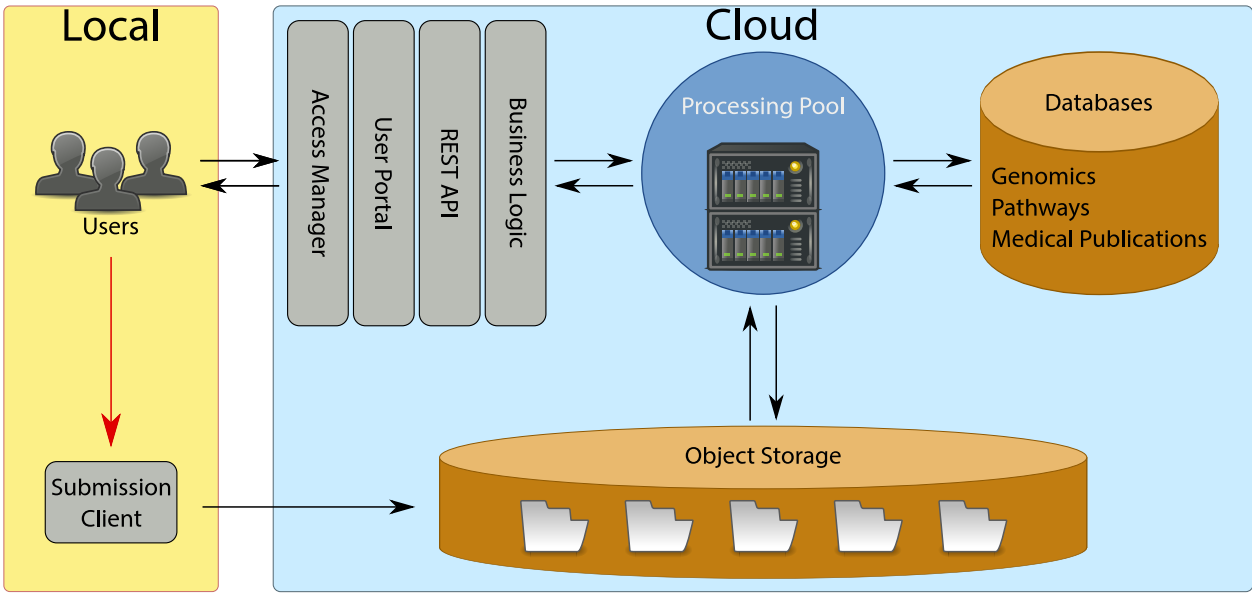

Figure 4. Watson Genomic Analytics Prototype (WGAP) Component Diagram. Red arrows represent unencrypted data movement. Black arrows represent encrypted data transfers. All data is stored encrypted in the Object Store, and is decrypted on an as-needed basis when processing must take place. Each folder in the object store represents a different patient case, stored in its own container.

The IBM Security Access Manager (ISAM)–represented in the figure as the "Access Manager" layer–authentication and authorization solution controls access to the service with credentials and role checking provided at this layer. ISAM also provides session management functions, critical to managing time-out of sessions as required by compliance or policy. There are multiple roles in place to assist in providing needed access for not only the intended target users of the service but also for administrators who manage the accounts and users (including assigning roles), and also for support personnel to be able to view the operational aspects of the service, and even configure them dynamically as required for adjusting elasticity of load capacity.

The primary flow of the WGAP service starts with users interacting with the portal to create a case, and then submitting DNA data of the cancer cell of a patient, as a sample to that case. Submission of the data sample includes encryption of the data before it enters the network, storing of the data into Object Store, and updating of the meta-data for this DNA sample submission into the WGAP operational database. The submission is picked up by the case manager and analysis is scheduled via the work queue manager which requests a pool member from the pool manager. The work queue manager requests analysis for the sample via the provided analysis node (analysis pool member) and continues to monitor progress through to completion.

The analysis nodes ("Processing Pool" in the figure) request the sample data to be copied from the object store into its local storage via the storage manager, and then performs analysis, persisting results of analysis back to the object store. When complete, the user is provided indication that visualization of the results are available. They then use the service portal to view the reports, and various interactive graphical views of the results. Reports including the evidence and graphical illustrations can be saved or printed and used in oncology review board discussions to assist the user in the determination of an effective treatment plan for the patient.

3.1.1 *HIPAA considerations*

The data samples submitted to WGAP are considered de-identified data. Each client also agrees no patient identifiable information need to be entered into WGAP. So, technically, WGAP does not contain any PHI. Nonetheless, the team followed the guidance of HIPAA in order to design, implement and operate a system right from the start of the project, with enough safeguards to handle PHI should a future client deem it necessary.

The WGAP team worked with the corporate in-house HIPAA Program Office (HPO) to work through a checklist of fifty-nine control points that are derived from the HIPAA and HITECH regulations. The HPO and the WGAP HIPAA Security Officer defined all the roles in the team and determined which roles were in scope for HIPAA. The WGAP HIPAA Security Officer then worked with the HPO to understand the goals of the control points and to come up with acceptable solutions to address each control point and collect evidence of solution implementation.

Of the four categories of HIPAA requirements (summarized in Figure 3), the team spent the most time to fulfilling the ones for Administrative compliance and Technical compliance. For Physical security compliance, WGAP is hosted in the SoftLayer [20] cloud data centers, which are already enabled for hosting HIPAA workloads. For Policies and documentation, we did not have to start from scratch, because WGAP inherited a solid set of privacy and security policies from the corporate Chief Information Officer (CIO) office for our team members to follow. Our infrastructure support and delivery teams were already familiar with the set of privacy and security policies which had already been updated by the CIO to address HIPAA.

Since the technical security measures the team took were described in detail throughout this paper, this section covers

the remaining category, Administrative compliance. This requirement has a lot to do with properly training and tracking access of the team members, from project managers to infrastructure support members. This requirement is addressed mainly by implementing a set of on-boarding, off-boarding and user credentials (ID) approval procedures. The on-boarding procedure includes collecting evidence of having completed training on HIPAA and properly encrypting the data on their workstation and having up-to-date antivirus protection. The off-boarding procedure ensures access to project areas is properly terminated in a timely manner.

The WGAP HIPAA Security Officer makes sure that the procedures are followed and the evidence collected then stored for the proper length of time. The on-boarding procedure includes HIPAA training for all team members in scope, whether regular employees or contractors. The training has to be completed before access to code repository or to the environments can be granted. Formal training has to be repeated every 12 months, informal training and updates from the HPO are circulated at least once a quarter. All team members need to follow the company's security and use standards. Off-boarding has to happen in a timely manner with previously granted access relinquished and any PHI on local drives shredded. WGAP uses an in-house service to check against the company Human Resources (HR) database nightly to alert of any changes in employment status. Infrastructure support team members are required to take additional security courses. Their activities on the servers are monitored by an in-house audit, compliance and assurance team who looks for non-compliance and suspicious activities. All evidence is collected and stored, following HPO guidelines and available for internal audits.

3.1.2 Securing user data

In this section, we describe the precautions taken and methods used for securing user data both in transfer and at rest in the SoftLayer Object Store. As depicted in Figure 4, we can see that DNA data is encrypted prior to entering the system. It is worthwhile to describe exactly how and why this is done.

The security design of WGAP follows a Defense-in-Depth strategy [21] for storing user data. As mentioned above, user data is stored in the SoftLayer Object Store. For each user in the system, a new user account is created in the object store, creating different name spaces for each user in the system for file storage. Although the object store is publicly accessible, the access credentials are never forwarded to the user: every time a file is to be uploaded to the object store, a new authentication token must be requested from the system via the Representational State Transfer (REST) API. With this, we can guarantee that, should an account be compromised, only files that belong to that account would be accessible, and for at most for 24 hours, the maximum lifetime of an authentication token–as of the implementation of the service, it was not possible to customize the lifetime of such a token.

Together with name space segmentation, we also encrypt and decrypt files in-memory, in the same process that does network operations: before adding files to the object store, when uploading, and after retrieving them from the object store, when downloading. We do so by encrypting each file with the Advanced Encryption Standard (AES) algorithm in counter mode (CTR) and using a keyed-hash message authentication code (HMAC) based on the Secure Hash Algorithm in 256 output mode (SHA-256) for file authentication. The encryption is done in the following order: first, we compute the HMAC, then we encrypt the whole file together with the HMAC. Decryption is done first by decrypting the file, computing the HMAC of the output, then comparing the computed with the stored HMAC to check whether it is valid. On top of the random keys, for each file a random number used once (nonce) is generated and input to the AES-CTR algorithm. Therefore, even if the key for a file is compromised, a correct nonce is still needed to decrypt it. Both the key and nonce are stored in an encrypted DB2 database (DB2 v10.5 supports native encryption with the key being stored locally or in a remote key management service) and are transferred to client programs over TLS. In fact, all communications between the components of the system are encrypted with TLS, and these communications happen over a VLAN, segmented from the rest of the SoftLayer networks, with Vyatta [22] firewalls in these VLANs. The just-described encryption scheme is used in all patient cases. Hence, each patient case has a pair of keys that apply to its related files, and even if a malicious agent had access to user files, knowing the keys of one case would not allow that agent to decrypt files that belong to another case. HMAC and encryption work together to provide a secure service: the encryption prevents unauthorized users from *reading* the contents of the encrypted files. The HMAC allows the service to *verify data integrity*, and stop processing potentially malicious files.

3.1.3 Data submission client

One must keep in mind that the primary users of WGAP are medical doctors and cancer researchers. As such, the design decision was for the solution to not require changes to the client machine of the users. Users are not expected to have the technical knowledge to make changes to their systems, nor are they expected to have administrator privileges to make system-wide changes. Our most basic assumption was that users would have access to a web browser--after all, the service is a web-based one.

An argument could be made that the Web Cryptography Standard API (WebCrypto) could be used for client-side encryption, but during the time the service was conceived, such a specification was still in a draft state. Therefore, a pure browser-based solution would be unattainable; the browser sandboxing features disallow JavaScript from reading system paths, and a direct read followed by encryption would be impossible.

The approach taken was to use Java applets for delivering the encryption code to users. This decision was made based on the market share of the Java virtual machine on enterprise environments and desktop computing for operations such as home banking, and due to the fact that Java is a widely used language. Therefore, there is consolidated expertise and availability of libraries. The argument for choosing Java applets was also made stronger due to its support for signing (and verifying signatures of) binaries, and the possibility of caching applets locally to remove the impact of downloading the submission client every time the service was accessed.

3.1.4 Deployment challenges

One of the biggest challenges of deploying such a system using Java applets is that one is unable to use the standard Java Cryptography Extensions (JCE) API for using AES-256 without making changes to the client systems of the users. This is due to the fact that enabling so-called strong encryption in Java requires users to download special files from the Java Virtual Machine (JVM) provider and replace system files with those. This goes against our requirement of modifying the user system as little as possible, and was solved by shipping an alternative encryption library [23]. Unfortunately, it was not evaluated under the Federal Information Processing Standard (FIPS) 140-2 for cryptographic libraries, so it did not meet all of our requirements.

Alongside BouncyCastle, a collection of APIs used in cryptography, our own internal libraries had to be shipped with the applet. The problem with indiscriminately including libraries and their dependencies for deployment is that the resulting deployment file gets big. This would not be a problem had the Java file download process been managed by the web browser. This is not the case, and while the JVM is downloading the files, the browser user interface has to wait for the files to download, effectively hanging the whole browser. To reduce the impact and client impact due to excessive delays [24], we used a combination of hand-optimization, automatic file compression and obfuscation by post-processing the output of the Java build process with the ProGuard [25] obfuscation tool.

4 Related work

IBM, Microsoft, and Amazon have published white papers and use cases at their Web sites to help their Infrastructure-as-a-Service clients have better guidance on implementing secure and compliant solutions [26–28]. These articles help people, new to cloud computing, in implementing such solutions. However, such white papers have not discussed managing the various key aspects of the lifecycle of such sensitive applications on the cloud in detail.

The authors of [29] discuss the suitability of using cloud computing for biomedical research consortia, presenting technical trade-offs that must be considered, security issues that might arise when using the cloud, and presents applications the authors deem as good targets for the cloud and targets that are considered poor.

The authors of [30] present a series of challenges public clouds face with regard to security. In that paper, its authors generalize the risks associated with all cloud business models. IaaS certainly introduces novel security issues related with virtualization technology, which is the focus of the application described in this paper. However, SaaS and PaaS models should not be discredited for the risks they face – for example refer to [31]. To mitigate the risk of sharing user information, we explicitly designed the service to not gather PII, and segmented the account namespace to prevent potential malicious cloud providers from tampering with user data.

Van Gorp and Comuzzi present in [32] an independent system for the management of personal health data that empowers the patient to present her information to caregivers with different levels of granularity. The system in question uses Virtual Machines in an IaaS cloud computing setting. In particular, the personalized medicine use case implemented for validating the research mentions an approach similar to the one taken in Watson Genomic Analytics Prototype: Virtual Machines that do processing do not have access to the Internet, but have their databases updated frequently for improving the quality and reach of proposed treatments.

The authors of [33] use ciphertext policy attribute-based encryption (CP-ABE) to implement a personal health record (PHR) cloud to provide privacy protection and fine-grained access control to patient data, allowing different actors to have access to different parts of the health record data. The main concern of that work was to protect a the privacy of patients by encrypting PHR data, and giving a patient control over which parts of the PHR data can be accessed. That work differs from ours in which it provides a means for general cloud-based PHR storage, while the work here described deals with a specific, focused cloud-based service that uses identity-based encryption for protecting patient data.

The authors of [34] state that "safety of data in the cloud is a key consumer concern, particularly for financial and health data", and discusses the case for taking privacy into account when designing cloud services, and its importance. Main privacy risks and privacy requirements are listed, with guidelines listed as well.

The authors of [35] present an auto-scaling mechanism that autonomously allocates virtual resources on an on-demand basis and that can be applied to healthcare applications deployed on the cloud with deadline-critical real-time clients, such as wireless electrocardiogram sensors. The nature of the system there described is different from ours, though. There, clients with real-time requirements access the servers, while in this paper, we do not have such hard time requirements.

5 Conclusions

One of the primary problems in cloud computing today is how to manage sensitive workloads running on the cloud. Sensitive workloads handle data that are often under a regulatory compliance regime such as healthcare and personal data. Therefore, management of such solutions in the public cloud needs to address the problem of development, deployment, security, privacy, and compliance in a holistic manner. While the literature has publications addressing individual aspects, our paper provides a comprehensive treatment of all these issues in a holistic manner, and has been based on our practical experience spanning more than a year of development, deployment, and assurance of security and compliance of a world-class solution, Watson Genomic Analytics Prototype. We have discussed how compliance regimes impact such solutions, what security controls need to be in place, the kind of data protection techniques to be employed, privacy techniques to be used and how to address network and system security. We have established a relationship between these aspects and how development and deployment of the solution are carried out. In the future, we plan to address some of the research challenges regarding cloud platforms and how to assure data protection and auditing.

6 REFERENCES

1. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4,pp. 50-58, 2010.

2. F. Gens, Worldwide and regional public IT cloud services 2014–2018 forecast – IDC Market Analysis. Document #251730. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=251730.

3. G. Petri, Three Factors Will Significantly Impact Enterprise cloud Use in the Near to Midterm Future. [Online]. Available: https://www.gartner.com/doc/2568317/factors-significantly-impact-enterprise-cloud.

4. C. Humer, Anthem says at least 8.8 million non-customers could be victims in data hack.[Online]. Available: http://www.reuters.com/article/2015/02/24/us-anthem-cybersecurity-idUSKBN0LS2CS20150224.

5. S. Sachin and S. Banerjee, Target nears $20 million MasterCard data breach settlement: WSJ. [Online]. Available: http://www.reuters.com/article/2015/04/14/us-target-settlement-idUSKBN0N52LV20150414.

6. R. Bao, L. Huang, J. Andrade, W. Tan, W. A. Kibbe, H. Jiang, and G. Feng. "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing." *Cancer Informatics*, vol. 13, no. 2 pp. 67-82, 2014

7. S. Faroult and P. Robson. *The art of SQL*. Sebastopol, CA: O'Reilly Media, Inc., 2006.

8. S. Tiwari, *Professional NoSQL*. Indianapolis, IN: John Wiley & Sons, 2011.

9. Health Insurance Portability and Accountability Act of 1996, [Online]. Available: http://www.hhs.gov/ocr/privacy/.

10. Family Educational Rights and Privacy Act. [Online]. Available: http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

11. PCI Security Standards Council, Payment Card Industry Data Security Standard v3.0, [Online]. Available: https://www.pcisecuritystandards.org/security_standards/documents.php.

12. D. Holtzman, J. Koenig, and T. LeSueur. HITECH & The cloud: Control and Accessibility of Data Downstream, [Online]. Available: http://csrc.nist.gov/news_events/hipaa-2013/presentations/day1/holtzman_david_koenig_james_lesueur_ted_day1_115_cloud_computing_vendor_assurance.pdf, Slides 5-6. May 21, 2013.

13. Office of Civil Rights, Summary of the HIPAA Privacy Rule. [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html.

14. IBM BlueMix, [Online]. Available: https://bluemix.net/.

15. D. Vogel, Top 10 HIPAA Data Breaches of 2014. [Online]. Available: https://www.datapipe.com/blog/2015/01/28/top-10-hipaa-data-breaches-of-2014/.

16. R. Kissel, M. Scholl, S. Skolochenko, and X. Li. *Guidelines for media sanitization(NIST Special Publication 800-88)*. Gaithersburg, MD: NIST, 2006.

17. T. J. Kobus III and G. S. Zeballos, 2015 International Compendium of Data Privacy Laws. [Online]. Available: www.bakerlaw.com/files/Uploads/Documents/Data%20Breach%20documents/International-Compendium-of-Data-Privacy-Laws.pdf.

18. PCI Security Standards Council, Payment Card Industry (PCI) Pin Transaction Security (PTS) Hardware Security Module Security Requirements v2.0, May 2012.

19. M. Mahalingam, D. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar, M. Bursell, and C. Wright, Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, RFC 7348. [Online]. Available: https://tools.ietf.org/html/rfc7348, August 2014.

20. IBM SoftLayer. [Online]. Available: http://www.softlayer.ibm.com.

21. M. R. Stytz, "Considering defense in depth for software applications," *Security & Privacy, IEEE* , vol.2, no.1, pp.72,75, 2004.

22. M. Iyer, R. Dutta, G. N. Rouskas, I. Baldine, "Network Virtualization: Technologies, Perspectives, and Frontiers," *Journal of Lightwave Technology*, vol. 31, no. 4, pp. 523-537, 2013.

23. Legion of the Bouncy Castle. [Online]. Available: http://www.bouncycastle.org.

24. D. F. Galletta, R. Henry, S. McCoy, and Peter Polak. "Web site delays: How tolerant are users?,"*Journal of the Association for Information Systems*, vol. 5, no. 1, pp. 1-28, 2004.

25. E. Lafortune, ProGuard Java class file shrinker, optimizer, obfuscator, and preverifier. [Online]. Available: http://proguard.sourceforge.net/.

26. Amazon, Creating HIPAA-Compliant Medical Data Applications With AWS. [Online]. Available: http://aws.amazon.com/about-aws/whats-new/2009/04/06/whitepaper-hipaa/.

27. M. Ayad, H. Rodriguez, J. Squire. Addressing HIPAA Security and Privacy Requirements in the Microsoft cloud. [Online]. Available: http://download.microsoft.com/download/8/4/8/8483B6A9-1865-4D17-B6F1-5B66D5C29B10/Windows%20Azure%20HIPAA%20Implementation%20Guidance.pdf.

28. IBM, IBM Security Services Client Reference Guide. [Online]. Available: http://ibm.biz/security-services-reference.

29. A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester, and P. Reynolds. "Cloud computing: a new business paradigm for biomedical information sharing," *Journal of biomedical informatics*, vol. 43, no. 2, pp. 342-353, 2010.

30. K. Ren, C. Wang, Q. Wang, "Security Challenges for the Public cloud," *IEEE Internet Computing*, vol. 16, no. 1, pp. 69-73, 2012.

31. D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M. Freire, P. R. M. Inácio, "Security Issues in Cloud Environments: a Survey," *International Journal of Information Security*, vol. 13, no. 2, pp. 113-170, 2014. 32. P. V. Gorp, M. Comuzzi, "Lifelong Personal Health Data and Application Software via Virtual Machines in the cloud", *IEEE Journal of Biomedical and Health Informatics*, vol. 18 no. 1. pp 36-45, 2014.

33. C. Wang, X. Liu; W. Li, "Implementing a Personal Health Record cloud Platform Using Ciphertext-Policy Attribute-Based Encryption," in *Proc. Intelligent Networking and Collaborative Systems*, Bucharest, Romania, 2012, vol. 4, pp.8,14.

34. S. Pearson, "Taking Account of Privacy when Designing cloud Computing Services," in *Proc. Software Engineering Challenges of Cloud Computing*, Vancouver, Canada, 2009, pp.44, 52.

35. A. M. K. Cheng, J. Baek, M. Jo, H.-H. Chen, "An auto-scaling mechanism for virtual resources to support mobile, pervasive, real-time healthcare applications in cloud computing," *IEEE Network*, vol. 27, no. 5, pp. 62-68, 2013.