Journal of Cloud Computing
a SpringerOpen Journal

RESEARCH Open Access

Implementation of a secure genome sequence search platform on public cloud-leveraging open source solutions

Vikas Saxena, Shyam Kumar Doddavula* and Akansha Jain

* Correspondence: shyamkumar_d@infosys.com
Infosys Labs, Infosys Limited, Bangalore, India

Abstract

With looming patent cliffs, resulting in no patent protections for several block buster drugs, several Life sciences organizations are looking at ways to reduce the costs of drug discovery. They are looking to change business models from having all drug discovery activities being done in-house to a more economical collaborative innovation model by forming ecosystems through consortiums and alliances with several other partners to collaborate especially in the pre-competitive areas of drug discovery. They are considering leveraging cloud computing platforms to create the collaborative drug discovery platforms needed to support these new drug discovery models. Another area of focus is to improve the success rate of drug discovery by creating more complex computer models and performing more data intensive simulations. Next generation sequence sequencers are also providing unprecedented amounts of data to work with. Cloud computing has proven to be scalable and capable of meeting the computation needs in life sciences domain but a key inhibitor has been security concerns. This paper is an extension of an earlier paper we had written that describes how to leverage a public cloud to build a scalable genome sequence search platform to enable secure collaboration among multiple partners. This paper describes a few additional techniques and open source solutions that can be leveraged to address security concerns while leveraging public cloud platforms for collaborative drug discovery activities.

Keywords: Genome sequence search, BLAST, Ensembl, Cloud security, Encryption, Federated identity, SAML, OpenVPN, ACL, Hadoop, Hadoop security

Introduction

Several block buster drugs will go off patent protection by 2015 [1]. This means several life sciences companies will have cost pressures and so will be looking at ways to reduce costs. Current drug discovery business models involve significant redundancies among the various life sciences organizations. They all duplicate effort is the early stages of drug discovery which are considered pre-competitive and non-differentiating. With the increasing cost pressures, several life sciences organizations have come together through industry alliances like Pistoia Alliance [2] to look at ways to increase collaboration among the various players, in the pre-competitive areas of drug discovery to reduce costs. As part of one such initiative, the members of the Pistoia Alliance [2]

have mutually agreed to define and document the standards for a *secured sequence service*. Ensembl [3] is chosen as one of the relevant public sequence services to be made available as a secure service that can be used by multiple life sciences organizations instead of duplicating the efforts in deploying it in-house, maintaining it and keeping it in sync with the constant new releases. Ensembl is a joint project between EMBL-EBI and the Sanger Centre. Ensembl produces genome databases for vertebrates and other eukaryotic species and makes them available for free on over the internet and also enables search leveraging BLAST [4] algorithm. Though Ensembl is available for free over the internet, several life sciences organizations are not able to use it because of security concerns. Ensembl doesn't currently offer adequate security for the search operations so there are concerns about competitors could eavesdrop on the sequence searches being performed by an organization's scientists and use that to infer several confidential and proprietary information. Another challenge with use of Ensembl is lack of SLAs around performance and support. The response times are not predictable and depend on the number of users currently performing searches and the complexity of the search operations they are performing. This can result in scientists wasting their precious time. As a result most life sciences organizations resort to hosting the Ensembl applications and the datasets in-house but that results in increased costs as each organization has to invest separately on the infrastructure and people needed to keep it operational. This problem is not jus with Ensembl, there are several other such popular life sciences applications and datasets that are available in public domain but there are hosted and managed internally by most organizations resulting in redundancies. Pistoia Alliance members therefore wanted a solution that offers a shared platform that is secure and offers several such applications that are used in the pre-competitive activities of drug discovery on-demand with predictable SLAs. They wanted to evaluate public cloud platforms for this with Ensembl as the pilot application to be hosted and made available on-demand with adequate security. Infosys is one amongst the IT vendors that have been invited to implement a proof of concept for developing a secured sequence search solution. This paper and the one we published earlier [5] are based on our experiences in implementing the proof of concept.

Our earlier paper [6] explained how to implement a Secure Next Generation Sequence Services business cloud platform that is highly scalable and can be shared by multiple life sciences companies securely. The earlier paper described a few techniques for securing web applications and data hosted on a public cloud such as Amazon AWS leveraging open source security components and how they have been leveraged to secure the Ensembl solution. In this paper we expand on the earlier paper and describe a few additional security solutions and best practices that can be leveraged in offering a secure life sciences business cloud platform.

Background and related works

Life sciences organizations have been forming several alliances and consortiums to enable collaboration in the pre-competitive areas of drug discovery. The Pistoia Alliance (www.pistoiaalliance.org), Open Source Drug Discovery (www.osdd.net) , the European Bioinformatics Institute (EBI) Industry Program, the Predictive Safety Testing Consortium, Sage Bionetworks, Innovative Medicines Initiative (http://www.imi.europa.eu) are

examples of such alliances. In the paper - Implementation of a Scalable Next Generation Sequencing Business Cloud Platform [7], the authors describe how to address the scalability of a Next Generation Sequencing solution and a strategy to port a pre-configured Sequence Search application such as BLAST [4] onto a scalable storage and processing framework like Hadoop framework to address scalability and performance concerns. Our previous paper [5] was an extension to the same and focused on several security aspects of such business cloud architecture. It gave an overview of Amazon AWS Cloud, what cloud security is about and a few open source security products like OpenAM, Truecrypt etc and then described how to implement firewalls in AWS cloud and techniques to address security of data at rest and in transit and how to implement Federated Identity and form Circle-Of-trusts to enable collaboration using OpenAM and WS-Federation Standards. This paper is an extension to previous papers [5,7] so, the authors assume that the reader has gone through those to get the understanding of the context, the solution overview and the solution components used.

In this paper we describe a few additional techniques that address a few more security aspects of business cloud architecture. This paper aims at achieving the following:

1. Describe how to secure a Hadoop Cluster to prevent impersonation and unauthorized access.
2. Provide controlled access to data residing in Hadoop Cluster through Access Controlled Lists (ACLs).
3. Demonstrate how to implement a virtual private cloud with secured access to the machines on a public cloud through virtual private networks (VPNs).

Introduction to hadoop and security concerns

Hadoop is a java based open source framework that supports data intensive distributed applications. Hadoop provides a highly scalable and fault tolerant distributed data storage and processing framework. It is designed to scale up from single server to a cluster containing thousands of servers, leveraging the compute and storage capabilities of the individual servers. Hadoop is used to store the genome databases and to implement a parallelized BLAST search. The reasons why Hadoop was chosen for this PoC were:

1. It is free and open source.
2. Hadoop Distributed File System (HDFS) offer a highly scalable storage solution that has been proven to scale to petabytes of data. So, it can be used to store the genome data.
3. HDFS besides being scalable also offers fault tolerance by replicating data across multiple machines there by addressing availability and reliability concerns.
4. Hadoop Map Reduce framework offers a highly scalable distributed processing solution that leverages several commodity servers to parallelize processing. It is therefore offer a good solution to parallelize BLAST search.

Another paper [7] describes how BLAST search has been parallelized using Hadoop.

The following are a few concerns with respect to security with the default Hadoop settings and configuration options:

1. *Impersonation:* Hadoop does not have any inbuilt authentication mechanism of its own. Hence a malicious user can easily impersonate as the superuser or any valid user of Hadoop Cluster and can access the HDFS cluster from any machine.
2. *Default permissions in HDFS file system:* The default permissions on HDFS file system are *-rw-r–r–* for files and *drwxr-xr-x* for directories. This gives sufficient privileges to users to view other users' files and directories. In some case such as a shared infrastructure between competitors, this may not be desired.
3. *Direct Access to Data Blocks:* DataNodes do not enforce any access control on access made to the data blocks they are storing by default. This allows an unauthorized user to read a data block by supplying the blockid. This also allows an unauthorized user to write arbitrary data to data blocks on the DataNodes.

Introduction to openVPN

OpenVPN is an open source solution that enables the implementation of virtual private network (VPN) for creating secure point-to-point connections in routed or bridged configurations for secured access to remote machines. It makes use of a custom security protocol that employs SSL/TLS protocol suite for key exchange and is capable of traversing firewalls and network address translators. It supports authentication amongst peers by means of a pre-shared secret certificates, key, or userid/password. When used in a multi client–server scenario, it allows server to issue an authentication certificate for each client. It utilizes OpenSSL encryption library and SSLv3/TLSv1 protocol suite, to provide multiple features pertaining to security and control.

Introduction to Access Control List(ACL)-based security model

In an ACL based-security model, ACLs are defined and enforced to control the access of subjects to objects. The term subject can refer to a real user, system process or daemon while the term object can refer to the resource(s) that a subject tries to access such as files, directories, system ports etc. When a subject requests an operational access on an object, the Access Control Manager (ACM) checks the ACL data for a matching ACL to decide whether the requested operation is permitted or not.

Given this background of Amazon AWS Cloud, OpenAM, Truecrypt, Hadoop Security, ACL-based security Model and OpenVPN we describe the security techniques in the next section.

Solution approach

In this section we describe how to implement a secure collaboration platform using a public cloud (Amazon AWS Cloud) and then describe how to migrate a popular genome sequence search application called Basic Local Alignment Search Tool (BLAST) and provide web based secure access to collaborating groups of life sciences organizations.

Next generation secure sequence search business cloud platform – solution design

A high level conceptual view of the Next Gen Business Cloud Platform is show below (Figure 1).

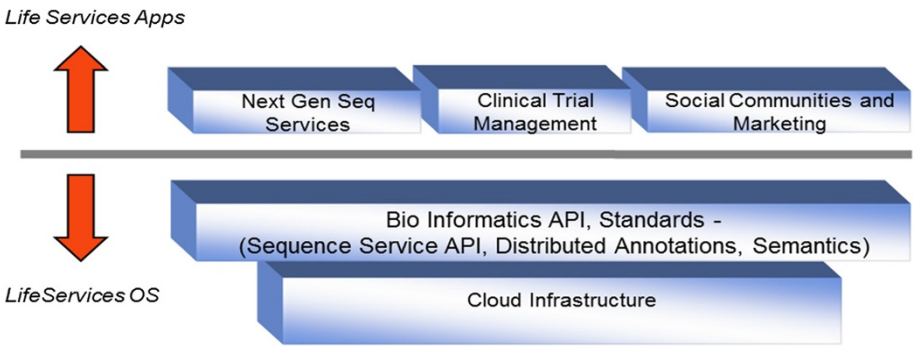

Figure 1 Life Scinces Business Cloud Platform – Conceptual View. The figure provides a conceptual view of a life sciences business cloud platform.

The solution consists of a 3 layers. At the bottom is Cloud infrastructure layer that provides scalable compute and storage capabilities. Over that is a Life Sciences Services platform which provides domain specific services like sequence search, distributed annotations management etc with API based on standards. The next layer is domain specific applications supporting business processes like target management, genomics, clinical trial management, drug sales etc in the various business areas like drug discovery, development and marketing.

A prototype of a scalable elastic, pay-per-use genome sequence search platform based on BLAST algorithm over public genome datasets with a web based genome browser application on Amazon public cloud infrastructure is developed to validate the concept. Hadoop was chosen as a scalable processing framework and BLAST processing was parallelized which was described in our earlier paper [7]. Ensembl is a popular genome sequence search application and it was used for the prototype. Our earlier paper [5] describes several security concerns with using public cloud infrastructure and some of the techniques we used to address those concerns.

The figure below shows various security layers of the "Next Generation Secure Sequencing Business Cloud Platform" solution:

To address concerns with network intrusions, a network firewall is used. To ensure security of data in transit while using web interfaces, Secure Socket Layer (SSL) based solution is used. Federated Identity solution based on OpenAM is used to provide Single-SignOn and enable creation of CirclesOfTrust for collaborations (Figure 2).

The components in red color have been described in our previous paper [5]. The components in blue color are the additions that will be described below.

The following components have been added to the Solution Design:

1. *Virtual Private Network (VPN):* A Virtual Private Network provides secure access to instances on Amazon AWS Cloud especially for scenarios where they are accessed through mechanisms that not based on HTTP. A secured direct access may be needed for the desktop based applications trying to acccess the services or data residing on Hadoop Cluster in the Amazon Cloud. Additionally, while uploading private data to Hadoop Cluster, one may again need a secured access to Hadoop Cluster. Also for maintainence, troubleshooting or upgradation

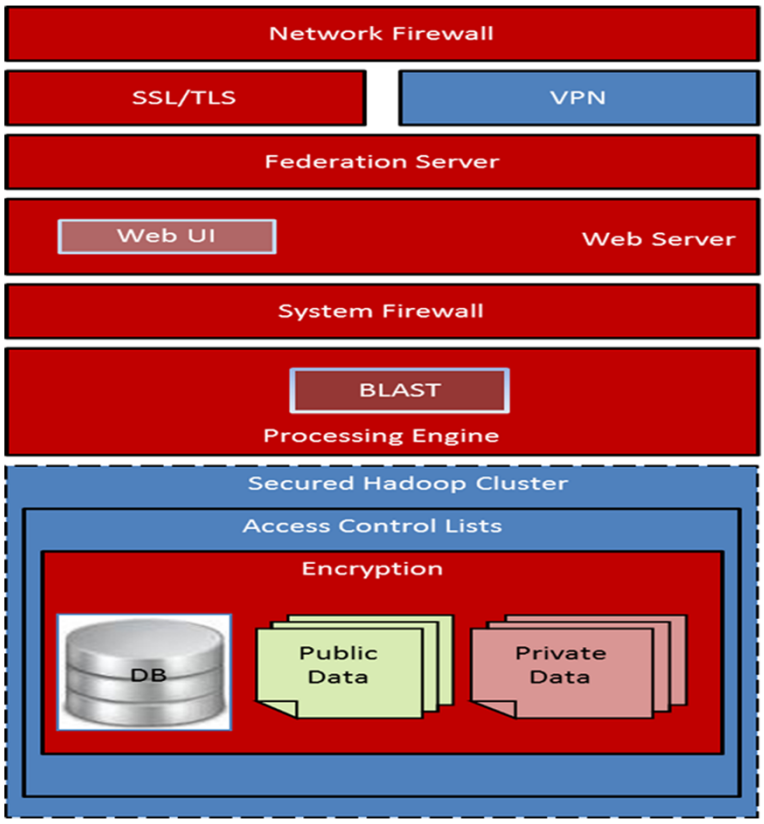

Figure 2 Next Generation Secure Sequence Search Business Cloud Platform – Security Solution Design Diagram. The figure describes various components of the Security solution design implemented for the solution. The components in grey have already been covered in original paper [5] while the colored components have been discussed in this chapter.

purposes one may need a secured acccess to Amazon instances. This helps in achieving security of data in transit

2. *Secured Hadoop Cluster:* Hadoop Cluster is used to store the genome data so, there is a need protect the data from unauthorized access from malacious users. This is required to take care of Hadoop related security issues such as impersonation (section II.A.1), unrestrictive file permissions on HDFS (section II. A.2) and direct access to data blocks (section II.A.3).
3. *Access Control Lists(ACLs)*: ACLs add another level of security for data at rest from unauthorized access (section II.C)

The subsequent sections explain the deployment architecture and implementation of various components described above.

Next generation secure sequence search business cloud platform – deployment architecture

The figure below gives an overview of the deployment architecture of the solution (Figure 3):

Each Amazon Elastic Compute Cloud (EC2) node provisioned has equivalent of 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform.

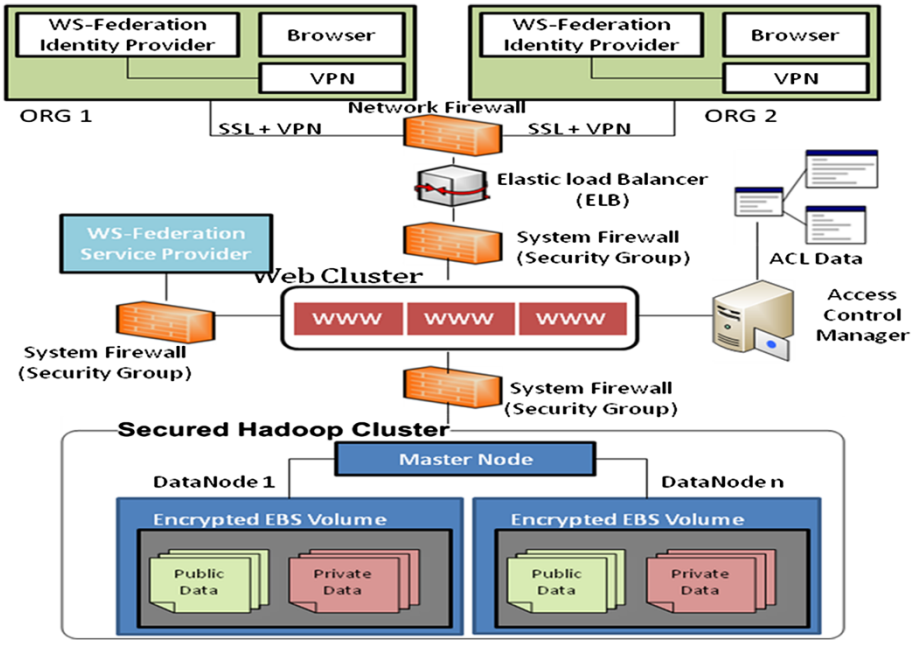

Figure 3 Deployment Architecture for Next Generation Secure Sequence Search Business Cloud Platform. The figure describes the deployment architecture for the solution.

The deployment architecture consists of the following Amazon EC2 instances:

1. *Web Clusters:* The VM has Apache web server and is load-balanced through the Amazon ELB.
2. *Federation Server*: This node runs OpenAM server deployed over Apache Tomcat. This machine additionally runs an OpenDS LDAP server which serves as OpenAM user store.
3. *Hadoop Cluster:* The cluster contains a master node and two datanodes.

Securing hadoop cluster

In order to address security issues with Hadoop default settings described earlier, the following security measures were implemented.

Impersonation

Prior to version 0.20, Hadoop had no in-built implementation of any authentication mechanism. It trusted the underlying operating system for user's authentication and used `whoami` utility to identify users , and `bash –c groups` for groups. This was the weakest link with respect to security as one can write his/her own *whoami* script or *groups* script and add it in his/her path to impersonate someone else including super user. This is a major threat.From version 0.20 onwards Hadoop supports Kerberos protocol for authentication as a security measure against impersonation. The complete step by step procedure for securing a Hadoop Cluster is described in Cloudera Security Guide [6].

Additionally, Single Sign-On (SSO) can be used to avoid manual ticket generation. The user can obtain his Kerberos ticket while logging in into the system. This reduces

the need to generate a Kerberos ticket manually by firing a command from the shell. For ease in management of user credentials, integration of Identity Provider and Kerberos authentication system is recommended. This can be done by two ways:

a) One can run a Kerberos KDC and a realm local to the cluster, create all service principals in this realm and then set up one-way cross-realm trust from this realm to the Active Directory realm. If this approach is used, there is no need to create any additional service principals in Active Directory and Active Directory principals (users) can be authenticated against Hadoop.
b) Alternatively, one can use Active Directory as the Kerberos KDC, create all service principals in the Active Directory itself and configure Hadoop/Linux machines as Kerberos clients to use Active Directory directly.

Cloudera provides an exhaustive doumentation [8] for both the approaches.
We used approach two because of the following reasons:

1. It is easy to configure.
2. There is no overhead of maintaining master and slaves KDC (as compared to approach 1).
3. Maintenance of user accounts becomes quite easier.

The step by step documentation for integration of Active Directory and Hadoop Security using Approach 2 is already provided by cloudera [8]. The documentation does not cover the procedure to import Active Directory UPNs to Linux Box This can be done by:

1. Using Samba and Winbind : Ubuntu wiki provides a complete documentation [9] for the same.
2. Using LDAP: There is a blog [10] by Scott Lowe which describes the steps.

The work flow for Kerberos integrated security mechanism is shown below (Figure 4):

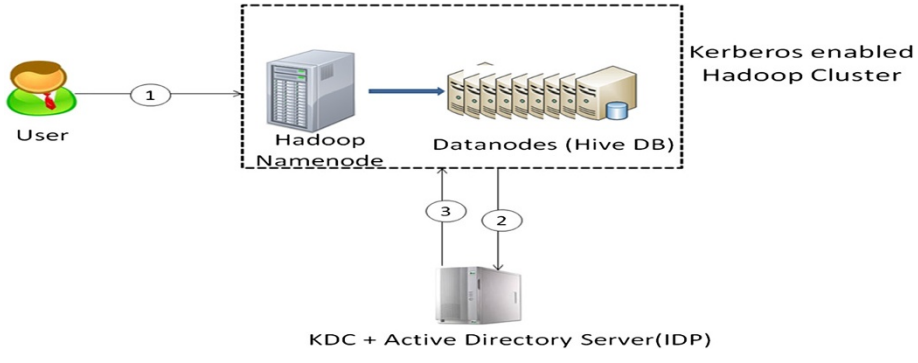

Figure 4 Workflow for Kerberos enabled Hadoop Cluster. The figure decribes the step by step procedure for authentication in a Kerberos enabled Hadoop Cluster.

Each of the machines on Hadoop Cluster is preconfigured to accept Kerberos based login. The authentication process occurs in three steps:

1) User tries to authenticate to the machine by supplying his/her credentials.
2) The machine contacts the Kerberos KDC (in our case the Active Directory server of the organization) for verification of credentials.
3) The KDC verifies the credentials and issues a Kerberos ticket to the user.

The user with the Kerberos ticket granted by the Kerberos KDC can access Hadoop services till his Kerberos ticket is valid.

For restrictive permissions on hdfs file system

The default permissions on HDFS file system are -rw-r–r– for files and drwxr-xr-x for directories. This gives sufficient privileges to users to view other users' files and directories. In some case such as a shared infrastructure between competitors, this may not be desired. To check this, the property dfs.umaskmode nedds to be added to hdfs-site.xml file. The value is to be set as per your requirements, for our case we set it as 700 which imparts -rw-——— permissions for files and drwx--——— permissions for directories (Table 1).

For controlling direct access to data blocks

For controlling direct access to data blocks, the following properties need to be added to *hdfs-site.xml* file (Table 2).

Implementation of a Virtual Private Cloud using OpenVPN

Another key requirement for the Proof of Concept (PoC) was to provide a secured remote access mechanism to the machines on the Amazon AWS Cloud including the Hadoop Cluster which contains public as well as company specific confidential private data from the client organization's data centers and their own private clouds. To ensure secured access to the Amazon instances, a VPN tunnel was created between the client machines and the Amazon instance. The workflow diagram for the implementation is shown below (Figure 5):

The process involves the following steps:

1) The user authenticates to the VPN access gateway using web interface or OpenVPN Connect Client.
2) The request reaches the OpenVPN Access Server.
3) The OpenVPN server queries the authentication source (ADFS Server).
4) On successful authentication, a VPN tunnel is created between the infrastructure on Amazon Cloud and the User's machine thus creating a Virtual Private Cloud.

Table 1 Properties to restirict permissions on HDFS filesystem

Property	Value to be set	Description
dfs.umaskmode	<depends on requirements>	umask value for HDFS file system

The table contains list of properties to restrict permissions of files and directories on HDFS file system.

Table 2 Properties to control direct acess to data blocks

Property	Value to be set	Description
dfs.block.access.token.enable	true	If "true", access tokens are used as capabilities for accessing datanodes. If "false", no access tokens are checked on accessing datanodes.
dfs.block.access.key. update.interval	Depends upon your requirement, default is 600	Interval in minutes at which namenode updates its access keys.
dfs.block.access.token.lifetime	Depends upon your requirement, default is 600	The lifetime of access tokens in minutes.

The table contains a list of properties required to check direct access to data blocks.

The VPN tunnel can be configured on Layer 2(Data Link Layer) and Layer 3 (Network Layer) of OSI model. The table below describes the pros and cons of both the implementation methods (Table 3):

In our solution VPN at OSI Layer 3 was implemented because:

1. It is more efficient and scalable
2. Provides better control over IP and routing configuration
3. Provides more granular control
4. Supported on all platforms such as Windows, Mac and Linux.

Implementation of ACLs-based Security Model

In an ACL-based security model, lists of permissions termed as ACLs are enforced to control access of subjects over Objects. In an ACL-based security model there are three key players:

1. *Subject:* The term subject refers to a real user, system process or daemon.
2. *Object:* The term object in an ACL based security model refers to the resource(s) that a subject tries to access.
3. *ACL Server:* This is the server which contains the list of permissions that controls the access of a subject over an object.

The purpose of using this model in our PoC was to ensure that a user is able to access only the data that he/she is authorized to access. Thus, limiting a user to use only the public data and only that part of his/her organization's private data that he/she is authorized to access.

The key principles of this security model were:

1) The enforcement of ACLs was de-centralized so that there are multiple checks across various layers.
2) The decision logic of ACLs was centralized: The ACL data resided on a centralized ACL server that determines whether the user is authorized to access the selected data or not.

The key functional requirements related to merging public and private data and providing access control were

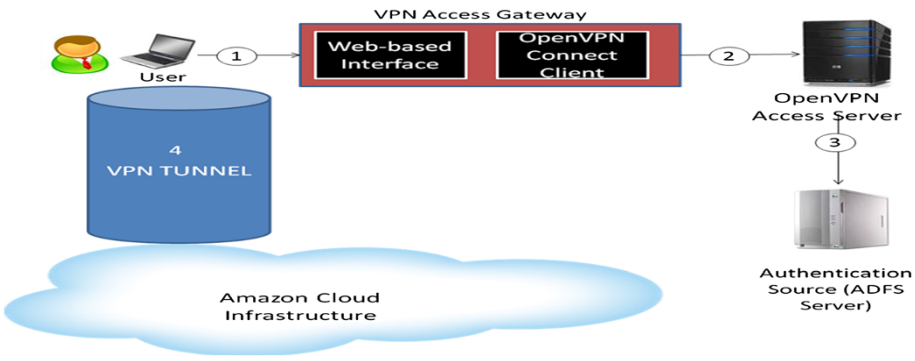

Figure 5 Formation of VPN Tunnel. The figure describes step by step procedure for formation of VPN Tunnel using OpenVPN for secured access to Amazon instances.

1) Enable user input a genome sequence and select public data stores and only those company specific private data stores that he/she is authorized to view, in order to find matching genomes
2) Present matching genome sequences with ranking that is combined across the public and private data stores
3) Enable providing company specific aliases to genome IDs so that it is easier to cross-link the genome annotations and information available in public domain with the confidential information available for the same genome but with a company specific internal ID.

Table 3 Pros and cons of implementations of VPN over various topologies

Topology	Pros	Cons
OSI Layer 2	• Most appropriate for smaller networks. • Easy to configure.	• Because LAN broadcasts are propagated to all VPN clients, this topology doesn't scale well to LANs that have a larger amount of broadcast traffic.
	• VPN clients receive their network • properties from the same DHCP server as machines that are • physically connected to the server-side LAN.	• Doesn't scale well with larger numbers of concurrent VPN clients. • Can only be used when the Access Server is connected to a LAN that provides DHCP services.
	• Works well with application-layer • protocols that depend on LAN broadcast resolution.	• Should only be used when the Access Server has a fixed IP address on a private LAN.
	• Can tunnel non-IP protocols.	• Currently only works with Windows Clients
OSI Layer 3	• More efficient and scalable. • Greater control over IP and routing configuration.	• More complex to configure. • Doesn't work well with application-layer protocols that depend on broadcast resolution.
	• Better fine-grained access control. • Works on all client platforms that support OpenVPN.	

The table provides a comparison between the implementation of VPN over Data Link layer and Network layer of OSI Model.

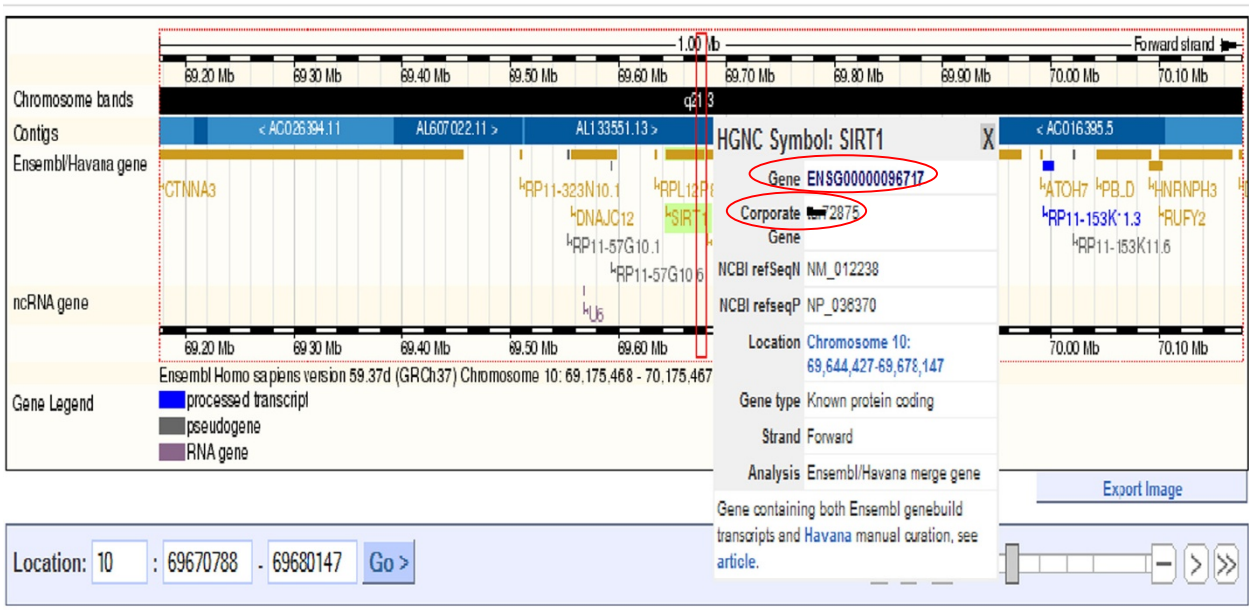

Figure 6 Linking private datasets with public datasets. The figure shows successful implementation of ACLs over the datasets to ensure that a user can only access the public data or the data pertaining to his/her organization only.

The following solution components were used

1) Access control enforcement was first done at the UI layer by enabling selection of only relevant genome data stores.
2) Next, there was again access control enforcement in the processing layer during the execution of the BLAST search to restrict search to only those data stores that the user is authorized to view. This is to ensure that if the processing layer is reached through any other channel, access controls are still enforced.
3) An Aliasing service was implemented that maintained a map of public gene IDs from National Center for Biotechnology Institute(NCBI) with those internal to each organization. The organization specific aliases that the user is entitled to shown in the search result along with the NCBI IDs for the public genome data as shown in the screenshot below (Figure 6).

Conclusion

The experience report along with the earlier reports [5,7] described how to use open source tools and solutions to create a secure drug discovery collaboration platform in a public cloud which provides several features like protection against possible Denial of Service attacks, security of data in transit, security of data at rest, implementation of Federated Identity and creation of secure Circle-Of-Trusts for collaboration, parallelization of processing using Hadoop and securing data stored in Hadoop, enabling secured access to Amazon AWS instances through VPNs, correlating public and private data sets and providing access controls. We believe our work can be leveraged by practitioners and researchers in life sciences domain who plan to use public clouds.

Our study, addressing a few security aspects through PoCs.These are first steps in building a foundational business cloud platform for collaborative drug discovery. In this experience report, we have described solutions for securing data stored in Hadoop, enabling secure access of genome data and services through non-HTTP channels also leveraging VPNs, enabling security of private genome data through application specific access control components. In future we look forward to expansion of this work to address other public clouds, addressing more cloud security vulnerabilities and enabling more drug discovery related applications on to the foundational platform and expanding the platform with additional layers of capabilities like high performance computing, collaborative workflows, social collaboration workspaces and addressing the security aspects of those components.

Abbreviations

BLAST: Basic Local Alignment Search Tool; ACL: Access Control List; VPN: Virtual Private Network; AWS: Amazon Web Service; NCBI: National Center for Biotechnology Institute; PoC: Proof of Concept.

Competing interests

The authors' declare that they have no competing interests.

Authors' contributions

The contributions of various authors are: SKD was responsible for the overall architecture and design of the solution. The PoC was executed under his technical guidance. VS contributed to the design and implementation of Hadoop Security solution, implementation of OpenVPN and implementation of federated identity solution. AJ contributed to the design and implementation of Hadoop Security solution. All authors read and approved the final manuscript.

Authors' information

Shyam Kumar Doddavula
Shyam works as a Principal Technology Architect and heads the Cloud Centre of Excellence at Infosys Ltd.

Akansha Jain
Akansha is a Technology Lead at Cloud Centre of Excellence at Infosys. She has around 5.5 years of experience in Java, Spring, Hibernate, Cloud Computing and Hadoop.
Vikas Saxena
Vikas is a Systems Engineer at Cloud Centre of Excellence at Infosys. He has around 2.5 years of experience in Web Security, Application Security, Network Security, Cloud Computing, Hadoop and Cloud Security.

Acknowledgment
Authors would like to thank the Pistoia Alliance Sequence Service team members – Simon, Claude, Ralf, Cary, John, and Nick for their help while defining the solution. Authors also wish to thank the Infosys sponsors and project team members – Arun, Subhro, Rajiv, Shobha, Kirti, Krutin, Ankit, and Nandhini.

Received: 30 January 2012 Accepted: 20 June 2012
Published: 19 July 2012

References

1. Barnes Michael R et al. Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery, http://www.nature.com/nrd/journal/v8/n9/abs/nrd2944.html.
2. Simon Thornber, Cary O'Donnell, Claus Stie Kallesøe, John Wise (2011) The Pistoia Alliance. The Sequence Service Project. GIT Laboratory Journal, Trends in Drug Discovery Business 1–3.
3. Ensembl Genome Browser, http://ensembl.org/index.html.
4. Basic Local Alignment Search Tool (NCBI), http://blast.ncbi.nlm.nih.gov/Blast.cgi.
5. Implementation of a Secure Genome Sequence Search Platform on Public Cloud: Leveraging Open Source Solutions by Shyam Kumar Doddavula and Vikas Saxena published at 2011 IEEE Third International Conference on Cloud Computing Technology and Science (IEEE cloudcom 2011).
6. Cloudera Security Guide, https://ccp.cloudera.com/display/CDHDOC/CDH3+Security+Guide.
7. Doddavula SK, Rani M, Sarkar S, Vachhani HR, Jain A, Kaushik M, Ghosh A (2011) Implementation of a Scalable Next Generation Sequencing Business Cloud Platform-An Experience Report, In: IEEE CLOUD: IEEE., pp S598–S605. ISBN 978-1-4577-0836-7.
8. Integrating Hadoop Security with Active Directory, https://ccp.cloudera.com/display/CDHDOC/Integrating+Hadoop+Security+with+Active+Directory.
9. Active Directory Winbind Howto, https://help.ubuntu.com/community/ActiveDirectoryWinbindHowto
10. Linux-AD Integration with Windows Server 2008, https://ccp.cloudera.com/display/CDHDOC/Integrating+Hadoop+Security+with+Active+Directory.

doi:10.1186/2192-113X-1-14
Cite this article as: Saxena *et al.*: Implementation of a secure genome sequence search platform on public cloud-leveraging open source solutions. *Journal of Cloud Computing: Advances, Systems and Applications* 2012 **1**:14.