PROCEEDINGS OF SCIENCE

eTRIKS Cloud Platform: a Platform for Knowledge Management

Pengfei Liu[1]
CC-IN2P3/CNRS
Lyon, France
E-mail: pliu@cc.in2p3.fr

Ghita Rahal
CC-IN2P3/CNRS
Lyon, France
E-mail: ghita.rahal@cc.in2p3.fr

In this paper, we present a cloud-based platform aiming at providing a core facility dedicated to eTRIKS, a European project for knowledge management in biomedical research.

Drug development and the study of populations for biomarker discovery require the treatment of large and diverse collections of heterogeneous data. The challenge is to offer a platform that satisfies the needs of security and availability of the data and offers the possibility of sharing common knowledge and tools.

With these constraints in mind, CC-IN2P3 has developed an original cloud-based storage-computing platform to meet the needs of the eTRIKS project. A standard eTRIKS platform contains three modules (i.e. Security module, Data curation and storage module, Data analysis and visualization module) which are composed of a set of virtual machines. This platform can be deployed for each new project joining the eTRIKS collaboration on the eTRIKS cloud or on a private cloud.

This paper will focus on the description of the eTRIKS platform and its utilization by the hosted projects.

International Symposium on Grids and Clouds (ISGC) 2015
15-20 March, 2015
Academia Sinica, Taipei, Taiwan

[1] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence. http://pos.sissa.it/

1.Introduction

Traditional approaches on biomedical research have very low success rate. Translational research[1] can improve the success rate significantly. The eTRIKS project aims at providing translational information and knowledge management[2] platform to facilitate translational research.

eTRIKS is a European project funded by Efpia[2] and IMI[3]. Currently IMI has forty seven projects on biomedical research. All these projects produce significant amounts of heterogeneous data every day. It is very difficult for all those projects to deal with these data by themselves. The data security is also a major issue, because clinical data and genomic data are of patients must be handled confidentially. To ensure data security and increase the efficiency of translational research, we must provide a translation information and knowledge management platform which is secure, flexible, scalable and easy to deploy.

With these constraints in mind, we designed and developed a cloud-based platform named the eTRIKS platform, which allows biomedical research projects to upload and store data. We also provide a data curation[3] environment and a database storage environment. We have dedicated virtual machines to host data analysis tools. The platform also provides authentication, authorization and access audit services. CC-IN2P3 provides also a private cloud (i.e. eTRIKS cloud) to host the eTRIKS platform for projects joining the eTRIKS collaboration. Currently two projects, abirisk[4] and oncotrack[5] are hosted in this cloud.

In this paper, we focus on the design and development of the eTRIKS platform. The data curation procedures and data analysis tools are not discussed in this paper.

2. The eTRIKS Platform

Figure 1 shows an overview of the eTRIKS platform. It contains three modules :

- Security module
- Data curation and storage module
- Data analysis and visualization module

The security module is designed to ensure the eTRIKS platform security. The data curation and storage module is designed to facilitate data upload, storage and curation. The data analysis and visualization module is designed to facilitate translational research. In the following sections, we discuss these modules in detail.

[2] European Federation of Pharmaceutical Industries and Associations - http://www.efpia.eu/

[3] Innovative Medicines Initiative - http://www.imi.europa.eu/

[4] http://www.abirisk.eu/

[5] http://www.oncotrack.eu/

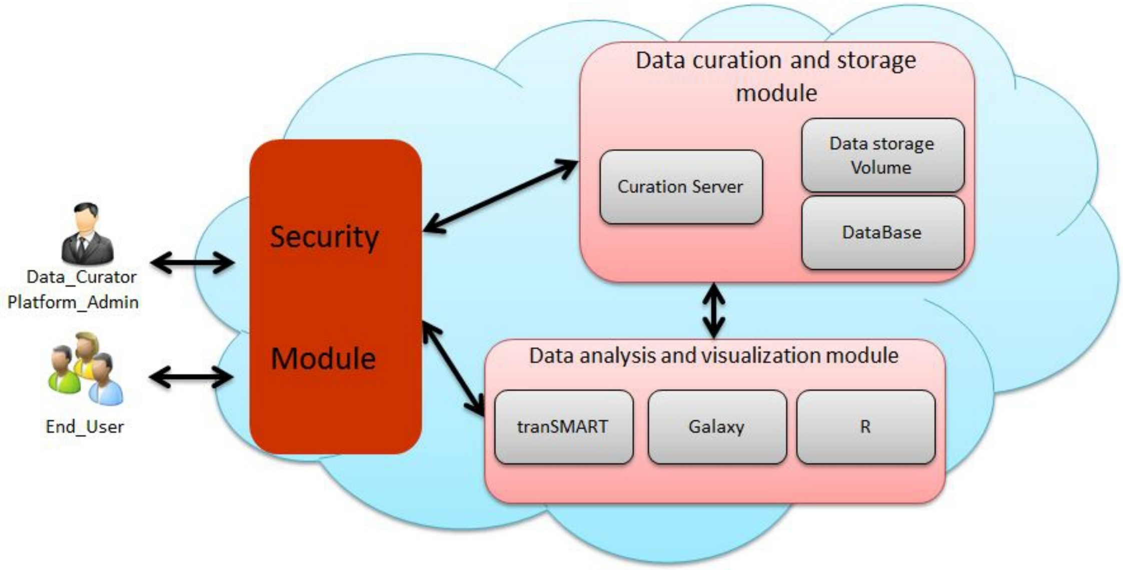

Figure 1 : eTRIKS platform overview

PoS(ISGC2015)001

2.1 The security module

During the years, many definitions of information security have been proposed. There is a continuous debate about the definition until now. But there is agreement that a security system[4,5] is the whole hardware and software solutions which can ensure the core security properties such as

- Authenticity,
- Confidentiality,
- Integrity,
- Non-repudiation,
- Privacy,
- Availability,
- Etc.

The key security requirement of the eTRIKS platform is to ensure the authenticity of user and platform, data confidentiality and data integrity. To fulfil these security requirements, we have implemented an authentication mechanism, an authorization mechanism and a log mechanism in eTRIKS platform.

2.1.1 Authentication mechanism

The authenticity of the eTRIKS platform is ensured by a certificate signed by a certification authority. All hosted projects in the eTRIKS cloud are covered by this certificate.

The authenticity of users are ensured by password and public key infrastructure[16]. Figure 2 shows the architecture of the authentication mechanism.

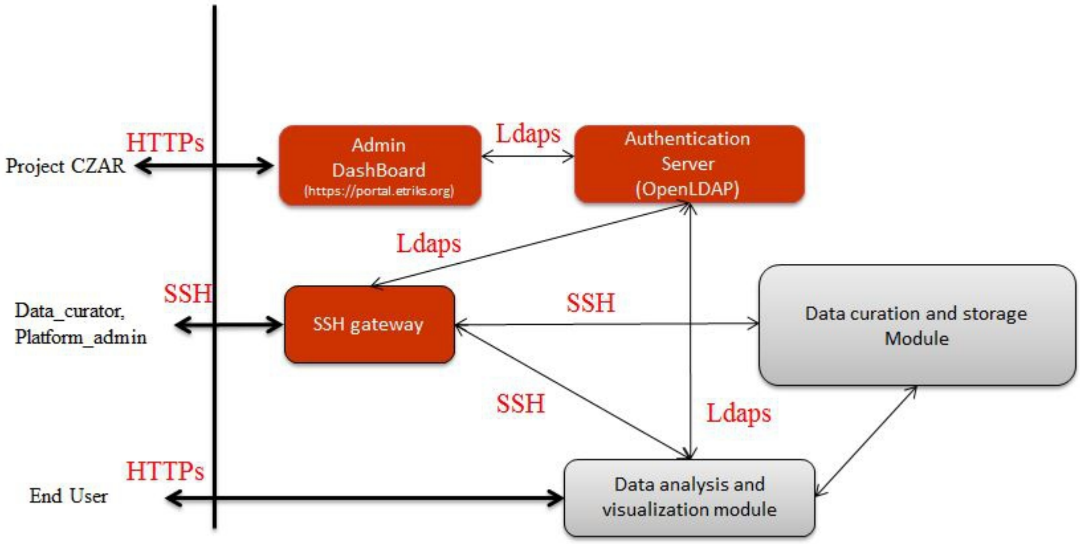
HTTPs
Project CZAR
Admin DashBoard (https://portal.etriks.org)
Ldaps
Authentication Server (OpenLDAP)
Ldaps
SSH
Data_curator, Platform_admin
SSH gateway
SSH
Data curation and storage Module
SSH
Ldaps
HTTPs
End User
Data analysis and visualization module

Figure 2 : Authentication mechanism architecture

In a typical eTRIKS platform, there are four kinds of users. The *end user* can only access data via data analysis and visualization tools. The *data curator* and *platform administrator* can access the whole platform via ssh gateway. The *project czar* can access the administration dashboard via a web interface. User accounts can be created via the administration dashboard by project czar. The authentication mechanism uses client-server architecture. The authentication server is implemented by openldap[6]. All the web interfaces and ssh gateway are authentication clients. For example, when a user wants to access data via a data analysis tool, the tool will check the user login and password via authentication server. If the user credentials are correct, the authentication server will send back a list of roles which the user has. Based on these rules, this tool can decide which data this user is granted access to.

We support different authentication clients for different services. For the virtual machines, we use Linux pluggable authentication modules (i.e. pam). For the administration dashboard, we developed a Java client by using JNDI. We also have authentication clients for data analysis tools. For example, in tranSMART[13], we use spring security ldap plugin[7], in galaxy[14], we use apache module mod_authnz_ldap. These clients can use ldaps or ldap with mod startTLS to connect to the authentication server. As a result, all the communications between authentication clients and the authentication server are encrypted.

PoS(ISGC2015)001

PoS(ISGC2015)001

2.1.2 Authorization mechanism

To ensure the confidentiality of the data hosted in the eTRIKS platform, we need to control access to the data. For this purpose, we implemented a role-based authorization mechanism. In section 2.1.1, we have seen that after successful authentication, the authentication server will send back a list of roles which the user has. In this section, we present how to determine what data a user can access given a specific role.

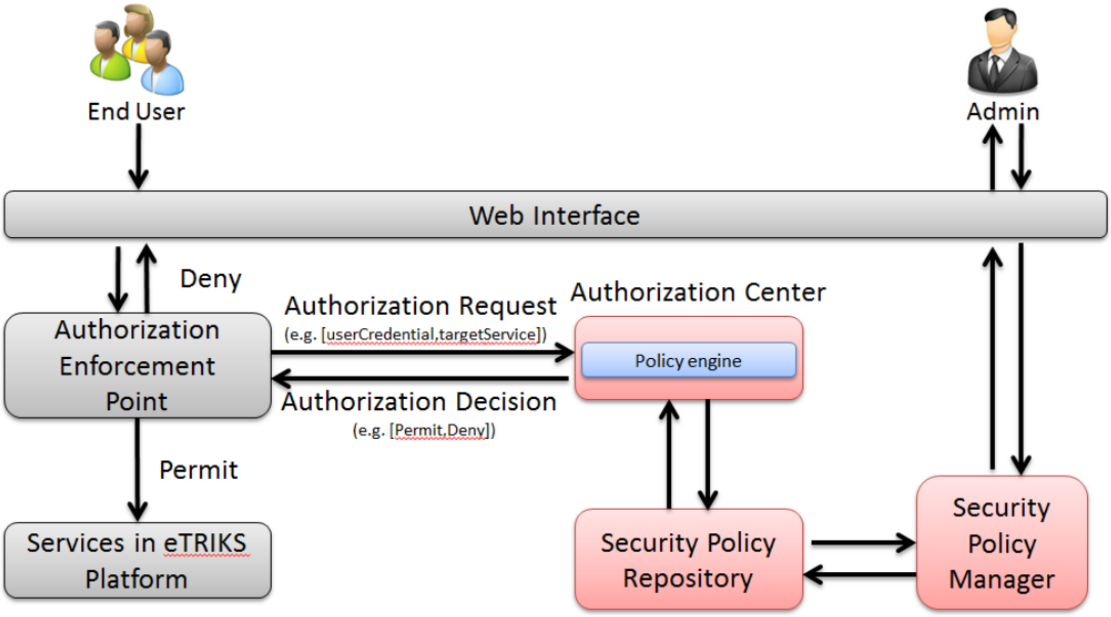

Figure 3 : Authorization mechanism architecture

To satisfy the security constraints of the eTRIKS platform, we need a security policy management system[11] to enforce these policies. Figure 3 shows the architecture of the authorization mechanism which can manage security policies and enforce them in the eTRIKS platform. To make the authorization mechanism more flexible, we use a client-server architecture to implement the authorization mechanism. The client side can send an authorization request to the authorization server. The authorization server can send back a decision based on the security policy which the admin user has specified.

The authorization server contains a security policy manager, a security policy repository and a policy engine. The security policy manager allows an admin user to specify security policy rules. The security policy repository stores security policy rules[6]. The policy engine is responsible for parsing the authorization request to a decision by comparing with security policy.

[6] A set of rules which define access control specification

The client side contains an authorization enforcement point. Whenever a user wants to access confidential data in the eTRIKS platform, the authorization enforcement point will intercept the request. Then the authorization enforcement point will form an authorization request. The authorization request contains user information (e.g. uid, role, etc.), actions which the user wants to perform (e.g. read, write, etc.) and the target object (e.g. data, services, etc.) which the user wants to access.

This architecture allows us to use different policy specification languages (e.g. XACML[8], SPL[9], Rei[10], etc.). In a standard eTRIKS platform, we use XACML as the policy specification language. Because XACML is an extension of XML which increases the expression power of the security policy. In a standard eTRIKS platform, the authorization server is accessible via restful web services. As a result, it is very simple to integrate new services into the eTRIKS platform authorization mechanism. To avoid conflict between security policies, we applied policy combing algorithm such as

- First applicable : The first applicable policy rule will be applied, if there are conflicts.
- Deny override : The policy with deny as decision will be applied, if there are conflicts.
- Permit override : The policy with permit as decision will be applied, if there are conflicts.

Figure 4 shows the policy combining mechanism architecture. The policy combining mechanism allows administrators to specify their own security policies for the services they manage without worrying policy conflicts.

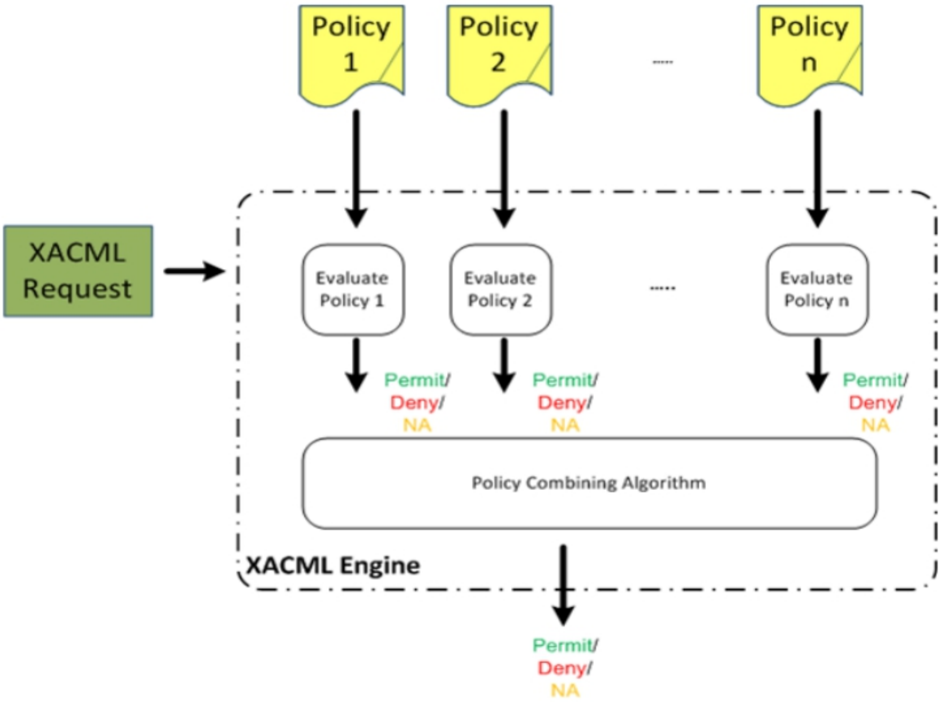

Figure 4 : Combining security policies

PoS(ISGC2015)001

In a standard eTRIKS platform, the policy engine is implemented by using WSO2-IS, which can be easily replaced by other XACML policy engine implementations.

2.1.2 Log mechanism

To ensure the integrity of the data, critical actions such as create, modify and delete must be restricted. All these critical actions which can modify data must be logged too. With these logs, the admin user can find out when the data is modified and who is responsible for that.

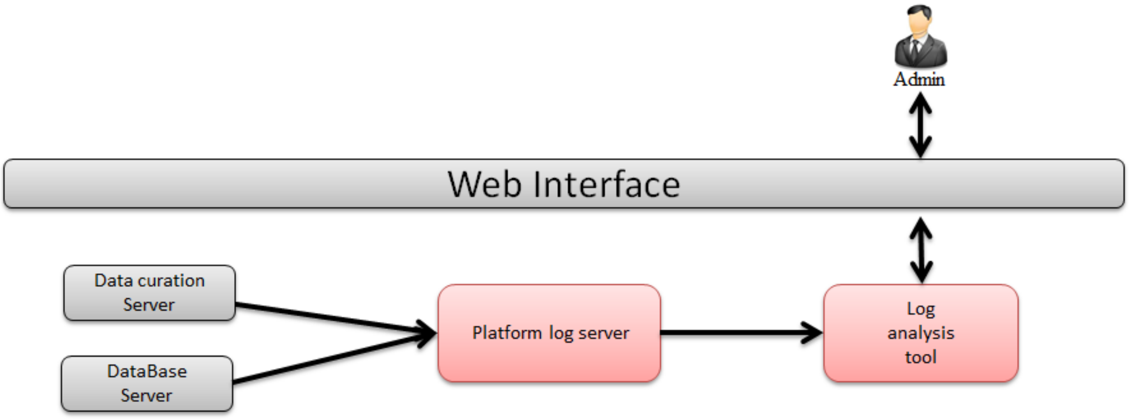

Figure 5 : eTRIKS platform log mechanism architecture

Figure 5 shows the architecture of the eTRIKS platform log mechanism. In a standard eTRIKS platform, we offer two types of data storage such as block storage and database storage. The data in the block storage can be only accessed via data curation server. All the data in the block storage are monitored by inotify[7] and auditd[8]. The data in the database storage are monitored by the database server. Log messages which are produced by these tools are sent to a centralized log storage server. Then admin user can use log analysis tools to view and analyse the log. In a standard eTRIKS platform, we use elasticsearch[9] and Kibana[10] to do the log analysis. We also log other events such as ssh login failure and network scan. But this is beyond the scope of this paper, so we do not discuss them here.

[7] Inotify is a linux kernel subsystem that acts to extend filesystems to notice changes to the filesystem, and report those changes to applications.

[8] Auditd can trace linux system call through auditd daemon based on the auditd rules which are specified by admin.

[9] Elasticsearch is a search server based on Lucene. It provides a distributed multitenant-capable full-text search engine.

[10] Kibana is a browser based analytics and search interface for Elasticsearch.

PoS(ISGC2015)001

2.2 The data curation and storage module

In section 2.1, we have seen that the eTRIKS platform is secure. In this section, we will present how eTRIKS platform handles the data curation and storage. To facilitate data storage and curation, eTRIKS platform provide a data curation and storage module.

This module contains a curation server, a block storage and a database storage. In a standard eTRIKS platform, the block storage is an Openstack cinder volume[12], the database storage is an Postgresql database server. In the data curation server, we use Pentaho Kettle[11] for data curation.

Data accesses to the block storage and to the database storage must be done via the data curation server. In other words, all the data uploading and curation must be done via the data curation server. For example, users can upload raw data to block storage via the data curation server. They can use the data curation tool to process these raw data and insert them into the database storage. The data analysis tool can also access the curated data via the database storage.

2.3 The data analysis and visualization module

To allow user to do the translational research, the eTRIKS platform provides a data analysis and visualization module. This module is designed to host various data analysis and visualization tools, and to provide data access to these tools.

A standard eTRIKS platform provides by default three data analysis and visualization tools such as tranSMART, galaxy and R[15].

tranSMART is a bioinformatics platform that allows user to do the statistical analysis on clinical data. Figure 6 shows an example of the tranSMART work flow. From a given study, we can build different cohorts of patients in tranSMART. Based on the analysis selection, users can generate different kinds of comparison between cohorts. The eTRIKS project participates the development of tranSMART.

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. The eTRIKS project does not participate in the development of Galaxy, but uses it as a tool for translational research.

R is a free software environment for statistical computing and data visualization. The eTRIKS project does not participate the development of R, but uses it as a tool for translational research.

[11] Kettle is a data integration tool for data extraction, transformation, and loading.

PoS(ISGC2015)001

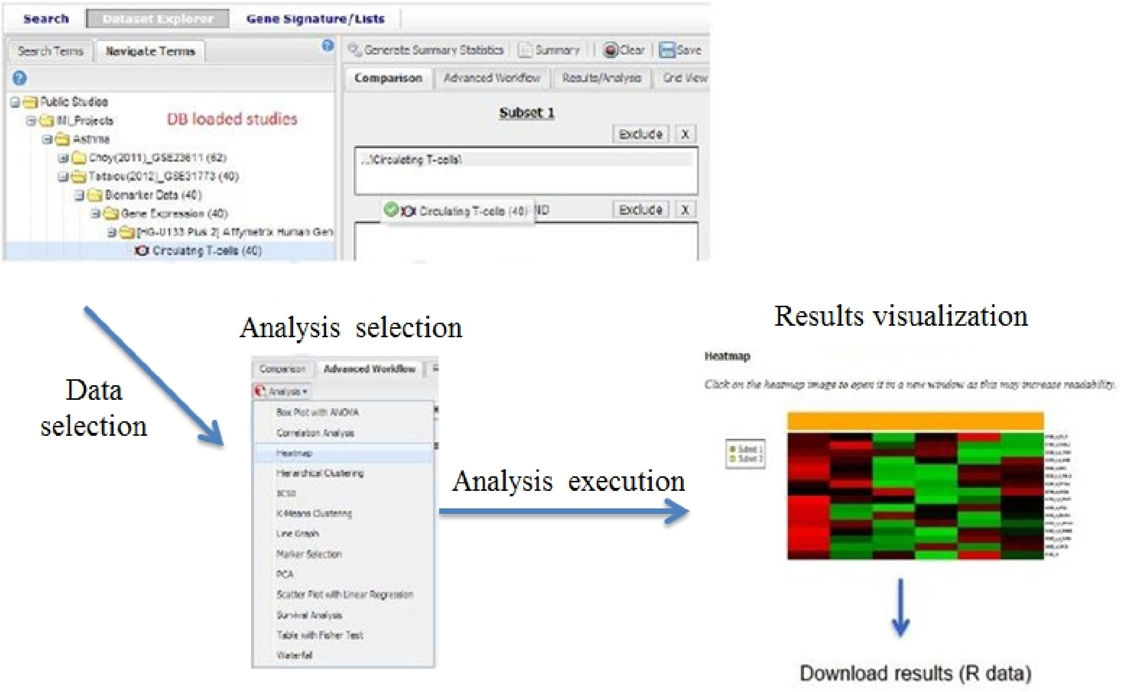

Figure 6 : An example of tranSMART work flow and web interface

3. The eTRIKS Cloud

In section 2, we have presented the cloud-based eTRIKS platform, which leverages propeties of cloud computing such as :

- Quick provisioning
- Horizontal scalability
- Resource utilization efficiency

In order to deploy eTRIKS platforms for other European projects, we built a private cloud at CC-IN2P3[12] named the eTRIKS cloud.

3.1 eTRIKS cloud infrastructure

The eTRIKS cloud currently have 2 controller (i.e. PE-R420) and 6 hypervisors (i.e. PE-R620). In total the eTRIKS cloud have 128 CPU cores and 768 GB memory in VM. We also have 100 TB block storage and 100 TB database storage in the eTRIKS cloud.

The eTRIKS cloud uses Openstack 2014.1 (i.e. IceHouse). All the virtual machines which are running on the eTRIKS cloud use Ubuntu 14.04.1 LTS (i.e. Trusty Tahr) as the operation system.

[12] Computing centre of the national institute of nuclear physics and particle physics of French national centre for scientific research

PoS(ISGC2015)001

3.2 Hosted project in eTRIKS cloud

Currently, the eTRIKS cloud hosts three eTRIKS platform instances. The public server instance is a place that allows all European projects to share their public data. It's also a demonstration platform for the eTRIKS project. Public visitor can access the public data via data analysis tools (e.g. tranSMART, Galaxy.). For the public visitor, the public instance is a software as a service. For those who are interested, they can visit https://public.etriks.org.

The abirisk instance is deployed for abirisk project, which works on anti drug-immunization for pharmaceutical procedures.

The oncotrack instance is deployed for oncotrack project, which works on identification of bio-marker for colon cancer.

The administrator users of abirisk and oncotrack, are allowed to install new tools on the eTRIKS platform. As a result, it is a platform as a service for them. The access to both instances are restricted.

4. Conclusions

The eTRIKS platform is a cloud-based translational information and knowledge management service. The eTRIKS platform is secure, flexible, scalable, and easy to deploy.

The eTRIKS platform is secure, because the eTRIKS platform ensures the authenticity of users and the platform, the confidentiality and integrity of the data. As the security module uses client-server architecture. It's easy to integrate new services into eTRIKS platform without losing the benefits of security module.

The eTRIKS platform is flexible, because the modular architecture of eTRIKS platform allows admin user to integrate new tools (e.g. data analysis tools, data curation tools, etc.) easily.

The eTRIKS platform is scalable, because the eTRIKS platform benefits all the advantages of cloud computing, such as virtual machine resizing to increase cpu or memory capacity, or creating new virtual machines to do load balancing.

The eTRIKS platform is easy to deploy, because all modules of the eTRIKS platform are one or several virtual machines. And we have virtual machine images which can be deployed into any Openstack based private cloud.

All these advantages increase the efficiency of translational research and decrease the cost of mass heterogeneous data hosting and analysis.

PoS(ISGC2015)001

References

[1]Woolf SH, The meaning of translational research and why it matters, JAMA vol.299 pages 211-213 Jan. 2008.

[2]Zerhouni EA, Translational research: moving discovery to practice, Clin Pharmacol Ther, 2007 Jan;81(1):126-8

[3]P. Lord, A. Macdonald, Liz Lyon and D. Giaretta, From Data Deluge to Data Curation, In Proceeding 3th UK e-Science All Hands Meeting, 3:371-375, 2004

[4]M. Bishop. Introduction to Computer Security. Addison-Wesley Professional, 2004

[5]M. Bishop. What is computer security? In Security & Privacy, IEEE, volume 1, pages 67-69, Davis, CA, USA, Jan 2003.

[6]OpenLDAP, http://www.openldap.org/

[7]Spring Security, http://docs.spring.io/spring-security/site/docs/3.0.x/reference/springsecurity.html

[8]M. T. et al. Extensible access control mark-up language (xacml) version 3.0. Technical report, OASIS, 2013

[9]C. Ribeiro, A. Zuquete, P. Ferreira, and P. Guedes. Spl: An access control language for security policies and complex constraints. In NDSS, 2001.

[10]L. Kagal, T.W. Finin, and A. Joshi. A policy language for a pervasive computing environment. In IEEE 4th International Workshop on Policies for Distributed Systems and Networks, pages 63-74, June 2003.

[11]D.C. Verma Policy-Based Networking: Architecture and Algorithms. New Riders Publishing, Thousand Oaks, CA, USA, 2000.

[12]OpenStack Block Storage Cinder, https://wiki.openstack.org/wiki/Cinder

[13]tranSMART foundation http://transmartfoundation.org/

[14]Galaxy, http://galaxyproject.org/

[15]The R Project for Statistical Computing, http://www.r-project.org/

[16]Adams, Carlisle and Lioyd, Steve, Understanding PKI: concepts, standards, and deployment considerations. Addison-Wesley Professional. pp. 11-15. ISBN 978-0-672-32391-1,2002.

PoS(ISGC2015)001