OHDSI Cloud Architecture

Lee D. Evans, LTS Computing LLC, West Chester PA

Abstract

The OHDSI cloud provides shared computing and collaboration tool services for the diverse global OHDSI community. It has been developed iteratively over a one year time frame based on expanding community needs. The OHDSI cloud has been operationally very stable and satisfies the performance and operational budget constraints.

Declaration of Interests

LTS Computing LLC delivers commercial IT projects/services for Life Sciences clients, including pharmaceutical companies.

Introduction

The OHDSI cloud supports five groups of services:

- OHDSI Community global collaboration tools
- Continuous integration build and automated deployment of OHDSI applications
- OHDSI applications sandbox/demonstration environments running the OHDSI WebAPI
- OHDSI LAERTES knowledge base ETL operational environment
- OHDSI research network data repository

Infrastructure

Ubuntu LTS 14.04 server images running on m3 medium EC2 servers in Amazon AWS. Windows 2012 server image running on an m3 medium server to support Windows based OHDSI tool continuous integration builds. PostgreSQL 9.3 databases running on Amazon EC2 and RDS.

OHDSI Community global collaboration tools

WordPress Web Site, Dokuwiki Wiki, Discourse Forums.

Continuous integration build and automated deployment of OHDSI applications

Linux builds are executed using Travis CI software as a service. Windows builds are executed on an AWS hosted Jenkins server. The builds and automated deployments are automatically triggered when new code is pushed to the OHDSI GitHub master repositories.

OHDSI applications sandbox/demonstration environments running the OHDSI WebAPI

The OHDSI web applications are automatically deployed to Tomcat and Apache servers running within Docker containers. Docker volumes are shared with the Amazon EC2 host servers for persistent storage. The WebAPI provides a set of web services that are used by the OHDSI tools and also provide a general purpose Application Programming Interface to OHDSI standard information.

The WebAPI services:

- Standardized Vocabularies
- Common Data Model
- LAERTES Evidence Knowledge base
- Cohort definition / analysis
- Study feasibility
- Therapy path reporting
- “SQLRender” RDBMS SQL translation, with support for Oracle, SQL Server, PostgreSQL and Redshift

OHDSI LAERTES knowledge base ETL operational environment

The LAERTES ETL is executed quarterly. LAERTES is also known as the OHDSI knowledge base[1]. There are multiple data feeds including SemMedDB, PubMed, UMLS metathesaurus, Standard Product Labels and FAERS adverse events. The data sources are downloaded as files from websites onto Amazon EC2 servers where the data is loaded into MySQL or PostgreSQL databases, transformed via Python programs and then loaded into the LAERTES PostgreSQL and Virtuoso semantic databases. Virtuoso provides a semantic SPARQL interface to LAERTES knowledge and PostgreSQL provides counts and statistics.

OHDSI research network data repository

A PostgreSQL database hosted in Amazon RDS is configured with SSL certificates to identify the database server to research database clients and ensure that all traffic between the database clients and the server is securely encrypted. Only de-identified aggregated results of OHDSI research network studies is stored in the database. PostgreSQL password protected read-only database accounts are provided to OHDSI community researchers who participate in the studies, so they can review and analyze the study results.

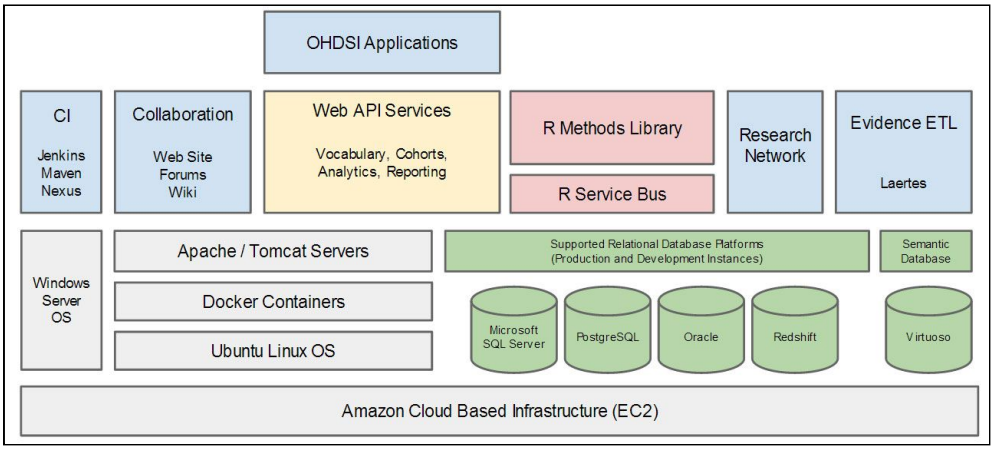

Figure 1. Logical overview of the OHDSI Cloud Architecture.

Conclusion

The OHDSI cloud is a flexible architecture that enables a large global research community to run analyses using a heterogeneous set of services, RDBMSs & programming languages. It leverages cloud infrastructure for flexibility and future scaling.

Future enhancements will likely include:

- Additional support for initiating and conducting OHDSI network based research studies and a centralized repository to coordinate the storage of network study aggregate summary results that are shared by sites to the OHDSI central coordinating center.
- Future deployment of R analysis server & R service bus for OHDSI R methods. See red boxes in (Figure 1).
- Deployment of multiple RDBMS OMOP Common Data Model 'sandbox' database instances with sample test data, so an OHDSI researcher can design a new analysis and test it across the 'development network' before dissemination to the participating research sites
- Expansion of OHDSI WebAPI services and deployment of additional OHDSI tools as they are developed.
- Implementation of a flexible security module using the Apache Shiro security framework.
- More dynamic allocation of cloud storage & compute infrastructure to increase server utilization.

References

1. Boyce RD1, Ryan PB, Norén GN, Schuemie MJ, Reich C, Duke J, Tatonetti NP, Trifirò G, Harpaz R, Overhage JM, Hartzema AG, Khayter M, Voss EA, Lambert CG, Huser V, Dumontier M. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Safety. 2014.