在 AWS 中规划基因组数据安全架构

执行概述

Angel Pizarro

Chris Whalley

Carina Veksler

2014 年 12 月

amazon web services

目录

概述

个人层面的基因型与表型研究已经成为解决多种健康问题的关键所在。然而，由于此类数据在体积与应用层面的持续臃肿，适当的数据处理、存储及安全性实现技术已经成为制约基因组研究的重要因素。

云计算提供一种简单方式以访问服务器、存储、数据库以及经由互联网实现的各类应用服务——而全球研究社区也意识到，Amazon Web Services（简称 AWS）能够实现多种实际收益。

在评估云平台时，研究人员需要考虑利用各类安全最佳实践，从而保障国家卫生研究院（简称 NIH）等人类基因组数据及相关数据集访问，具体包括基因型与表型数据库（简称 dbGaP）以及全基因组关联分析（简称 GWAS）。作为最为基本的 dbGaP 架构合规性考量，我们需要在 AWS 当中确定其架构系统全部运行在 AWS 之上，抑或是以混合部署方式将 AWS 与非 AWS 资源加以结合。本份白皮书将着重探讨 AWS 资源在架构混合型部署场景时的控制议题。

人类研究中的基因组数据隐私性与安全性

研究人员通常对于利用 AWS 运行基因组序列数据的安全性与合规性遵循能力抱有怀疑态度。具体来讲，研究人员需要了解如何满足诸如美国国立卫生研究院等政府拨款资助机构提出的指南与最佳实践要求。着眼于科学调查人员、机构签约官员、IT 主管、伦理委员会以及数据访问委员会等群体，此类用户的常见问题包括：

- 数据在安全服务器之上是否得到保护？
- 数据存储于何方？
- 如何对数据访问加以控制？
- 数据保护机制是否适用于数据使用认证要求？

这些注意事项并非新鲜要求，且非仅存在于云环境当中。人类基因组数据的基本数据保护要求是一致的，无论数据具体存储在哪些位置——包括实验室、机构网络、机构托管数据存储库或者 AWS 云。

在规划研究系统时，数据保护与安全性控制机制必须得到明确定义，而后再将架构设计引入系统构建。这一点对于评估责任分担模式时非常重要。

AWS 的共享安全模型方案

AWS 提供一套强大的 Web 服务平台，其能够帮助世界各地的研究团队在 AWS 云中创建并控制自身专有资产，同时实现灵活的快速访问与低成本 IT 资源。而利用云计算，我们不必再承担沉重的前期硬件投入或者消耗时间维护系统及设施。然而，由于 AWS 并不会对客户数据所在的专有 AWS 环境进行访问或者管理，因此客户自身需要负责 AWS 的配置与安全控制实现工作。这种由客户分担部分责任的设定，源自 AWS 对于客户专有相关数据的保护要求与人类基因组数据安全操作的深刻理解。

AWS 云内的安全性责任由客户与 AWS 共同承担。

- AWS 负责保护底层基础设施。
- 研究人员需要保护一切引入该基础设施的资产（例如操作系统、平台与数据），同时负责满足特定监管与数据合规性要求。

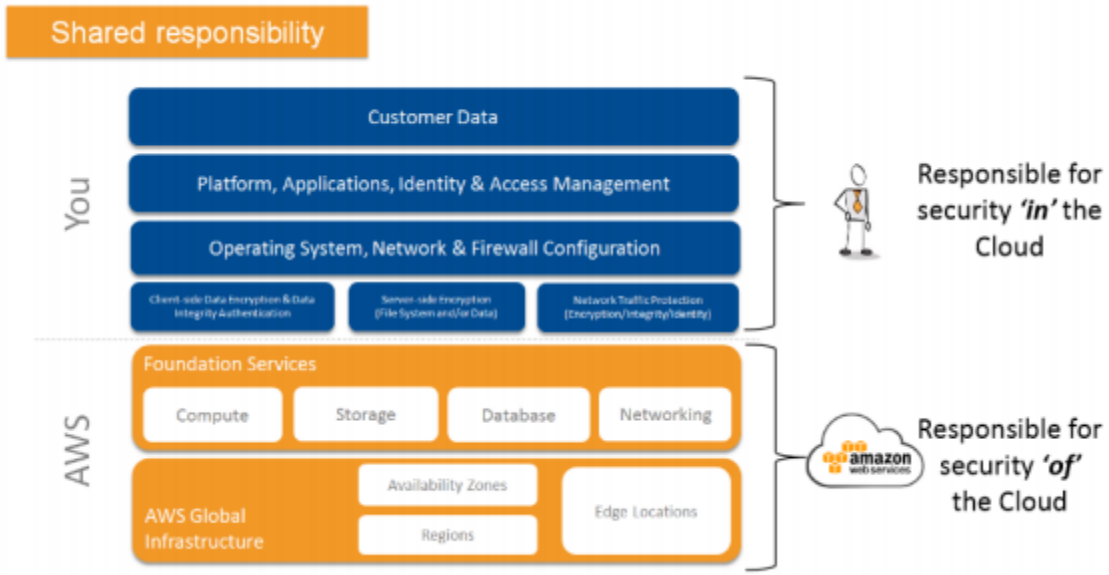

Figure 1: Shared Responsibility Model

这种责任分担机制还为研究人员提供充分的灵活性，允许其在应用程序部署工作中充分满足行业特定认证要求。另外，研究人员亦能够借此强化安全性并利用主机防火墙、主机入侵检测/预防、加密以及密钥管理等 AWS 技术手段巩固合规性要求。

在 AWS 中规划安全性与合规性架构

根据 dbGaP 安全最佳实践的要求，研究人员应当将数据下载至安全计算机或者服务器，同时避免使用非安全网络驱动器或者服务器。[1] 另外，dbGaP 安全最佳实践可分为三种 IT 安全控制领域，用于满足具体原则。

AWS 全球基础设施

AWS 被组织为多个服务区与可用区，允许在不同区域之间实现高能量与低延迟通信。研究人员还能够立足于本地特定要求或者区域内数据隐私性政策，选择合适的位置建立并维护自己的专有 AWS 环境。另外，客户还能够选择在多个服务区之间进行内容复制与备份，并由研究人员进行具体配置。

物理服务器访问

AWS 的物理服务器与网络硬件处于高安全性、业界领先的数据中心之内。这意味着 AWS 的独立第三方安全评估面向 ISO 27001、服务组织控制 2（简称 SOC 2）、NIST 的联合信息系统安全标准与其它安全认证资质。对 AWS 数据中心及硬件的访问基于最小特权原则。

在云计算操作环境当中，访问操作只授权给拥有经验的必要人员，并由其负责物理环境的维护。

互联网、网络与数据传输

尽管 AWS 云在本质上属于通过互联网交付的一组 Web 服务，但客户专有 AWS 账户内的数据将仅在客户专门配置的情况下放公布于互联网之上。这一设定符合 dbGaP 安全最佳实践的基本要求，同时 AWS 云亦拥有多项内置功能以预防经由互联网对基因组数据的直接访问。

- Amazon Elastic Compute Cloud (即 Amazon 弹性计算云，简称 EC2)。当研究人员创建新的 Amazon EC2 实例以下载并处理基因组数据时，这些实例将只能由专有 AWS 账户之内的授权用户进行访问。具体来讲，相关数据无法通过互联网被发现或者直接访问，除非研究人员进行专门配置。

- 数据存储。基因组数据通常存储于 Amazon Simple Storage Service（即 Amazon 简单存储服务，简称 S3）或者 Amazon Elastic Block Store（即 Amazon 弹性块存储，简称 EBS）当中。其它存储服务还包括 Amazon Relational Database Service（即 Amazon 关系型数据库服务，简称 RDS）、Amazon Redshift、Amazon DynamoDB 以及 Amazon ElastiCache 等。与 Amazon EC2 类似，这些存储及数据库服务同样默认为最低权限访问且无法通过互联网进行直接发现或者访问——除非客户进行特殊配置。

- Amazon Virtual Private Cloud (即 Amazon 虚拟专有云，简称 VPC)。研究人员可以在 AWS 云当中创建专有隔离网络，并以此为基础保留对虚拟网络环境的全面控制能力。

安全注意事项

在设计一套混合部署模式时，研究人员还需要在架构设计中考虑以下事项。

安全控制事项	描述
便携式存储介质	当数据被下载至便携式设备，例如笔记本电脑或者智能手机，该数据应当进行加密并控制打印等硬输出机制。
用户账户、密码与访问控制列表	根据最低权限原则立足 dbGaP 要求进行用户访问管理，从而确保个人及/或流程仅具备执行相关任务及功能的必要权限。
数据加密	在 AWS 当中，提供多种选项以利用自动化 AWS 加密解决方案（服务器端）、手动与客户端机制对基因组数据进行加密与排列。由于研究人员需要设计系统架构以控制数据集访问，因此需要严格确保每项 AWS 服务及其使用的加密模式适用于基因组数据。
文件系统与存储分卷	在专有 AWS 账户内，研究人员需要配置存储服务及安全功能，从而确保仅由授权用户实施访问。
操作系统与应用程序	研究人员需要负责对各类相关 AWS 服务中的操作系统与应用程序进行配置与维护，具体包括 EC2 与 Amazon S3。而其配置与维护依据则包括 NIST 800-53、dbGaP 安全最佳实践附录 A 或者其它地区性标准。

审计、登录与监控	dbGaP 安全建议要求利用安全审计与入侵检测软件以定期扫描并检测潜在数据入侵活动。在 AWS 生态系统当中，研究人员能够利用内置监控工具——例如 Amazon CloudWatch 或者 AWS CloudTrail——外加丰富的合作伙伴安全与监控软件生态系统支持 AWS 云服务。AWS 合作伙伴网络清单中包含多家系统集成商与软件供应商，能够帮助研究人员满足各类安全性与合规性要求。
数据访问授权	研究人员必须从数据访问委员会（简称 DAC）或者现有数据使用认证（简称 DUC）条款以获取访问批准，从而控制数据集的访问活动。
清理数据并保留结果	在 AWS 当中，数据检测与保留操作由研究人员进行控制。研究人员能够遵循 dbGaP 安全建议，具体包括将数据加密与其它标准操作规程（例如资源监控与安全审计）相结合。

客户实例

贝勒医学院

位于德克萨斯州休斯顿的贝勒医学院正是人类基因组测序中心（简称 HGSC）的所在地，亦是由联邦政府进行资助的三大测序中心之一。HGSC 的当前项目之一为基因组流行病学内以及与老化疾病项目（简称 CHARGE），来自世界各地 5 家机构的 200 多名科学家正在努力查明老化与心脏疾病在基因层面的根源。过去一个世纪以来，已经有大量研究观察患者终生以确定特定病症的演变方式。利用 DNA 测序工具以及大规模数据集的管理能力，这些研究的成果开始在 CHARGE 项目当中接受重新分析。世界各地的 CHARGE 科学家们都在利用数据研究疾病的发生原因并借此找到预防手段。

目前 CHARGE 项目中的数据总量已经超过 430 TB，贝勒医学院需要一套具备成本效益且易于维护的解决方案以实现安全且高效的全球范围协作，同时摆脱由物理基础设施带来的延迟问题。

通过迁移至 AWS，贝勒医学院得以在十天之内完成首批分析——速度相当于本地基础设施的五倍——且快速实现成果共享。

AWS 的可扩展性优势帮助 CHARGE 的科学家们获得了更为强大的预测能力，这要归功于云服务几乎无限的计算与数据存储资源。另外，他们还得以快速安全地发现“保护”基因，即负责帮助人类抵御疾病的护盾。

宾夕法尼亚州立大学生物工程系

宾夕法尼亚州立大学生物工程系希望为生物技术研究人员们提供一种便捷的研究方法与数据共享途径，旨在运行计算资源密集型 DNA 模拟实验。通过将研究负载迁移至 AWS，该大学得以简化了全球 6000 名研究人员利用其设计方法与优化算法处理超过 5 万种合成 DNA 序列的流程。

- 不再需要采购设备，从而节约大量时间与资金。

- 能够灵活地开启及关闭计算节点，从而满足不断变化的计算能力要求。
- 研究人员不再需要等待访问设计算法，各服务可根据需求随时使用。
- 利用一套 Web 界面，现在研究人员能够直接访问信息内容。

克拉瑞塔斯基因组

克拉瑞塔斯拥有最高标准的基因诊断实验室，旨在为儿科疾病诊断提供检测服务。作为原本波士顿儿童医院的基因检测实验室，克拉瑞塔斯于 2013 年 2 月作为独立实体开始运营。

然而，在脱离波士顿儿童医院之后，克拉瑞塔斯不再能够共享该医院及哈佛医学院的数据中心。在评估了各类现有选项之后，克拉瑞塔斯方面决定采用 AWS 云。

- 节约了 500 万美元数据中心构建费用。
- 实现更高成本效益，每月支出降低 36%。
- 将临床周围周期由 4 到 6 个月缩短至 4 到 6 周。

AWS 允许克拉瑞塔斯立足于 HIPAA 合规性要求处理自己的系统，且全部数据在输入及输出 AWS 云基础设施时皆进行加密。通过采用 AWS，克拉瑞塔斯得以为开发人员提供一套灵活的基础设施，其轻松易行的扩展能力可满足各类计算需求。

总结

各项安全最佳实践使得研究人员能够满足国立健康研究院提出的具体要求，从而以安全、可扩展且极具成本效益的 Amazon Web Services 环境控制对基因组序列数据的访问。

更多信息

欲了解更多信息，请参阅以下资料：

- 《安全流程概述》白皮书。[2]
- AWS 生命科学合作伙伴页面 [3]。
- IAM: IAM 说明文档 [4], IAM 最佳实践 [5] 以及多因素验证 [6]。
- VPC: Amazon VPC 白皮书 [7], VPC 说明文档 [8] 以及 VPC 连接选项白皮书 [9]。

- 加密: 利用加密机制保护闲置数据安全白皮书 [10], AWS CloudHSM[11].
- A3 安全性: 访问控制 [12], 使用数据加密 [13], Amazon S3 开发者指南 [14].
- AWS 安全性概述: Amazon Web Services: 安全流程概述 [15]
- Amazon EBS 安全功能: Amazon EBS 加密 [16], Amazon Elastic Block ,Store[17]

[2] http://media.amazonwebservices.com/pdf/AWS_Security_Whitepaper.pdf

[3] http://aws.amazon.com/partners/competencies/life-sciences/

[4] http://aws.amazon.com/documentation/iam/

[5] http://docs.aws.amazon.com/IAM/latest/UserGuide/IAMBestPractices.html

[6] http://aws.amazon.com/iam/details/mfa/

[7] https://d36cz9buwru1tt.cloudfront.net/Extend_your_IT_infrastructure_with_Amazon_VPC.pdf

[8] http://aws.amazon.com/documentation/vpc/

[9] https://media.amazonwebservices.com/AWS_Amazon_VPC_Connectivity_Options.pdf

[10] https://media.amazonwebservices.com/AWS_Securing_Data_at_Rest_with_Encryption.pdf

[11] https://aws.amazon.com/cloudhsm/

[12] http://docs.amazonwebservices.com/AmazonS3/latest/dev/UsingAuthAccess.html

[13] http://docs.amazonwebservices.com/AmazonS3/latest/dev/UsingEncryption.html

[14] http://docs.amazonwebservices.com/AmazonS3/latest/dev/

[15] http://awsmedia.s3.amazonaws.com/pdf/AWS_Security_Whitepaper.pdf

[16] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSEncryption.html

[17] http://aws.amazon.com/ebs/

通告

本文档所提供的信息仅供参考，且仅代表截至本文件发布之日时 AWS 的当前产品与实践情况，若有变更恕不另行通知。客户有责任利用自身信息独立评估本文档中的内容以及任何对 AWS 产品或服务的使用方式，任何“原文”内容不作为任何形式的担保、声明、合同承诺、条件或者来自 AWS 及其附属公司或供应商的授权保证。AWS 面向客户所履行之责任或者保障遵循 AWS 协议内容，本文件与此类责任或保障无关，亦不影响 AWS 与客户之间签订的任何协议内容。