

Research on Unified Representation and Storage of User Log in eID Based Virtual Property Preservation System

Xiang Fu[#], Weihui Zhu, Weihong Han

School of Computer, National University of Defense Science and Technology, Changsha Hunan 410073, China

[#]Email: xjtusefox@163.com

Abstract

The Virtual Property Preservation System^[1] based on eID involves lots of different types of application platforms, such as taobao.com, jd.com, weibo, etc. Users log will be generated in these application platforms. Different platforms have different log formats and management systems. Besides, these management systems are completely isolated. But in reality, the user's behavior is often complex, users can operate on multiple application platforms in a short period of time. To obtain a user's complete network behavior from user log is of great importance to virtual property audit traceability, abnormal finding of the Virtual Property Preservation System. So the unified representation and storage of the cross-platform user log is necessary. We provide a unified representation method to put user logs on different platforms together. Based on the early research on eID Virtual Identity Data Storage^[2] of Han, one of the authors of this paper, We analysis the data characteristics of cross-platform user log and provide a storage model which can meet the storage demand of mass cross-platform user log data.

Keywords: eID; Virtual Property; Cross-platform; MongoDB

基于 eID 虚拟资产保全系统的海量用户日志统一描述与存储的研究^{*}

傅翔, 朱伟辉, 韩伟红

国防科技大学 计算机学院, 湖南 长沙 410073

摘 要: eID 虚拟资产保全系统^[1]涉及到多个不同种类的应用平台, 如电子商务类的淘宝, 京东, 社交网络类的新浪微博等。用户在这些应用平台上会产生不同的日志数据, 且不同平台的用户日志管理系统各不相同, 现有的建立在用户日志上的工作都局限在了某一特定应用之上。而现实中, 用户的行为往往更加复杂, 一个时间段内, 用户可以同时在多个应用平台上进行操作。获得一个用户完整的网络空间的行为, 对虚拟资产保全系统的审计追溯, 异常发现等功能的实现必不可少。我们通过 eID 这个虚拟身份唯一标识来将不同平台的日志整合到一起, 提出了一个海量跨平台用户日志的统一描述方法。在本文工作者之一前期对海量虚拟身份数据存储研究^[2]的基础上, 分析了跨平台用户日志的数据特点, 提出了符合海量跨平台用户日志数据特点及现实需求的存储模型。

关键词: eID; 虚拟资产; 跨平台; MongoDB

引言

eID(electronic IDentity)全称为公民网络电子身份标识, eID 是网络上远程证明个人真实身份的权威性电子

^{*}基金资助: 国家高技术研究发展计划(863 计划)[2012AA01A401]、国家自然科学基金[60933005,91124002]、国家科技支撑计划[2012BAH38B04, 2012BAH38B06]、国家 242 信息安全计划[2011A010]。

信息文件。当 eID 在网上远程使用时，使用基于公安人口数据库以及 eID 服务平台完成真实身份的验证，可在实现个人身份的真实性和有效性确认的同时保护公民身份隐私，具有权威性、安全性、可追溯、方便易用等特点。在互联网中，用户与各种应用、平台下的虚拟身份之间存在着一对多的关系。而在基于 eID 的网络环境中，上述这些对应关系都可基于 eID 这个唯一标识。

虚拟资产是指在网络世界中存在的具有竞争性、持久性以及可以交换或者买卖的物品包网上银行、网络帐号、网游装备武器、虚拟货币等等。目前，针对虚拟资产的犯罪和破坏活动日益猖獗，虚拟资产保全是指为维护虚拟资产的完整性，防止流失而采取的一系列措施。虚拟资产保全系统是国家 863 重大项目。因此，研究虚拟资产保全系统相关技术具有重要意义。

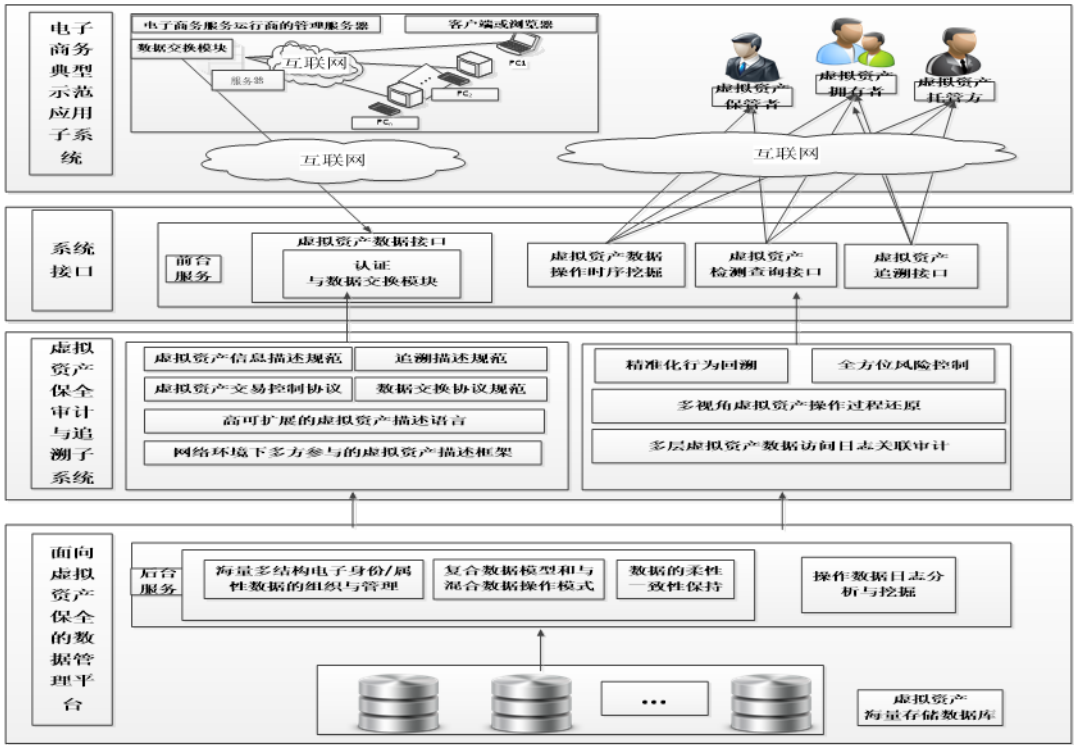


图 1 虚拟资产保全系统总体架构图

虚拟资产保全系统的体系架构如图 1 所示，包括面向虚拟资产保全的数据管理平台、虚拟资产保全审计与追溯子系统、网吧网游典型示范应用子系统以及虚拟资产保全安全防护子系统等四个部分。其中，面向虚拟资产保全的数据管理平台支持 eID 系统中海量虚拟资产及其交易日志的存储、管理、数据集成和查询、分析的统一管理平台，遵循相关数据访问与数据集成的国际/国家标准和规范。本文作者之一前期对基于 eID 虚拟身份数据存储进行了研究^[2]，在总结了现有数据库优势的基础上，分析了基于 eID 虚拟身份数据的特点以及现实中的应用需求，然后提出了基于 eID 虚拟身份数据的存储模型，最后进行了性能测试，获得了很好的效果。在此基础上，本文主要研究基于 eID 的虚拟资产保全系统海量用户日志统一描述与存储。为了该系统的其他基于日志文件的子课题提供支撑。

1 用户日志统一描述

在网域空间，用户根据自己的需求在不同的应用平台上注册账号，这些应用平台包括电子商务，社交网络，网络游戏等。传统的浏览器用户行为日志会包含用户的每一次操作，几乎记录了用户的所有行为，但虚拟资产保全系统却不需要记录用户的所有行为，我们只关心与 eID 相关的部分，那些不需要登录 eID 就能进行的操作，我们不做记录。在传统的用户日志的基础上，对其原有的用户日志进行过滤，筛选出虚拟资产保全系统所需要的重要字段，这些日志记录将包含用户的一系列操作如登录，收藏，购买，发布消

息，评论，注销等，并对每条记录加上 eID 这一唯一标识来进行统一存储。

表 1 平台 A 的日志示例表

用户 ID	用户名	登录时间	登录 IP	登录地点	操作类型	...
...						
100013	丁香花	2013/10/23 20:03:54	221.208.129.117	湖南长沙	登录	...
...						

表 2 平台 B 的日志示例表

用户 ID	用户名	登录时间	登录 IP	登录地点	操作类型	...
...						
798102	丁丁	2013/10/23 20:00:12	222.35.127.83	上海	发布消息	...
...						

基于以上的日志统一描述方法，跨平台日志系统不仅能实现传统单一平台日志系统的所有功能，还能发挥一些其特有的优势。举个异常发现方面的简单例子：每个平台的日志文件不相关，不同平台同一个用户的虚拟身份不能关联起来，在这种情况下，发生在平台 A、平台 B 分别发生如上表 1、表 2 中所示操作行为被视为正常操作行为。而若使用虚拟资产保全系统统一平台，则可利用 eID 轻易地做到身份关联，如下表 3 所示，发现在平台 A 中用户名为“丁香花”的用户和平台 B 中用户名为“丁丁”的用户是同一个用户，但他们却在很近的两个时间点在不同的两个地点分别在平台 A 和平台 B 进行了操作行为，这种情况下我们完全有理由相信其中一个操作是异常行为，可以进行相应的异常警报。

表 3 虚拟资产保全系统统一平台日志示例表

	平台	用户 ID	用户名	登录时间	登录 IP	登录地点	操作类型	...
1r39s3ph(eID)	平台 A	100013	丁香花	2013/10/23 20:03:54	221.208.129.117	湖南长沙	登录	...
	平台 B	798102	丁丁	2013/10/23 20:00:12	222.35.127.83	上海	发布消息	...
...					...			
...					...			

2 用户日志的存储

2.1 数据特点及现实需求

基于 eID 的虚拟资产保全系统海量用户日志按照前文所提出的统一描述方法进行存储，可以发现这些数据具有以下特点：

- 1) 数据规模大。虚拟资产保全系统的用户量是亿级的，每个用户都有各自在不同平台的日志数据。
- 2) 易扩展。用户日志不停在产生，必须随时增加节点来缓解存储压力。
- 3) 不具有统一格式。不同类型的操作日志字段不同，如用户电子商务网站中的购买操作与社交网络中的发状态操作。
- 4) 多处理。可以实现并行插入不同记录，同时访问相同记录等。
- 5) 一次写入多次读取。所有数据是记录用户操作的数据，因此是不可更改的，一次写入多次读取。

6) 支持精确查询。可以通过某一用户的 eID 和查询时间范围，查找到该用户的所有日志数据；或通过某个平台的用户名和时间范围，查询该用户在该平台下的日志信息等。

7) 对数据的插入不要求实时性，可以周期性批量插入。

8) 针对查询操作，不需要严格的数据一致性。

2.2 用户日志存储模型

针对上述所提到的用户日志数据的特点，结合本文工作者之一前期对基于 eID 虚拟身份数据存储的研究，我们将目光转向了 NoSQL 数据库，经过比较，我们放弃了前面工作中所使用的 Cassandra 数据库，而是使用 MongoDB 来对日志进行存储。本文对 Cassandra 数据库不做过多研究，主要研究 MongoDB 的特点以及分析选择它的原因。

不同于 Cassandra 这种列存储数据库，MongoDB 是一种面向文档(Document-Oriented)的数据库，以文档的方式存储数据，一个文档类似于关系数据库中的一行，一系列文档构成一个集合，类似于关系数据库中的表，如表 4。其文档存储格式为 BSON（一种 JSON 的扩展），文档的每一个数据就是一条日志记录。

表 4 MongoDB 与关系型数据库对比表

Mongodb	关系型数据库
文档(document)	行(row)
集合(collection)	表(table)
数据库(database)	数据库(database)

在存储海量数据时，MongoDB 会使用预分配方式来保证写入性能的稳定性，预分配在后台进行，并且每个预分配的文件都用 0 进行填充，这会让 MongoDB 始终额外的空间和空余的数据文件，从而避免了数据增长过快而带来的分配磁盘空间引起的阻塞。MongoDB 提供自动分片以支持云计算层次的扩展性，支持水平的数据库集群，可动态增加额外的机器，实现真正的分布式存储。在查询方面，MongoDB 作为 NoSQL 中最像关系型数据库的数据库，支持动态查询，支持丰富的查询表达式，查询指令使用 JSON 形式的标记，可轻易查询文件中内嵌的对象及数组。由于其支持完全索引，包含内部对象，并且 MongoDB 的查询优化器会分析查询表达式，生成一个高效的查询计划。在灾备和故障恢复方面，MongoDB 支持主/从模式及服务器之间的相互复制数据，不仅能提供数据备份和自动故障转移，还能用来读扩展以提高系统性能。除此之外，MongoDB 使用的是内存映射存储引擎，它会把磁盘 I/O 操作转换成内存操作，从而能够体现出内存数据库的一些优势，例如在读操作时，内存中的数据起到了缓存作用，如果是写操作，内存还可以把随机写操作转换成顺序的写操作。

虽然 MongoDB 默认使用不安全的操作以换取性能，其插入，删除和更新等都不需要等待数据库响应，客户端将文档发送给服务器就完成了操作，不去等待也不会受到返回码，哪怕失败也不会回滚，但由于用户日志不要求事务性和安全操作，MongoDB 的以上缺点可以被掩盖掉。MongoDB 目标是提供一种高性能、可扩展的数据存储方案，是最适合替代 MySQL 的 NoSQL 数据库。此外，由于日志文件无需更新而需要批量插入和删除，MongoDB 限制容量的固定集合和限制保存时间的 TTL(Time-To-Live)集合功能天然适用于日志的存储。

3 性能测试

针对前面的介绍，我们使用 MongoDB 来对日志文件进行简单的查询性能的测试。在一台服务器上搭建 MongoDB 数据库，服务器配置为：处理器 Intel(R) Xeon(R) CPU E5-2403 0 @ 1.80GHz，内存 32G，硬盘 2T。MongoDB 数据库都存储有 300 万条某电子商务平台用户行为日志记录。分别在数据库上提交 5 万次，10 万

次，20 万次，30 万次，50 万次随机查询。测试结果如图 2。

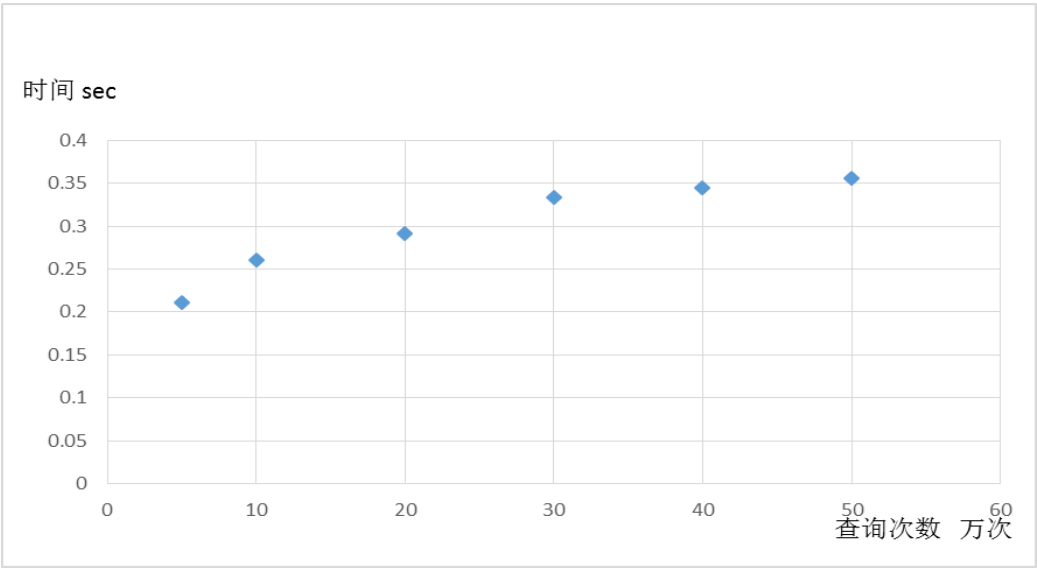


图 2 MongoDB 查询性能测试图

通过实验可以看，MongoDB 随机读数据的能力较强，用它来存储海量日志文件在实际应用中是可行的。综合前文的分析和实验结果，我们最终选用 MongoDB 来存储海量跨平台日志数据。

4 结语

在本文作者之一前期基于 eID 虚拟身份数据存储研究的基础上，我们提出了基于 eID 的虚拟资产保全系统海量用户日志统一描述方法与存储模型。该统一描述方法将所有用户日志统一起来，不仅方便管理，在异常发现等领域也提供了新的数据基础。在存储模型上，通过对海量跨平台日志数据特点和需求以及 MongoDB 的分析，测试了 MongoDB 随机查询的效率，最终选择了 MongoDB 这一面向文档存储的 NoSQL 来存储基于 eID 的虚拟资产保全系统中的海量用户日志。

REFERENCES

- [1] Bai Z H, Li S D, Li A P, et al. Research and Implementation Virtual Property Preservation System Based on Electronic Identity [J]. Telecom market, 2013, (4): 109-114
- [2] Deng L, Han W H, Liu D, et al. Research on eID Virtual Identity Data Storage [J]. Netinfo Security, 2013, (10): 101-103
- [3] Han J, Haihong E, Le G, et al. Survey on NoSQL database[C]//Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE, 2011: 363-366
- [4] Deng L, Han W H, Liu D, et al. The Model of Fuzzy Retrieval Based on External Index[J]. Applied Mechanics and Materials, 2013, 380: 1605-1608
- [5] Rosenblum M, Ousterhout J K. The design and implementation of a log-structured file system[J]. ACM Transactions on Computer Systems (TOCS), 1992, 10(1): 26-52
- [6] Ray I, Belyaev K, Strizhov M, et al. Secure Logging as a Service---Delegating Log Management to the Cloud[J]. IEEE systems journal, 2013, 7: 323-334
- [7] Okman L, Gal-Oz N, Gonen Y, et al. Security issues in nosql databases[C]//Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on. IEEE, 2011: 541-547
- [8] Chodorow K. MongoDB: the definitive guide[M]. " O'Reilly Media, Inc.", 2013
- [9] Han J, Haihong E, Le G, et al. Survey on NoSQL database[C]//Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE, 2011: 363-366

[10] Michael K, Miller K W. Big data: New opportunities and new challenges [guest editors' introduction][J]. Computer, 2013, 46(6): 22-24

【作者简介】



¹ 傅翔（1990-），男，汉族，硕士，研究方向：数据挖掘与数据存储，国防科技大学。
Email: xjtusefox@163.com

² 朱伟辉（1991-），男，汉族，硕士，研究方向：数据挖掘与信息安全，国防科技大学。Email: 591283456@qq.com

³ 韩伟红（1973-），女，汉族，博士，研究员，研究方向：数据库与数据挖掘，国防科技大学。
Email: hanweihongnudt@139.com