高速 IP 网络入侵检测系统的安全数据组织和管理

陈浩　　丁伟•
（东南大学计算机科学与工程系，南京，210096）

The Security Data Organization and Management of IDS over High-Speed IP Network

Chen Hao，　Ding Wei
（Dept. of Computer Science and Engineering, Southeast University, 210096 Nanjing）

【摘要】　本文对高速 IP 网络入侵检测系统的设计和实现中在安全数据的存储和管理中遇到的一些实际问题进行了研究总结。通过分析这些问题的特征和性质，本文阐述了一些可行的数据存储子系统性能改进策略，并结合实践对这些改进策略进行了分析。
【关键词】　入侵检测系统（IDS），安全数据，数据存储，数据库

【Abstract】 This paper is a summary of the research work on some problems encountered in the storage and management of security data throughout the design and implementation of an intrusion detection system over high-speed IP network. Through analyzing the features and characteristics of these problems, this paper has expounded several practicable strategies of improving the performance of the data storage subsystem, and has carried out some detail analysis of these strategies based on our experiments.
【Keywords】intrusion detection system(IDS), security data, data storage, database

1. 引言

IP 网络是指以 TCP / IP 协议为基础的通信网络。因特网是 IP 网的一种，也是具代表性的 IP 网络。1994 年以来，因特网在中国进入正式使用和不断发展阶段，国内各大运营商都建设有自己的互联网，组成了中国互联网的基本构架。随着网络的不断普及和广泛应用，各类基于网络的攻击事件大幅度增加。据中国教育科研网华东（北）地区网络中心统计，管区内每天所检测到的攻击事件达到数万起，实际发生的至少为数百起。运用入侵检测系统来监控、追踪和响应各类攻击事件，维护网络的正常运作和用户的信息安全，对于网络的管理和运营方来说，已经成为一种必然的选择。

入侵检测系统是近十几年来发展起来的一种主动计算机网络安全防范技术，入侵检测系

• 作者简介：陈浩，东南大学计算机科学与工程系硕士研究生，主要研究方向为网络安全。丁伟，工学博士，东南大学计算机系教授、硕导，主要研究方向包括网络管理、网络安全、网络体系结构、开放分布式处理等。
• 定稿日期：2002/6/19

统通过各种手段和技术对系统进行实时的监测，发现来自系统外的入侵者和系统内部的滥用者，为计算机系统提供完整性、可用性以及可信性的主动保护。

入侵检测系统必须完成的一个重要任务是安全数据的存储和管理。这些安全数据包括：系统产生的安全事件，安全结论，安全状况统计报告，会话还原数据，系统日志，辅助信息等。安全数据是入侵检测系统为使用者提供用户可见信息的数据来源，是对安全状况进行分析以及对安全事件进行处理的重要依据。安全数据存储和管理机制的好坏，直接关系到系统的性能好坏、系统能够给用户提供的信息质量和系统的可用性，是入侵检测系统的设计中相当重要的一个环节。一个前端采集和分析能力强大的系统，如果后端处理能力底下、无法给用户提供及时、准确、有用的信息，也称不上是一个好的系统。

2. 问题分析

对于高速 IP 网络入侵检测系统来说，其安全数据具有一个十分显著的特点，就是数据量大、流量高、存储压力巨大。根据 Cernet 主干信道上的入侵检测系统时机运行结果显示，在峰值时刻，每秒钟需要记录入库的安全数据记录经常达到 500~1000 条甚至更多。

在这种高速环境下，安全数据的组织和管理所要解决的一个首要问题就是，系统必须要具备足够的处理速度来存储前端报送的数据。否则，数据将会丢失，导致信息的有效性下降，和用户的使用困难；如果牺牲前端性能来满足相对较慢的安全数据存储瓶颈，就将拖垮整个系统的效率。所以，如何高速地实时存储数据，是我们第一个要关心的问题。

安全数据的组织和管理所要解决的第二个问题是存储效率问题。这个问题是由第一个问题所带来的。在入侵检测系统中，其它模块的功能是“通过数据”，“处理数据”，而数据存储模块的功能是“储存数据”，具有对空间要求苛刻的特点。高速网络环境下，安全数据的量非常巨大，如果存储效率不加考虑，就会极大地浪费存储空间，增加存储成本。另一方面，即使能够把数据的存储效率提高一点点，那么总的空间节省也会非常可观。因此，如何尽最大可能提高数据存储效率，是我们第二个要关心的问题。

安全数据的组织和管理所要解决的第三个问题是面向用户需求的，也就是数据的组织结构要有利于用户的查询。这个目标意味着数据库的设计、文件和目录的结构应该有利于查询程序（界面）的开发，能够满足用户的需求。一个简洁、有效、可靠、功能强大的数据存储结构，是我们第三个要关心的问题。

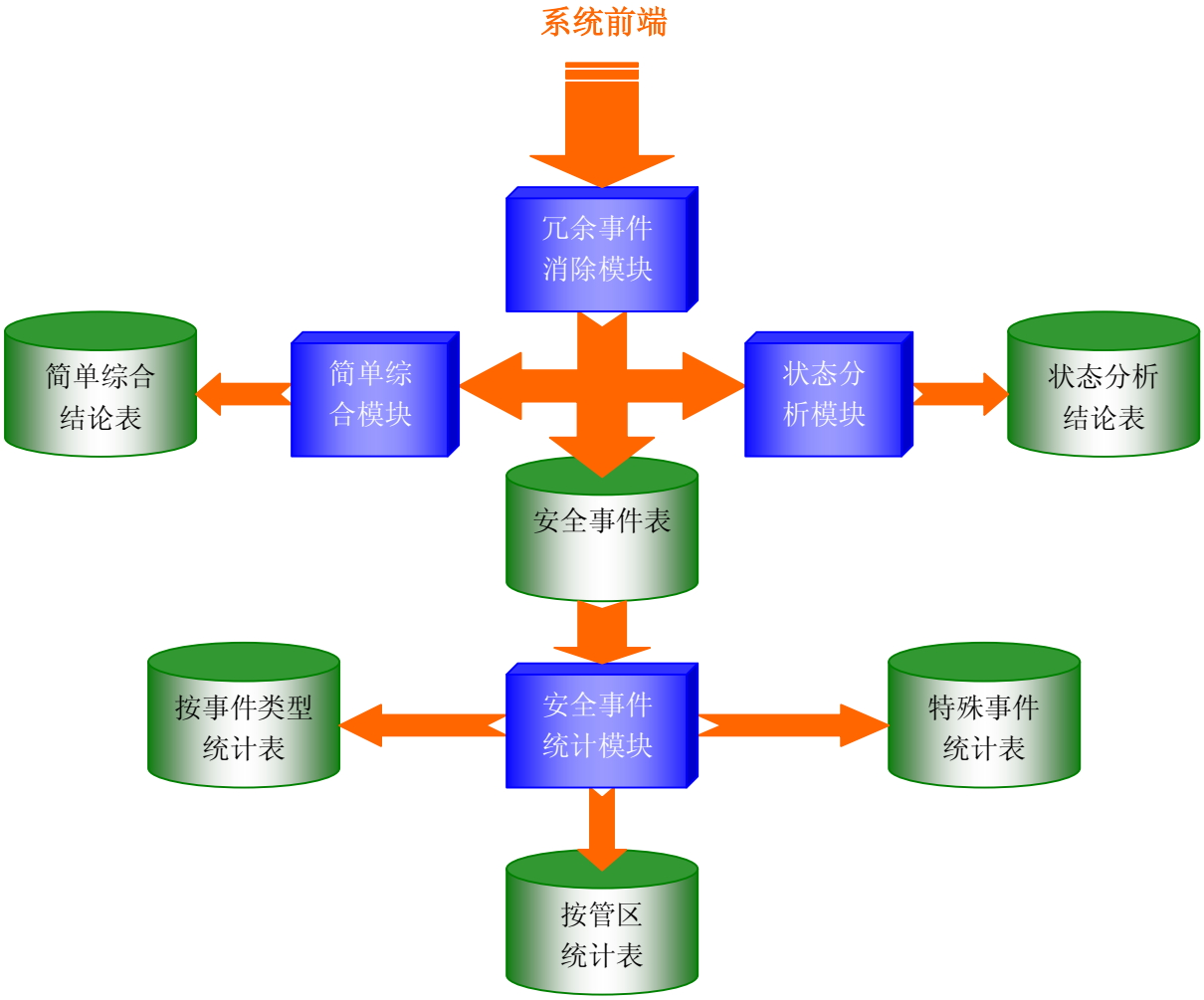
安全数据的组织和管理所要解决的第四个问题也与用户需求有关，就是系统要具备足够快的查询速度。在大数据量的环境下，如果没有一个好的组织结构和查询策略，则检索速度很可能是无法被用户接受的，特别是在检索原始数据的时候。为了给用户提供快速便捷的查询，在数据库的设计、文件和目录结构的设计方面，有很多需要权衡的因素。如何为用户提供高效的查询和管理接口，是我们第四个要关心的问题。

3. 组织和管理策略

3.1 数据组织机制的设计

入侵检测系统的原始记录，即前端报送的安全事件，是大量的、孤立的、缺乏可读性的。我们要把这些记录转化为用户所关心的数据，包括各种安全结论、状态结论、统计报表等。这样，用户所面对的，就是经过分析整理的高级别安全数据，这些数据更加接近客观世界，更符合人脑的思维形式，从而使用户更容易从中获取有用的信息。

我们实现的一个高速 IP 网络入侵检测系统的后端处理流程如下图所示：



图一　　入侵检测系统后端的数据组织和流向

箭头的粗细表示各阶段所处理的数据量的大小。

前端报送的安全事件经过冗余消除模块的处理后，大批量的原始记录存入安全事件表中；同时，对冗余消除模块所输出的数据，生成以下两种安全结论：由简单综合模块进行简单综合分析，对一系列相关的安全事件进行关联，分析其攻击特征与性质，生成简单综合结论，存入简单综合结论表中；另外，由状态分析模块对被攻击的对象当前的安全状态进行状态分析，生成状态分析结论，存入状态分析表中。到达安全结论这一级，数据量已经大幅度减少，并且含义更加清晰明确（发生了一次什么样的攻击？目标受侵害程度如何？），更有利于用户分析得出结论。

通过调查我们发现，用户对一段时间内所发生的安全事件的管区分布和类型分布比较感兴趣。针对这一需求，我们设计了一些对原始记录所作的统计表：统计各管区每天发生的安全事件数量的管区统计表，统计各种安全事件每天发生次数的事件类型统计表，以及对我们特别关心的个别类型的安全事件所做的特殊事件统计表，例如 Nimda 蠕虫的大规模爆发期间，专门对 Nimda 蠕虫所做的攻击情况统计。这些统计表的数据量比原始数据有了大规模的减少，差别达到两个数量级或更高。因此，用户对安全结论和安全事件统计报告所进行的

查询，可以以很快的响应速度得到查询结果。

3.2 数据查询机制的设计

对应于以上组织结构，后端提供给用户的查询接口也应该体现出一定的层次性：首先，提供给用户按管区或者按攻击类型的统计报告的查询接口，以及简单综合结论和状态分析结论的查询接口；此时，用户所面对的是精简过的数据，可以以较快的速度对所关心的安全状况得到一个大概的了解。

当用户对某一些具体范围内的安全事件产生兴趣时，可以设定查询条件，进一步在安全事件表中查询原始安全数据。由于用户必须输入严格的查询条件：准确日期，加上管区或攻击类型之一，加上时间段，才能查询安全事件表，因此用户所能够看到的就仅仅是他所关心的很小一部分数据，这就避免了大量安全数据杂乱无章地显示在用户面前，造成用户阅读的困难。

3.3 数据管理机制的设计

以上的数据组织形式虽然给用户的查询带来了方便，但是数据量大、存储空间占用高的矛盾仍然没有消除。并且，安全事件表的记录太多，尺寸太大，查询速度仍然难以得到保证。为此，我们专门针对安全事件表做了以下优化设计。

首先，把安全事件表按日切分成日表。日表的表名为：表名+日期（yyyymmdd），每张日表中只记录起始时间发生在当天（00：00：00~23：59：59）的安全事件。日表的尺寸适中，查询速度较快。用户在查询安全事件时必须首先指定日期，这样就保证了一次检索只需要对一张表进行操作，不需要进行跨表查询。

其次，系统运行一个守护进程，这个进程于每天定时进行以下操作：建立后一天的日表（无索引）；为前一天的日表建立索引；对前一天的日表进行统计，生成按日的统计记录并插入统计表中；删除不必要的数据库系统日志。这些操作一般选在系统每天最空闲的时刻，也就是凌晨 3-4 点运行，以减小对冗余消除进程和安全结论分析进程的影响。

再次，为日表建立查询索引。考虑到用户最经常使用的查询条件是时间范围，其次是 IP 地址或地址段、攻击类型，我们在安全事件的起始时间和事件 ID 字段上建立聚集索引，在源宿 IP 地址、攻击类型字段上建立非聚集索引。经实践，这种索引结构可以使用户查询数据时效率达到比较可以接受的程度。对日表的索引采取以下维护策略：每天的安全事件表刚被生成的时候不建立任何索引，也没有任何关键字、限制条件，这样数据入库的速度可以达到最大；等当天的数据已经全部入库，即次日凌晨，再建立所有的索引结构。由于主索引的顺序（事件 ID）就是事件记录入库的次序，因此一般情况下不需要对表中的记录位置做任何调整，在很短的时间内就可以建立起全部的索引。这样，只有查询当天的安全数据的查询效率会低一些，但整个系统的运行效率因此而得到了保证。

最后，后端提供了过期数据自动清除和数据导入导出的功能，以节省磁盘空间。对于超过一定保存期限的安全数据，用户可以设置为每日自动删除，也可以由用户手工以文件形式备份到大容量外部存储设备上，例如磁带机或刻录光盘。在用户需要查询过去的数据时，可以插入相应的磁带或光盘，再把数据导入回数据库中，然后执行查询。

4. 将来的改进方向

4.1 数据集的淘汰机制

高速网络入侵检测系统的数据存储模块面临着巨大的性能压力，一个根本的原因就是前端所设置的规则没有能够很准确地刻画和描述各种攻击行为的特征。这分为三种情况：第一

是事实上发生了攻击行为却没有被系统检测到，即漏报；第二种是事实上没有发生攻击行为而系统却认为发生了攻击，即误报；第三种是规则本身没有问题，但是产生的安全事件价值太低，不能给用户提供所关心的信息。对于数据存储模块来说，主要关心的和能够有所作为的是第三种情况：如何处理低价值或无价值安全数据。

为了节省存储空间，必须建立一个有效的数据淘汰机制，也就是以一定策略减小数据集的组成范围。这里有两个问题值得考虑：一个是淘汰范围的选择，另一个是淘汰时机的选择。淘汰范围的选择是指，什么样的数据不能淘汰，而什么样的数据可以淘汰；淘汰时机的选择是指，在哪个时间和处理环节把数据淘汰掉。

淘汰范围应该根据安全事件的类型来选取。对于某一类安全数据来说，把它们全部永久存储下来需要一定的存储代价；而如果不进行存储，就会导致安全事件被遗漏或无法追踪，也会带来一定的损失代价。对于重要程度较低的安全事件，例如主干网络上常见的 TCP SYN 扫描和 Proxy Hunter 扫描，数量很多，存储代价很大；而这类攻击如果被忽略，所带来的损失代价却低到几乎可以忽略不计。对两种代价进行比较，我们认为这种安全数据可以列入待淘汰的范围。

淘汰时机由用户需求而定。第一种方案是根本不生成低价值安全数据（由前端在规则级上进行淘汰），前提条件是用户对该类型事件完全不关心，不认为是安全事件，这不属于本文的讨论范围。第二，可以在生成安全结论和统计记录的时候忽略低价值数据。这样，存储空间的节省很有限，但 CPU 的处理负担可以大幅度减轻，而当用户确实有查询这些数据的需求时，仍然可以从原始数据中查询得到。第三个策略是低价值数据仍参与生成安全结论和统计记录，但是在数据库中保留一定期限后即由系统自动删除。这样，低价值数据在一定时间内对用户仍然是可见的。由于用户一般最关心的是近期内发生的安全事件，所以这种策略可以较好的平衡用户需要和系统存储能力之间的矛盾。这个期限可以根据系统存储容量和应用环境数据流量来估算确定。

4.2 数据的存储方式

众所周知的一个事实是，在一般情况下，直接对文件系统进行操作，其效率都会比对数据库进行操作要高得多。那么，如果我们直接把冗余消除模块所输出的原始安全数据写入文件系统，而不是写入数据库的表中，会对系统的性能产生什么样的影响呢？

首先，对文件进行写操作，其效率要远远高于，通过数据库系统所提供的编程接口，构造 SQL 语句来对数据库进行插入操作。因此，在数据被存储下来这个环节，写文件比写数据库具有更大的潜力，更能适应高速实时应用环境。

其次，从存储效率来看，由于数据库需要维护表索引、控制信息、日志、存储间隙等结构，而在文件中可以以最紧凑的格式存储各种类型的数据，所以总的存储效率是文件系统略高一些。但并不具有实质性的显著差异。

第三，从后期处理的效率来看，由于在生成各类统计表的记录时，需要把一天内的所有记录无条件地全部读取一遍，进行处理，所以数据库也不能表现出更好的效率来。采用文件存储，顺序读取时稍具优势。

最后，在用户需要查询安全事件时，数据库中数据的查询效率显著高于文件。在数据量非常大的应用环境下，由于文件的无结构特征，按照指定条件进行查询需要对所有的数据进行扫描，因此而造成的延迟通常是用户所无法接受的；而在建立了恰当索引的数据表中进行查询时，效率的提高是十分显著的（通常可以在 15 秒以内得到查询结果）。

综上所述，为了保证用户查询的效率，采用数据库来存储安全事件是比较好的存储策略。如果在更大流量的实际环境下，单一数据库设备无法处理高流量的前端数据时，可以考虑采用分布式数据库的组织形式。分布策略的制定要充分考虑到负载平衡的问题，这需要根据具

体的问题性质来进行研究。至于文件系统，由于查询上的局限性，可以只作为不需要用户直接查询的数据，例如中间数据，的存储方式。

5. 结束语

高速 IP 网络入侵检测系统是国内外的研究热点，在这个领域内有很多新问题需要被考虑。数据存储模块作为数据流的最终汇集地和直接面向用户需求的子模块，通常是重要的性能瓶颈，其性能的好坏直接制约着整个系统的表现。

通过一系列探索和尝试，我们设计和实现了一个完整的高速 IP 网络入侵检测系统的后端，其数据处理能力达到了实际应用环境的需要。在开发的过程中，积累了一些解决问题的经验，也产生了一些思路。本文就是对这些经验和思路的总结和探讨。

随着网络传输技术的发展，在不久的将来，因特网主干信道的带宽会进一步提高，达到 10G 甚至更高。根据经验规律，CPU 的主频每 18 个月翻一番，其它外部设备技术发展的速度更慢；而通信信道的带宽每 6 个月左右就会翻一倍。这种发展速度的不平衡所带来的矛盾使入侵检测系统的方方面面在未来都将面临更加严峻的考验。入侵检测系统势必会越来越难做。因此，需要更多的研究工作来推动入侵检测技术的发展，保障网络的运行安全。

参考文献:

1. 北京大学计算机系 刘宇虹 胡建斌 段云所 陈钟， *入侵检测技术*， 《网络安全技术与应用》2001 年第 9 期
2. 中国科技大学 陆殿军 张强 中科院高能所计算中心 李首杰，*入侵检测和漏洞检测系统*
3. 杨孝如 徐任 李立 彭立军 *Sybase 原理*、*高级系统管理与性能调优*，中国水利水电出版社
4. 陶浦洲 李强，*Sybase 数据库技术大全*，科学出版社
5. Cisco Inc. *Intrusion Detection System Director for UNIX Configuration and Operations Guide, Version 2.2.3*