ICS 01.140.20
A 14

C A D A L 项 目 标 准

CADAL 20901—2012

数字图书馆数字对象存储安全规范

Digital Library Digital Object Storage Security Specification

第一稿

2012-05-08 发布　　2012-05-09 实施

CADAL 项目管理中心　　发　布

目　　次

前言…… 115
引言…… 116
1　范围 …… 117
2　规范性引用文件 …… 117
3　术语和定义 …… 117
　3.1　黑客…… 117
　3.2　防火墙…… 117
　3.3　互联网…… 117
　3.4　内联网…… 117
　3.5　补丁…… 118
　3.6　Unix …… 118
　3.7　Linux …… 118
　3.8　独立磁盘冗余阵列…… 118
4　数字资源存储安全规范 …… 118
　4.1　防黑客攻击…… 118
　4.2　物理多副本…… 119
　4.3　异地容灾…… 120
　4.4　硬盘级数据安全存储…… 120

图 1　异地备份的结构 …… 120

前　言

《数字图书馆安全标准规范集》包括以下 4 个部分：

——第 1 部分：数字图书馆数字对象存储安全规范；

——第 2 部分：数字图书馆访问控制规范；

——第 3 部分：数字图书馆数字资源长期保存规范；

——第 4 部分：数字图书馆安全传输标准。

本标准是其中的第 1 部分。

本部分的制定依据标准化工作导则第 1 部分(GB/T 1.1—2009)。

本部分是由大学数字图书馆国际合作计划(CADAL)项目管理中心提出并归口。

本部分起草单位：数字图书馆教育部工程研究中心。

本部分起草人：尹彦飞、张寅、边科。

引　　言

数字资源的存储安全对于数字图书馆的建设至关重要。本标准在CADAL数字图书馆建设的实践基础上，规范了防攻击、物理多复本、异地容灾、硬盘级数据安全存储相关的策略，提供了切实可行的具体措施，以从技术和管理角度确保数字资源存储的安全。

数字图书馆数字对象存储安全规范

1 范围

本标准规定了CADAL项目数字对象存储安全规范。

本标准适用于CADAL项目中对数字资源的安全管理。

本标准适用的数字资源包括印刷文献的数字化衍生物。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

RFC 3767 2004.6

RFC 3760 2004.4

RFC 3157 2001.4

3 术语和定义

3.1 黑客 Hacker 缩写：HACKER

利用系统安全漏洞对网络进行攻击破坏或窃取资料的人。

3.2 防火墙 Firewall 缩写：FIREWALL

防火墙指的是一个由软件和硬件设备组合而成，在内部网和外部网之间、专用网与公共网之间的界面上构造的保护屏障，是一种获取安全性方法的形象说法，它是一种计算机硬件和软件的结合，使Internet与Intranet之间建立起一个安全网关，从而保护内部网免受非法用户的侵入，防火墙主要由服务访问规则、验证工具、包过滤和应用网关4个部分组成，防火墙就是一个位于计算机和它所连接的网络之间的软件或硬件。

3.3 互联网 Internet 缩写：INTERNET

互联网指由多个计算机网络相互连接而成，而不论采用何种协议与技术的网络。

3.4 内联网 Intranet 缩写：INTRANET

内联网又称企业内联网，是用因特网技术建立的可支持企事业内部业务处理和信息交

流的综合网络信息系统，通常采用一定的安全措施与企事业外部的因特网用户相隔离，对内部用户在信息使用的权限上也有严格的规定。

3.5 补丁 Patch 缩写：PATCH

补丁是生产商会为了维护消费者的合法权益和公司的信誉，而对存在质量问题的软件进行修理所使用的程序或者是代码。

3.6 Unix 缩写：UNIX

Unix 是一个强大的多用户、多任务操作系统，支持多种处理器架构，按照操作系统的分类，属于分时操作系统，最早由 KenThompson、DennisRitchie 和 DouglasMcIlroy 于 1969 年在 AT&T 的贝尔实验室开发。

3.7 Linux 缩写：LINUX

Linux 是一套免费使用和自由传播的类 Unix 操作系统，是一个基于 POSIX 和 UNIX 的多用户、多任务、支持多线程和多 CPU 的操作系统，该操作系统内核由林纳斯·托瓦兹于 1991 年 10 月 5 日首次发布。

3.8 独立磁盘冗余阵列 Redundant Array of Independent Disks 缩写：RAID

独立磁盘冗余阵列是把相同的数据存储在多个硬盘的不同地方的方法。通过把数据放在多个硬盘上，输入、输出操作能以平衡的方式交叠，改良性能。因为多个硬盘增加了平均故障间隔时间，储存冗余数据也增加了容错。

4 数字资源存储安全规范

为了保证数字资源能够安全的存储，数字资源在存储过程中应该符合以下要求。

4.1 防黑客攻击

黑客作为利用系统安全漏洞对网络进行攻击破坏或窃取资料的人，对于数字资源的安全存在很大的威胁。因此，为了严防黑客的攻击，数字资源存储系统应该具有以下的保护。

4.1.1 密码保护

密码保护的主要功能就是提高账户安全性，除密码外又加了一把“锁”。传统的账号加一组密码的账户保密方式存在很大的安全隐患，当今互联网木马猖獗，这种账户极易被不法者使用盗号木马利用系统漏洞轻易盗取或破解，如键盘钩子这一类的木马工具。密码保护就是为增强账号安全性而诞生的，是账号的二级密码，相当于为账户再加了一把锁，安全性大大提高。

在数字资源存储系统中，密码保护可分成两个部分：一个是对数字资源存储系统管理者的密码保护，这个层面的保护主要是为了避免恶意攻击者利用管理员账户在系统中进行

恶意的破坏活动；另一个是在数字资源传输过程中需要对使用到的密码进行加密等保护措施，这主要是防止恶意用户通过已知的密码进行数字资源的下载。

4.1.2 防火墙

防火墙指的是一个由软件和硬件设备组合而成，在内部网和外部网之间、专用网与公共网之间的界面上构造的保护屏障，是一种获取安全性方法的形象说法，它是一种计算机硬件和软件的结合，使 Internet 与 Intranet 之间建立起一个安全网关(security gateway)，从而保护内部网免受非法用户的侵入。防火墙主要由服务访问规则、验证工具、包过滤和应用网关 4 个部分组成，防火墙就是一个位于计算机和它所连接的网络之间的软件或硬件。该计算机流入流出的所有网络通信和数据包均要经过此防火墙。

在网络中，所谓“防火墙”，是指一种将内部网和公众访问网(如 Internet)分开的方法，它实际上是一种隔离技术。防火墙是在两个网络通信时执行的一种访问控制尺度，它能允许你“同意”的人和数据进入你的网络，同时将你“不同意”的人和数据拒之门外，最大限度地阻止网络中的黑客来访问你的网络。换句话说，如果不通过防火墙，公司内部的人就无法访问 Internet，Internet 上的人也无法和公司内部的人进行通信。

在数字资源存储系统中，使用防火墙对访问的 IP 进行严格的限制能够对系统起到非常好的帮户作用，能够确保系统资源的访问 IP 是系统所允许的。

4.1.3 杀毒软件

存放数字资源的服务器需要安装杀毒软件，主要对入侵到系统中的病毒等恶意攻击程序进行有效的清除，确保系统的安全防护机制能够平稳地运行。

4.1.4 系统升级

系统要进行不断的升级，新系统或者是升级之后的系统往往在防止外界攻击的能力上比老系统好很多，所以对于系统发布的补丁或者是新的版本，要进行版本的升级。同时，在系统的选择上应该选经过长时间验证的安全系统，例如 Unix、Linux。

4.2 物理多副本

数字资源应当存储在多个设备上，确保稳定地对外服务的同时数据不会丢失。多副本存放首先能够对较大的访问量进行负载均衡，能够保证服务的质量。同时，多副本也能够在一个或者多个服务出现丢失或者损坏的情况下实现快速的数据替换，确保数据不丢失。

对于多副本的个数应该包括以下三个方面：

(1)在线可快速访问备份：此种备份主要用来对系统中的数字资源进行利用，用户在线的实时访问，确保能够根据数字资源的唯一标识符快速地定位资源。

(2)离线可快速访问备份：此种备份是作为在线可快速访问备份的一个替代，当在线备份出现异常或者是损坏的情况下能够暂时地使用该备份进行对外的服务，其存储的方式应该与在线部分相同，并要确保自动的切换。

(3)离线存储不可快速访问备份：此种备份完全用于对数字资源进行长期的保存而提供，使用例如磁带等方式来完成，不要求能够快速地定位和访问。主要用于当前两种备份

发生丢失时，能够通过该备份进行完全的恢复。

另外，对于多个副本的数据应当考虑数据传输的便捷性，对于大数据而言，如果数据的拷贝效率太低，将大大影响多副本作用。在数字资源的存储中，要求至少满足以上三种多副本中的第一种和第三种，推荐提供第二种副本方案以确保系统长期对外服务。

4.3 异地容灾

通过互联网 TCP/IP 协议，备特佳容灾备份系统将本地的数据实时备份到异地服务器中，可以通过异地备份的数据进行远程恢复，也可以在异地进行数据回退和备份，如果想做接管，需要专线连接，只有在同一网段内才能实现业务的接管。异地备份的结构如图 1所示。

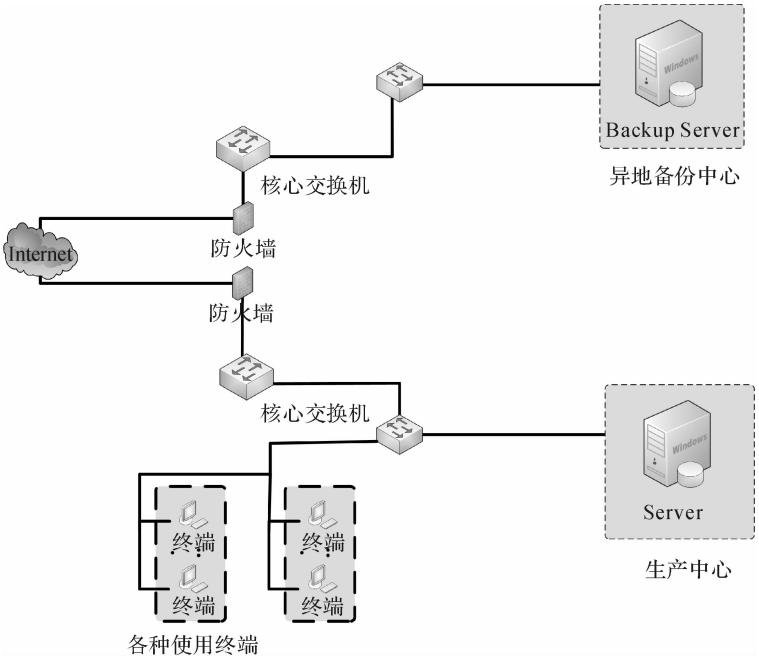

图 1 异地备份的结构

使用异地备份方式进行容灾是为了确保在一地或者多地出现不可抗力等因素导致数据丢失的情况下能够快速地重新恢复原有的数据。在数字资源存储过程中，异地备份应当不少于 3 份，并且选择的地点应该具有一定的广度。

4.4 硬盘级数据安全存储

本规范使用磁盘阵列(redundant array of independent disks，RAID)机制来保证数字资源的安全，根据分条、数据镜像以及奇偶校验技术的不同，RAID 可分为以下几个级别：

(1)RAID 0：在 RAID 0 中，数据是分带存储在 RAID 集的各个硬盘上的，因此利用了全部的存储空间。读取数据时，控制器会将各分条带数据重新组合起来。随着阵列中磁盘数目的增加，它便能够并发地读写更多的数据，因而性能也随之提高。RAID 0 特别适用于那些对 I/O 带宽需求很大的应用程序，然而，如果这些应用同时要求提高可用性，RAID 0 就无法提供数据保护功能和应对磁盘故障的高可用性。

(2)RAID 1：RAID 1 通过数据镜像来提高容错性。一个 RAID 1 组至少由两块硬盘构成。RAID 1 的数据恢复代价是所有 RAID 级别中最小的。这是因为 RAID 控制器利用镜像磁盘中的数据进行数据恢复，并同时继续对外提供服务。RAID 1 适用于那些对高可用性有需求的应用。

(3)嵌套 RAID：许多数据中心对 RAID 阵列的数据冗余和性能都有需求。RAID 0＋1 和 RAID 1＋0 集成了 RAID 0 的性能优势和 RAID 1 的冗余特征，将镜像和分条的优点组合起来。这类 RAID 需要由偶数数量的磁盘构建，且至少需要 4 块磁盘。

(4)RAID 1＋0 通常被称作“分条的镜像”：RAID 1＋0 的基本构成是镜像对。也就是说，数据首先被镜像，然后再将两个副本分别分条存储在 RAID 集的多个硬盘上。当替换故障磁盘时，我们只需要建镜像。换句话说，阵列控制器利用镜像组中的幸存磁盘来完成数据恢复，并继续提供服务。幸存磁盘中的数据将被复制到新替换的磁盘上。RAID 0＋1 也被称作“镜像的分条”：RAID 0＋1 的基本构成是条带。这意味着数据将首先分条存储到各个硬盘上，然后再对条带生成镜像。当一块磁盘失效时，整个条带都将失效。重建操作必须复制整个条带：从幸存条带的各磁盘中将数据复制到失效条带的相应磁盘上。这将给幸存磁盘带来额外的和不必要的 I/O 负载，并且 RAID 集很容易引发二次磁盘失效。

(5)RAID 3：通过存储分带提供性能，并利用奇偶校验提升容错性。它将奇偶校验信息存储在一个专用的磁盘上，因此，如果有一个磁盘损坏，数据是可以被重新构造的。例如，5 块磁盘中的 4 块将用于存储数据，而另外一块存储校验和。因此，总的磁盘开销将是数据磁盘的 1.25 倍。在 RAID 3 中，数据总是以整个条带为单位读写的，从而所有磁盘能够并发地执行操作，而不存在只更新同一条中某些存储带的部分写操作。RAID 3 为传输大量数据提供了很高的带宽，因而常应用于视频流服务等涉及大量顺序数据访问的场景中。

(6)RAID 4：与 RAID 3 类似，RAID 4 也通过存储分带提供高性能，并利用奇偶校验提升容错性。数据被分带存储在除了校验磁盘以外的其他磁盘上。奇偶校验信息存储在一个专用的磁盘上，以备磁盘损坏时重构数据。分带是在磁盘块层次完成的。与 RAID 3 不同的是，RAID 4 的数据磁盘支持独立访问。因此，某个数据单元可以从单块磁盘中读写，而无须访问整个条带。RAID 4 提供了很好的读吞吐率和较好的写吞吐率。

(7)RAID5：是一种适用性很强的 RAID 实现。它与 RAID4 类似的地方在于它也采用了分带技术，而且不同磁盘上的存储带是可以单独存取的。二者的不同点在于它们存储校验值的方法。由于 RAID 4 将校验值存储在一个专用的磁盘上，这就使得校验磁盘成为写性能瓶颈。在 RAID 5 中，校验值是分布存储在所有磁盘上的，这种方法克服了校验值写性能瓶颈的缺陷。

(8)RAID 6：RAID 6 的工作模式和 RAID 5 的基本相同，但它引入了第二个校验元素以应对 RAID 组中的两块磁盘同时失效的情况。因此，RAID 6 至少需要 4 块磁盘。RAID 6 也将校验值分布在所有磁盘上。由于 RAID 6 的写代价要比 RAID 5 大，因此 RAID 5 的写性能要比 RAID 6 好。此外，RAID 6 有两个校验集，因此它的重建操作要比 RAID 5 更耗时。

在数字资源的存储过程中，我们规定，对于一般的数字资源，例如普通图书、视频等，所满足的最低标准是 RAID 3，而对于一些特殊资源，例如拷贝较少的古籍、数字资源制作费时的书画等需要满足的最低级别是 RAID 4。