访问日志解析与敏感数据安全

■中国移动通信集团广东有限公司南方基地 严敏 李冠道 周乐坤

编者按：当前，大数据平台的应用层面、基础层面存在大量安全隐患，现有的安全防护手段远没有跟上大数据时代独有的安全需求，大数据平台、各应用系统之间存储、共享、流转的敏感数据正时时刻刻受到信息窃取、数据篡改、黑客攻击、病毒侵袭等诸多安全威胁。

数据的访问、流转途径很多，当前针对每种途径的监控及防护手段基本成熟，出现了诸如4A、字符堡垒、文件堡垒、桌面终端管控系统、DLP、入侵防护等安全设备。虽然每类技术大都能监控、记录数据的访问操作过程，发挥自己应有的作用，但大部分不具备甄别对敏感数据操作行为，更不能全面分析、呈现敏感数据流转过程。

针对当前大数据平台下的访问日志解析与敏感数据流转视图展现能力的不足，本文提出一整套解决方案：首先采集大数据组件以及数据访问操作行为日志，结合敏感数据分类分级、日志格式化基础信息等进行析取和格式化，并标识敏感数据访问。然后横向关联所有敏感数据访问操作类日志，综合分析敏感数据访问、流转途径。最后通过可视化工具，绘制出敏感数据访问操作流转视图。

敏感数据分类、分级定义

对敏感数据进行分类分级定义，根据数据敏感程度采取与数据安全风险相适应的管理措施，为敏感数据访问行为识别提供依据。

1. 敏感数据分类

敏感数据分类包括：用户身份和鉴权信息、用户数据及服务内容信息、用户服务相关信息三大类，每项大类又可分为多项子类。

用户身份和鉴权子类信息包括：自然人身份标识、网络身份标识、用户基本资料、实体身份证明、用户私密资料、用户密码及关联信息。

用户数据及服务内容子类信息包括：服务内容数据、联系人信息。

用户服务相关子类信息包括：业务订购关系、服务记录和日志、消费信息和账单、位置数据、违规记录数据、终端设备标识、终端设备资料。

2. 敏感数据分级

敏感数据分为四个级别，分别为极敏感级、敏感级、较敏感级、低敏感级。

极敏感级数据分类包括：实体身份证明、用户私密资料、用户密码及关联信息。

敏感级数据分类包括：自然人身份标识、网络身份标识、用户基本资料、服务内容数据、联系人信息、服务记录和日志、位置数据。

较敏感级数据分类包括：消费信息和账单、终端设备标识、终端设备资料。

低敏感级数据分类包括：业务订购关系、违规记录数据。

日志采集、解析、处理及敏感数据操作标识

采集大数据平台组件、平台访问控制设备、应用程序接口、网络设备、安全设备等日志信息，通过标准化引擎对日志属性进行解析、格式化，输出统一格式的安全日志，并对标准化后的日志进行过滤、归并等降噪处理，结合敏感数据分类、分级信息，通过正则表达式等模式匹配方法识别并标识敏感数据操作类安全日志，为敏感数据流转过程分析提供数据支撑及依据。

日志采集

数据采集范围广、采集方式丰富、采集能力强，确保采集数据的全面性、及时性及准确性。

1. 采集范围

采集范围覆盖敏感数据生成、传输、存储、使用、共享、销毁各个环节，包括：

大数据平台组件（HDFS、HBase、Hive、Sqoop）；

访问控制设备（4A、字符堡垒、FTP 堡垒，桌面终端管控系统）；

安全设备（网络 DLP、终端 DLP、入侵防护、网络嗅探、数据库嗅探）；

网络设备（交换机、路由器）；

主机（各类操作系统）；

数据库（各类关系型数据库）；

中间件（各类中间件系统）。

2. 采集方式

采集方式包括被动及主动两种方式，被动采集支持 SYSLOG、TRAP 两种方式，可实时接收日志信息；

主动方式支持 FTP、SFTP、JDBC、webservice 协议，可周期、准实时采集设备日志信息。

日志解析

结合日志知识库信息，通过日志标准化引擎对设备原始日志信息进行解析和抽取，根据日志分类识别表达式识别出原始日志对应的安全日志分类，并按该分类所适用的属性字段析取相应的值。然后由标准化引擎根据特征值公式进行计算、精准匹配出安全日志，最终输出的安全日志信息包括日志类型、操作对象、操作命令、账号、源 IP、目的 IP、报送设备 IP、日志内容、日志级别、发生时间等相关属性。

1. 日志分类识别

通过正则表达式规则匹配算法的一系列特殊字符构建日志分类的匹配模式，然后依据匹配模式对原始日志进行匹配，匹配成功后析取正则表达式中的分组变量或特殊变量值，并将属性变量及其值以 key-vale 形式缓存在 map 里，通过日志分类识别表达式及其已经析取出的变量值计算出所属的日志分类。

原始日志样例：<86> sshd[12915]: Accepted password for root from 186.31.27.53 port 4991 ssh2

日志分类识别规则样例：<(\d+)>([^;]+):\s*(Accepted password) for\s+(\w+)\s*from\s*([\w|\W]+)\s+port\s+\d+\s+(\w+)

日志分类识别规则样例：if {原始数据析取变量：主账号}="" then {BM 数据库事件：BM 数据库绕行操作

事件｝ else ｛BM数据库事件:BM数据库堡垒操作事件｝

2. 安全日志识别

通过安全日志的特征值公式(逻辑表达式)和已析取出的变量值计算出特征值。将计算出的特征值和所有日志分类下安全日志(操作细项)的特征值进行匹配,匹配成功后生成对应安全日志对象,完成安全日志识别。

特征值公式是由解析的日志基础属性变量、逻辑关系符等逻辑表达式组成,最终通过内部表达式规则引擎算出结果。

特征值公式样例:如果｛日志属性:日志ID｝=527并且｛日志属性:源地址｝<>'128.12.17.10'则'绕行登录'否则｛日志属性:日志ID｝。

安全日志分类特征值样例:25。

日志过滤、归并,降噪处理

对不具备分析意义的安全日志进行过滤,减少不可信、不重要的安全日志,析取出真实有价值的日志。对重复发生或大部分属性相同的安全日志,在不影响后续事件分析的前提下对个体进行合并,减少事件总数量。

安全日志过滤、归并规则以日志类型、源地址IP、目的地址IP、日志发生时间、操作命令等日志属性值作为条件参数,支持等于、不等于、大于、小于、包含、LIKE、IN等数学函数表达式,如:｛事件属性:目的地址｝like'192.132.12.%' and ｛日志 属性:操作命令｝in'rm,vi'。

识别并标识敏感数据操作日志

结合敏感数据分类、分级定义的敏感数据特征属性,利用正则表达式等模式匹配方法对安全日志相关属性如操作对象、操作内容等进行匹配,匹配成功,则标识该安全日志为敏感数据操作日志。

敏感数据流转路径分析

基于以上步骤分析出的不同设备产生的敏感数据操作类标准日志,通过对日志相关属性如日志类型、操作对象、操作命令、操作内容、时间、源IP、目的IP等进行多维、综合关联分析,输出敏感数据在各流转节点之间的流转关系。

1. 收集敏感数据源信息,确认敏感数据传播扩散起始点,收集的数据源信息包括敏感数据源设备类型、数据源IP、访问方式、访问策略、开放的服务等。

2. 获取所有敏感数据对象,保存至敏感数据对象列表SL中。敏感数据对象信息包括敏感数据源IP、敏感数据名称、敏感数据形态、敏感数据存储路径、敏感数据分类、敏感级别、敏感数据生成时间等。

3. 遍历敏感数据对象列表SL中的对象,找出敏感数据对象流转的第一级节点。以对象属性敏感数据源IP、对象名称、数据形态、存储路径为条件,与有敏感数据操作标识的标准化日志相关属性(如:源IP、操作对象名称、操作内容)进行匹配,匹配成功,则根据标准化日志相关属性信息生成过程敏感数据对象,并存储在过程敏感数据对象列表PL中,同时

生成敏感数据访问或流转路径节点对象，存储在流转路径节点对象列表 TL 中。

重复以上步骤直至遍历完 SL 中的所有对象。流转路径对象信息包括上一级节点 IP、当前节点 IP、流转方式、流转时间、敏感数据名称、账号。

4. 遍历过程敏感数据对象列表 PL 中的对象，找出该敏感数据对象访问、流转的下一级节点。以该过程敏感对象属性如敏感数据源 IP、对象名称、数据形态、存储路径为条件，与有敏感数据操作标识的标准化日志相关属性（如：源 IP、操作对象名称、操作内容）进行匹配，匹配成功，则将该对象移除 PL 列表，根据匹配的标准化日志相关属性信息生成过程敏感数据对象，并存储在过程敏感数据对象列 PL 表中，同时生成敏感数据访问或流转路径对象，存储在流转路径对象列表 TL 中。匹配失败，则将该对象移除 PL 列表。重复以上步骤直至遍历完 PL 中的所有对象。

视图生成及展现

根据输出的敏感数流转关系，利用可视化工具生成敏感数据流转视图。

敏感数据流转视图包括两个要素，分别为流转节点和节点之间有方向流转路径。流转节点信息包括：节点 IP、敏感数据名称、敏感数据级别等。流转路径信息包括：流转时间、流转方式、操作账号、操作命令。

1. 首节点及一级流转节点信息生成。以敏感数据源对象部分属性值如敏感数据源 IP、敏感数据名称等为参数，遍历查找流转路径节点对象列表所有节点对象数据，查找条件：敏感数据对象．敏感数据源 IP= 流转路径节点对象．上一级节点 IP 并且 两个对象敏感数据名称属性值相等。匹配成功，则以当前敏感数据对象属性值为准生成首节点信息，同时以所有与之匹配成功的流转路径节点对象属性值为准生成所有一级流转节点信息。

2. 更多流转节点信息生成。以流转路径节点对象 A. 当前节点 IP= 流转路径节点对象 B. 上一节点 IP 且两个对象的敏感数据名称相同为条件进行匹配，匹配成功，即表示敏感数据由 A 流转到 B，生成相应流转节点信息。同理，遍历分析所有流转路径节点对象，直至生成最后一个流转节点信息。

3. 通过可视化工具，利用敏感数据首节点、中间流转节点、最后一个流转节点之间的纵横向关系，绘制出敏感数据流转视图。

本方案具有如下几方面优点：

用于分析的数据源广，确保发现任何访问、操作敏感数据的蛛丝马迹。参与分析的数据源包括大数据平台组件、访问控制设备、安全设备、网络设备、主机、数据库等。

采用全量、多层次综合分析方法，全面、准确的发现敏感数据所有访问流转路径，并利用可视化工具，生成敏感数据访问视图，清晰再现敏感数据范围流转轨迹。

充分利用现有网络环境中网络设备、安全设备的功能及价值，从而能更好的降低企业运营成本。N