CAICT 中国信通院

大数据平台安全研究报告

中国信息通信研究院安全研究所

2021 年 1 月

版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本研究报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

前　言

2020 年全球遭到了新冠肺炎疫情的打击，据经济合作组织 2020 年 9 月 16 日发布的《经济展望中期报告》，受疫情影响，2020 年全球 GDP 预计萎缩 4.5%，各国面临着经济逆行的巨大挑战。为应对经济压力，各国纷纷将目光聚焦于数字经济，将其作为经济逆境中的新动力。据中国信通院发布的《全球数字经济新图景（2020 年）》预计，2020 年全球数字化转型技术和服务支出将增长 10.4％，达到 1.3 万亿美元，是 2020 年中为数不多的经济增长亮点之一。

大数据是数字经济的重要基础，蕴含着巨大的潜在价值，为探索客观世界新规律、减少战略决策中的主观因素影响提供了可能。据 2020 年 9 月 10 日国家互联网信息办公室印发的《数字中国建设发展进程报告（2019 年）》，2019 年我国大数据产业规模超过 8100 亿元，同比增长 32%，依托大数据，人工智能、区块链、工业互联网等数字经济产业得到快速发展。

大数据平台为大数据提供了计算和存储的能力，这使得海量的静态数据“活动”起来，并释放出自身价值。然而，一旦缺少了平台安全这个前提，数据价值的释放将受到阻碍。如果将大数据平台比作大厦，其价值的释放能力和安全能力分别是大厦的地面建筑和地基，地基的深度决定了大厦地面建筑的高度，地基不稳的大数据平台，注定只能是“空中楼阁”。为了“测量”并“夯实”大数据平台的安全“地基”，我院在 2020 年发起了卓信大数据平台安全专项

行动。行动开展过程中，我们发现大数据平台在建设和运维方面确实存在一定的安全问题，本报告对此进行了深入分析。

报告全文以本次专项行动中积累的安全检测数据为基础，从平台配置安全隐患和安全漏洞的分布规律、产生原因、危害影响、修复难度等维度分析了大数据平台的安全现状。同时，详细分析了形成该安全现状的问题根源，并给出了相应的解决方案建议。最后，从监管、标准、技术研究等方面提出了大数据平台安全未来的工作方向。

“工欲善其事，必先利其器”，在数字经济蓬勃发展的大背景下，我们深刻认识大数据平台安全，不仅需要技术手段上的安全防护，还需要安全意识和管理运维水平的同步提升，亟需多方共同协作努力。希望本报告能够为企业大数据平台安全建设提供参考，为行业大数据的健康发展提供助力。

目 录

一、大数据平台概况.....1
(一)大数据产业蓬勃发展.....1
(二)大数据平台应用模式多样化演进.....2
二、大数据平台安全现状.....4
(一)组件配置类安全隐患.....6
(二)组件安全漏洞.....9
(三)组件安全隐患统计分析.....12
三、大数据平台安全问题分析.....18
(一)基于 Hadoop 的开源大数据平台安全配置复杂度较高.....19
(二)安全漏洞修复对平台运行影响较大.....19
(三)大数据平台建设过程中安全投入不足.....21
(四)大数据平台重视边界防护忽视内部安全.....21
(五)企业大数据平台安全管理制度滞后.....22
(六)企业技术人员安全能力不足.....23
四、大数据平台安全解决方案建议.....23
(一)加强大数据平台安全基线管理.....24
(二)对大数据平台安全进行整体规划.....24
(三)大数据平台边界防护与内部安全建设并重.....25
(四)建立完善的大数据平台安全制度流程.....25
(五)增强企业技术人员安全能力.....25
五、大数据平台安全未来发展建议.....26
(一)加强企业大数据平台安全防护工作的监管.....26
(二)强化大数据平台安全防护技术研究.....27
(三)推动大数据平台安全产品和服务市场发展.....27
(四)构建大数据平台安全生态.....28

图 目 录

图 1 中国大数据市场规模预测 2
图 2 大数据平台架构 2
图 3 大数据平台类型分布占比 5
图 4 大数据平台组件配置安全隐患统计 6
图 5 大数据平台安全漏洞统计 10
图 6 单个组件安全隐患占比 13
图 7 单个组件配置安全隐患占比 13
图 8 单个组件安全漏洞占比 14
图 9 各类组件安全隐患占比 15
图 10 各类组件配置安全隐患占比 15
图 11 各类组件漏洞安全隐患占比 16
图 12 传输类组件安全隐患等级分布 17
图 13 存储类组件安全隐患等级分布 17
图 14 计算类组件安全隐患等级分布 18
图 15 平台管理类组件安全隐患等级分布 18
图 16 漏洞修复方式选择流程图 20

一、大数据平台概况

当今时代大数据产业持续发展，对于挖掘新的经济增长点大有益处，大数据成为了推动各行业发展的新动力。大数据平台作为大数据的载体起到了关键作用。

（一）大数据产业蓬勃发展

根据 IDC 最新预测，2020 年中国大数据相关市场的总体收益将达到 104.2 亿美元，较 2019 年同比增长 16.0%，如图 1 所示，增幅领跑全球大数据市场。

2020 年，大数据硬件在中国整体大数据相关收益中将继续占主导地位，占比高达 41.0%；大数据软件和大数据服务收入比例分别为 25.4%和 33.6%。而到 2024 年，随着技术的成熟与融合、以及数据应用和更多场景的落地，软件收入占比将逐渐增加，服务相关收益占比将保持平稳，而硬件收入在整体的占比则将逐渐减少。硬件、服务、软件三者的比例将逐渐趋近于各占三分之一的比例。IDC 预计，在 2020-2024 年的预测期间内，中国大数据相关技术与服务市场将实现 19.0%的年均复合增长率。

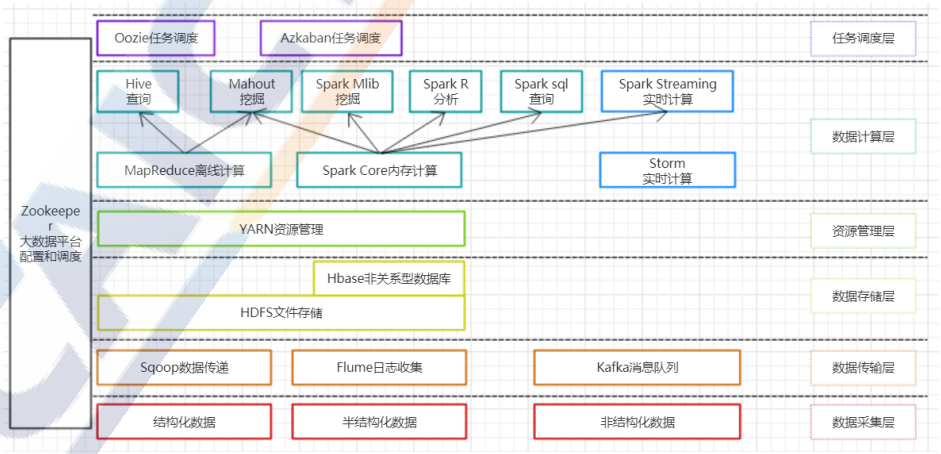


数据来源：IDC 中国，2020

图 1 中国大数据市场规模预测

（二）大数据平台应用模式多样化演进

大数据平台通常使用 Apache Hadoop 架构进行搭建，由存储、计算、平台资源管理、传输交换等类型组件组成整个平台，如图 2 所示。目前，大数据平台主要有基于开源技术的自建平台和商业化平台两种。自建平台灵活性强、自主性高，商业化平台安全性强、使用便捷。



数据来源：中国信息通信研究院

图 2 大数据平台架构

1. 自建大数据平台灵活性强、对企业技术水平要求较高

自建平台使用开源大数据平台技术架构，其优势在于有较高的自主性和灵活性，企业可以根据业务需求选择不同的功能组件，同时所有数据都在自己的管控范围内，避免第三方参与造成的数据泄露。除此之外，虽然自建大数据平台建设成本较低，但需要足够的人力资源和较高的技术水平，才能满足平台正常运行维护的需求。运维人员需要关注各个组件的配置与漏洞情况，根据最新的风险情报对平台进行加固，防止风险事件的发生。参与卓信大数据平台安全专项行动的企业中，只有少数几家企业采用自建的方式，这些企业都有自己的大数据技术团队作为支撑。

2. 商业化大数据平台易管理、安全性高

商业化大数据平台可分为委托式大数据平台和自主控制式大数据平台。委托式大数据平台由供应商进行日常维护，用户通过服务调用方式使用大数据平台。自主控制式大数据平台搭建在用户的服务器或私有云上，用户对大数据平台具有控制权，且自主进行日常维护。二者都采用成熟的商用大数据平台产品，产品安全性均经过供应商的专业测试，用户的日常使用亦可得到供应商的技术支持。

（1）委托式大数据平台成本低，供应商一般提供全托管式服务，但定制化程度不高

委托式大数据平台部署于供应商侧，其提供的服务通过接口等

方式供使用者调用。平台通常采用虚拟化技术，大数据平台服务被多个使用者共享。供应商对该种大数据平台拥有控制权，并负责日常运维。对用户而言，此种方式无须自行搭建大数据平台，前期使用成本较低。然而，该种方式的大数据平台对企业的定制化程度不高，企业部署的系统及软件需符合大数据提供商的相关技术标准。提供委托式大数据平台的厂商有：亚马逊、阿里云等。

（2）自主控制式大数据平台成本较高，功能全面、保密性强

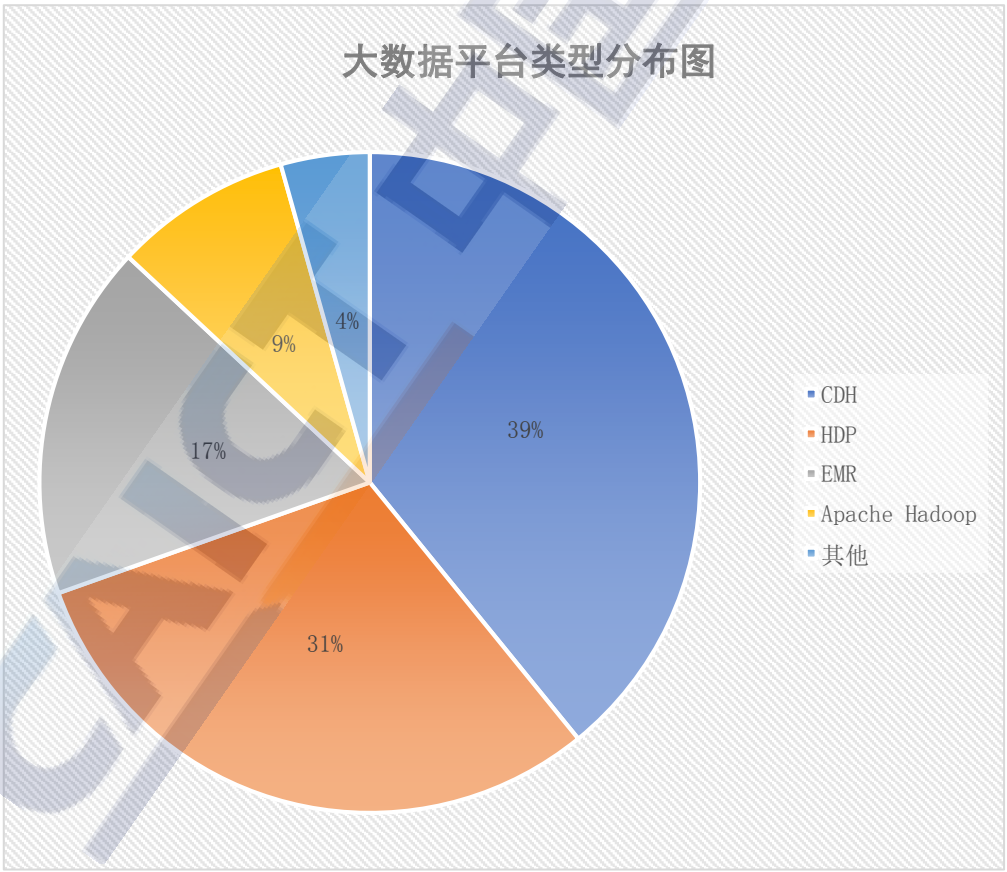
自主控制式大数据平台是企业将具备完善功能的大数据平台部署在自己的服务器或私有云上，并自行负责日常的大数据平台运维。此种大数据平台服务为企业独占，用户可根据自身需求进行定制化，因此该模式的平台保密性强，且功能全面。同时，自主控制式大数据平台可在一定范围内进行大数据组件的自主选择和搭配使用。相对于委托式大数据平台，自主控制式大数据平台的硬件使用和日常运维由企业自行负责，成本较高。

二、大数据平台安全现状

随着企业大数据平台对内运营的支撑能力不断提升，数据来源不断丰富，数据分析挖掘功能不断创新，数据安全问题与挑战日益增加，企业对大数据平台的安全保障要求也不断提高。目前，大数据平台组件往往独立设计、开发，并根据不同的业务需求进行组合搭建，若是对平台组件的安全管控不当，极易造成非法访问、敏感数据泄露等安全风险。

本章以参与卓信大数据平台安全专项行动的企业大数据平台安全检测初测结果作为数据来源，统计各类安全隐患的分布、排名和影响等级情况，同时分析了主要安全隐患产生的根源和可能的危害影响。以下统计数据不代表参与行动企业当下的大数据平台保障安全工作现状。

首先看到，本次专项行动检测的大数据平台类型基本涵盖了主流平台应用类型，具体分布如图 3 所示，CDH 平台占 39%，HDP 占 31%，EMR 占 17%（其中阿里云 EMR 占 9%，腾讯云 EMR 占 8%），原生 Apache Hadoop 集群占 9%，优刻得 Ucloud 和星环 TDC 等其他平台占 4%。

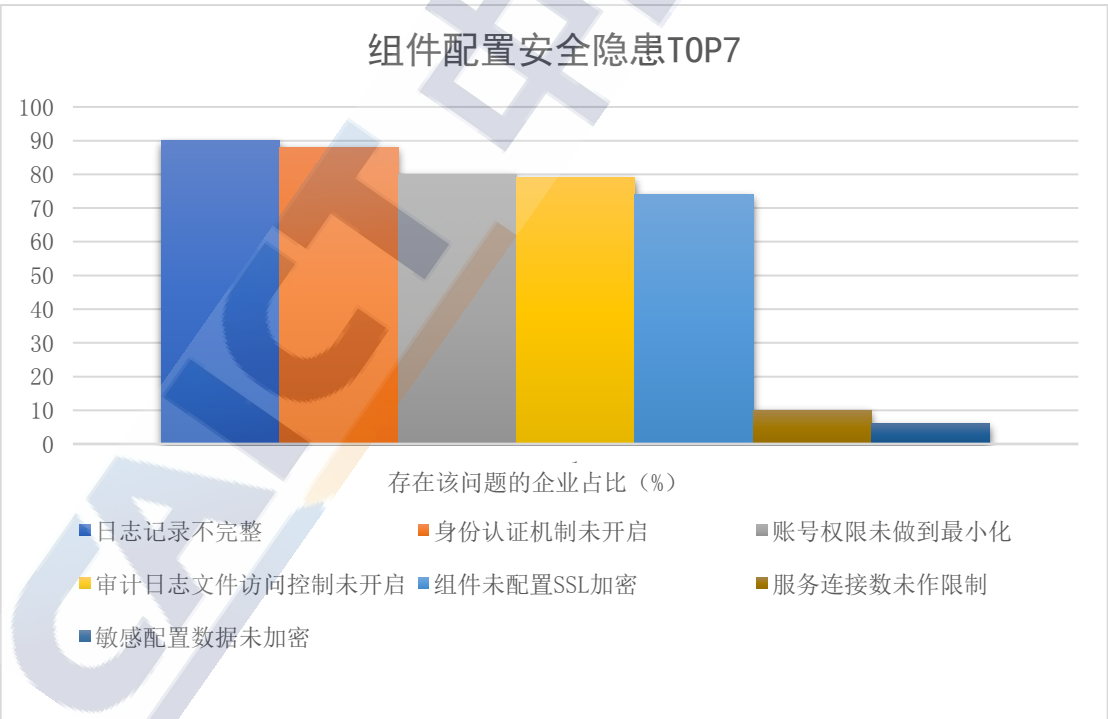


数据来源：中国信息通信研究院

图 3 大数据平台类型分布

（一）组件配置类安全隐患

从本次专项行动检测初测结果来看，日志记录不完整、身份认证机制未开启等组件配置安全隐患在受检企业大数据平台中普遍存在，如图 4 所示，其中排在前五位的是日志记录不完整、身份认证机制未开启、账号权限未最小化、审计日志文件权限未最小化和组件间数据传输未加密，出现以上配置安全隐患的企业在所有受检企业中的占比分别为 90%、88%、80%、79%、74%。这些配置管理上的安全隐患会给大数据平台带来数据安全审计日志不全、组件访问权限管理失控、敏感数据泄露或被篡改、集群拒绝服务等安全危害。本章对常见的组件配置类安全隐患出现的根源及可能引发的后果进行解析。



数据来源：中国信息通信研究院

图 4 大数据平台组件配置安全隐患统计

1. 组件配置安全隐患一：日志记录不完整

出现此类隐患的主要原因是组件的日志记录文件使用了默认的配置参数——ERROR。这种情况下，组件的日志文件只记录了组件已发生错误事件的日志信息，而未记录潜在的错误信息、组件访问记录等信息。潜在的错误信息往往隐含着口令爆破等攻击行为信息，组件访问记录包含了组件被访问的时间、操作者、事件类型、被访问资源记录等要素，这些信息的缺乏会导致后续数据安全审计因缺乏相应数据源，而无法及时发现数据滥用、内部泄露等风险。

2. 组件配置安全隐患二：身份认证机制未开启

出现此类安全隐患的主要原因是开启 Kerberos 认证机制会提高集群的消耗、降低工作效率。开启 Kerberos 认证机制后，所有的新节点加入集群时，都要向 KDC 申请身份凭证；每隔一段时间，集群中所有节点都要再次向 KDC 申请身份凭证；且所有外部应用（Client）对集群（Service）的访问都需要通过 Kerberos 认证后才能进行，相比 Client 到 Service 的点对点直接访问，开启 Kerberos 认证机制后，集群增加了额外的认证过程消耗。集群不开启 Kerberos 认证机制时，采用默认的 Simple 认证方式，通过简单 ACL（访问控制列表）方式执行外部用户的访问控制。若攻击者通过字典等攻击方式获取了系统用户名密码，则可以获得集群资源的控制权。不仅对集群的正常运行带来极大危害，也会使得平台上的数据完全暴露给攻击者。

3. 组件配置安全隐患三：账号权限未做到最小化

出现此类安全隐患的主要原因是账号权限分配不规范，权限配置方面未做到最小化，给组件运行账号分配的权限过高。有的企业给运行组件程序的账号分配了 root、administrator 或 supervisor 等高级别权限，有的企业把运行组件程序的账号放到了 sys、root、administrator、super 等超级管理员组中。权限过高的账号加重了运维人员误操作、内部人员窃取数据或破坏系统的安全风险，同时增加了外部攻击可能造成的破坏和影响程度。

4. 组件配置安全隐患四：审计日志文件权限未最小化

出现此类安全隐患的主要原因是未修改平台审计日志文件的访问权限，使用了默认参数值，致使审计日志文件的访问权限向平台所有用户公开，安全审计组和超级管理员组之外的用户也可以访问审计日志文件，进行修改、删除等操作。攻击者通过非法手段获取到任意一个账号，就可以对审计日志进行操作，隐藏数据窃取、篡改等攻击行为，使得集群安全审计和溯源失去准确的数据源，无法达到及时发现安全威胁、处置安全事件的目的。

5. 组件配置安全隐患五：未设置组件间传输加密

出现此类安全隐患的主要原因是未对平台服务与客户端间的数据转移实施加密，导致数据在传输过程中存在截取、篡改等安全风险，无法保证数据的保密性。

6. 组件配置安全隐患六：服务连接数未限制

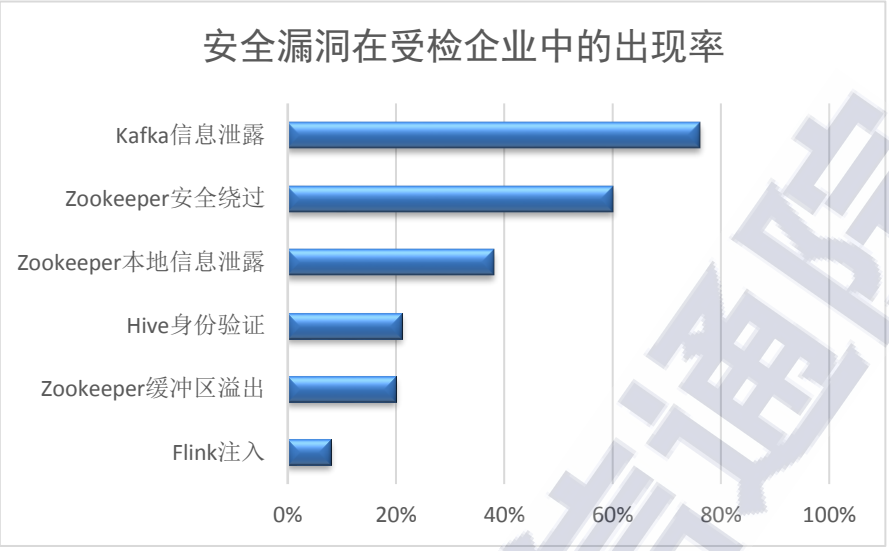
出现此类安全隐患的主要原因是未对组件服务进行连接数限制，出现此类安全隐患的主要有 HDFS、HBase、Zookeeper 等组件。若不对单一用户可发起的最大服务连接数以及多个用户同一时间的并发服务连接数进行限制，无论是正常的服务请求，还是恶意攻击，在请求数量过多的情况下都会导致集群服务瘫痪，影响大数据平台的正常运行。

7. 组件配置安全隐患七：敏感配置数据未加密

出现此类安全隐患的主要原因是未对配置文件中的敏感数据（如口令数据）进行加密，当攻击者窃取到敏感配置数据时，可根据数据内容获取关键线索，进而执行提权等操作，最终实现非法获取敏感数据的目标。

（二）组件安全漏洞

从本次专项行动检测初测结果来看， Kafka 信息泄露、Zookeeper 安全绕过、Zookeeper 本地信息泄露、Hive 身份验证等安全特征管理方面的安全漏洞占所有检出漏洞的大多数，出现以上安全漏洞的企业在所有受检企业中的占比分别为 76%、60%、38%和 21%。如图 5 所示，安全漏洞 TOP6 中，Zookeeper 缓冲区溢出、Flink 注入等输入验证类安全漏洞出现率较低。本章对主要安全漏洞所影响的组件版本及相应安全风险进行解析。



数据来源：中国信息通信研究院

图 5 大数据平台安全漏洞统计

1. 安全漏洞一：Kafka 信息泄露

此类漏洞主要在 Kafka 组件（Apache Kafka 2.0.0、2.0.1、2.1.0、2.1.1、2.2.0、2.2.1、2.3.0 版本）中被发现。该漏洞源于组件在运行过程中存在配置等错误时，未授权的攻击者可利用漏洞获取组件敏感信息。例如，在 Apache Kafka 的 Connecter 的配置文件中，若系统使用默认配置，则 Connector 的敏感变量信息值如 token、key 等，就以简单 key-value 映射的形式明文存储，那么攻击者就可以利用该漏洞向集群发送请求以获取 Connector 的任务信息，在响应信息中就会包含上述明文的敏感变量信息。

2. 安全漏洞二：Zookeeper 安全绕过

此类漏洞主要在 Zookeeper 组件（Apache Zookeeper 1.0.0-3.4.13，3.5.0 alpha-3.5.4 beta 版本）中被发现。该漏洞源于身份验证/授权检

验过程中存在漏洞，可使任意节点加入集群，修改内容。若存在该漏洞，因 Zookeeper 未能强制执行身份验证/授权检测，攻击者可以远程加入集群，并获取 Leader（集群中数据同步的参考节点）身份，从而利用数据同步机制，在集群数据更新时注入伪造的数据信息，影响集群的运算结果；攻击者也可以利用该漏洞，制造多个 Leader 导致服务器崩溃，同时控制数据的同步，破坏集群数据的一致性。

3. 安全漏洞三：Zookeeper 本地信息泄露

此类漏洞主要在 Zookeeper 组件（Apache Zookeeper 3.4.6 之前的版本）中被发现。在这些版本的 Zookeeper 单独事务 log 中，客户端密码等敏感信息是以明文存储的。攻击者通过查看配置文件 zoo.cfg 中的配置参数 dataLogDir 的数值，即可得知本集群是否设置了单独事务 log 以及该 log 的存储位置，那么直接读取单独事务 log，即可获取以上敏感信息。

4. 安全漏洞四：Hive 身份验证

此类漏洞主要在 Hive 组件（Apache Hive 0.110-1.00，1.1.0 版本）中被发现。在这些版本的 Hive 组件中，LDAP 服务默认配置为允许未经身份验证的绑定，当 HiveServer2 配置为使用 LDAP 身份验证模式，会允许没有凭证的用户通过身份验证，从而带来数据泄露的风险。

5. 安全漏洞五：Zookeeper 缓冲区溢出

此类漏洞主要在 Zookeeper 组件（Apache Zookeeper 3.4.9 之前的版本和 3.5.0-3.5.3 版本）中被发现。这些版本组件的 C cli shell 存在缓冲区溢出漏洞。当使用‘cmd: ’批处理时，如果命令字符串超过 1024 个字符，则会发生缓冲区溢出，造成拒绝服务。

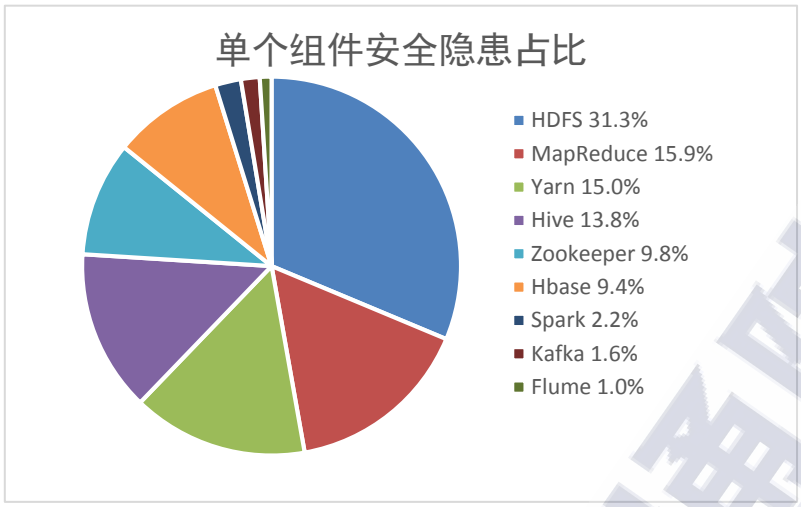
6. 安全漏洞六：Flink 注入

此类漏洞主要在 Flink 组件（Apache Flink 1.1.0-1.1.5、1.2.0-1.2.1 等版本）中被发现。本地攻击者可借助特制请求利用该漏洞进行中间人攻击，入侵通过 JMX 与进程建立的连接，获取传递的数据。

(三) 组件安全隐患统计分析

1. 单个组件安全隐患分布排名

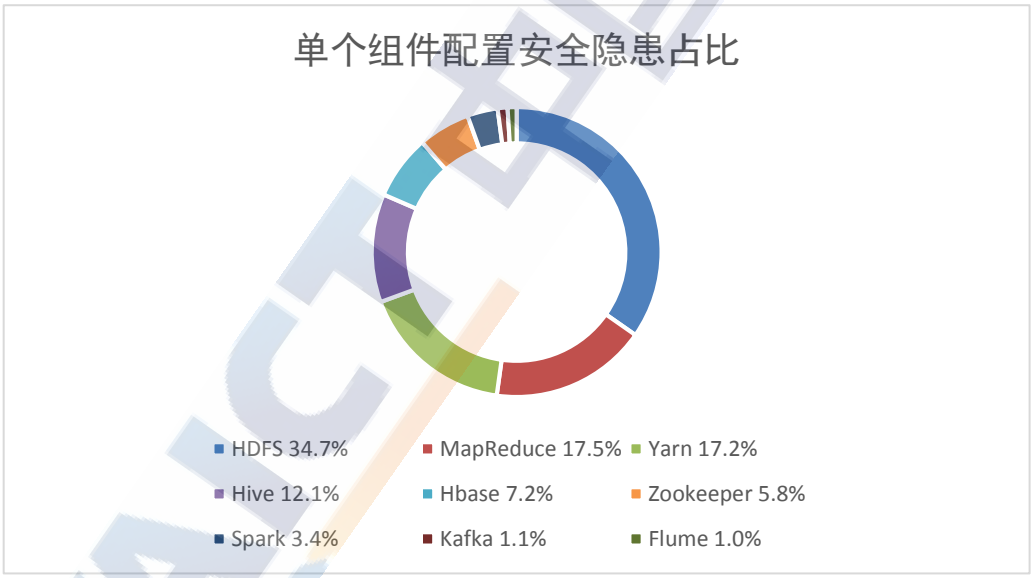
从单个组件的安全隐患（组件配置安全隐患和安全漏洞）数量占比排名来看，如图 6 所示，处于 Top3 的组件为 HDFS、MapReduce 和 Yarn，分别为 31.3%、15.9%和 15%，这些组件为产生最早的 Hadoop 组件，代码是完全开源的，攻击者对这些组件的了解程度较高，因此此类组件中安全隐患出现的较多。



数据来源：中国信息通信研究院

图 6 单个组件安全隐患占比

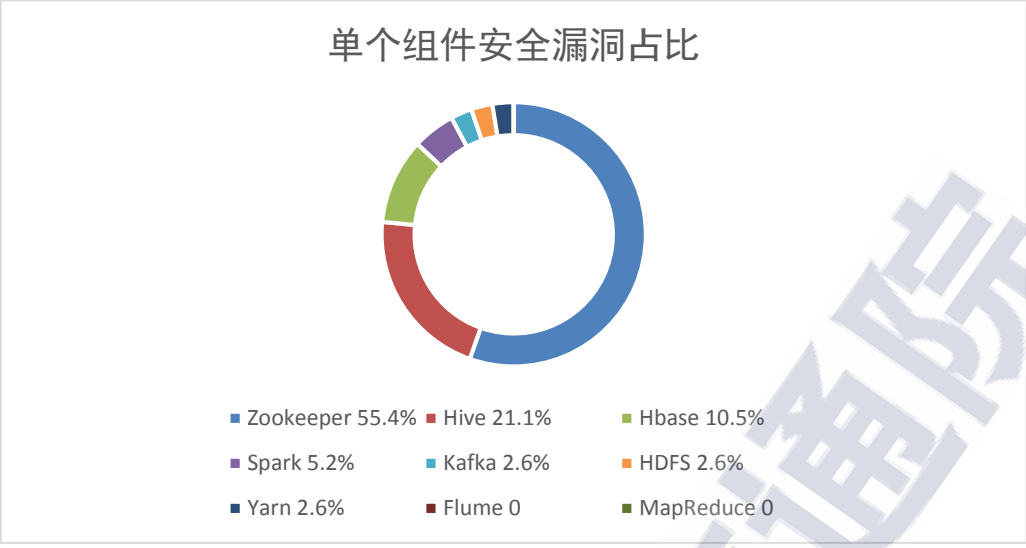
从单个组件配置安全隐患数量占比排名来看，处于 Top3 的组件为 HDFS、MapReduce 和 Yarn，占比分别为 34.7%、17.5%和 17.2%，如图 7 所示。



数据来源：中国信息通信研究院

图 7 单个组件配置安全隐患占比

从单个组件安全漏洞数量占比排名来看，Top3 的组件为 Zookeeper、Hive 和 Hbase，占比分别为 55.4%、21.1%和 10.5%，如图 8 所示。

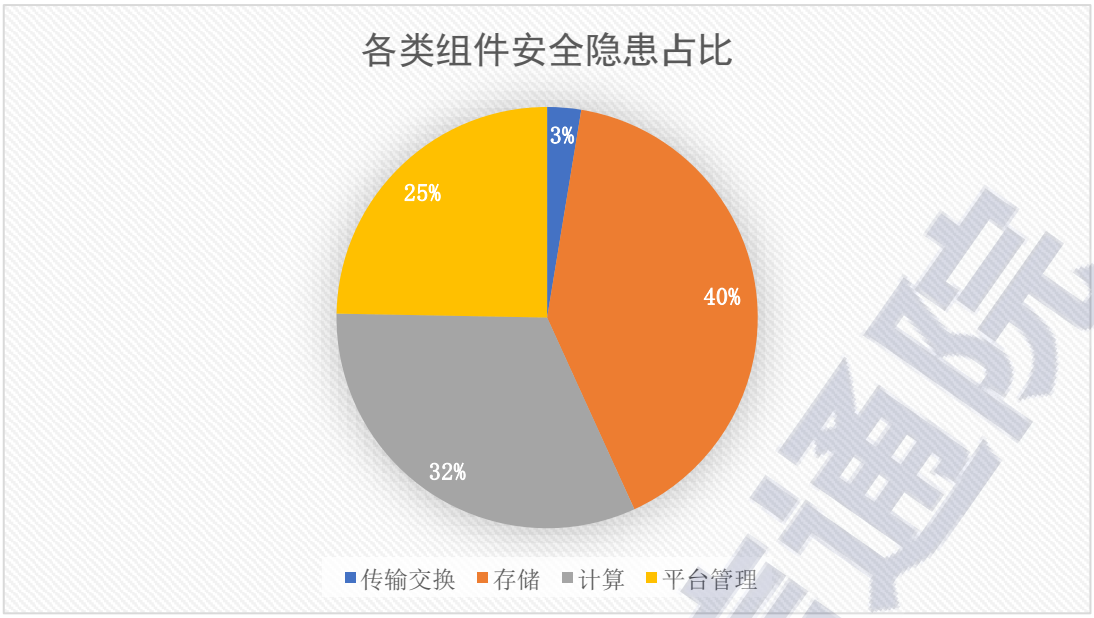


数据来源：中国信息通信研究院

图 8 单个组件安全漏洞占比

2. 各类组件安全隐患分布排名

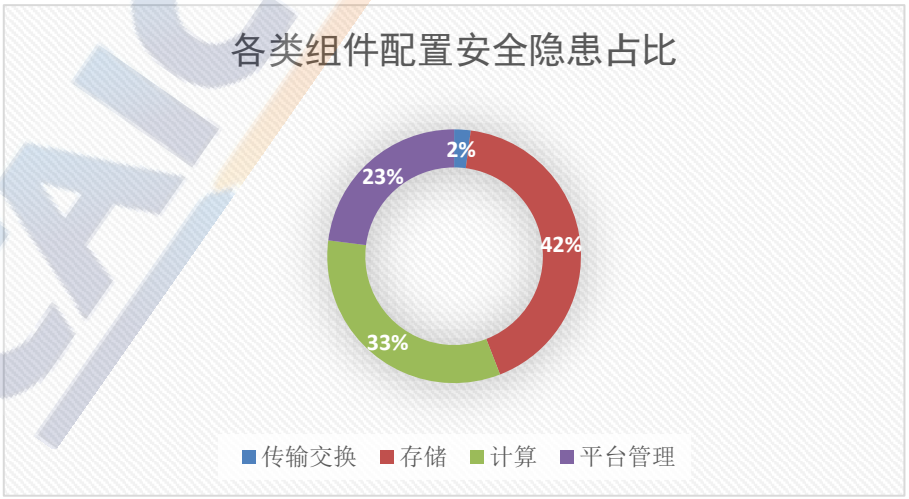
从安全隐患（组件配置安全隐患和安全漏洞）总数量上来看，4 类组件的安全隐患数量在所有安全隐患中的占比由高到低依次是存储类组件（如 Hbase、HDFS）、计算类组件（如 Hive、Spark、MapReduce）、平台管理类组件（如 Yarn、Zookeeper）和传输交换类组件（如 Flume、Kafka），如图 9 所示。存储类组件的安全隐患数量占比最高，约为 40%，就组件功能而言，由于存储类组件承载着大量用户敏感信息和商业机密数据，这些数据很容易成为恶意攻击者和黑客的目标。



数据来源：中国信息通信研究院

图 9 各类组件安全隐患占比

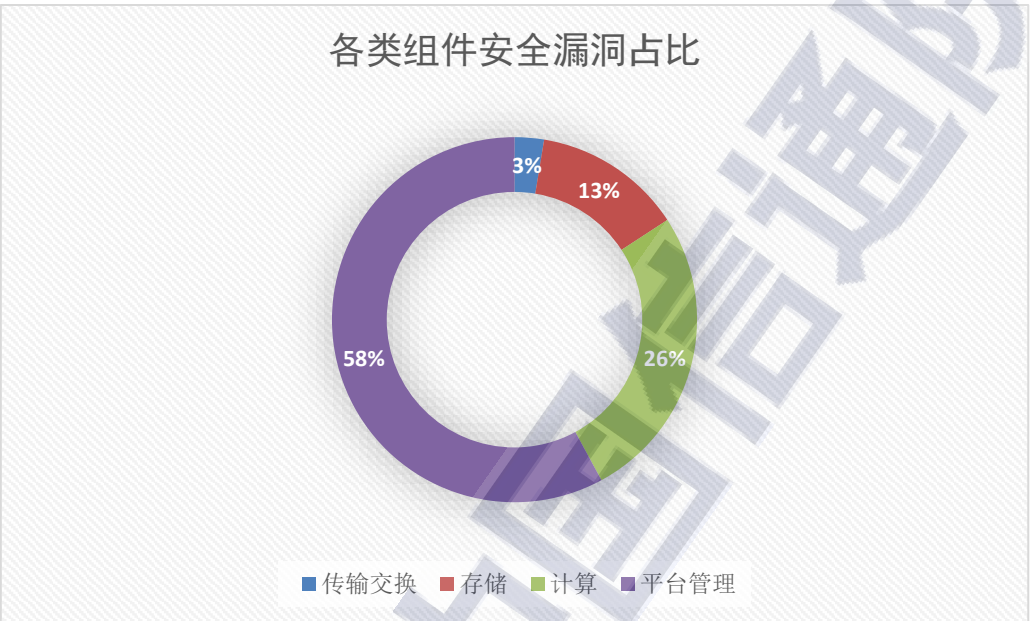
从各类组件的配置安全隐患比例上来看，4 类组件的配置安全隐患数量在所有安全隐患中的占比由高到低依次是存储类组件、计算类组件、平台管理类组件、传输交换类组件，如图 10 所示。其中，存储类组件的配置安全隐患占比最高，约为 42%。一般情况下集群中存储类组件的数量最多，一个集群可能有几百个存储节点甚至更多，配置和管理复杂度高。



数据来源：中国信息通信研究院

图 10 各类组件配置安全隐患占比

从各类组件的安全漏洞比例上来看，4 类组件的安全漏洞数量在所有安全隐患中的占比由高到低依次是平台管理类组件、计算类组件、存储类组件、传输交换类组件，如图 11 所示。其中，平台管理类组件的安全漏洞占比最高，约为 58%。



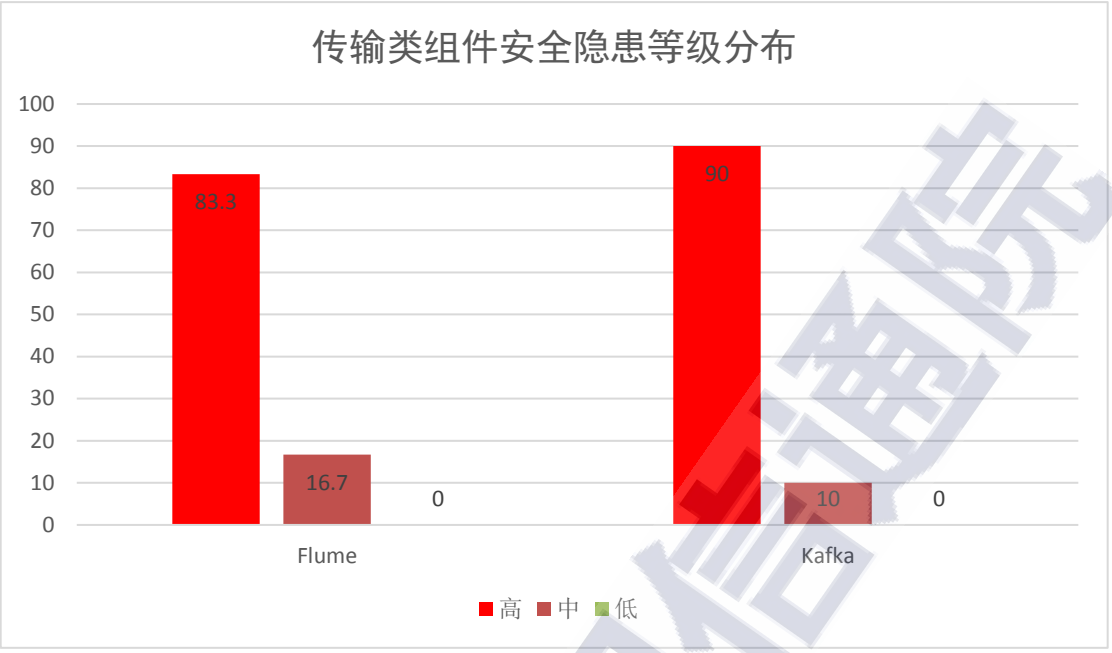
数据来源：中国信息通信研究院

图 11 各类组件漏洞安全隐患占比

3. 各类组件的安全隐患等级分析

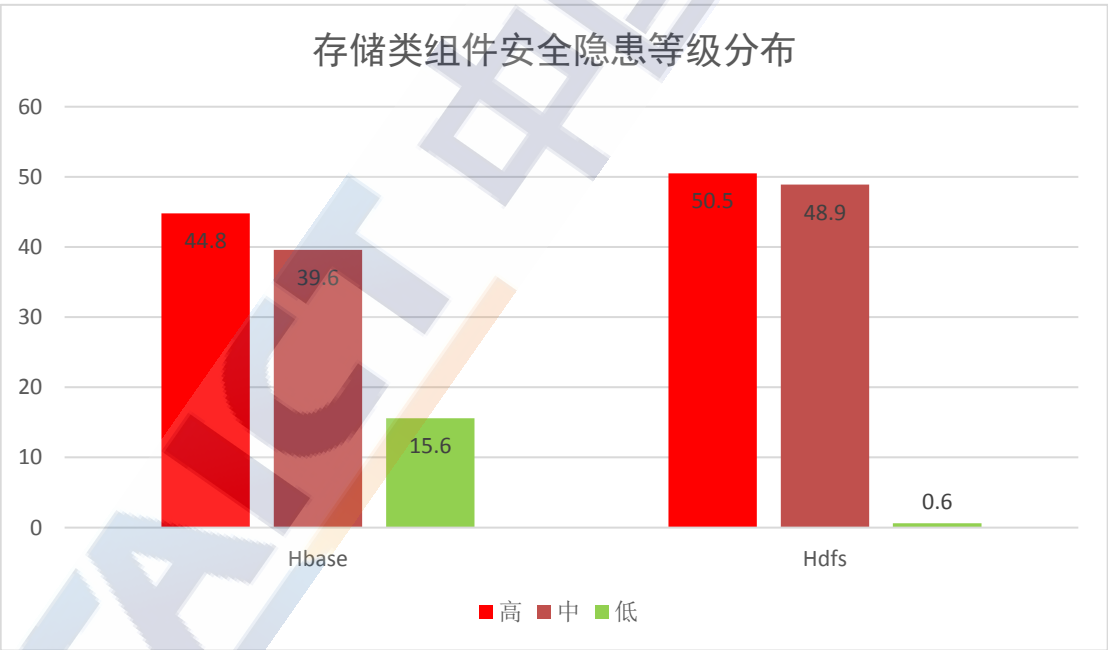
从安全隐患等级分布的角度来看，传输交换类组件的高危隐患数占比最高，Kafka 和 Flume 的高危隐患分别为 90%和 83.3%。如图 12 所示。存储类组件中高危安全隐患占比排在第二位，HDFS 和 Hbase 的高危安全隐患数分别占 50.5%和 44.8%，如图 13 所示。计算类组件中，Spark 的高危安全隐患数占 50%，MapReduce 与 Hive 的高危安全隐患数别占 33.3%和 19.7%，如图 14 所示。平台管理类组件中，Yarn 的高危和中危安全隐患数量占比相似，分别为 45.1%与 44.0%；Zookeeper 中危安全隐患占比最大，为 73.4%，如图 15

所示。



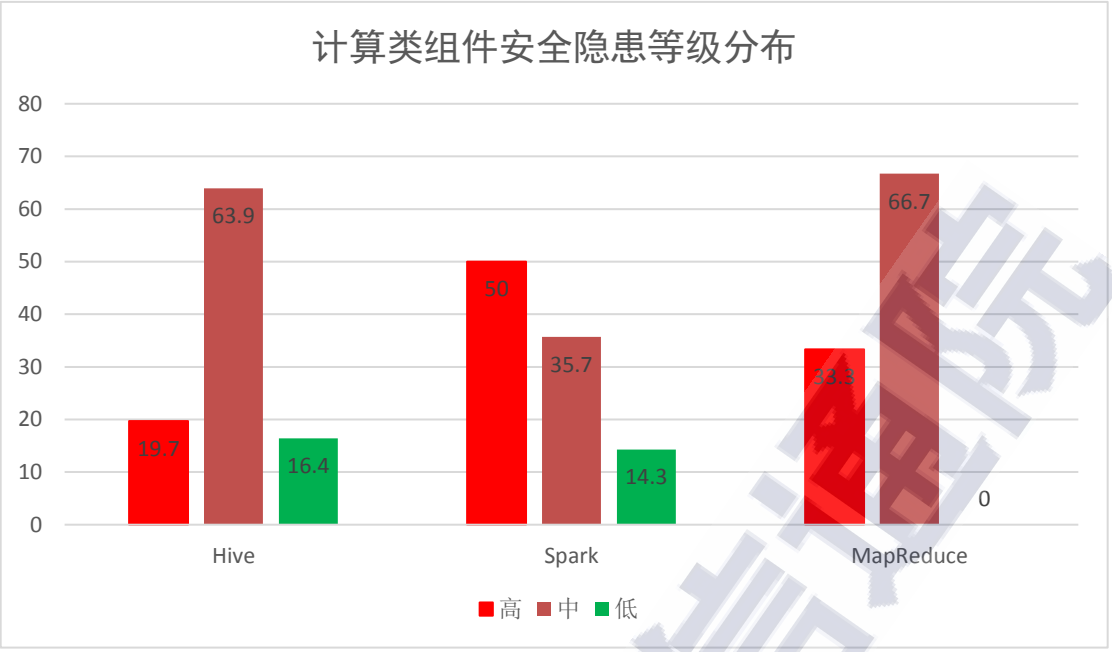
数据来源：中国信息通信研究院

图 12 传输类组件安全隐患等级分布



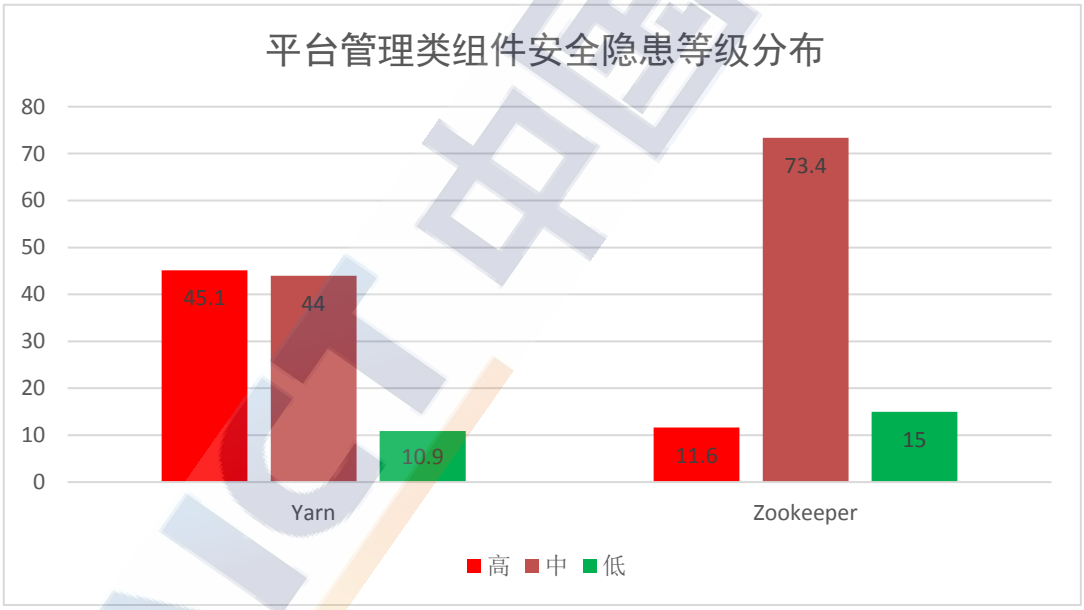
数据来源：中国信息通信研究院

图 13 存储类组件安全隐患等级分布



数据来源：中国信息通信研究院

图 14 计算类组件安全隐患等级分布



数据来源：中国信息通信研究院

图 15 平台管理类组件安全隐患等级分布

三、大数据平台安全问题分析

从以上分析可知，大数据平台普遍存在组件安全配置和安全漏洞方面的风险，本章从大数据平台的安全配置、漏洞管理、建设规

划、防护策略、管理制度等方面，分析造成以上安全风险的原因。

（一）基于 Hadoop 的开源大数据平台安全配置复杂度较高

基于 Hadoop 的开源大数据平台已日渐成熟，企业可根据自身需求对大数据平台进行增加或修改功能组件。这些组件的使用和管理均依赖于其配置文件，加之基础设施、网络系统方面的配置，基于 Hadoop 的开源大数据平台在配置管理方面复杂度较高，主要体现在三个方面。**一是配置文件集中管理难。**大数据平台采用分布式部署方式，各组件的配置文件分散在分布式集群的各节点上，管理人员需要记录各配置文件所存放的位置与文件中的配置选项，集群规模越大，组件种类、配置文件集中管理难度越大。**二是排查配置问题困难。**大数据平台一旦出现配置问题，需要工作人员登录各集群节点查找配置文件，排查问题，这使得因配置问题造成的业务损失持续时间较长，无法及时止损。**三是不安全配置易出现。**缺省配置、错误配置均可能造成配置不安全的问题。同时，大数据平台组件的版本更新可能会带来新的安全配置特性，如若未及时更新组件安全配置，亦可能带来新的安全风险。

（二）安全漏洞修复对平台运行影响较大

与组件配置安全隐患相比，安全漏洞的修复对平台运行的影响更大。大多数组件配置安全隐患的修复可以根据检测结果直接对错误的配置参数进行修改，而安全漏洞类安全隐患的修复要复杂的多。

一方面，安全漏洞的修复可能对集群的正常运行带来不确定因素。由本报告对安全漏洞的分析可以看出，目前大部分的安全漏洞是在特定版本的组件中出现的，修复这些安全漏洞可以使用的方式主要包括关闭相关端口或服务、打补丁、升级版本和使用相似组件替换等四种。以对集群的影响程度最低为出发点，修复方式的选择流程如图 16 所示。无论采用哪种修复方式，都会给整个平台兼容与稳定性带来不确定性，尤其是大版本的升级、组件替换更是如此。**另一方面，企业对安全漏洞的修复需要一定的时间周期。一是**安全漏洞的修复补丁主要依赖于平台服务商、社区等机构，在相关方没有发布解决方案之前，企业是很难自主进行修复。**二是**修复安全漏洞一般需要先在测试集群中进行版本变化后的兼容性、稳定性等测试还需与使用该服务的用户同步更新时间后，再进行统一的维护升级。



数据来源：中国信息通信研究院

图 16 大数据平台漏洞修复方式选择流程图

大数据平台在整个企业 IT 系统中往往处于核心地位，需要 24

小时正常运行，服务中断或服务器瘫痪给企业带来的损失难以估量，所以运维人员对集群组件的改动一般相当谨慎，最终有可能选择接受该风险，采用管理手段或加强外围安全管控技术的方式，降低风险发生的概率。

（三）大数据平台建设过程中安全投入不足

平台建设初期，首先要解决的问题是从零到一、从无到有的问题，此时建设团队成员往往人数较少、经费不足，只能将有限的精力集中到亟待解决的平台功能性问题，这导致了平台安全规划工作相对欠缺，安全防范措施也未能同步建设。

在大数据平台功能迭代时，随着服务的业务越来越多，平台性能往往是多数企业优先考虑的问题。在这个阶段，平台的集群节点不断扩充，经费和人力被集中投入到集群建设中去，安全性可能再次被忽视。当大数据平台的功能建设日渐成熟时，来自安全方面的风险逐渐严峻，而此时再进行安全方面的建设则面临着安全基线模糊、人员安全能力弱、安全配置修改困难等问题，也因为缺少整体的安全规划，此时大数据平台建设者仅能通过“补丁”的方式，对平台进行修补，甚至会在不能打“补丁”的情况下，推翻之前的建设成果重新进行设计实施，此时的平台安全建设成本，将远大于具有整体安全规划的大数据平台安全建设成本。

（四）大数据平台重视边界防护忽视内部安全

有观点认为大数据平台是运行于公司内部，在业务架构上，处

于最后端；在网络架构上，隐藏于内网之中，只要做好边界防护，其安全就处于相对可控状态。因此，实践中，企业往往借助传统的网络安全设施，例如，WAF、IDS 等，对网络边界进行加固，忽视了大数据平台自身存在的安全问题。这种做法一般会带来两方面的安全隐患，**一方面**，大数据平台虽然隔离于外网，却不能排除“内鬼”带来的风险。“内鬼”是企业内部的恶意攻击者，边界防护手段对其无效，此时大数据平台将暴露于“内鬼”的威胁之下，数据泄漏、损毁风险加剧。**另一方面**，开源的大数据平台组件存在着大量已知或未知的漏洞，攻击者可以采用社工、穿透等手段渗透进内网，针对组件的漏洞进行扫描，进而加以利用，从而使大数据平台陷入“防御空心”的状态。

（五）企业大数据平台安全管理制度滞后

目前，大部分企业缺少针对大数据平台的安全管理制度，或者安全管理制度不完善。其具体的问题表现在两方面。**一方面**，由于大数据平台的组件复杂，包含了采集、存储、使用各个方面，每种组件均有其独特的安全基线。在企业尚未明确大数据平台的安全基线情况下，无法将其体现到制度中去。同时，开源大数据组件或商用大数据平台的安全风险在不断变化，漏洞频发，安全基线也会随之改变，如何制定好平台的安全基线并体现到制度中去，对企业而言是一项具有挑战性的工作。**另一方面**，为了保证平台使用效率，在管理制度中未制定完整、闭环的平台使用流程。一般来说，使用

流程包括使用者申请、层级审批、批复使用等主要环节，对于平台的使用者，完整的使用流程管理必然增加工作量，降低工作效率。因此，如何平衡使用效率和流程安全制度之间的矛盾，是企业建设大数据平台面对的另一个问题。

（六）企业技术人员安全能力不足

大数据平台开发和运维人员对大数据平台的开发和运维会有比较深入的了解，但对平台的安全防护工作了解不足，从而导致两方面的安全问题。**一方面**在平台研发过程中，开发的代码缺少安全审计，可能会存在潜在的漏洞被黑客利用。**另一方面**在平台的维护过程中，由于维护人员对安全配置不熟悉可能会随意配置组件，从而导致一些安全配置的缺失。

同时，安全管理人员对大数据平台了解也不够深入。在很多企业中，大数据平台的防护通常由安全部门与业务部门合作进行，没有设置专门的大数据平台安全专职岗位。大数据平台的安全防护需要比较专业的大数据组件知识，对大数据平台的不了解使得安全部门难以有效开展大数据平台安全保护工作。

四、大数据平台安全解决方案建议

大数据平台自身的复杂性以及企业对大数据平台安全建设和管理的不到位造成了大数据平台的安全隐患。为帮助企业弥补在大数据平台安全保护上的短板，本章针对以上大数据平台安全问题提出解决方案建议。

（一）加强大数据平台安全基线管理

对于组件配置管理，应加强三个方面的建设。**一是，**明确安全基线具体内容，建立安全基线更新机制。企业需要梳理自身大数据平台的技术特点，针对性的建立适合自身的安全基线。同时，针对各开源组件或商业产品的漏洞爆发，应有预警和预案，及时进行处理，并且更新现有的安全基线，形成预警－基线更新－安全配置更新的良性循环。**二是，**建立组件配置统一管理平台。以自动化部属配置的方式替代人工修改配置的方式，解决人工修改配置易出错、效率低的问题。同时组件配置统一管理平台可进一步接入态势感知系统中，实现配置自动化更替，及时防范来自内部和外部的安全风险。**三是，**加强大数据平台组件的安全漏洞管理工作。理清平台各组件的版本和各组件间的版本兼容关系，一旦发现安全漏洞，根据组件版本兼容关系及时选择安全补丁或升级组件。

（二）对大数据平台安全进行整体规划

针对大数据平台建设应有整体的安全规划。在大数据平台建设的初始阶段，在解决平台功能方面从无到有的问题的同时，亦要考虑到平台安全防护的从无到有。在大数据平台功能迭代的同时，应考虑到当下的防护手段是否能够满足当下的防护需求。当大数据平台的规模和性能从量变到质变的时候，亦要考虑到安全技术手段能够满足当下的规模和性能，如果不能应及时更新安全技术手段。

（三）大数据平台边界防护与内部安全建设并重

杜绝大数据平台防护只讲边界防护而忽视内部安全的问题。大数据平台的安全规划应该是具有全面性的，既要守好“家门”，又要管好“家人”。两者的防护应是相辅相成的关系，任何一方的薄弱，都会带来“木桶”效应，让非法入侵者有机可乘。

（四）建立完善的大数据平台安全制度流程

在大数据平台安全流程制度的建设方面，有如下两方面建议。**一方面**制定针对大数据平台的专项安全制度，实现“对症下药”。安全制度方面宜包括组织架构、人员规范和技术要求。在组织架构方面，应明确大数据平台对应的各组织的安全责任关系；在人员规范方面，应对大数据平台的使用者制定基本的访问、操作规范，使人员“有法可依”；在技术要求方面，应针对企业大数据平台自身的防护需求，提出安全技术要求，以明确平台安全技术建设的方向。**另一方面**建立大数据平台的使用管理规范。对于大数据平台管理、开发部门内部和外部人员宜有不同的使用流程规范，在使用效率和流程安全之间做出权衡。

（五）增强企业技术人员安全能力

在大数据平台人员安全能力方面，有以下两个方面建议。**一方面**设置大数据平台的专职安全人员。专职的安全人员应具备大数据平台的基础知识以及安全防护知识，对大数据平台的安全防护有深入的理解。同时，专职的安全人员可以配置较高的大数据平台权限，

定期进行大数据平台的组件配置安全基线检查以及进行安全漏洞扫描。**另一方面**开展大数据平台使用人员的安全教育培训。大数据平台存储有海量的用户数据和业务数据，平台的使用人员在使用这些数据的时候应有职业操守和相应的安全意识。安全意识教育是加强人员安全意识的有效手段，通过安全意识教育可以让平台的使用者明确哪些可为哪些不可为，同时反面案例的宣传教育要能够对平台使用者起到警示作用。另外安全技能的培训可以增强大数据平台使用人员对安全技术的理解，并可使其真正应用在开发和运维工作当中去。

五、大数据平台安全未来发展建议

近年来，大数据产业蓬勃发展的同时，也暴露出了一些安全问题。基于所梳理的大数据平台安全现状与问题，我们对大数据平台安全发展提出如下几点建议:

(一) 加强企业大数据平台安全防护工作的监管

大数据平台存储的数据量巨大，数据一旦泄露或者被破坏，将导致严重的后果，政府部门需要加强对企业大数据平台安全的监管指导。**一方面**，强化大数据平台安全管理制度，明确平台管理主体责任和工作要求，加强对大数据平台和应用的监督指导，定期对相关企业开展监督检查。**另一方面**，推动大数据平台安全标准落地实施。目前，行业标准《电信网和互联网大数据平台安全防护要求》已经发布，为平台厂商、安全服务提供商、大数据服务提供上等开

展大数据平台安全防护工作提供了技术要求和参考规范，下一步建议加快推动标准落地实施，全面提升平台安全防护能力。

（二）强化大数据平台安全防护技术研究

平台安全是大数据系统安全的基石，目前无论是自建大数据平台还是商业化大数据平台，都处在高速发展阶段，平台安全防护却依然依赖边界防护和操作系统安全机制，需要产业各方在大数据平台安全技术研究方面加大投入。**一方面**，提升大数据平台本身的安全防御能力，引入组件身份认证、细粒度的访问控制、数据操作安全审计、数据隐私保护机制，从机制上防止数据的未授权访问和泄露，同时加强对平台紧急安全事件的响应能力；**另一方面**，从攻防两方面入手，密切关注大数据攻击和防御两方面的技术发展趋势，建立适应大数据平台环境的安全防护和系统安全管理机制，构筑更加安全可靠的大数据平台。

（三）推动大数据平台安全产品和服务市场发展

目前，大数据平台的核心技术机制仍存在很大的完善空间，且Hadoop集群模式及主流大数据平台技术均发源于国外。因此，可以鼓励国内重点企业、科研机构、高校等加强合作，加快对大数据平台核心关键技术的公关，重视大数据平台安全问题，推动大数据平台安全集中管理、平台安全代码审计、平台应用安全监测等相关安全产品和服务的开发应用，提升大数据平台应用安全水平和抗攻击能力，不断优化大数据平台。

（四）鼓励并促进大数据平台安全行业自律工作

大数据技术发展日新月异，大数据安全问题也需要各方共同关注和解决。建议相关企业同步开展行业自律，作为政府管理的补充力量，充分发挥专业性、经济性、灵活性等优势，共同营造大数据安全良好生态。**一方面**，鼓励第三方安全供应商、评估机构等自发或依托相关行业协会、社会组织平台，根据法律法规要求、国家标准和行业标准规范，共同制定大数据平台安全准则，签订行业自律公约，形成行业自治。**另一方面**，推进第三方评估机构和人员资质认证等配套机制建设，为推进大数据平台安全提供服务保障和人员保障。

免 责 声 明

本报告提供给媒体、公众和相关政府及行业机构作为涉及大数据平台安全状况介绍和研究资料，请相关单位酌情使用。如若本报告阐述之状况、数据与其他机构研究结果有差异，请使用方自行辨别，中国信息通信研究院安全研究所不承担与此相关的一切法律责任。因研究团队能力有限，报告仅作为参考研究，多有纰漏与不足之处，欢迎各领导专家批评指正，我们持续改进。

CAICT 中国信通院

中国信息通信研究院 安全研究所

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62305770

传真：010-62300264

网址：www.caict.ac.cn