

Hadoop安全架构

ADD YOUR POWERPOINT TITLE HERE

演讲人：安琪

跨界互联
数聚未来

第四届中国数据分析师行业峰会
CHINA DATA ANALYST SUMMIT

北京 中国大饭店 2017.07

Hadoop安全现状

 Hadoop认证授权

 Hadoop网络访问安全

 Hadoop数据安全

 Hadoop安全审计与监控

 Hadoop安全技术架构总结

Hadoop安全漏洞

- 部分漏洞:

- **CVE-2012-3376**
- **CVE-2014-085**
- **CVE-2015-1775**
- **CVE-2015-1776**
- **CVE-2015-1836**
- **CVE-2015-5210**
- **CVE-2016-0707**

Apache Hadoop爆信息泄漏漏洞 Apache ZooKeeper信息泄露漏洞 Apache Ambari服务器端请求伪造漏洞

Apache Hadoop MapReduce信息泄露漏洞 Apache HBase多个远程漏洞

Apache Ambari 开放重定向漏洞

Apache Ambari信息泄露漏洞

- 部分事件:

- 2013-Hive任意命令/代码执行漏洞+多家厂商被渗透实例
- 2015-乌云曝光国内知名安全厂商某个Hadoop的HDFS信息泄漏

身份认证

- 没有密码验证的账户体系
- 没有分权的账户管理功能

访问控制

- 继承了LINUX的权限体系
- 授权方式为自主授权

数据加密

- 数据明文保存
- 密钥管理

多租户

- 不同用户的磁盘空间没有隔离
- 不同用户的计算任务没有隔离

节点通信

- 传输没有加密
- 网络访问无限制

客户端交互

- 直接和服务节点通信
- 直接和资源管理节点通信

分布式节点

- 数据可以在任何可利用的节点进行处理
- 很难验证分散的平台集群的一致性和安全性

配置和补丁管理

- 配置文件，加密密钥，证书等众多管理工作
- 如何保持开源库是最新

软件包

- 开源组件的管理
- 容器的管理

应用程序和节点的验证

- 节点被恶意冒充风险
- 客户端的多样性

审计和日志

- 没有主客体访问行为的详细日志
- 单一的日志记录，无法分析出安全事件

监控

- 没有实时监控用户访问行为功能
- 应用种类繁多，无法统一监控

API安全性

- 漏洞攻击，注入、溢出等。
- API无自主边界防范能力

脱敏

- 敏感数据输出无控制
- 加密后的数据无法进行挖掘

Hadoop关键安全因素





Hadoop安全现状



Hadoop认证授权



Hadoop网络访问安全



Hadoop数据安全

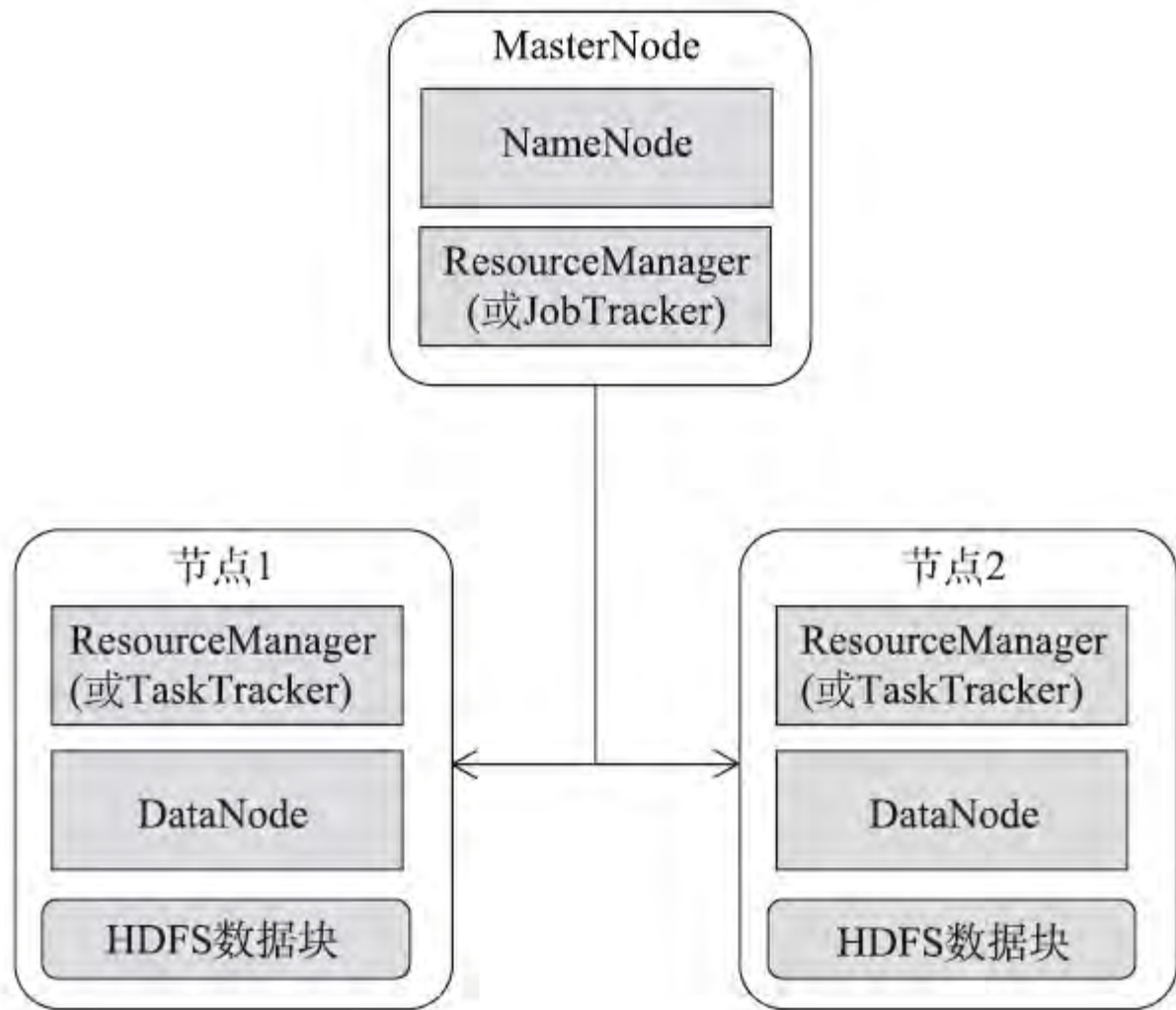


Hadoop安全审计与监控



Hadoop安全技术架构总结

Hadoop自身认证模型的缺陷



Hadoop没有进行用户及服务的安全认证

- 1、Chmod与chown改变文件及目录权限
- 2、编写whoami脚本模拟超级用户
- 3、NameNode注册恶意机器收到复制hadoop的数据块
- 4、篡改配置文件（hdfs.site.xml、dfs.hosts）连接DataNode，读取数据块，甚至添加恶意数据块

Hadoop需要实现的安全认证

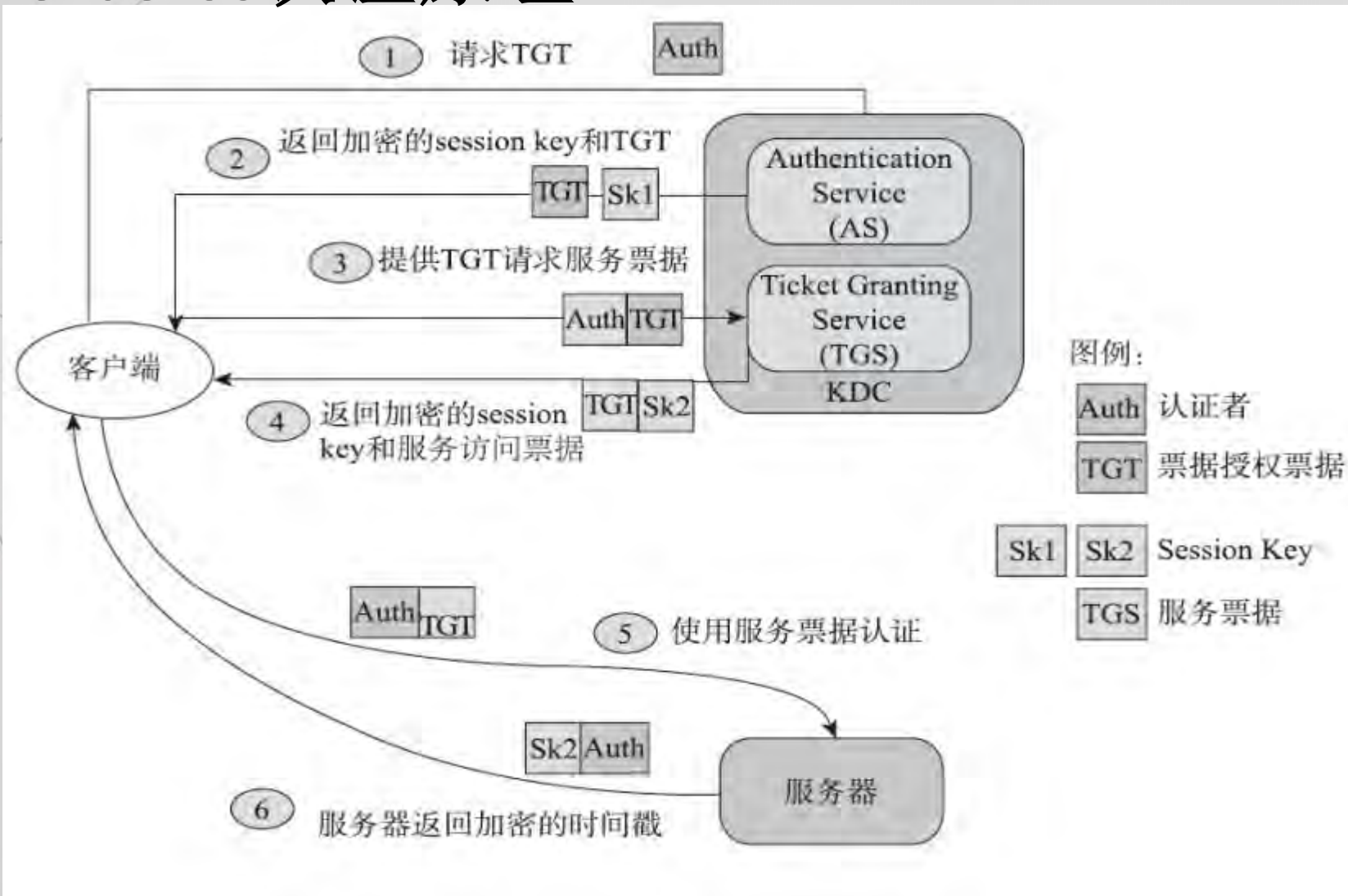
用户层次 访问控制

- 1、Hadoop用户只能访问授权的数据。
- 2、只有认证的用户可以向Hadoop集群提交作业。
- 3、用户可以查看、修改和终止他们的作业。
- 4、只有认证的服务可以注册为DataNode或TaskTracker。
- 5、DataNode中数据块的访问需要保证安全，只有认证用户才能访问Hadoop集群中存储的数据。

服务层次 访问控制

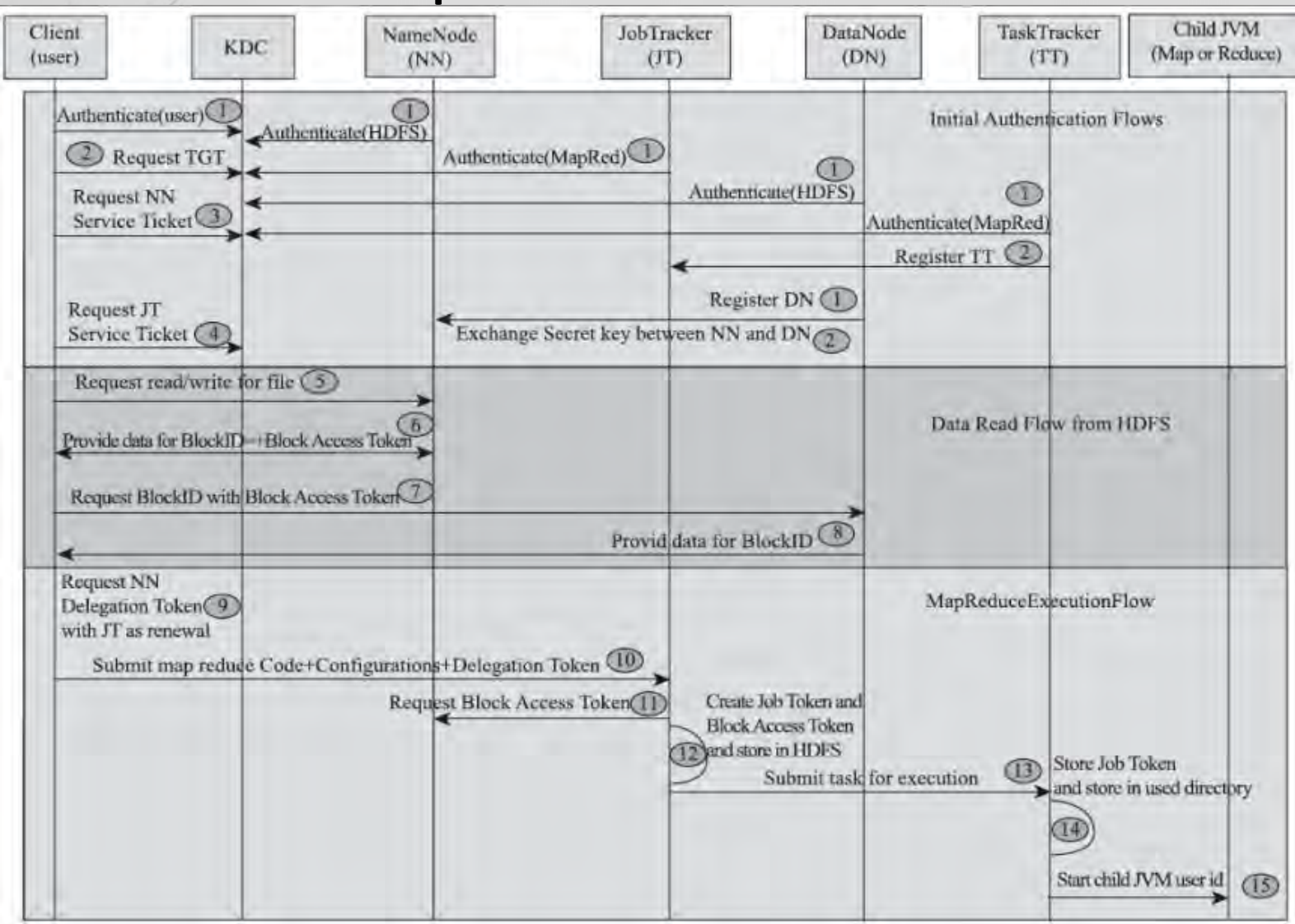
- 1、可扩展的认证：Hadoop集群包括大量的节点，认证模型需要能够支持大规模的网络认证。
- 2、伪装：Hadoop可以识别伪装用户，保证正确的用户作业隔离。
- 3、自我服务：Hadoop作业可能执行很长时间，要确保这些作业可以自我进行委托用户认证，保证作业完整执行。
- 4、安全的IPC：Hadoop服务要可以相互认证，保证它们之间的安全通信。

Kerberos认证原理



- 1、建立KDC数据库
- 2、设置KDC管理员标识
- 3、启动kerberos守护进程
- 4、设置kerberos管理员
- 5、添加用户或服务标识
- 6、配置LDAP作为kerberos数据库
- 7、为hadoop创建keytab文件
- 8、向所有slave分发keytab文件
- 9、为每一个hadoop生态组件配置kerberos

安全hadoop交互行为



- 用户和服务认证
- 授权令牌
- 作业令牌
- 数据块访问令牌



Hadoop安全现状



Hadoop认证授权



Hadoop网络访问安全



Hadoop数据安全

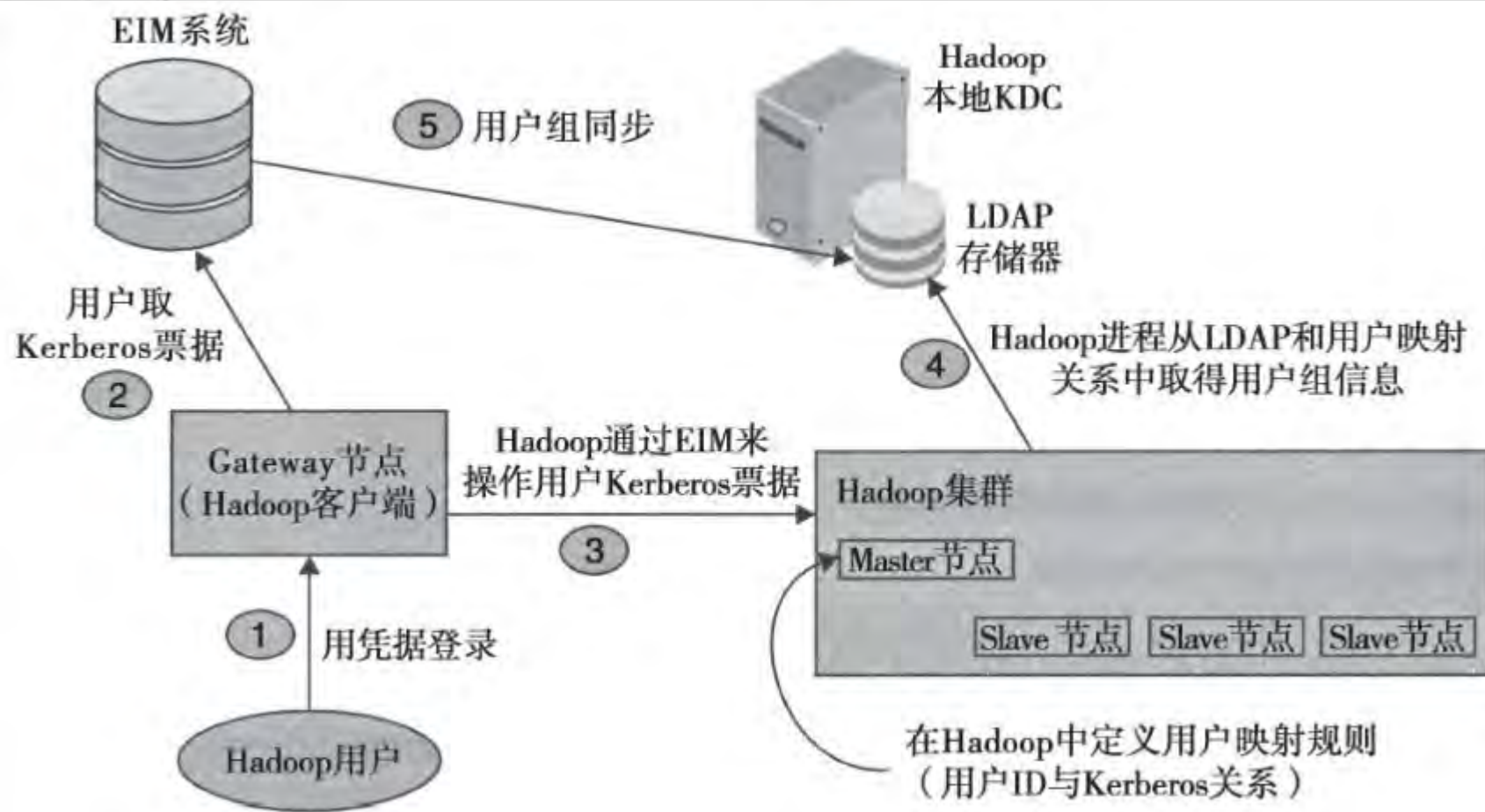


Hadoop安全审计与监控



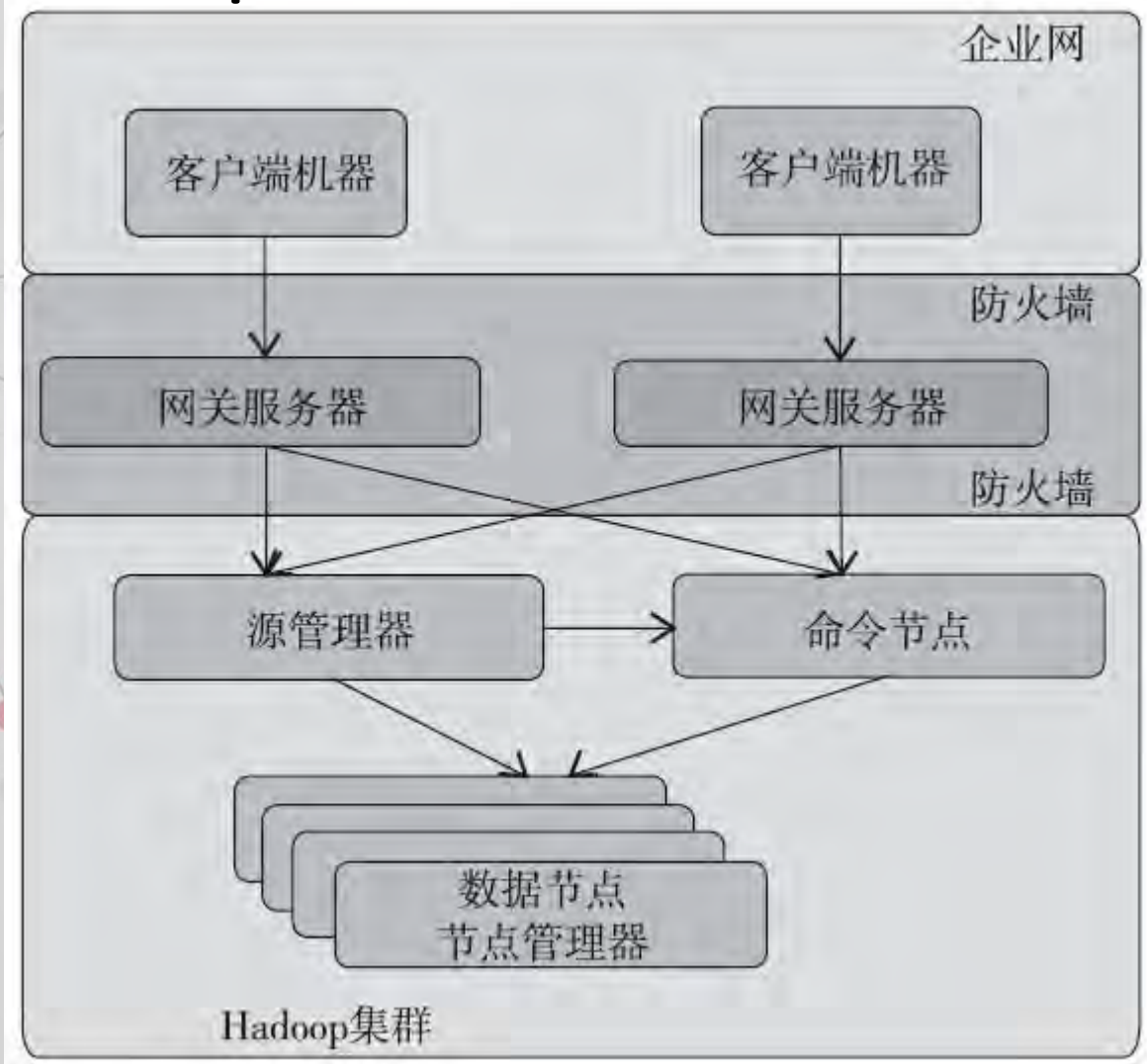
Hadoop安全技术架构总结

统一身份认证管理机制



- 1、设置EIM系统和本地KDC单项信任连接
- 2、设置EIM域信息和集群中每个节点上的本地KDC域信息，确保跨域认证
- 3、在EIM系统设置终端用户、用户凭据、用户角色
- 4、在本地KDC设置所有hadoop服务的标识和凭据
- 5、设置EIM系统和本地KDC直接的角色/用户组同步。可以采用LDAP同步连接器实现
- 6、配置规则转换标识为kerberos中对应的用户，设置core-site.xml中属性
`hadoop.security.auth_to_local`

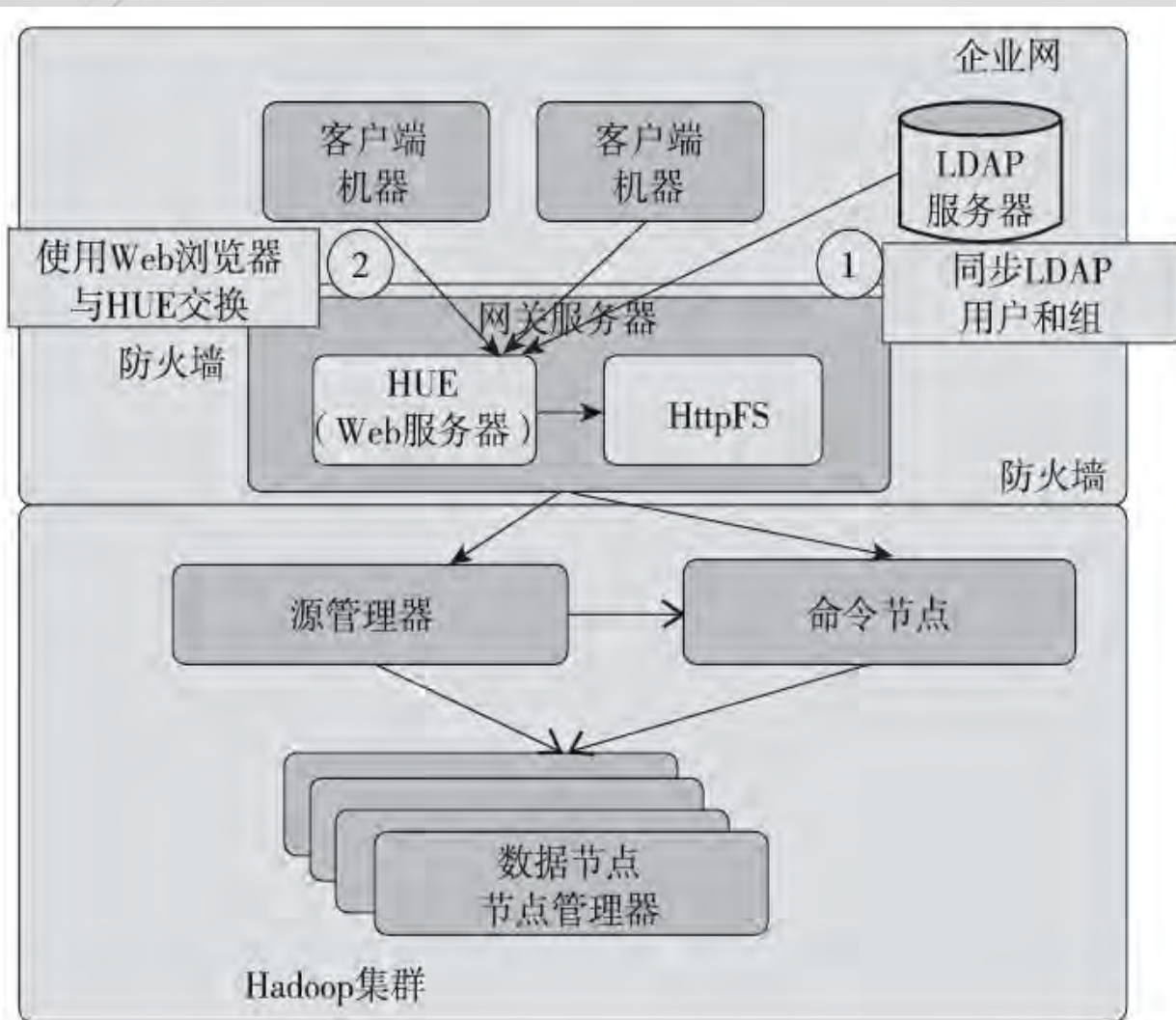
Hadoop网络访问安全



客户端工具统一安装至网关上

- 访问控制
- 策略增强
- 日志及网关服务

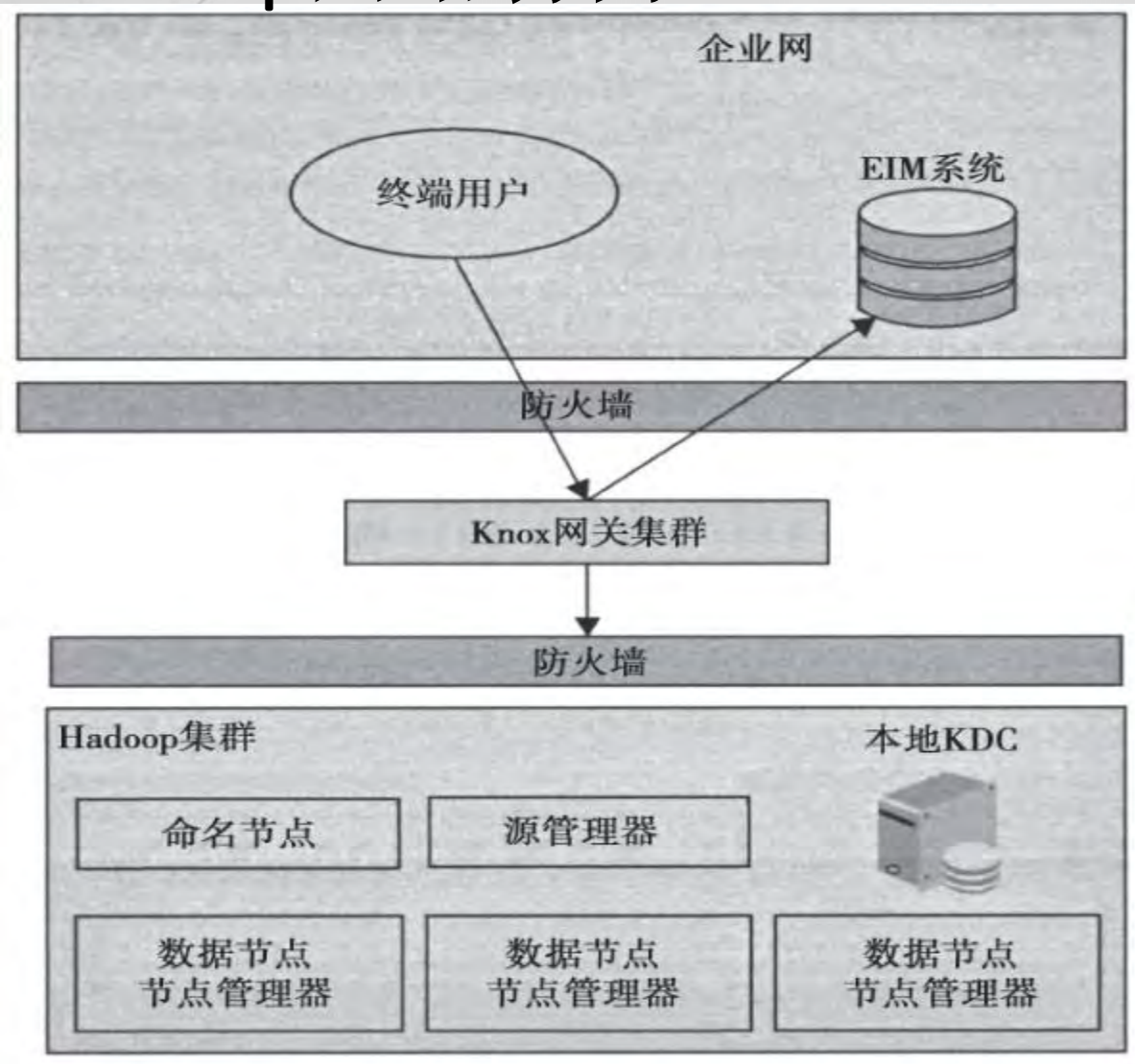
Hadoop网络访问之HUE



- 1、采用HttpFS代理
- 2、支持统一身份认证
- 3、支持LDAP同步用户和用户组信息
- 4、细粒度用户访问控制
- 5、SPANEGO-Base认证协议
访问SSL加密

- 浏览HDFS文件系统。
- 开发工作流，部署到Oozie。
- Pig编辑器和执行器。
- 浏览安全Hadoop集群中执行的MapReduce作业。
- 用户接口来提交Hive和Impala查询。
- Scoop命令执行器。

Hadoop网络访问之Know Gateway Server



- 1、隔离hadoop集群，限制访问集群终端
- 2、提供Hadoop REST API支持
- 3、提供单点认证和令牌验证
- 4、支持统一身份管理系统
- 5、支持HUE的代理访问



Hadoop安全现状



Hadoop认证授权



Hadoop网络访问安全



Hadoop数据安全



Hadoop安全审计与监控



Hadoop安全技术架构总结

Hadoop数据安全的关键因素

- 生成的商业决策支持数据应该分类并加密，只有授权用户可以访问。
- 只有有限的用户可以访问这些敏感数据。非业务用户要限制访问这些敏感数据。
- 数据分析过程生成的中间数据也要被保护，不再需要时立即删除。
- 需要跟踪哪些用户下载了敏感数据，控制敏感数据下载的生命周期。
- 当用户不再有权限访问数据集时，对敏感数据或任何本地副本访问权限要被删除。
- 在Hadoop中存储和获取数据，在数据流动过程中也要保证数据安全。

Hadoop数据传输的通道加密

客户端连接服务器请求认证。

服务器返回支持的认证机制列表

客户端选择一种认证机制，比如DIGEST-MD5。

服务器始于客户端交互认证信息，只要认或失败。

一旦认证成功，客户端和服务端开始使用会话密钥，加密传输的数据

SSL使用公开密钥算法进行认证。客户端和服务端需要共享密钥来进行认证

SASL

- Hadoop RPC协议连接 NameNode
 - Hadoop客户端使用基于TCP的HTTP协议将数据传输至DataNode。
- 1、配置core-site.xml中属性hadoop.rpc来开启SASL加密，确保NameNode之间的传输加密
 - 2、设置hdfs-site.xml中dfs.encrypt.data.transfer为true来使SASL包装器生效，确保datanode传输加密
 - 3、开启JDBC的SASL保护
 - 4、Flume AVRO-RPC开启SSL支持

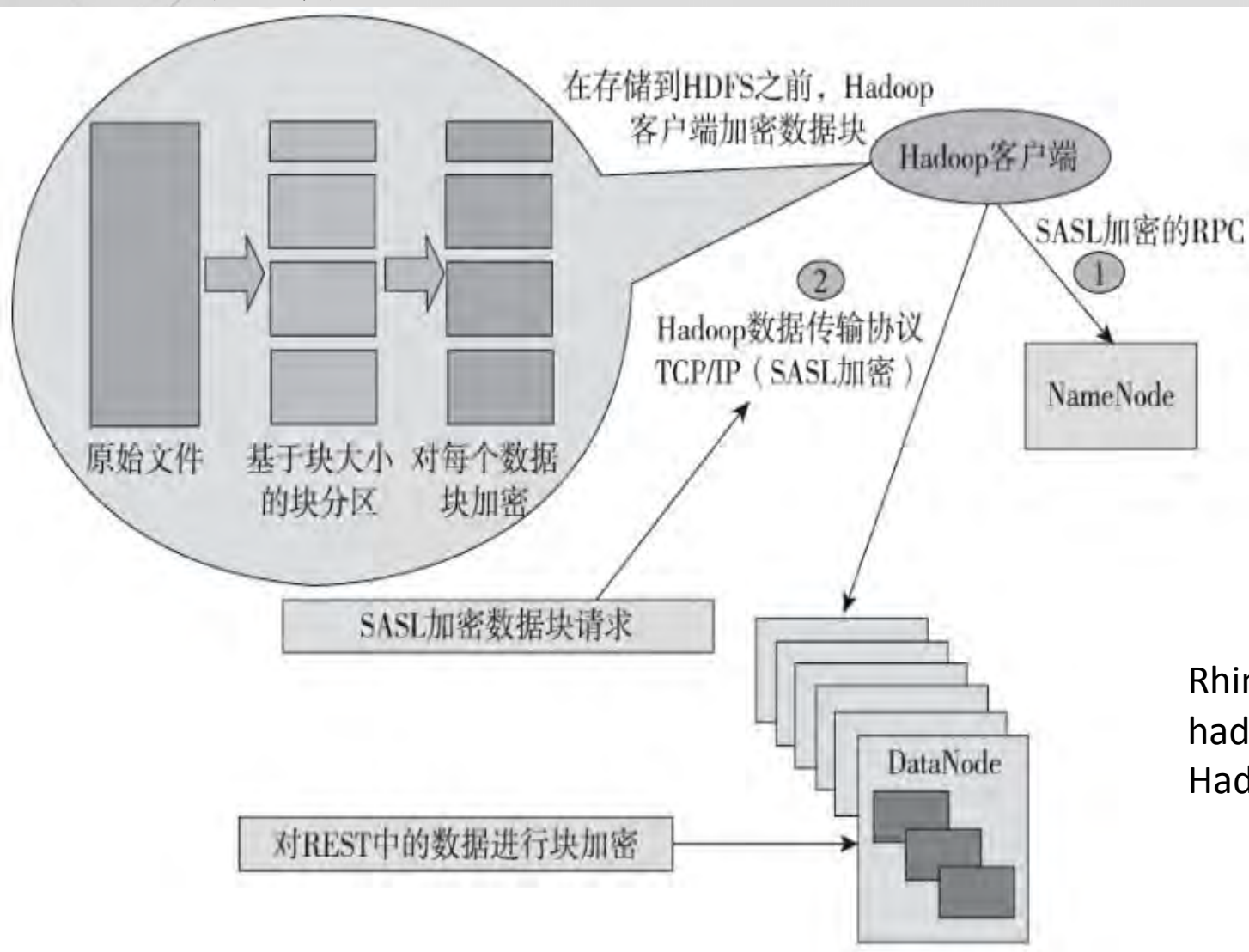
Hadoop静态数据的加密

加密思路

先加密后存储
在向Hadoop存储文件时，首先把整个文件进行加密，然后存储。这样，每个DataNode中数据块无法被解密

加密数据块 确保MapReduce程序可以独立的访问每个数据块，解密逻辑可以在MapReduce作业中进行。解密密钥需要告诉MapReduce作业。这种方案是可扩展的

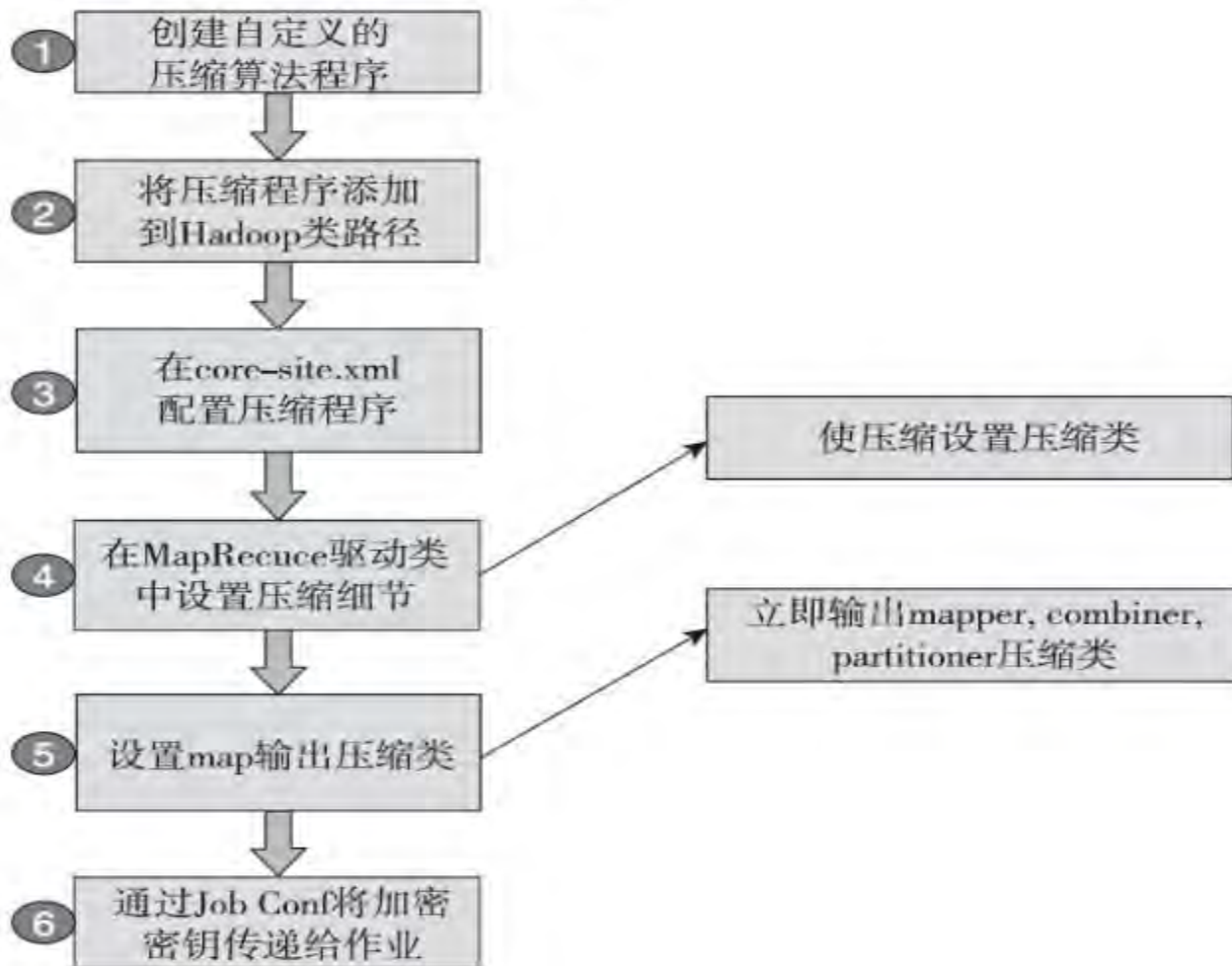
加密数据块



Rhino (<https://github.com/intel-hadoop/project-rhino/>) 目标是为Hadoop提供端到端的加密框架

借用压缩文件处理能力实现加密

Support Encryption in MapReduce



- 1、创建开发压缩解码器
让AVRO和序列文件使用压缩解码器
- 2、为mapper配置输入格式为压缩文件及对应的编解码器
- 3、解密密钥作为作业配置信息传输给mapper,或者直接从keystore读取
- 4、处理的中间结果在TaskTracker本地存储
- 5、由reduced读取来生成最终结果

企业用户的安全分级

每个安全级别定义唯一加密密钥，要存储在凭证保险箱中。只有授权响应安全级别的用户可以请求对应的密钥

敏感数据基于安全级别选择加密密钥。

当用户运行MapReduce作业访问分类的数据集，要从凭证保险箱获取对应的密钥发送给Map Reduce作业。只有用户通过对应安全级别授权，才可以获取到密钥。

用户分级安全



Hadoop安全现状



Hadoop认证授权



Hadoop网络访问安全



Hadoop数据安全



Hadoop安全审计与监控



Hadoop安全技术架构总结

用户登录和授权事件

当用户或服务标识在KDC或EIM系统进行认证时会生成用户登录事件，在集中EIM系统（活动目录或相似系统）将记录用户授权事件。用户向Hadoop进程每次请求服务票据都会生成日志

HDFS文件操作错误

当用户访问HDFS，Name Node会验证用户的访问权限。当存在越权访问时会在hadoop日志文件中产生错误事件，Hive或Pig作业遇到任何访问HDFS权限问题时都会产生相同的错误。

RPC授权错误

任何对Hadoop进程未授权的访问请求，异常会记录至Hadoop安全日志文件中。监控这些异常可以识别未授权访问

RPC认证错误

Hadoop RPC使用Java SASL APIS进行验证。这个交互过程可以设置质量保护，确保客户端可以安全的联机Hadoop服务，任何中间人攻击导致的验证失效都可以被记录下来

Hadoop安全监控要点

HDFS敏感文件下载

Hadoop支持记录每一个文件系统操作到HDFS审计日志文件。该审计文件，可以识别哪些用户访问或下载了敏感文件

MapReduce作业事件

Hadoop支持在日志中记录所有MapReduce作业提交和执行事件。审计日志会记录作业的提交、启动、查看和修改行为。因此该审计文件可以用来识别哪个用户访问和运行了集群上的作业

Oozie、HUE和WebHDFS的访问

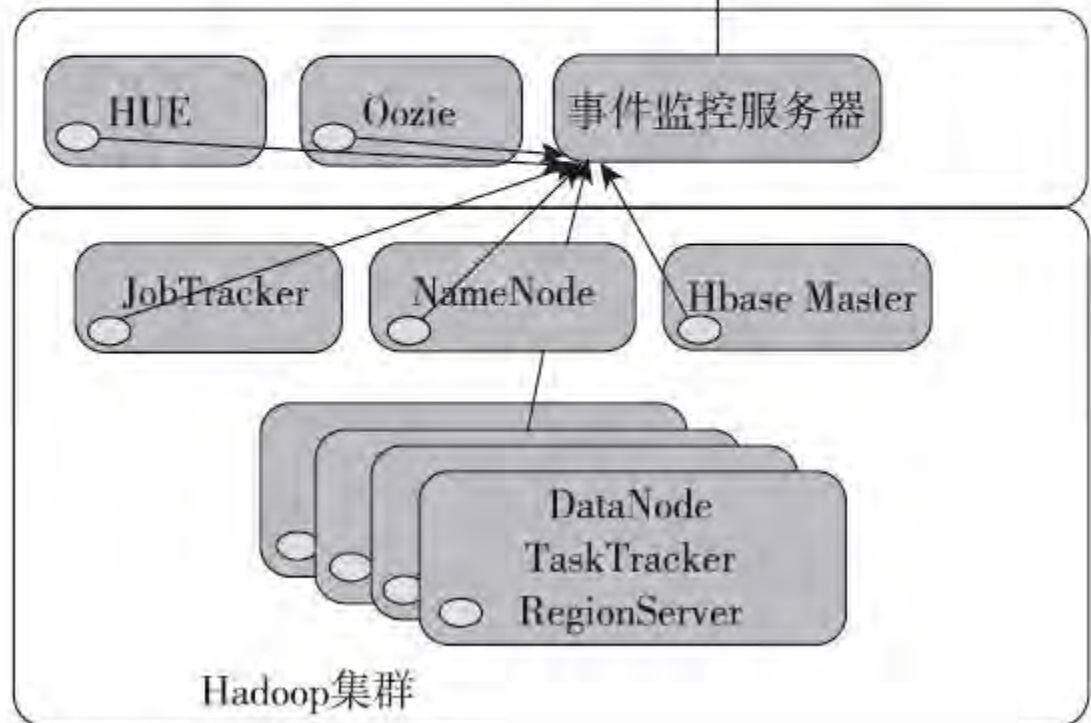
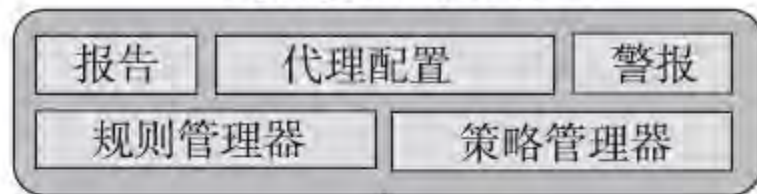
用户访问Oozie并进行工作流提交都会记录到Oozie的审计日志。所有用户与Oozie的交互也会记录到日志，可以用来跟踪执行特定工作流的用户信息

其他异常

除了用户认证和授权产生的异常，记录Hadoop中任何其他类型的异常也很有用。这些异常提供潜在讯息发现系统的脆弱性，也可以识别潜在的安全事故

Hadoop安全监控与审计系统

事件监控和审计界面

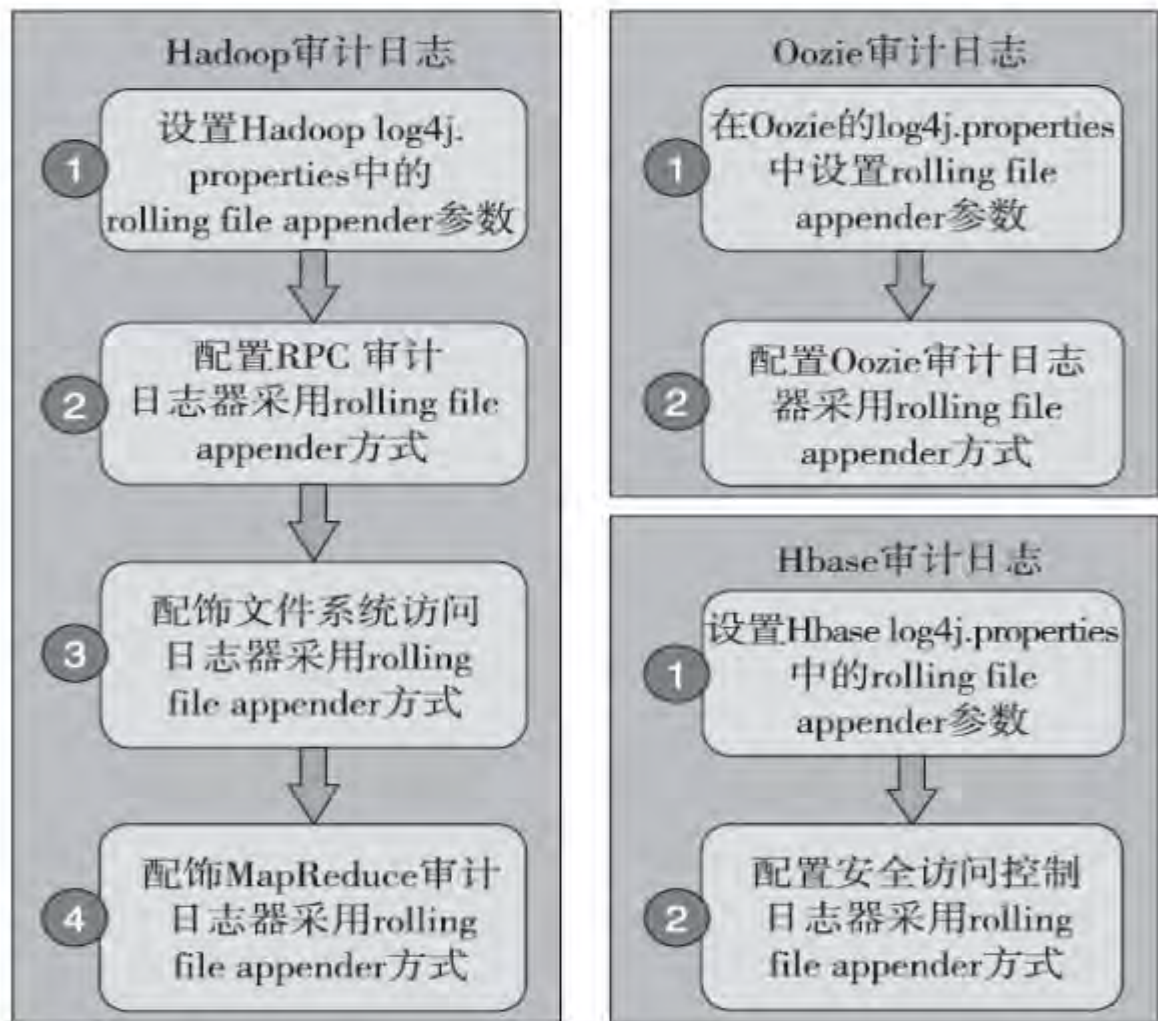


○ ← 事件日志收集代理

OSSEC是一个基于主机入侵检测系统的开源项目，支持收集hadoop集群中的各种日志和事件

Hadoop审计日志的配置

在Hadoop生态系统中审计和安全日志的配置步骤



- 1、Hadoop rpc事件日志、Hadoop文件系统访问日志、MapReduce日志、Oozie日志、Hbase日志：对etc/Hadoop/log4j.properties进行相应的设置
- 2、HUE审计日志：获取/var/log/hue目录下的access.log获取日志信息
- 3、KDC审计日志：配置kadmin和krb5kdc进程记录访问日志



Hadoop安全现状



Hadoop认证授权



Hadoop网络访问安全



Hadoop数据安全



Hadoop安全审计与监控



Hadoop安全技术架构总结

Hadoop安全技术架构总结





CDA 数据分析师
www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会
CHINA DATA ANALYST SUMMIT