大数据基础组件的安全解决方案

Security solutions based on big data component

2016 BDTC

北京观数科技有限公司

目录
CONTENTS

1

PART ONE

公司介绍

北京观数科技有限公司

成立于2015年，总部位于北京，在贵阳大数据安全产业园建立了产品中心

大数据基础框架安全解决方案提供商

团队骨干成员从事信息安全十余年

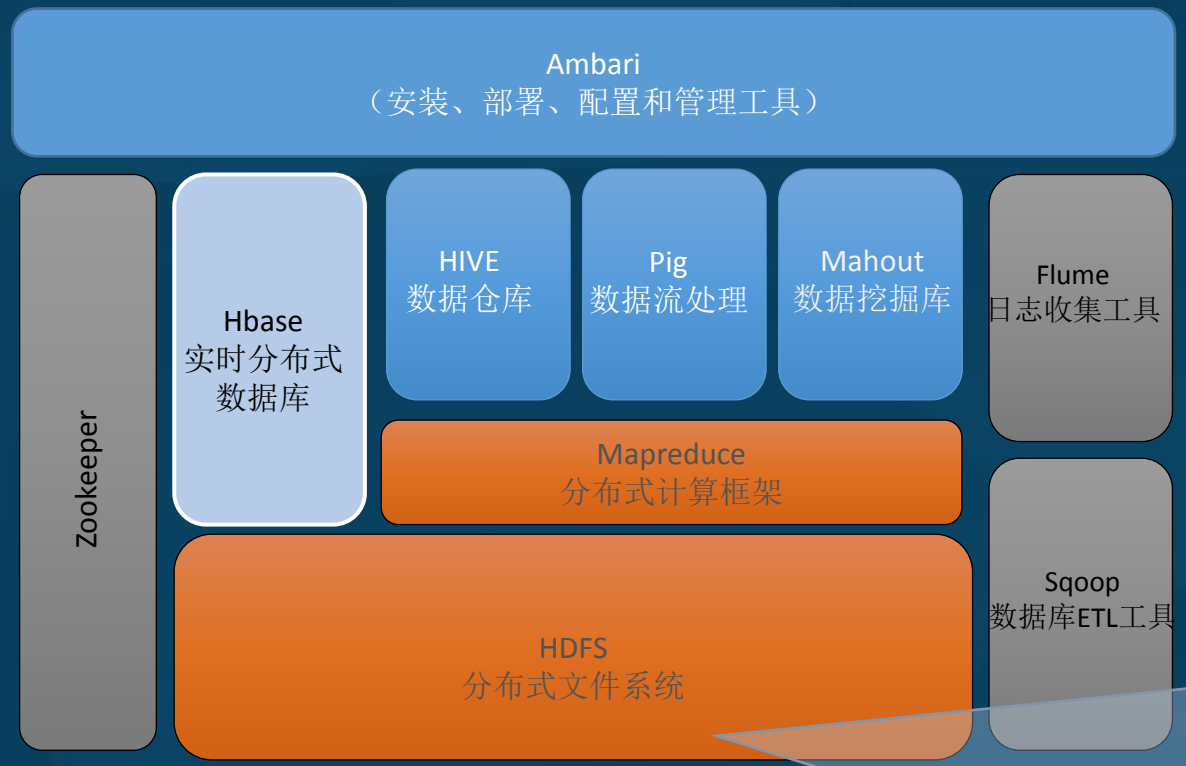
精通领域：操作系统安全加固、攻防技术、等级保护

2

PART TWO

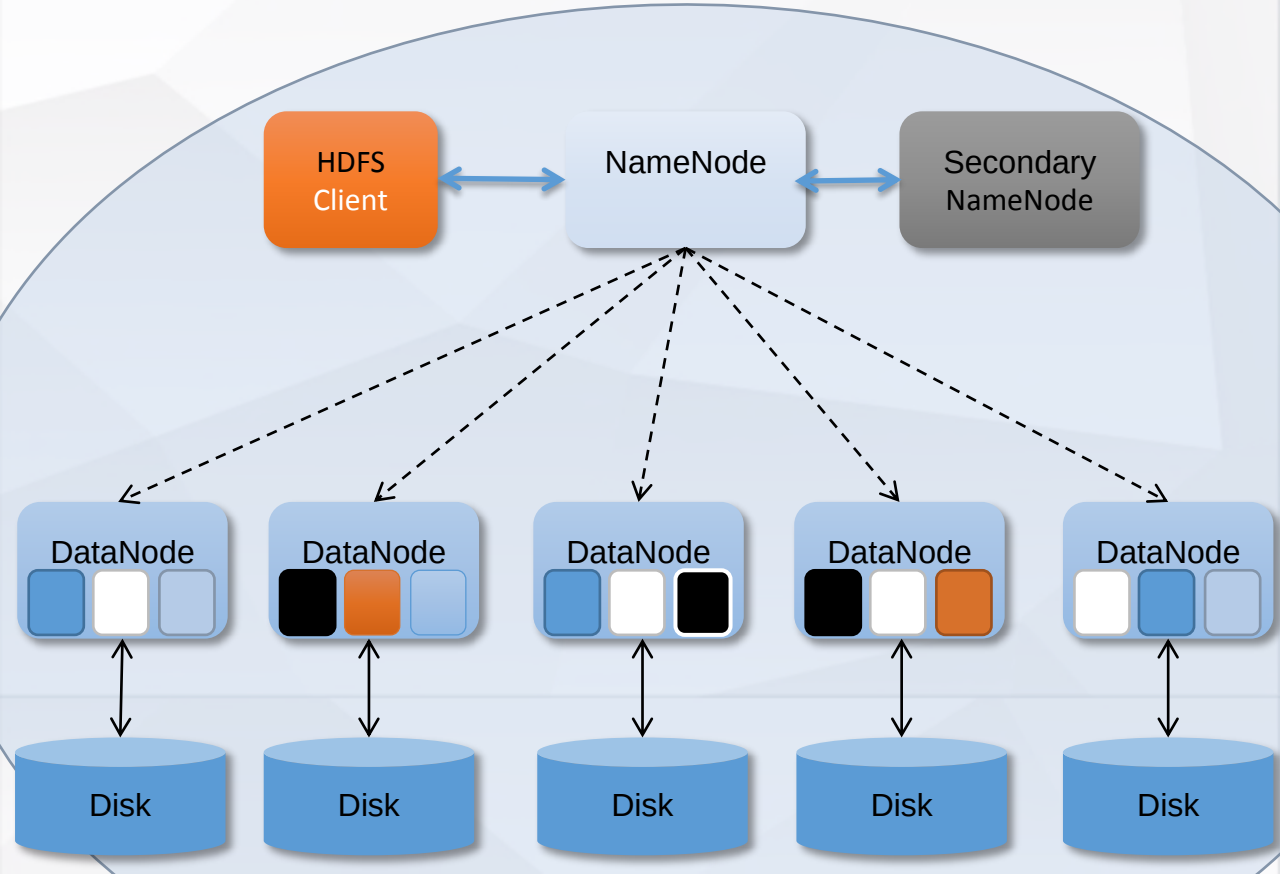
大数据基础框架生态

Hadoop 生态圈

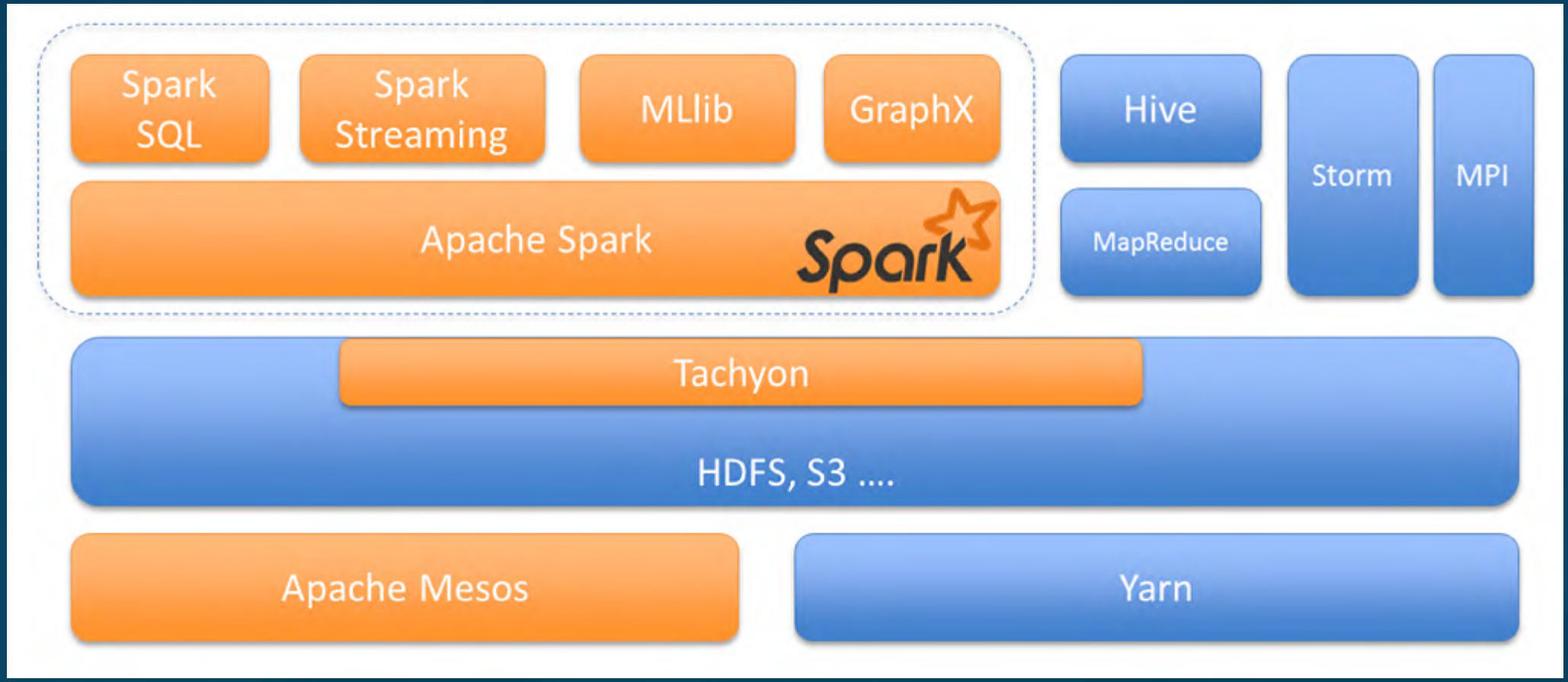


大数据基础框架

分布式文件系统



Spark 生态圈



Spark是开源基于内存的通用并行框架

3

PART THREE

大数据安全防护的必要性

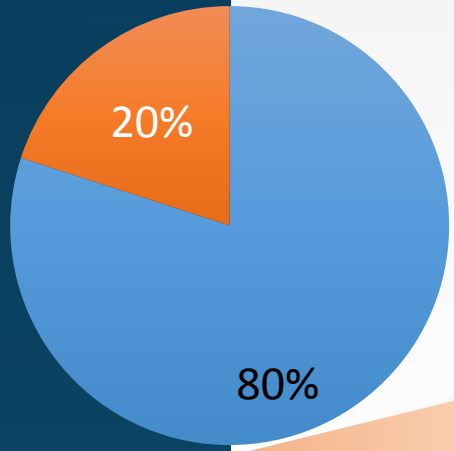
风险来源

内网

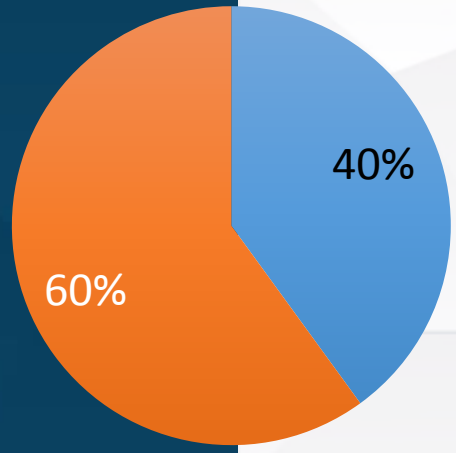
数据泄漏

越权访问

管理风险



损失程度



外网

APT

病毒

Web攻击

应用漏洞

离线

Mapreduce

hive

pig

……

私有云

solr

impala

Samza

……

在线

storm

spark

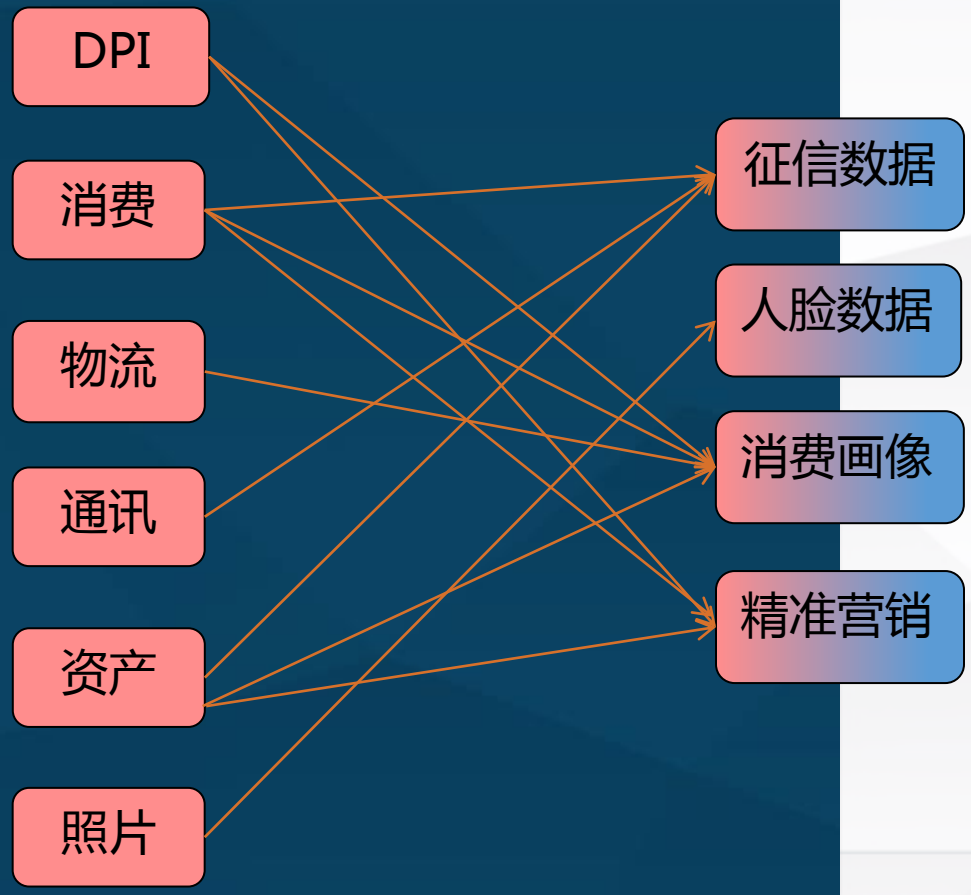
……

AWS

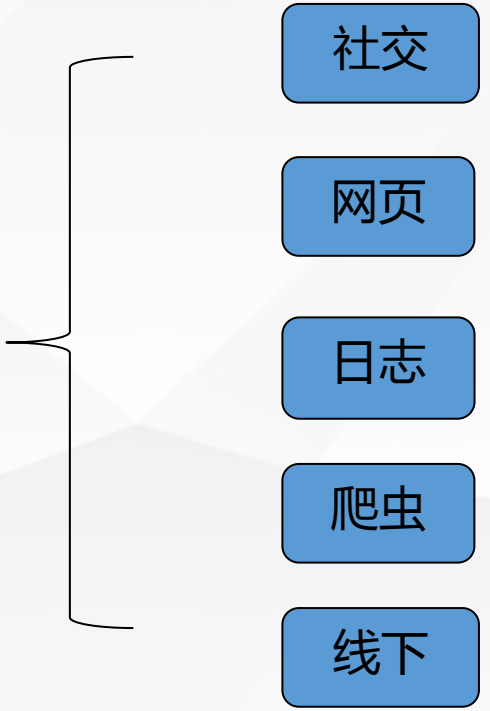
E-mapreduce

Uhadoop

敏感数据



非敏感数据



大数据安全的必要性

防的不是内网、外网，防的是风险

护的不是在线、离线，护的是业务

保的不是敏感、不敏感，保的是价值

4

PART FOUR

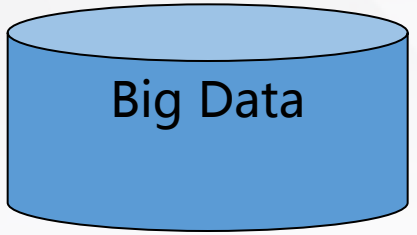
基本的组件防护思路

历史遗留问题

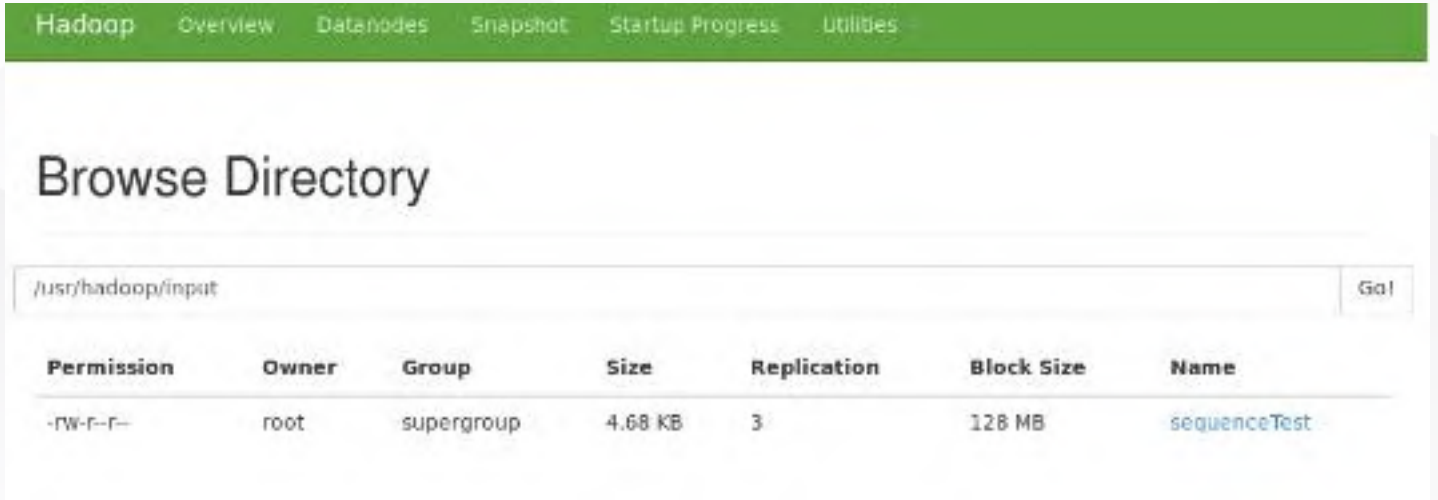
Root

Change Permissions

Change Owner



-rwxrwxrwx root file
drwxrwxrwx root filedir
-文件 –r读 –w写 –x执行
查看linux用户cat /etc/passwd
查看linux组 cat/etc/group



Hadoop安全现状

最初的Hadoop中没有安全模型

它不对用户或服务进行验证，也没有数据隐私。因为Hadoop被设计成在分布式的设备集群上执行代码，任何人都能提交代码并得到执行

恶意开发人员能轻易假冒其他用户

比如写一个新的TaskTracker并将其注册为Hapdoop服务，或者冒充hdfs或mapreduce用户，把HDFS里的东西全删掉等等

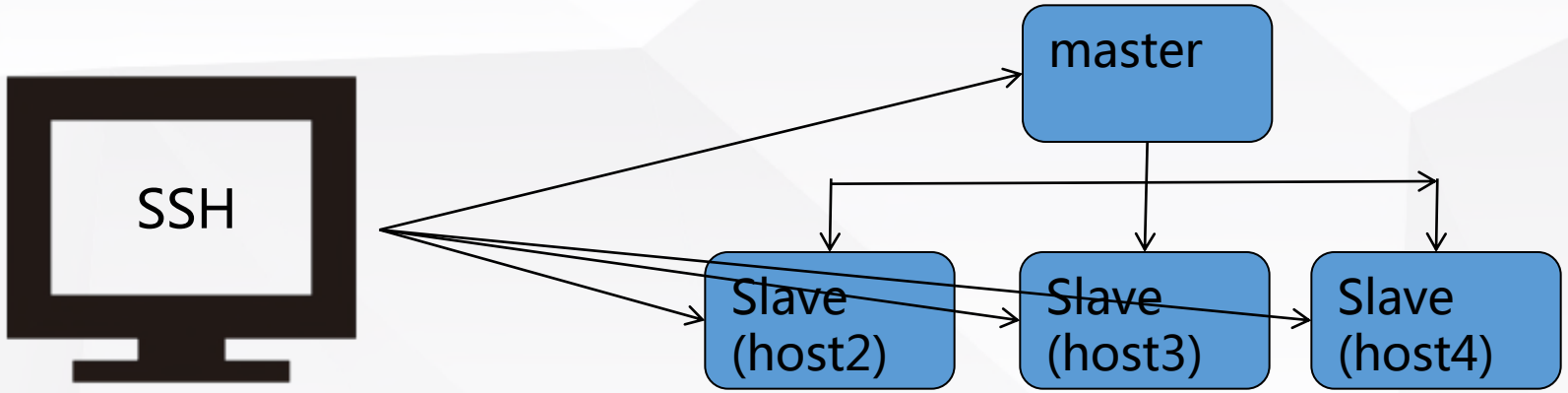
MapReduce没有认证或授权的概念

某个顽劣的用户可能为了让自己的任务更快完成而降低其他Hadoop任务的优先级，甚至更坏，直接杀掉其他任务

DataNode没有访问控制

恶意用户可以绕过访问控制从DataNode中读取任意数据块，或将垃圾数据写到DataNode中破坏目标分析数据的完整性

Hadoop安全现状




```
hadoop@box:~$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup
drwxr-xr-x   - hadoop supergroup
hadoop@box:~$ hdfs dfs -cat /output/*
1       dfsadmin
1       dfs.replication
1       dfs.namenode.name.dir
1       dfs.datanode.data.dir
hadoop@box:~$
```

NetWork
IPAddress:
master:10.10.11.191
host2:10.10.11.192
host3:10.10.11.193
host4:10.10.11.194
NetMask:
255.255.255.0
GateWay:
10.10.11.1

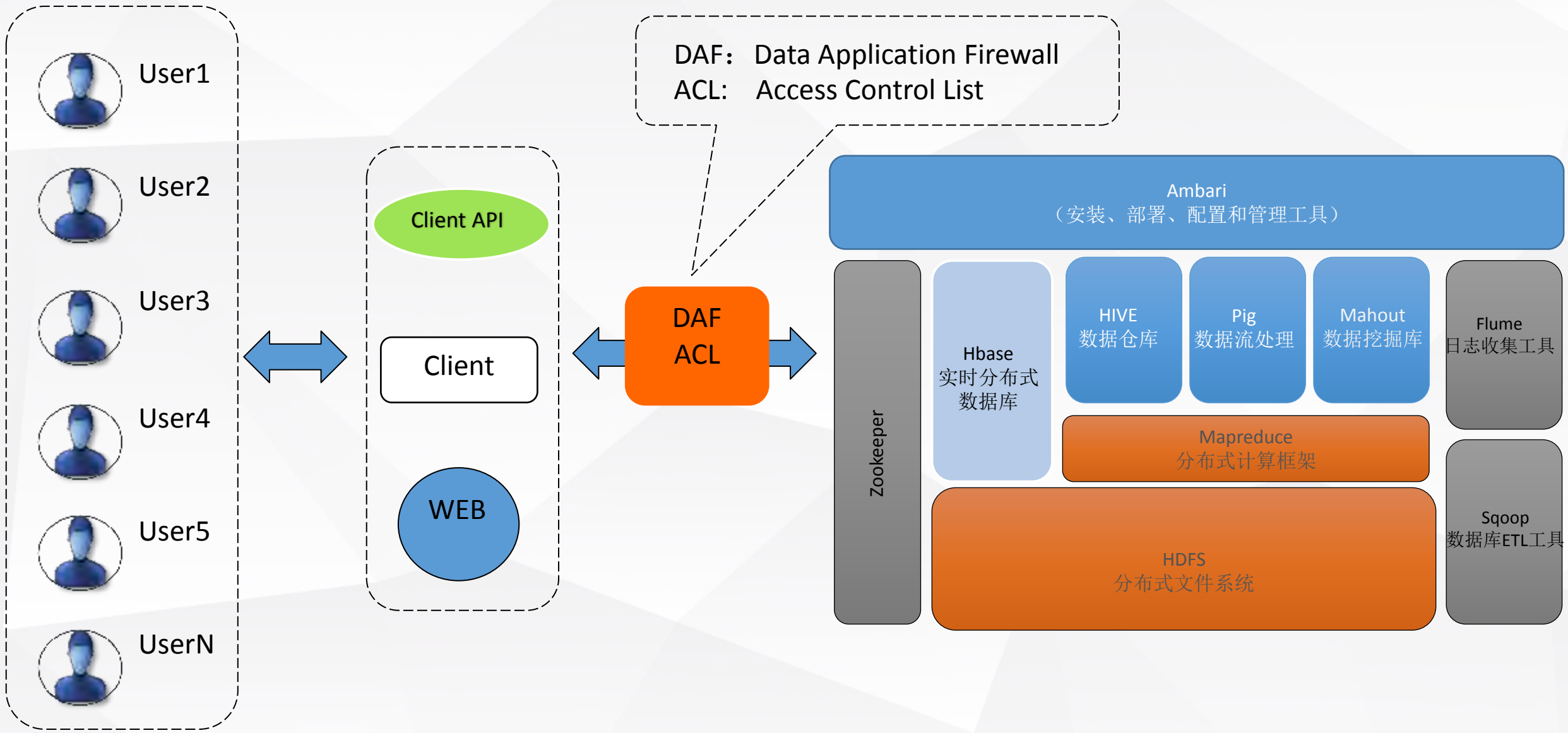
HDFS Roles
NameNode:
master
DataNode:
host2
host3
host4

MapReduce Roles
JobTracker:
master
TaskTracker:
host2
host3
host4

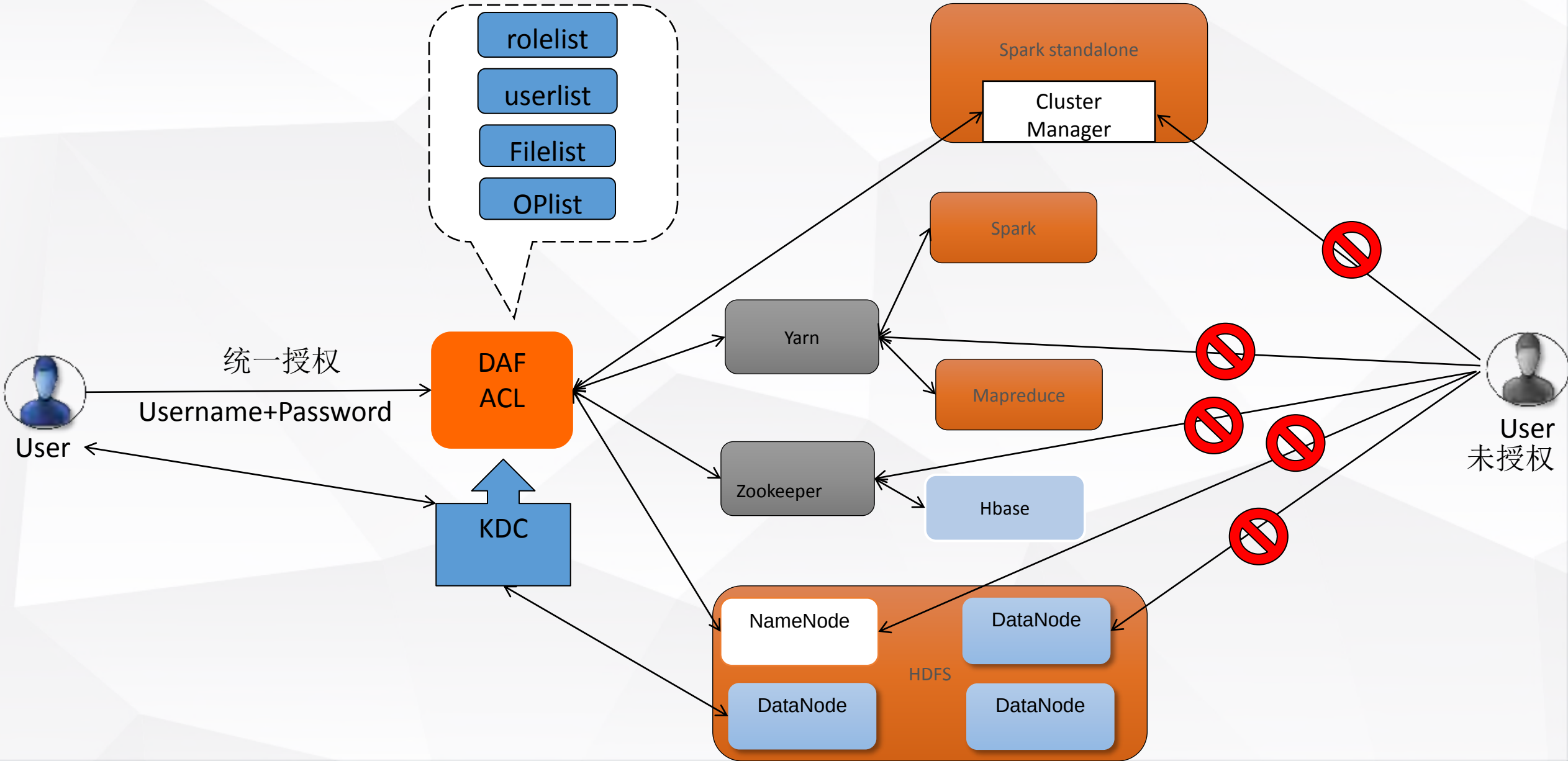
不安全的操作方式：利用SSH登陆到任何节点主机上本地用客户端操作集群，本地权限甚至可以关闭整个集群。

安全的操作方式：通过IP和端口远程操作集群。

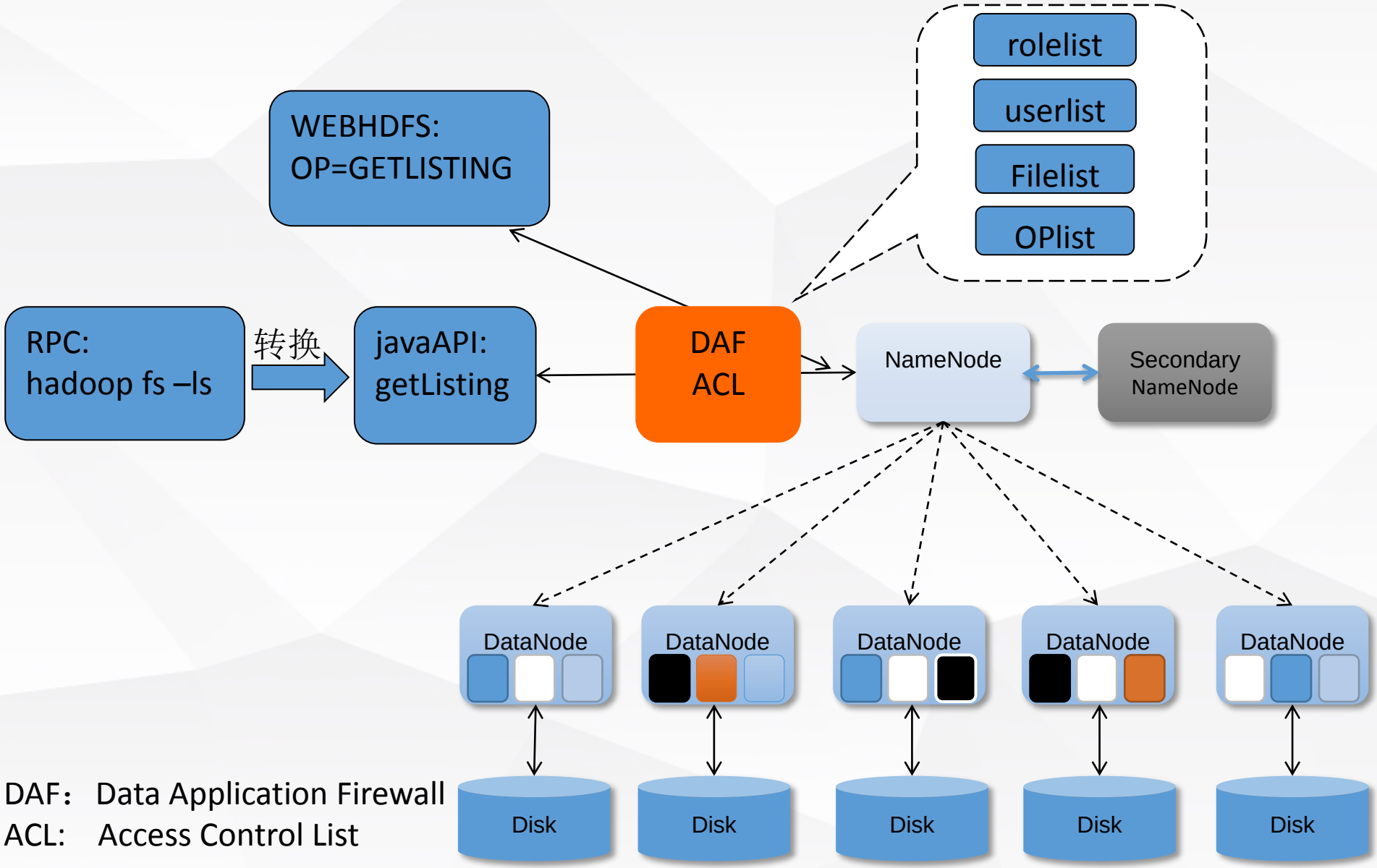
Hadoop安全防护整体思路



DAF控制器与各组件的关系



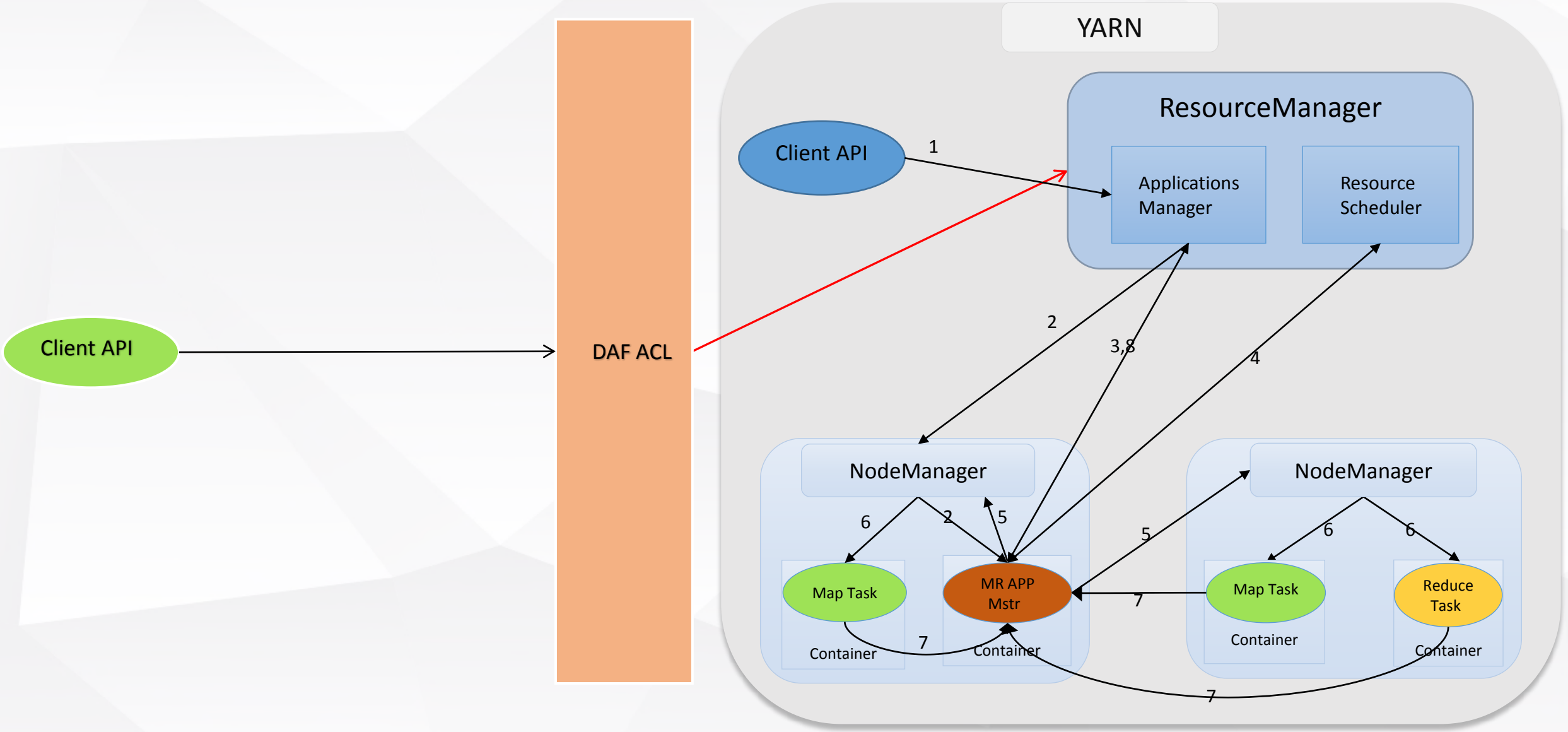
HDFS风险、防护方法



DAF： Data Application Firewall

ACL： Access Control List

MR2风险、防护方法


Client API
DAF ACL
YARN
ResourceManager
Client API
1
Applications Manager
Resource Scheduler
2
3,8
4
NodeManager
NodeManager
6
2
5
5
6
6
Map Task
MR APP Mstr
Map Task
Reduce Task
7
7
7
Container
Container
Container
Container

Spark

1、本地单机部署

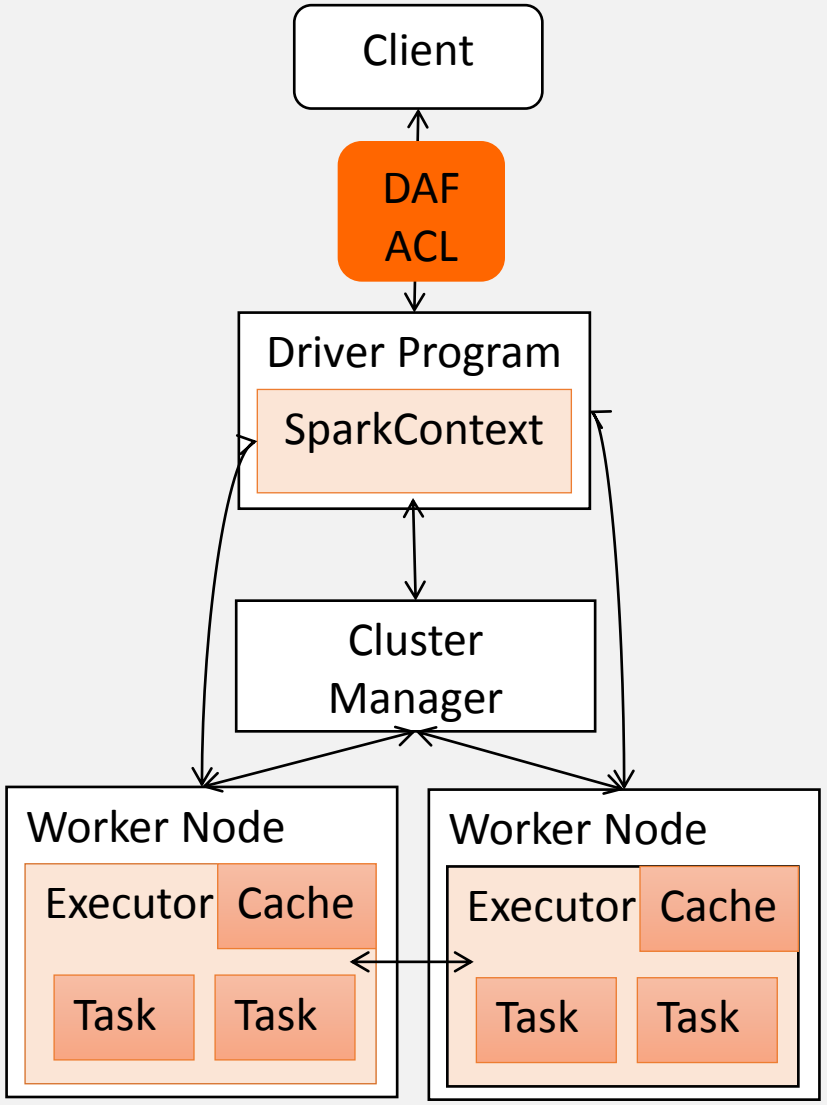
Local node

2、本地伪分布式

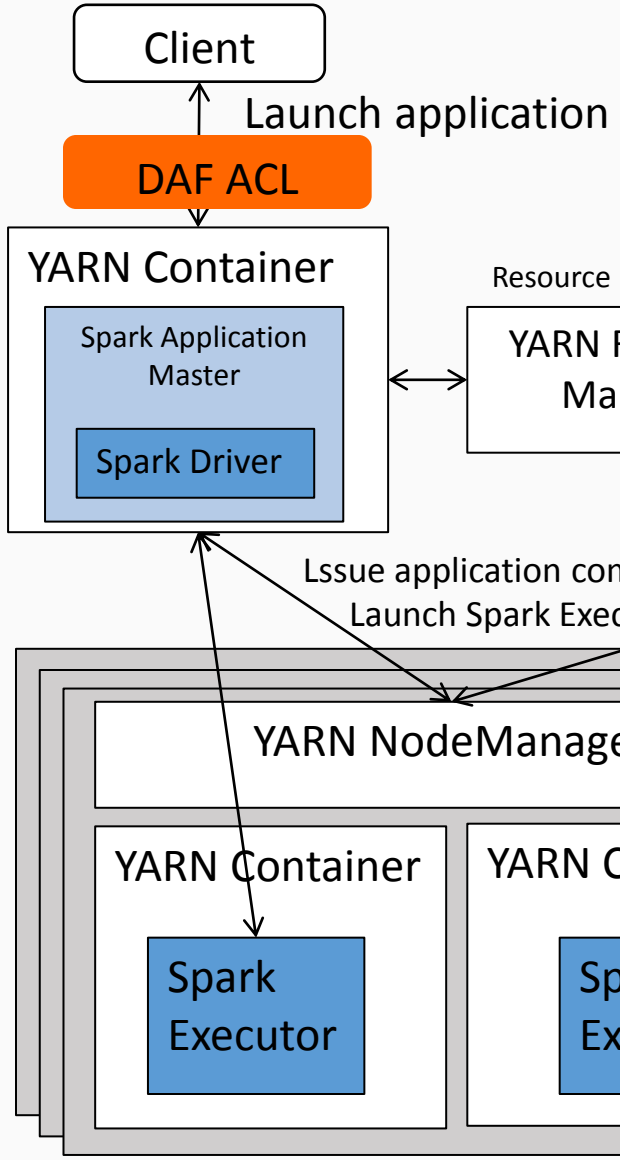
local cluster

node node

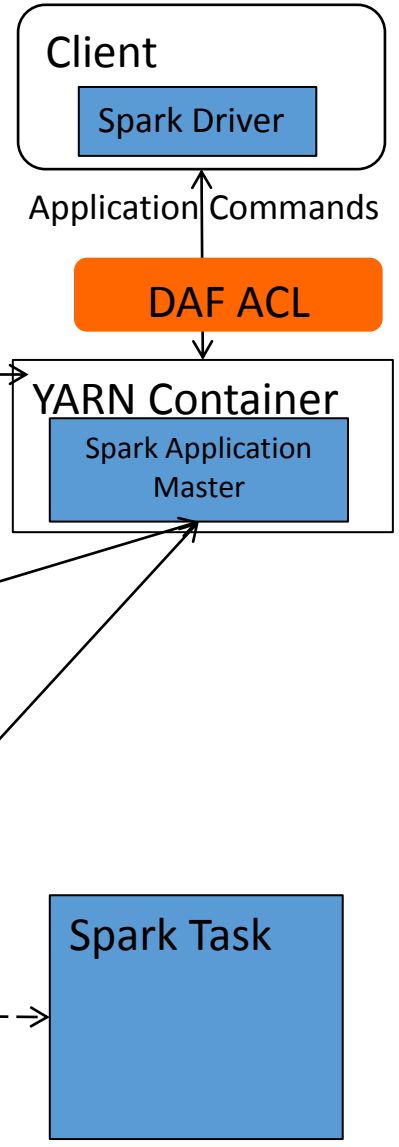
3、standalone



4、Spark on YARN cluster



5、Spark on YARN Client



Hbase风险，防护方法

user-1

user-2

user-3

user-n

裸奔

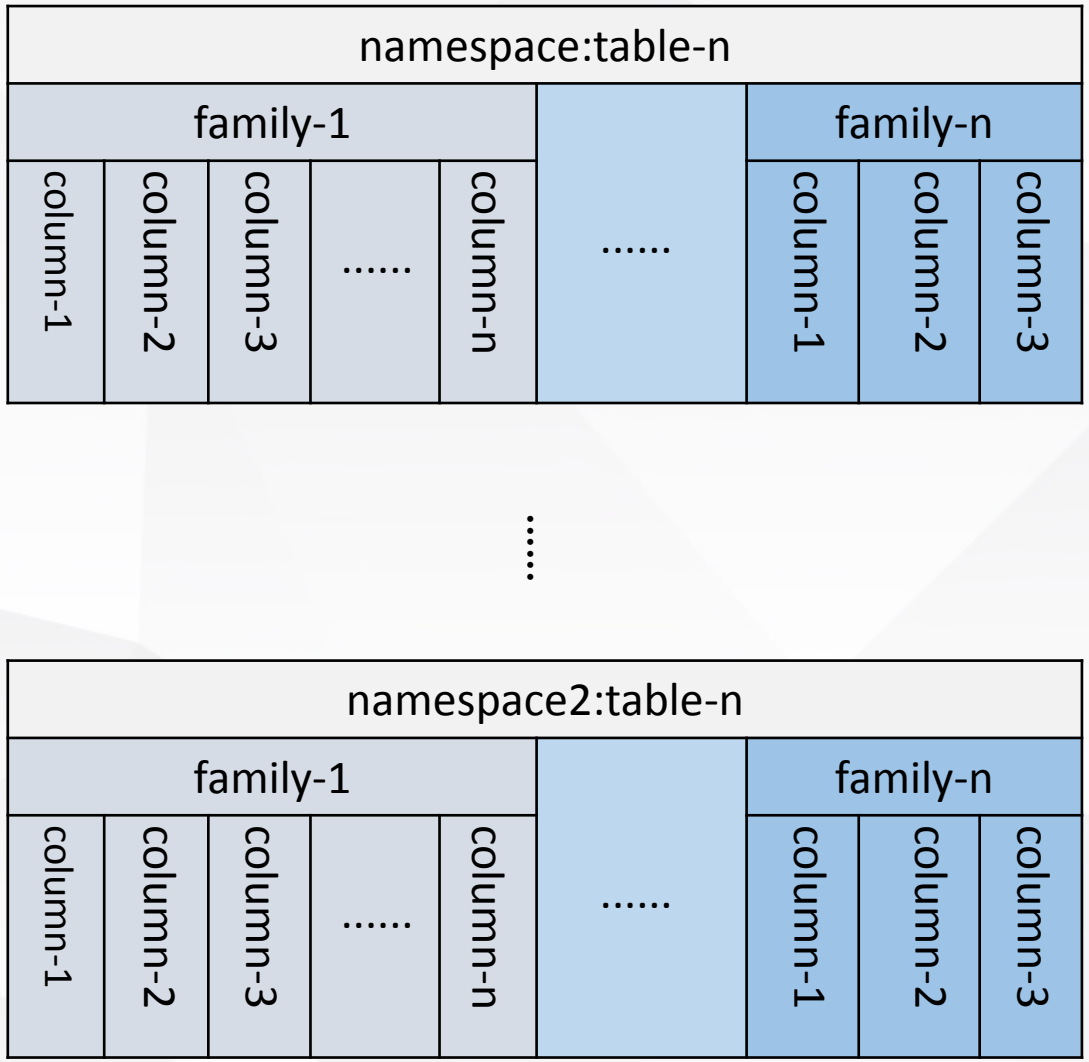
1. 任何用户都能访访问操作任何表里的数据
2. 数据库跟数据都没有任何安全保障

原生ACL

1.支持RWCA权限管理
2.登录到集群主机可以随意修改权限

DafACL

1.只有Daf后台管理员才能分配权限
2.支持RWCA权限，未来提供细粒化权限管理



还有哪些组件存在风险？

kafka的话题操作可以被恶意删除、恶意创建话题造成拒绝服务，通过访问描述和配置造成信息泄漏。

Strom无授权概念，可随意提交拓扑、文件上传、获得Nimbus配置、获取群集信息、文件下载、结束拓扑任务、Rebalance命令、启用、停用、获取拓扑配置、获取拓扑结构、获取用户拓扑结构、获取拓扑信息、上传新证书”。

Solr的文本检索无授权概念，可能出现恶意：查询、更新、管理。

Hive用户体系与hadoop相同，主要可能出现针对关系型数据库的表和字段的恶意访问：增删查改。

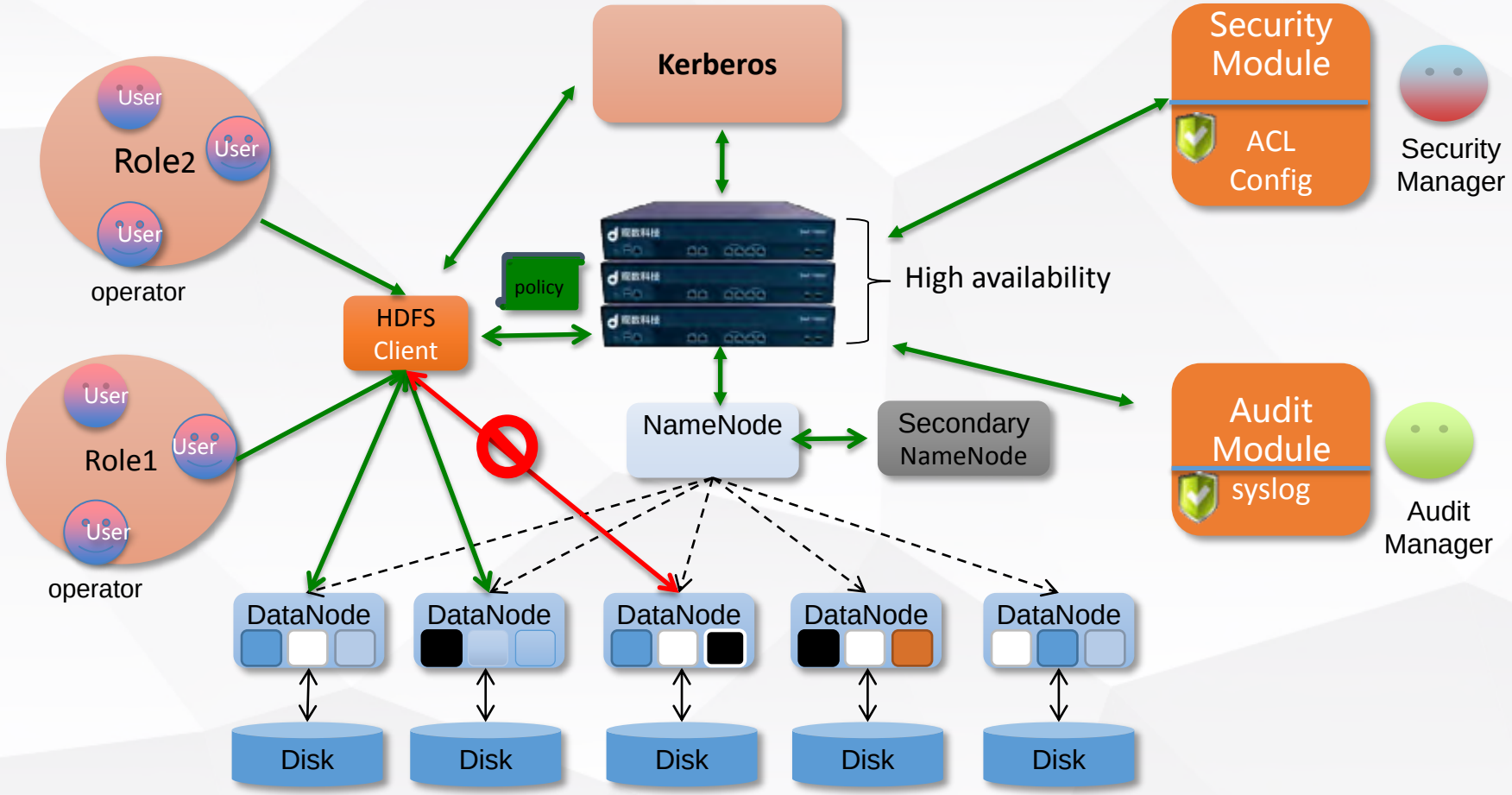
其他。。。。

5

PART FIVE

案例分享

BIG DAF部署逻辑



三权分立（安全、审计、操作）

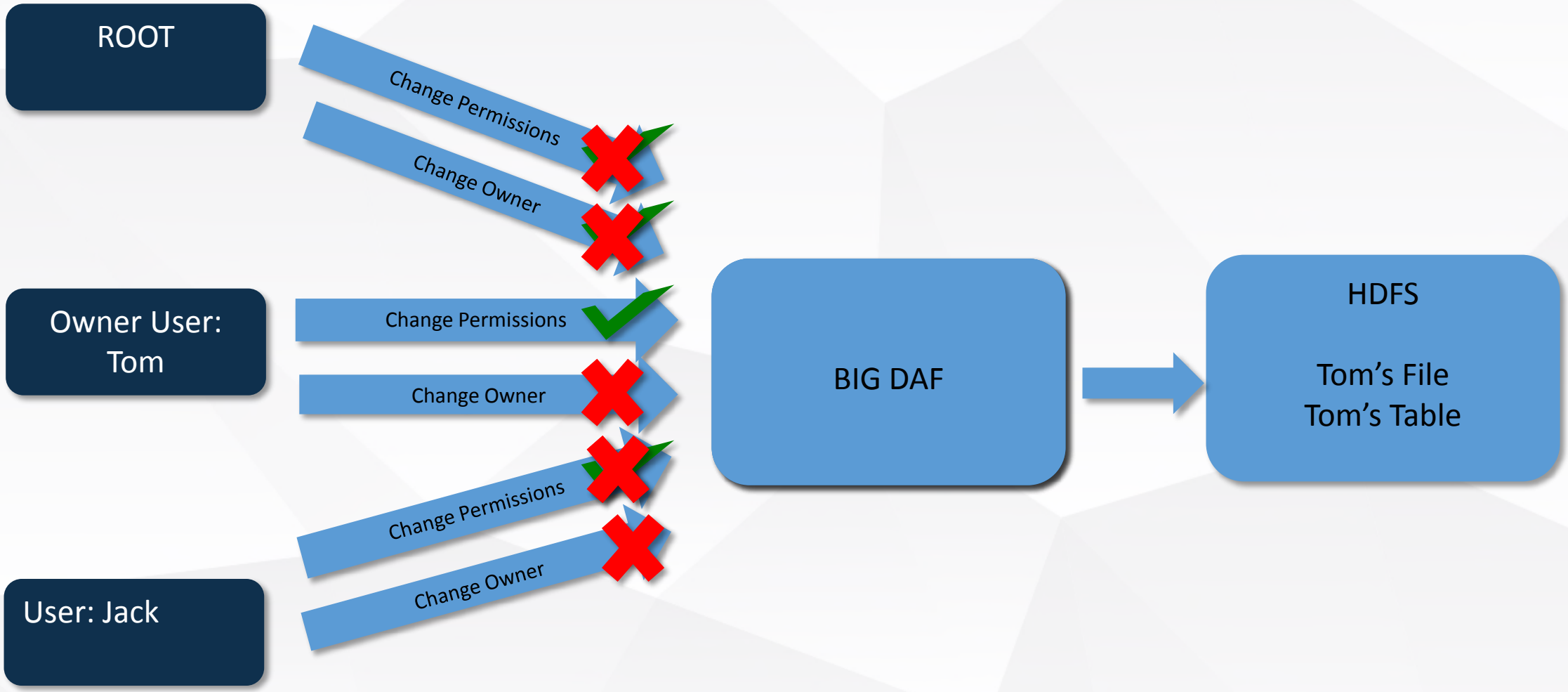
基于多种访问控制模型的Policy：DAC、MAC(BLP\Biba)、RBAC等

不需要Kerberos，采用自主认证机制，但兼容Kerberos凭证体系，无缝对接

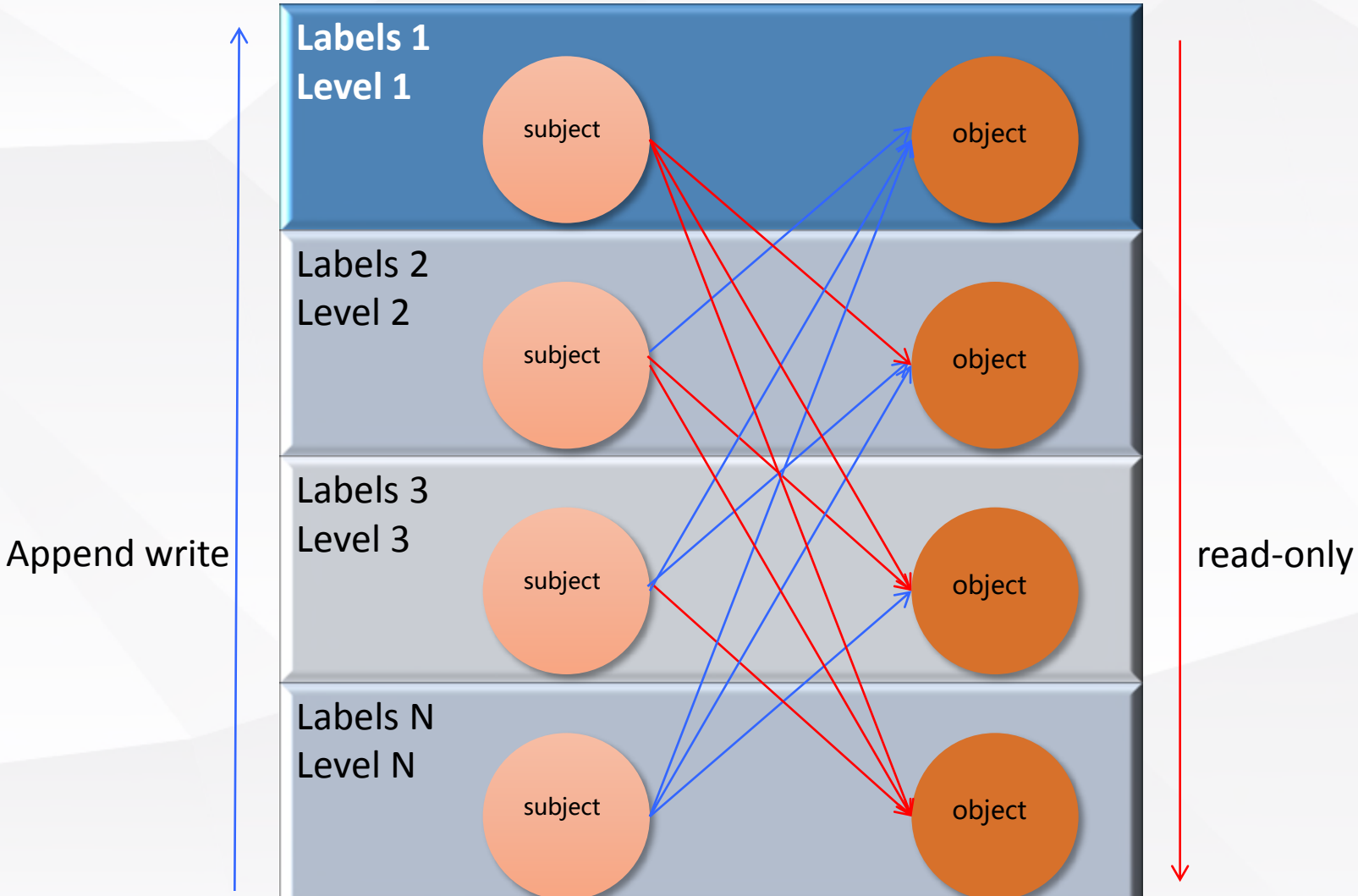
分权管理


安全管理员
配置安全规则
查看审计日志
访问数据资源
审计管理员
配置安全规则
查看审计日志
访问数据资源
数据
数据库管理员
配置安全规则
查看审计日志
访问数据资源

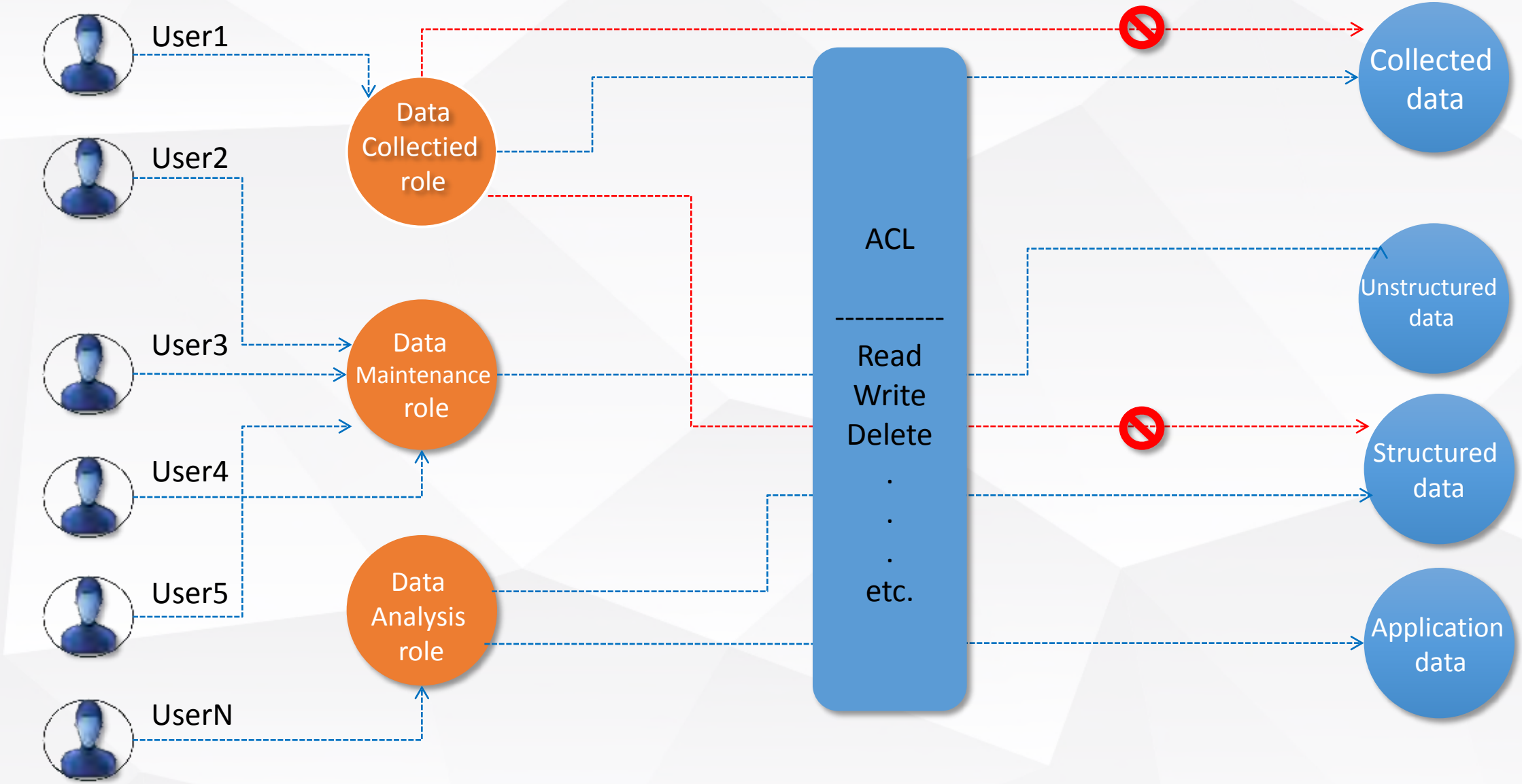
DAC


ROOT
Change Permissions
Change Owner
Owner User:
Tom
Change Permissions
Change Owner
Change Permissions
Change Owner
User: Jack
BIG DAF
HDFS
Tom's File
Tom's Table

MAC


Append write
Labels 1
Level 1
subject
object
Labels 2
Level 2
subject
object
Labels 3
Level 3
subject
object
Labels N
Level N
subject
object
read-only

RBAC


User1
User2
User3
User4
User5
UserN
Data Collectied role
Data Maintenance role
Data Analysis role
ACL

Read
Write
Delete
.
.
.
etc.
Collected data
Unstructured data
Structured data
Application data

Big DAF技术难点

技术难点

性能

并不转发流量，ACL与审计只针对主客体访问行为，几乎没有性能消耗。而且Big DAF支持多机高可用方案，不会存在性能瓶颈。

兼容性

非侵入式技术，并未修改hadoop源码，独立于hadoop系统之外的防护体系，因此对通用版、发行版都具有良好的兼容性，并且厂商会持[illegible]版本的特性，升级、适[illegible]

安全性

Big DAF为独立体系，无法从hadoop上配置Big DAF的安全策略，可形式化语言验证的多种安全模型，配置、日志可备份保存，格式加密，后台防破解，全程SSL加密。

易用性

操作上手容易、逻辑清晰简单，支持非拦截的监控模式，只记录违规不做阻拦，适用于新规则的适配。可一键暂停防护，或还原到初始无功能状态，方便排查故障。

Big DAF支持越来越多主流Hadoop的新版本：

Apche Hadoop 2.7
Cloudera cdh 5.8.1
Hortonworks Sandbox 2.4
Hortonworks HDP 2.4
星环Transwarp Data Hub
红象云腾Redoop Enterprise CRH

Big DAF 相关资质


中华人民共和国国家版权局
计算机软件著作权登记证书


公安部计算机信息系统安全产品质量监督检验中心


中关村
Z-Park
ZHONGGUANCUN SCIENCE PARK
中关村高新技术企业
企业名称：北京观数科技有限公司
编　　号：201620103446O1
发证时间：2016年10月13日
有 效 期：三年


CMA
CNAS
CNAS L0663
报告编号：公计检160939
检 验 报 告
样品名称　BIG DAF Hadoop安全防护系统
型号规格　V1.0
产品类型/级别　访问控制（网络-增强级）
受检单位　北京观数科技有限公司
检验类别　委托检验
国家网络与信息系统安全产品质量监督检验中心
公安部计算机信息系统安全产品质量监督检验中心
公安部信息安全产品检测中心

观数科技

谢谢

作者：李科
微信号：wuxiulike
2016.12