深圳竹云科技有限公司

竹云风险引擎产品白皮书

Bamboocloud Risk Engine Products White Paper

编号：BBC-WP-2018001

深圳竹云科技有限公司

2018 年 4 月

目录

1 背景

近年来，伴随着互联网的爆炸式发展、网络应用加快深入到经济社会各个领域，网络安全形势也面临着更加严峻挑战。现代化运营、生产力提升和操作效率提高的压力，促使公司企业纷纷引入联网技术和产品，但同时也抹去了工业网络与 IT 网络之间一贯的物理隔离。很多企业网络应用上缺乏防护，不能确保设备不存在安全风险。在现在，迫切需要建立与之相适应的保障体系。

在当前安全领域的复杂形势下，传统的安全团队不得不面对百花齐放的业务场景和大规模的数据情况。如何在传统攻防对抗之外，寻找更有效、可落地的对抗方式，已经成为了各大企业安全团队思考的重点。安全领域的发展已经从专家模式逐渐演变成为了现在的系统化、平台化，而随着机器学习和大数据技术的发展，未来安全将逐渐智能化。

因此，基于一些成熟的网络安全防护方案，建设基于识别用户可信身份的风险引擎也成了我们保障体系中的一环。风险引擎的整体概念是基于大数据环境、基于数据挖掘和机器学习技术，快速的对入侵行为进行识别。

对于入侵行为发生时，要做到能够实时检测，及时的报警或者响应；同时快速通知用户本人，立即采取措施防范；建设风险规则的规则库，规则库要做到能够及时更新，以应对新的入侵行为；对于入侵行为要把相应的数据包或者信息存入日志中，以供管理员进行处理。

提供基于用户行为数据的风险评分机制，所谓风险评分，就是根据用户信息使用特定的公式规则算法计算出用于标识客户风险级别的分值。评分机制可以根据情况随时调整，包括评分参数，计算公式，风险因子等。

基于风险分析等数据，可以进一步的对其数据的进行扩展分析，并提供其它数据分析服务。

2 产品介绍

2.1 概念和术语

图计算：图计算是以“图论”为基础的对现实世界的一种“图”结构的抽象表达，以及在这种数据结构上的计算模式。图数据结构很好的表达了数据之间的关联性，因此，很多应

用中出现的问题都可以抽象成图来表示，以图论的思想或者以图为基础建立模型来解决问题。

ETL：是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。

IP 画像：IP 地址画像的简称，根据用户 IP 的访问频率、访问信息等，对用户 IP 地址进行多维度的综合性的分析。

设备画像：对用户访问应用的硬件设备进行信息采集，根据唯一设备号，以访问频率、访问信息等多维度的指标做综合性分析。

风险引擎：全称是风险规则引擎，可依据一定的可设置的规则，对用户风险进行分析和判断。

分布式计算平台：分布式计算是一种计算方法,和集中式计算是相对的。主要指利用分布式计算技术，对数据进行分析的软件平台。

知识图谱：知识图谱（Knowledge Graph/Vault）又称为科学知识图谱，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

撞库：撞库是黑客通过收集互联网已泄露的用户和密码信息，生成对应的字典表，尝试批量登陆其他网站后，得到一系列可以登录的用户。很多用户在不同网站使用的是相同的帐号密码，因此黑客可以通过获取用户在 A 网站的账户从而尝试登录 B 网址，这就可以理解为撞库攻击。

暴破：账号暴力破解，通过不断尝试不同的密码，访问应用以达到暴力修改程序的代码来达到破解的。

2.2 产品描述

竹云风险引擎产品基于用户的部分行为数据，通过对数据的收集、清洗，对数据进行实时分析和离线分析，最终以实现用户潜在风险的分析、评估，将提前预估用户可能存在的风险，因此而起到将用户风险扑灭在初期、对用户安全防患于未然的作用。

2.3 特点和优势

- 提供基于用户数据的实时风险分析，数据基于用户访问应用时产生的行为数据。如 IP、访问地域、访问设备等；

- 提供基于用户数据的用户风险评分，用户评分基于灵活的调整策略；
- 提供基于规则引擎的风险策略调整，灵活的判断用户风险；
- 提供数据收集功能，对实时数据和非实时数据提供多种灵活的采集手段；
- 提供数据收集适配功能，通过配置灵活的采集不同数据源的不同数据；
- 提供数据收集监控功能，监控数据收集情况；
- 提供基于用户行为数据的报表展示，为企业提供辅助决策；
- 支持高并发请求、支持大数据的存储和计算；
- 支持基于机器学习、知识图谱的风险检测技术；风险检测算法支持如下：IP 画像检测、疑似被盗号检测、疑似撞库攻击检测、疑似账户暴力破解检测。
- 算法模型及行为分析报表会随着业务数据的不断增加而及时有效的更新。

2.4 系统架构

2.4.1 整体架构

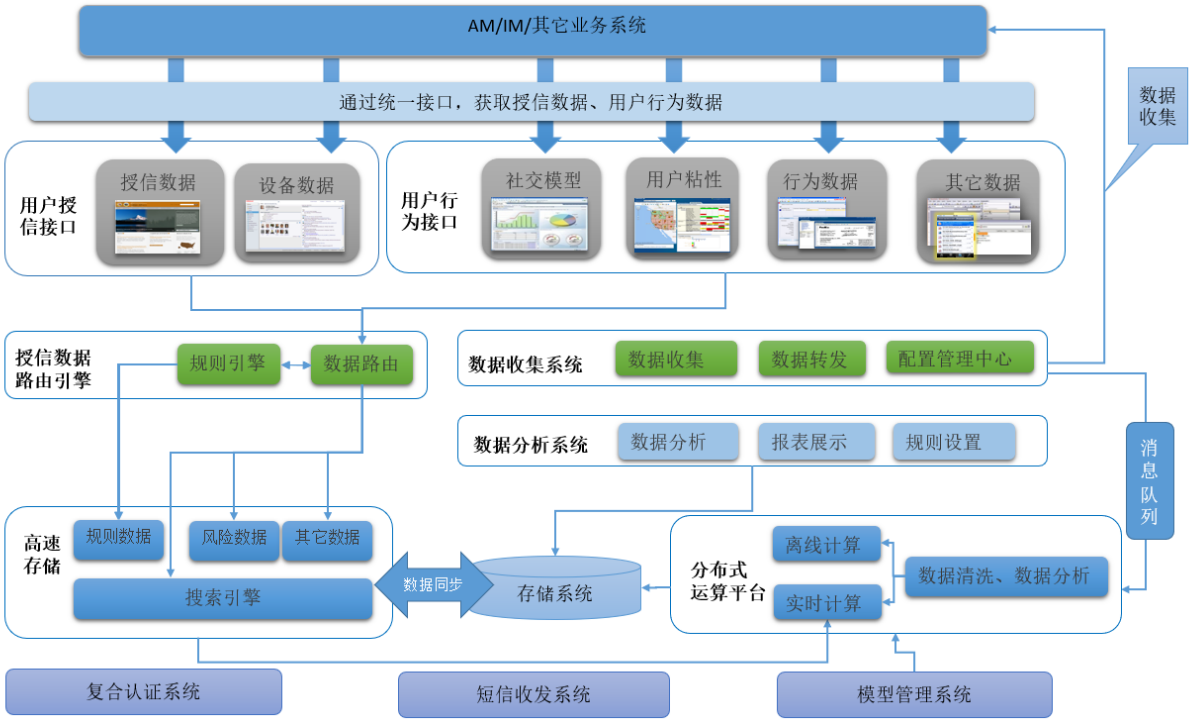

图 1 系统逻辑架构图

系统整体逻辑架构如图，具体模块描述如下：

- **用户授信接口：**用户授信接口是基于 http 协议的 restful 接口服务，主要提供用户授信、风险等信息。因为考虑到较为频繁的认证操作，因此用户接口属于高并发、大数

据量的应用，用户授信接口本身只负责接口通讯的工作，其数据通过调用授信路由引擎来获取返回的数据；

- **用户行为接口：**用户行为接口与用户授信接口一样是基于 http 协议的 restful 接口服务，主要提供用户行为及其相关的经过分布式计算平台分析后的数据。其数据根据频次不同可灵活的在架构上来调整数据源，数据源可以基于高速存储也可以基于存储系统。
- **授信数据路由引擎：**授信数据路由擎主要负责将收到的数据进行分类、逻辑处理，并通过信息来调用缓存中已经存在的数据，例如：接口调用者传递数据：姓名、年龄、ID，则授信引擎会通过 ID 找到用户已经存在的信用信息，并返回给接口调用者。授信路由引擎在进行数据路由时会先访问规则引擎，根据规则引擎来判断不同数据的路由规则。规则引擎的信息则来源于用户数据分析系统中的设置；
- **高速存储：**因为授信系统本身属于高并发、大数据的需求，因此高速存储负责的主要工作是在高并发、大数据的场景下解决高并发请求与快速的数据查询问题，高速存储本身属于底层的技术基础架构，其具体设计可以根据业务不同情况来处理，现阶段主要采用分布式缓存系统。
- **数据存储：**数据存储主要解决的问题是对用户风险信息及其它分析后的信息进行持久化的存储。其属于底层技术基础支撑系统，目前阶段主要采用 mysql 数据库。
- **用户数据收集系统：**用户数据采集系统做为最关键的一环，应提供多种数据采集渠道，采集的数据质量直接影响到数据分析的准确性和效率，采集数据要尽量做到全面，减少数据丢失和漏采。用户行为挖掘的数据来源于系统记录的各种类型的日志、数据库、用户访问中产生的数据。主要用于离线的收集用户信息，模式有两种：主动模式和被动模式。主动模式下可以通过 agent 的方式远程采集用户日志、数据库信息等，也可以通过 JDBC 的方式访问数据库主动采集信息。被动模式下会通过开发接口，提供给调用者，由调用者写入用户数据。用户数据收集系统的数据写入到分布式计算平台。数据收集系统本身提供基于不同采集模式的配置、数据采集监控功能。
- **用户数据分析系统：**主要是针对用户数据的分析系统，提供功能如下：

·用户风险数据的查看、修改、增加、删除功能；

·提供远程调用接口，供应业务系统调用，如：用户上次登录设备、日期等信息。远程调用接口部分采用分布式架构，直接调用查询系统的数据；接口部分集成在用户授信接口中）

·用户征信系统直接访问数据存储系统，提供针对数据存储系统的增删改查操作；

·提供可视化界面给用户及管理员使用；

·提供针对用户、角色的增删改查功能；

·集成规则引擎设置，提供对风险规则、风险评分对修改；

- **分布式计算平台：**主要提供对数据的分析功能，如排序、聚类、分类、文本挖掘等大数据算法操作，其数据源主要来源于用户数据采集系统。

2.4.2兼容性

软件名称	类型	作用	名称	兼容性
大数据分析平台	数据分析中间件	用户大数据存储、分析	Horton Works Hadoop	支持 centos、redhat、ubuntu
Web 界面组件	Web 开发界面组件	用于 web 系统前端	RiskEngine-UI 1.0	
消息队列	消息队列中间件	用于大数据的传输	Kafka 0.11	支持 centos、redhat、ubuntu
集群组件	分布式集群中间件	用户 hbase、kafka 等分布式集群环境	Zookeeper 3.4.5	支持 centos、redhat、ubuntu
数据采集组件	数据采集组件	用户 linux 系统的 log 日志、文本采集	Flume 1.8.0	支持 centos、redhat、ubuntu、windows
数据分析组件	数据分析组件	用于数据分析、挖掘	Spark 2.0.2	支持 centos、redhat、

				ubuntu
大数据存储组件	大数据存储组件	用于大数据存储	Hbase 1.2.2	支持 centos、redhat、ubuntu
Web 中间件	Web 容器中间件	用于风险引擎 web 系统的运行环境		支持 tomcat、jboss、weblogic、websphere
负载均衡组件	Web 负载均衡	用于风险引擎 web 系统运行环境	Nginx1.12	支持 centos、redhat、ubuntu、windows
数据库	关系型数据库	用于风险引擎 web 系统以及存储已经分析后的数据		支持 mysql5.5 以上 Oracle10g 以上
浏览器	用于界面展示			支持 chrom、ie10

2.4.3运行环境

名称	最低配置	最低数量（台）	备注
Hadoop 集群	16 核 CPU　32G 内存	4	视实际业务确定机器台数
Tomcat 集群	16 核 CPU　32G 内存	1	视实际业务确定是否需要集群
Kafka 集群	16 核 CPU　32G 内存	1	视实际业务确定是否需要集群
Nginx	16 核 CPU　32G 内存	1	视实际业务确定是否

			需要集群
数据库	16 核 CPU　32G 内存	1	视实际业务确定是否需要集群及集群方案

2.5 主要功能介绍

2.5.1 大数据分析平台

大数据分析平台采用了开源版的大数据集成环境 HortonWorks Hadoop，内置多种大数据开发组件，如图：

HDFS	2.7.3	Apache Hadoop Distributed File System
YARN + MapReduce2	2.7.3	Apache Hadoop NextGen MapReduce (YARN)
Tez	0.7.0	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
Hive	1.2.1000	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
HBase	1.1.2	A Non-relational distributed database, plus Phoenix, a high performance SQL layer for low latency applications.
Pig	0.16.0	Scripting platform for analyzing large datasets
Sqoop	1.4.6	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
Oozie	4.2.0	System for workflow coordination and execution of Apache Hadoop jobs. This also includes the installation of the optional Oozie Web Console which relies on and will install the ExtJS Library.
ZooKeeper	3.4.6	Centralized service which provides highly reliable distributed coordination
Falcon	0.10.0	Data management and processing platform
Storm	1.0.1	Apache Hadoop Stream processing framework
Flume	1.5.2	A distributed service for collecting, aggregating, and moving large amounts of streaming data into HDFS
Accumulo	1.7.0	Robust, scalable, high performance distributed key/value store.
Ambari Infra	0.1.0	Core shared service used by Ambari managed components.
Ambari Metrics	0.1.0	A system for metrics collection that provides storage and

图 2 开发组件示意图 1

Atlas	0.7.0	Atlas Metadata and Governance platform
Kafka	0.10.0	A high-throughput distributed messaging system
Knox	0.9.0	Provides a single point of authentication and access for Apache Hadoop services in a cluster
Log Search	0.5.0	Log aggregation, analysis, and visualization for Ambari managed services. This service is **Technical Preview**.
Ranger	0.6.0	Comprehensive security for Hadoop
Ranger KMS	0.6.0	Key Management Server
SmartSense	1.4.2.2.5.2.0-298	SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
Spark	1.6.2	Apache Spark is a fast and general engine for large-scale data processing.
Spark2	2.0.0	Apache Spark 2.0 is a fast and general engine for large-scale data processing. This service is **Technical Preview**.
Zeppelin Notebook	0.6.0	A web-based notebook that enables interactive data analytics. It enables you to make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
Mahout	0.9.0	Project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification
Slider	0.91.0	A framework for deploying, managing and monitoring existing distributed applications on YARN.

图 3 开发组件示意图 2

其平台除了内置组件、简化安装等功能，还提供有界面维护功能，可以通过界面进行比较方便的监控和维护工作，如图：

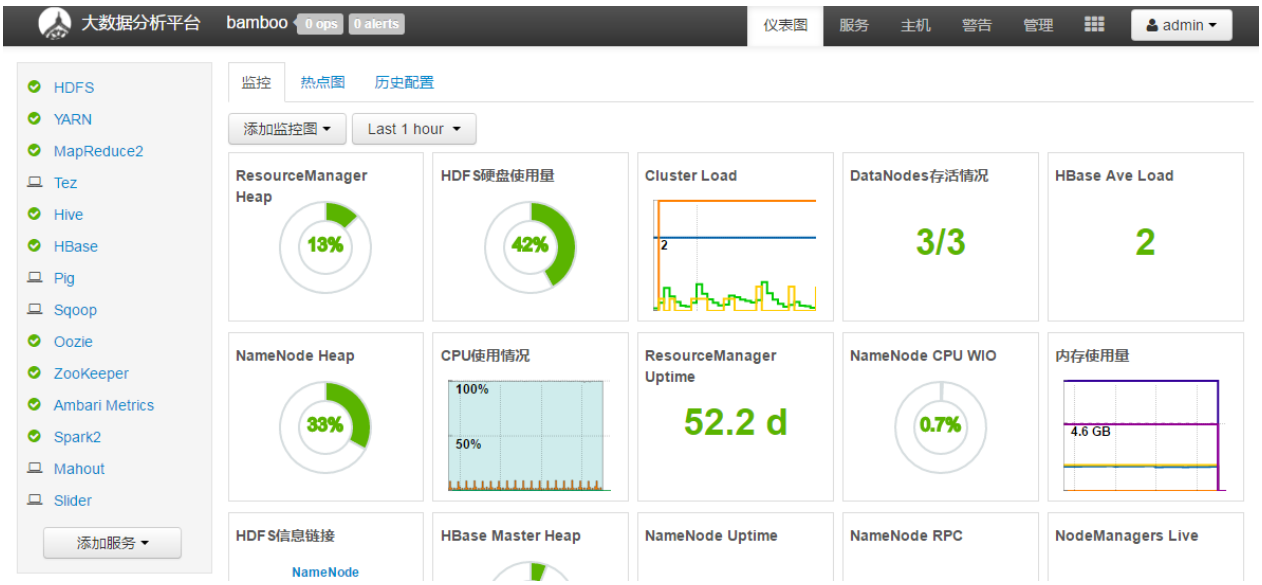

图 4 维护界面示意图

2.5.2 风险引擎管理系统

风险引擎管理主要对风险引擎进行风险信息监控、风险规则管理等功能。其主要功能包括如下：

➢ **监控面板**

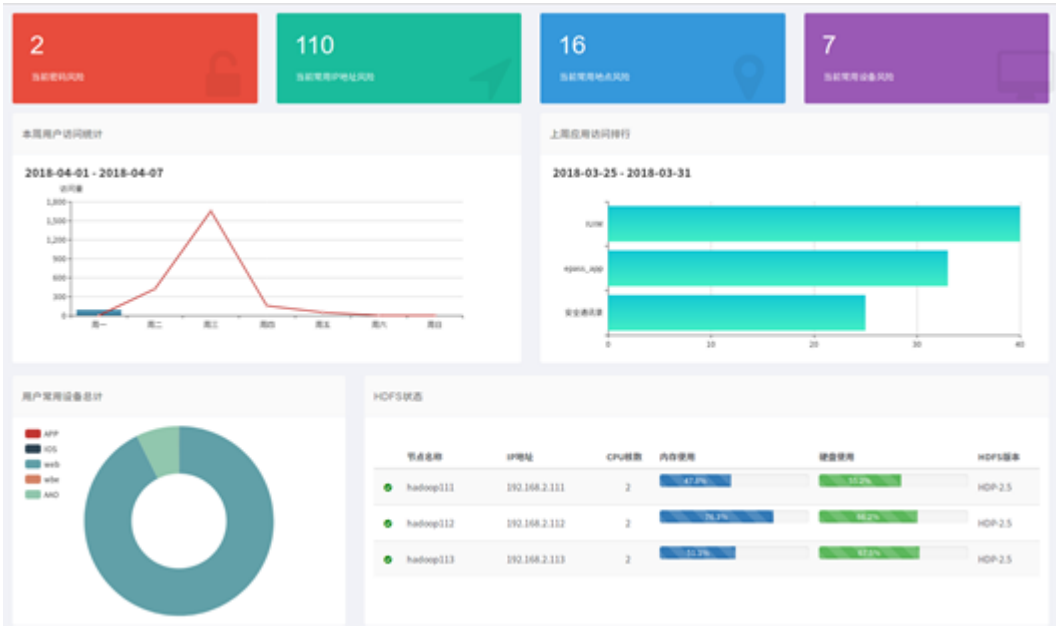

图 5 监控面板

控制面板包括了最上方当天风险报警次数；中间本周用户访问统计报表和上周应用访问排行榜；以及下方用户设备统计饼图和 HDFS 服务器健康监控。

➢ **风险监控**

图 6 风险控制界面示意图 1

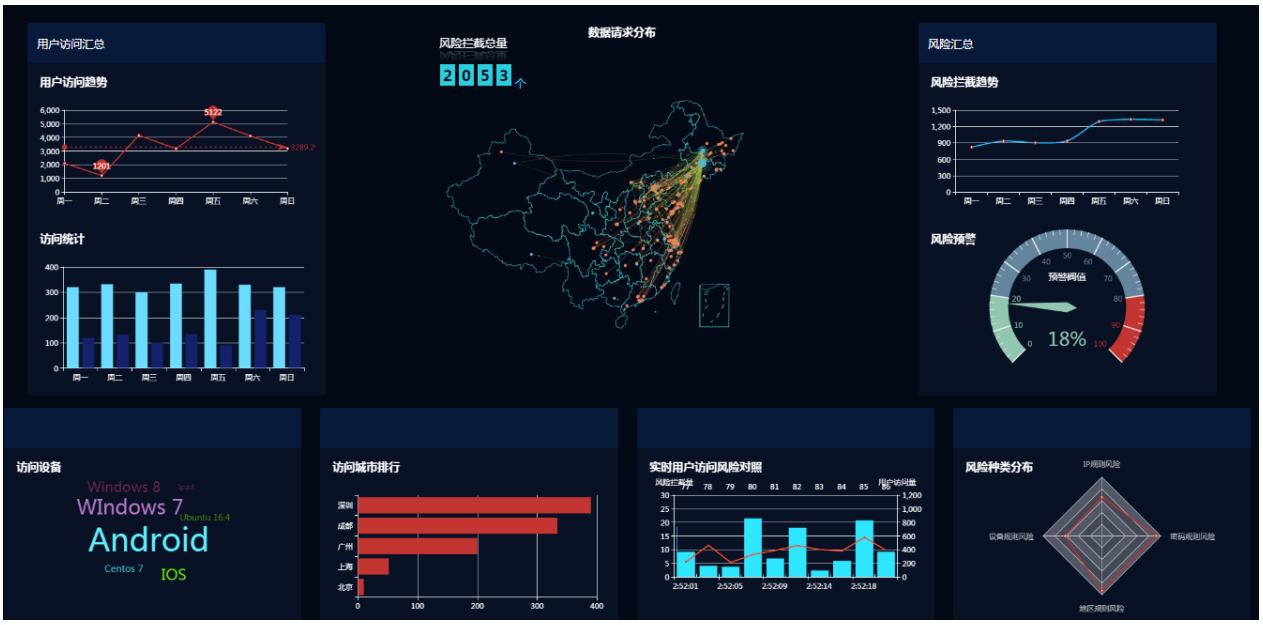

图 7 风险控制界面示意图 2

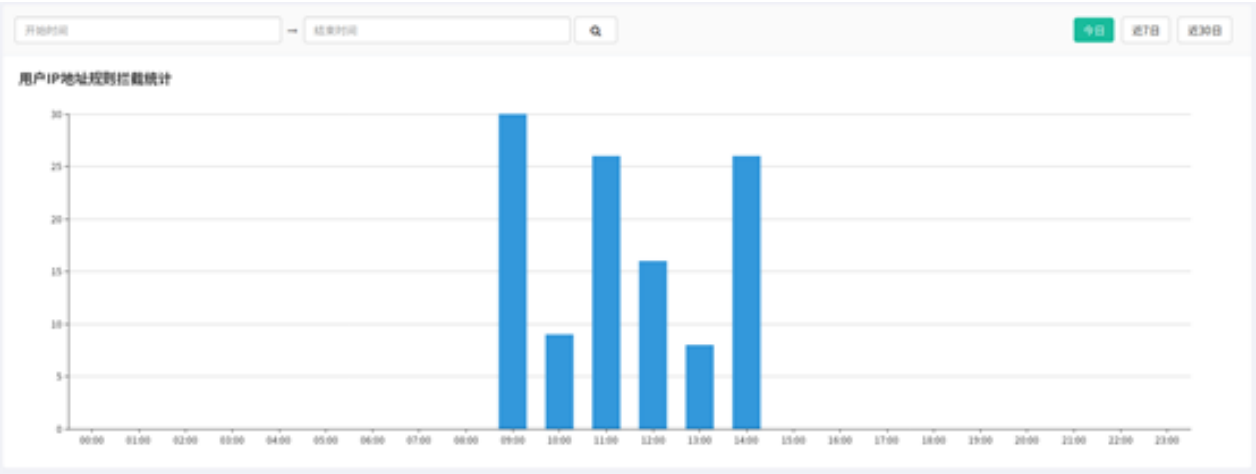

图 8 风险控制界面示意图 2

风险监控包括对用户认证趋势、规则拦截统计率、风险统计率、用户访问趋势等多维度信息进行综合监控。

3、规则设置

图 9 规则设置

规则管理将管理系统内置的拦截规则逻辑。目前风险引擎提供 4 中内置规则：

- **IP 规则：**

IP 规则将限制用户访问所处 IP 地址是否匹配该用户常使用的 IP 地址，规则值指定可允许历史常用 IP 地址个数，如请求 IP 地址不为此有效个数范围内的 IP 地址则命中风险；

- **密码规则：**

密码规则将限制用户输入错误密码次数，规则值为次数，如达到该次数上限，则命中规则。该规则应在配置的时间范围内重置，如 3 0 分钟；

- **设备规则：**

设备规则将限制用户访问所使用设备是否匹配该用户常使用的设备，规则值指定可允许历史常用设备个数，如访问设备不为此有效个数范围内的设备则命中风险；

- **地址规则：**

地址规则将限制用户访问所处地址是否匹配该用户常使用的地址，规则值指定可允许历史常用地址个数，如请求地址不为此有效个数范围内的地址则命中风险。

4、用户管理

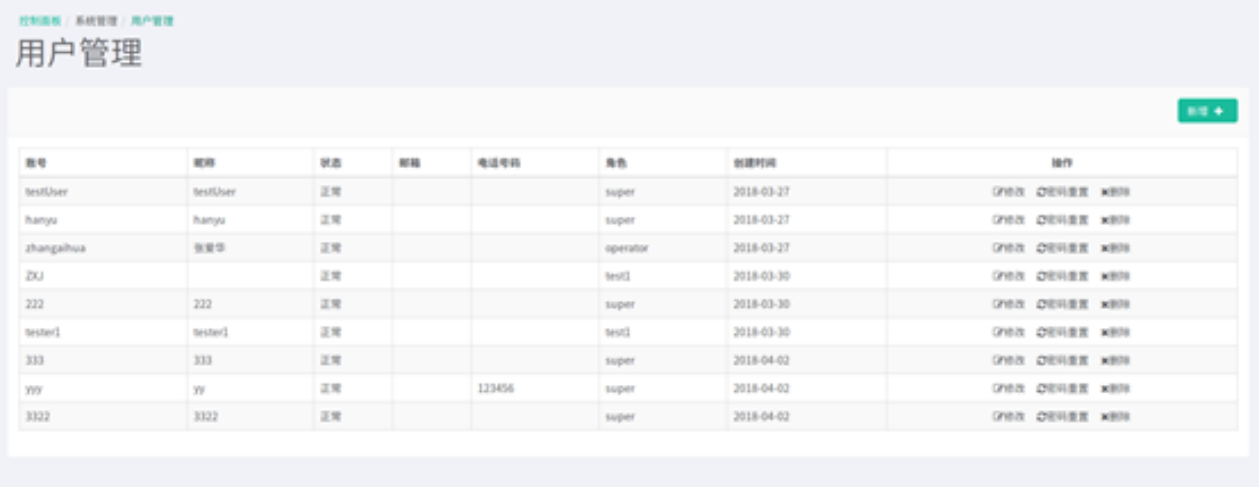

图 10 用户管理界面

图 11 角色管理界面

用户管理主要是针对系统用户、角色、权限的增加、删除、修改以及密码重置等管理服务。

3 产品部署

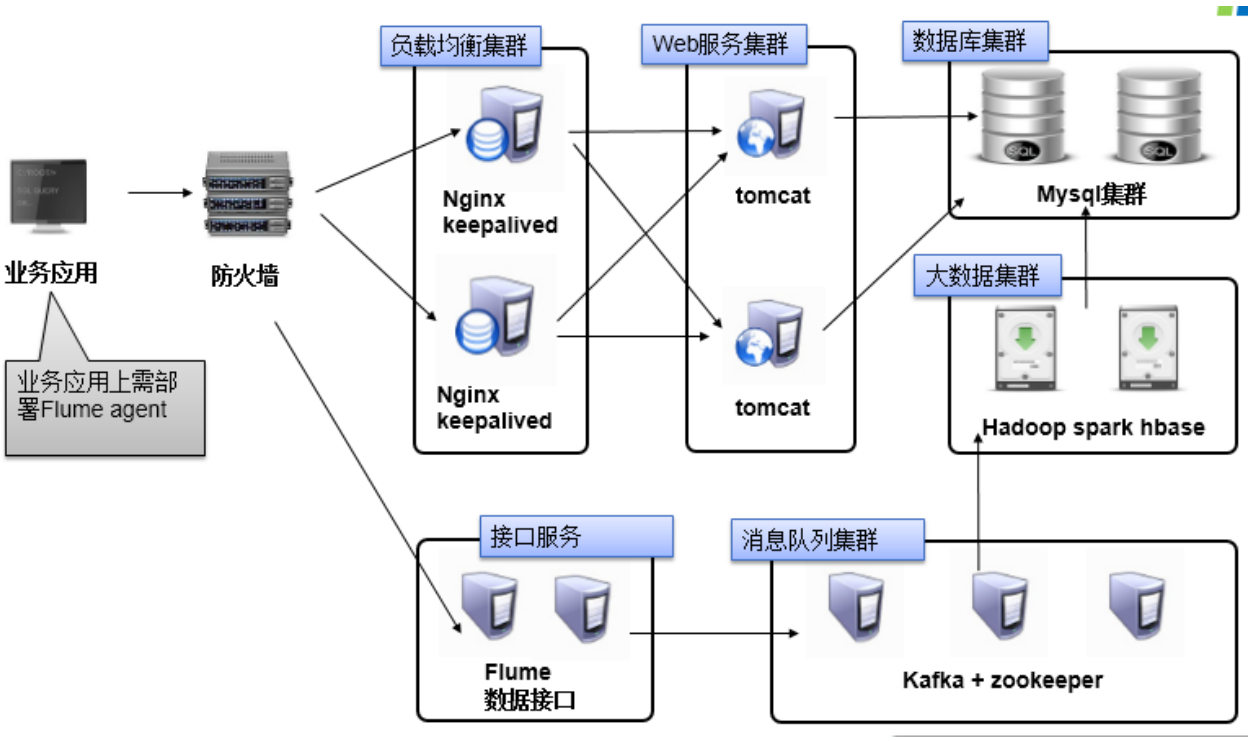

图 12 产品部署

部署架构上主要分为：

- **负载均衡集群：**采用 nginx+keepalived 双主架构；
- **Web 服务集群：**采用 web 容器单机部署，由 nginx 负责请求分发；
- **数据库集群：**支持多种数据库，目前采用 Haproxy+Keepalived+Mysql；
- **大数据集群：**采用 Horton Works Hadoop，内置包括 spark、hdfs、hbase 等多种大数据相关组件；
- **消息队列集群：**采用 Kafka+zookeeper 实现消息队列；
- **接口服务：**由 flume 组件和定制化接口服务共同组成。

4 产品集成

风险引擎产品主要是采用数据集成的方式，主要分为主动采集和被动采集；

主动采集由风险引擎采用 Flume 组件去远程服务器，通过读取日志的方式完成，被动采集由第三方产品调用风险引擎的基于 Restful 的服务接口；具体接口内容请参考《风险引擎接口文档》。