# 天融信数据防泄漏系统

# 技术白皮书

北京市海淀区西北旺东路 10 号院西区 11 号楼 1 层 101 天融信科技集团 100193

电话：010-82776666

传真：010-82776677

服务热线：4007770777

http://www.topsec.com.cn

## 版权声明



## 商标声明



## 信息反馈

http://www.topsec.com.cn

# 变更记录

| 版本 | 修订日期 | 修订人 | 修订类型 | 修订章节 | 修订内容 |
|---|---|---|---|---|---|
| V23.2.1 | 2023/05/10 | 张志颖 | A | 整体内容修订 | 整体内容修订 |
| V23.2.2 | 2023/05/11 | 张志颖 | M | 3.3.5 | 增加聚类技术说明 |
| V23.2.3 | 2024/2/20 | 张志颖 | M | 首页 | 修改地址信息 |

*修订类型分为 A- ADDED M- MODIFIED D -DELETED

注：对该文件内容增加、删除或修改均需填写此记录，详细记载变更信息，以保证其可追溯性

# 目录

# 1 产品概述

## 1.1 背景介绍

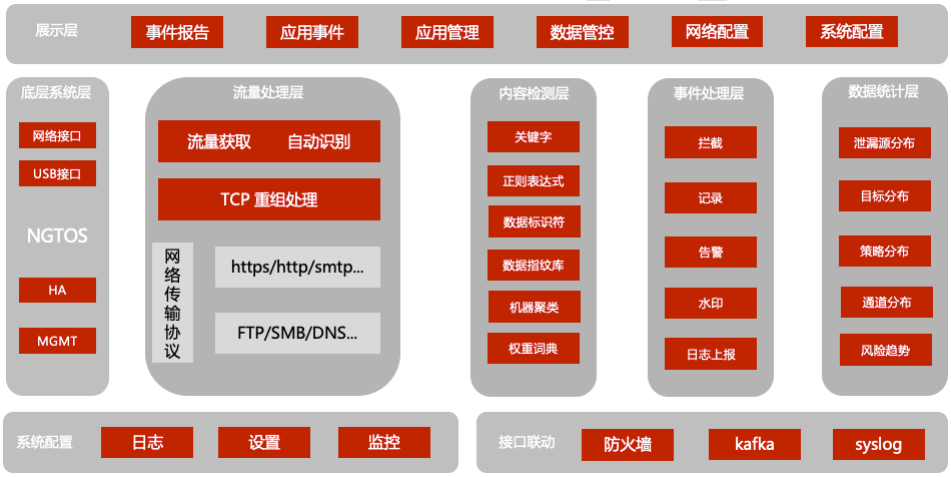
随着信息科技不断发展，各种信息化技术和系统的广泛应用，信息量成几何级增长。现阶段信息安全的重心，不再局限于系统本身的安全，而应更多关注信息自身的安全。与个人用户相比，企业中存在大量的内部信息，信息泄漏不仅会给企业带来严重的直接经济损失，而且对自身品牌价值以及社会公众形象等多方面造成损害。

目前，造成企业信息泄漏的原因主要有两类：黑客入侵窃密与内部主动泄密，其中企业内部员工有意或无意的泄密更为常见。因此，企业应该在完善信息安全管理制度的同时，融入信息泄漏防护技术，更有效地提高其内部体系的安全性。为了避免内部资料、用户信息等高机密信息外泄，建立一套完善的信息泄漏防护系统已迫在眉睫。据国家计算机信息安全测评中心数据显示，互联网接入单位重大损失的事件中,只有 1%是被攻击者窃取造成的,而 99%都是由于内部员工有意或无意的泄密行为所导致。由于泄密行为的隐蔽性,我们无法预知数据泄漏何时发生,甚至数据泄漏正在发生时无法感知、无法管控、无法溯源。针对数据的泄漏防护工作任重而道远。

从政策来看三大上位法《网络安全法》、《数据安全法》、《个人信息保护法》明确指出相关单位、人员需要履行的重要数据保护义务以及采取相关的技术措施来防止数据被泄漏、被窃取；《关键信息基础设施安全保护条例》、《网络安全等级保护》等辅助法规，更是细化数据安全领域网络防泄漏等安全防护设备的相关防护能力、防护方案。在不同行业内部，如电信、能源、金融、医疗等更是衍生出大量的贴合自身行业场景、特征的相关技术规范、要求，如《电信网和互联网数据安全要求》、《电力行业网络安全等级保护管理办法》、《金融行业网络安全等级保护 2.0》等文件中更是不缺数据泄漏防护技术的关注和相关安全能力的要求。

天融信昆仑系列数据防泄漏系统是基于国产化处理器和国产化操作系统自主设计开发的网络安全产品，支持纯透明代理、正反向代理、路由、探针等多种部署模式。系统基于深度内容识别技术，通过内置关键字、指纹库、数据标识符、权重词典、机器聚类等多种敏感数据定义手段，对抓取到的传输数据进行过滤，发现并监控敏感数据，确保敏感数据的合规使用。同时还适用于 IPV4、IPV6、云等多种场景，对敏感数据实施事中、事后的全面保护与审计，保障数据全生命周期的安全流转。

## 1.2 系统架构




天融信网络数据防泄漏（网络 DLP）系统由展示层、系统层、流量处理层、内容检测层、事件处理层、数据统计层、系统配置、接口联动这几部分组成。

1、展示层：系统的可视化操作界面，可对事件报告、应用事件、应用管理、策略配置、网络配置、系统配置进行界面配置，简单易用；

2、底层系统层：基于天融信自研系统 NGTOS 平台研发，系统稳定安全可靠；

3、流量处理层：基于端口流量进行检测，采集 HTTP/HTTPS/SMTP…等协议的流量内容；

4、内容检测层：基于关键字、正则表达式、数据标识符、指纹库等内容检测算法对流量中传输数据进行识别、过滤；

5、事件处理层：对违规、异常行为进行拦截、阻断、告警等多种处理，全方位保证用户敏感数据的安全；

6、数据统计层：对安全事件按照泄漏源分布、目标分布、策略分布、风险趋势分布等；

7、系统配置：进行网络配置、监控、日志设置等基础配置；

8、接口联动：提供与其他平台进行联动、对接等能力的相关配置，提高解决方案灵活性。

# 2 产品特点

## 2.1满足合规场景

网络 DLP 支持多种部署模式，包括透明代理、正向代理、反向代理、旁路镜像、MTA 邮件代理模式、防火墙联动对接、探针模式等，覆盖多维度应用场景，满足新业务场景如传统 IT 流转场景、工业互联网场景、云数据防护场景等数据防泄漏要求，满足《网络安全等级保护》相关防护要求，适用于不同数据泄漏防护技术应用场景，如等保三级数据防泄漏安全防护要求等。

## 2.2 协议识别覆盖广

凭借多年的协议解析技术积累，网络 DLP 具备多种网络传输协议、数据库协议以及 restful 接口的解析能力，支持网页类（HTTPS、HTTP）、邮箱类（SMTP、POP3、IMAP）、共享类（FTP、SMB）、运维类（TELNET、DNS）、数据库类（MYSQL、ORACLE、HBASE、HIVE、HDFS ）等。

## 2.3 检测算法精度高

网络 DLP 内置天融信独有的深度内容识别引擎，通过关键字、正则表达式、数据标识符、IDM、EDM 多种检测算法，提供基于关键字、文件属性、用户等不同维度的内容检测，支

持多种语言、多种编码、多种位置（页眉/页脚/正文/主题/信封），检测数据内容精准度高达 99%以上，满足不同场景下应用需求。

## 2.4 异常行为无遗漏

针对异常的办公场景、办公行为、办公文件网络 DLP 内置了多种检测模型进行识别发现，包括文件加密、邮件密送、修改文件后缀、word 嵌套、文件多层嵌套、多层压缩、缓慢泄漏、分片压缩等，全量审计员工恶意的藏匿数据行为。

## 2.5 违规响应选择多

违规响应处理，支持通过 Syslog、Kafka 等方式向第三方平台发送日志及告警信息。基于不同行业、不同场景的防护力度，网络 DLP 可提供多种违规响应措施，如拦截、记录、审计、钉钉告警、微信告警、邮件告警、短信告警、文档水印、syslog 上报、kafka 上报、违规事件留存、注释等，满足客户强管控、弱管控需求。

## 2.6 开箱上线易操作

网路 DLP 基于合规政策、技术规范、能力要求，挖掘不同行业数据特征，形成策略知识库，开机即用，节省配置周期。

策略知识库包括金融行业、医疗行业、通用行业、政府行业以及满足个人金融信息保护规范、个人信息安全规范、GDPR 等策略模版；

系统旁路部署即插即用，只需引入镜像流量，即可产生对传输数据进行检测，生成敏感数据泄漏事件，节省运维人员、业务人员时间成本。

## 2.7 产品资质认证全

网络 DLP 通过各类合规检测，具备软著、 销许、IPv6 Ready Logo 认证、 IT 产品信息安全认证证书 CCRC、 IT 产品信息安全认证证书 EAL3+（独家）、军证（独家）等多种产品认证资质，是目前市场中资质最全的数据防泄漏厂商。

# 3 产品功能

## 3.1网络协议解析

捕获网络传输流量，提取传输协议报文，解析具体内容，包括文件大小、文件类型、发送者、接收者、主题、正文、附件等。

支持 HTTP 协议、HTTPS 协议、FTP 协议、SMTP 协议、SMTPS 协议、POP3 协议、POP3S 协议、IMAP 协议、文件共享类协议、Restful 接口、HDFS 以及 Mysql、Oracle、Hbase、Hive 数据库。

## 3.2文件属性识别

网络 DLP 系统支持对文件类型、文件大小、文件名和修改文件后缀、文件作者、文件创建时间、文件修改时间进行识别；为提供文件类型识别的准确性，采用识别文件头特征进行判断；针对文件大小的检测，基于数据量大小对内容或文件进行识别；文件名检测支持精确和模糊检测。

支持的 Microsoft、Apple、Adobe、HP、IBM、金山等多种常见文件、自有文件、自定义文件类型的内容提取（包含单个应用不同版本产生的文件）。

## 3.3内容发现检测

### 3.3.1 关键字检测

系统根据预先制定的关键字列表，通过检测数据流中是否含有列表中的关键字来判断数据是否敏感。支持关键字模糊检测、精确检测、部分关键字检测、多关键字、临近关键字匹配、权重关键字等多种检测方式，支持“*”和 “？”通配符、多种语言（英文、中文、维吾尔、蒙古、藏 、日、韩文）、多种编码（UTF8、UTF16、LATIN-1）、忽略大小写。

### 3.3.2 正则表达式检测

正则表达式是一种基于有穷自动机实现的字符串匹配技术，能够快速检测识别结构清晰的文本数据。

系统支持 PCRE 等常见语法规则的正则表达式检测，如电话号码、地址等；支持基于正则表达式识别少量多次泄漏敏感信息行为。

### 3.3.3 数据标识符检测

数据标识符是比正则表达式更精准的检查方式，采取二层识别方式，首先识别符合字符串规则的内容，其次对数据的有效性通过算法进行验证，判断数据的真伪，如身份证、中国人、组织机构代码、银行卡号、月结卡号、海关单号；系统提供了相应的接口，用户可以基于实际情况自行编辑数据标识符校验器，自定义脚本匹配。

### 3.3.4 数据指纹库（结构化、非结构化、文档格式、图片）

指纹识别首先上传一组已知敏感文件的集合。系统自动捕获传输中的文件与与集合中的文件进行比较，如果相似度较高，则认定其为敏感文件。指纹识别主要通过 hash 件和被检测文件计算 hash 值 ，在根据 hash 值进行逐比特的比较，计算相似度。当相似度达到一定阈值时则认定被检测文件为敏感的。

系统支持对文档（所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像）学习生成敏感数据指纹，通过外发文档与受保护的文档指纹特征进行比对，精确识别敏感内容。

系统支持对数据库(Oracle、MySQL)表字段学习生成敏感数据指纹，通过外发数据库表字段与受保护的结构化指纹特征进行比对，精确识别敏感内容。

系统支持对内容格式相似的文档类型进行识别，通过定义固定内容格式的文档类型，发现传输文档中相同的格式文件进行告警。

### 3.3.5 机器聚类

自动聚类是一种典型的无指导机器学习方法：使用特定算法将不同文档分别映射成特征向量空间中不同的点，然后根据这些点的聚集程度，将对应文档聚集成某些特定类别。在一个特征空间中，同一类文本对应点的集合，往往集聚在一个空间区域中，机器即可通过计算点与点的相似度，将属于同一类的文档寻找出来。

支持通过机器聚类技术，自动获取网络中传输文件或者手动上传无序文件样本，按照文件内容特征自动进行聚类，实现文档分类，且根据分析结果生成推荐性权重关键字，依据权重词典创建检测规则，辅助更有效得 DLP 策略制定。

### 3.3.6 权重词典

支持依据多个关键词的重要程度判断整体内容或者文件的重要程度，通过手动定义多个敏感关键词，赋予不同的关键词不同的权重比，防泄漏依据不同的关键词的重要程度识别、判断传输内容的重要性，判别是否涉敏。

## 3.4异常行为判断

系统支持常见的主动泄密行为，如将敏感文档压缩后加密码、使用特殊程序将文档变换成不可识别的格式等；系统支持多类常见加密文件类型，如：PST (Windows)、PST (non-Windows)、ZIP、7-Zip 等；支持常见压缩文件类型的解压和多层压缩文件解压，包括 7-Zip、BinHex、Microsoft Backup、Bzip2、Cab、Linux 等；同时支持多层嵌套压缩；支持检测缓慢泄漏行为，通过统计某时间段内，风险内容传输的频次，发现点滴时泄漏行为，进行防护；支持对分片压缩中包括的风险内容进行发现和检测；

## 3.5API 接口识别

系统支持业务系统的用户越权行为进行审计，通过备案应用系统的访问账户、账户角色、API 接口信息三者关联关系，网络 DLP 对 API 传输接口数据进行发现，匹配账户、角色、接口信息，对违规访问行为进行监控；针对 API 的接口参数、访问量、访问用户、访问

报文、访问时间等字段可通过 kafka 接口进行外发，为第三方平台侧产品提供数据分析支撑。

## 3.6违规风险处置

系统支持多种响应策略，可进行组合配置：添加事件注释；保留事件待检数据；记录日志到 Syslog 服务器；发送电子邮件通知；设置事件属性；设置事件状态；阻断 HTTP 请求 ；阻断 FTP 请求；阻断 SMTP 消息；修改 SMTP 消息；邮件审核；水印；kafka 上报等。

## 3.7策略灵活定义

网络 DLP 系统提供大量内置模板与规则，包括个人基础信息类、行业模版类、财务数据、机密文档、技术方案、简历、电话号码、各类源代码检测标识等。用户可直接使用，无需再次收集、指定。同时可自定义策略模板内容快速构建满足业务和和法规需求的策略；提供策略批量导入导出功能。可以直接导入策略，支持自定义生成策略模版，可以进行策略“与”、“或”、“非”的逻辑运算组合；灵活配置策略，提升检测精度，降低误报率。

## 3.8违规日志可视

提供多维度风险展示、风险事件报告，包括按天、周、月、季度进行违规数据统计展示，泄漏源与目的 top 排行榜、策略分布图、数据外发通道分布、事件风险等级分布、按等级汇总等风险视图展示，违规事件日志颗粒度包括协议类型、检测时间、发送方、接收方、风险与处理、策略、用户名、部门、端口号、原始信息、概要信息等。

## 3.9分权分域管理

网络 DLP 可自定义分域管理账户，提供分域管理风险内容检测和风险事件处置能力，按照 IP 段对安全域、安全区、部门、业务等分管场景进行管理权限的划分，形成不同的 IP 域，针对 IP 域分配自定义管理账户，自定义管理账户仅支持对权限内的 IP 域进行相关检测策略的条件配置以及对产生的事件进行查看、处理、修改、备注、删除等操作。

# 4 部署模式

天融信网络数据防泄漏设备支持级联、串联、旁路方式接入网络，支持网桥模式，路由模式、正/反向代理模式、旁路模式、ICAP 接口检测模式、防火墙联动模式、探针模式。

支持双机部署：提供解决单点故障的双机功能，一旦设备出现故障，及时进行切换，避免业务中断。提供硬件 Bybass 与软件 Bybass，可手动指定在内存、CPU 使用率过高或者硬件故障、电源故障时，自动切换至直通状态，不再进行策略防护，避免成为单点故障，造成业务中断。

## 4.1 串联方式

产品串接部署方式。该方式为将产品串接在网络中，无需部署 Web Proxy、MTA 等服务器。数据流经系统时，对数据包进行识别、转发或阻断。该方式既能发现敏感数据并能主动阻截非法数据。

可配置网桥模式、路由模式。网桥模式不需要修改现有网络架构，采用透明代理技术对数据流进行处理；路由模式需要更改现有网络网关地址与架构，由网络 DLP 提供三层数据转发能力。

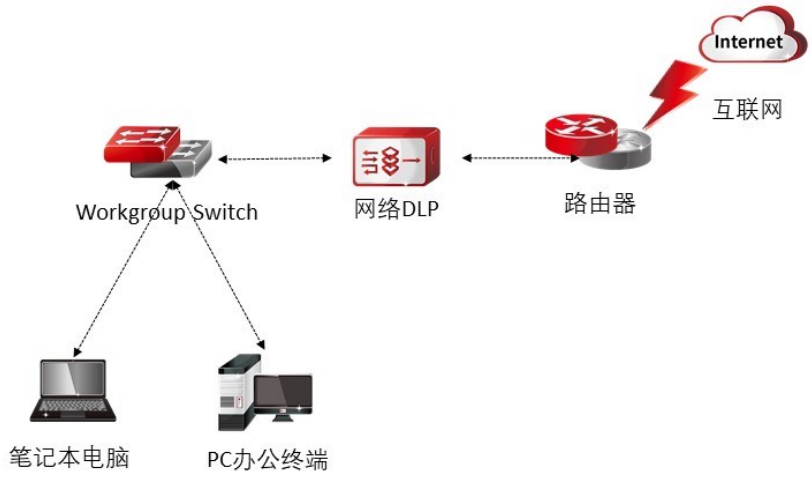


图 1：网络 DLP 串接部署示意图

## 4.2 旁路方式

产品旁路部署方式。该方式系统旁路部署于所在网络任意节点处，由交换机、防火墙等设备把数据流量镜像或分流到产品上进行审查。该方式目的在于监控与审计，不对数据进行阻断。 这种组网方式对于原有网络的变动最小，网络切换简单。

可配置正/反向代理模式、旁路模式、ICAP 接口检测模式、防火墙联动模式、探针模式、MTA 邮箱代理模式。

正/反向代理模式针对需要单独强管控的 PC、区域进行代理配置，传输数据经网络 DLP 代理后进行转发，实现针对性数据、业务、区域监管；

旁路模式，直接引入镜像流量即可对网络传输数据进行识别；

ICAP 接口检测模式，可与支持 ICAP 接口的第三方产品进行对接，第三方产品提供数据抓取、数据转发功能，网络 DLP 只提供内容检测能力与检测结果反馈；

防火墙联动模式，可与同品牌防火墙联动部署，由防火墙将传输数据捕获后转发给 DLP 设备，DLP 设备进行检测，并返回违规五元组信息，生成防火墙访问控制策略；

探针模式，可对接第三方平台侧产品，接收平台侧下发数据检测策略，全量审计现网流动数据，并上报数据流转日志；

MTA 模式，DLP 系统旁路接入邮箱服务器网络，由邮箱代理服务器配置 MTA，邮件数据转发至 DLP 设备进行检测，DLP 对邮件内容进行合规性判别后，将邮件转发至目标处或邮箱服务器，此方式针对邮件检测场景，无需改动邮箱现有网络架构，只需修改邮箱服务器配置。

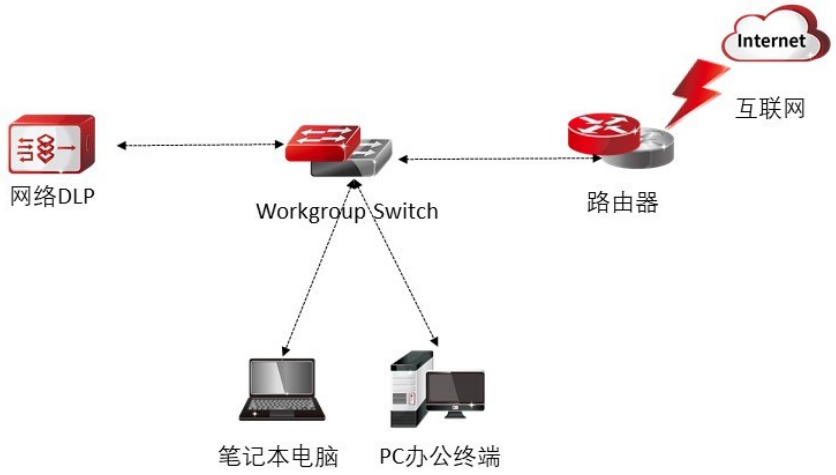



图 2：网络 DLP 旁路部署示意图

## 4.3 级联方式

产品级联部署方式。该方式应用于在多个区域部署多台网络 DLP 系统（旁路、串接）时，提供防泄漏系统集中管理系统，可同时对所有的网络 DLP 系统进行纳管，实现策略统一配置与下发、事件统一收集与分析、设备状态监控等多种能力，方便运帷侧、业务侧、监管侧对多台 DLP 系统进行管理。

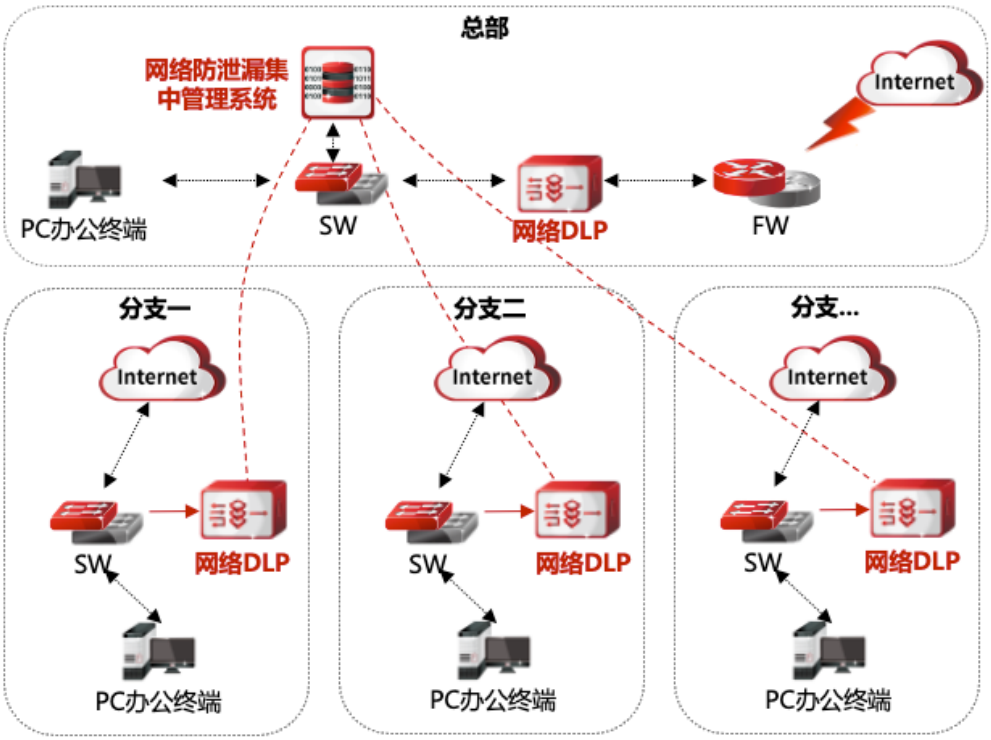



图 3：网络 DLP 级联部署示意图

# 5 产品规格

网络 DLP 为软硬一体件产品，提供 B/S 架构的系统管理平台供用户对系统进行全方位配置与管理。

| 型号 | 天融信数据防泄漏系统 V3 | |
|---|---|---|
| 说明 | 平台，级联部署 | 探针，单节点部署 |
| CPU | 鲲鹏 920 7260（2.6GHz、64 核）*2 颗，共 128 核 | |
| 操作系统 | 银河麒麟 | |
| 内存 | 256GB | |
| 系统盘 | 480G SSD 硬盘 | |
| 数据盘 | 6TB SAS 硬盘，支持空间扩展 | |

| | |
|---|---|
| 网络接口 | 4 个 10/100/1000M 以太网电口、4 个 10Gb 光口（带多模光模块） |
| 吞吐 | 3Gbps |
| USB 接口 | 2 个 |
| 产品形态 | 硬件，2U 标准机架式设备 |
| 冗余电源 | 双电源 |
| 电压 | 100-240V AC |
| 频率 | 50~60HZ |
| 电流 | 7A MAX |
| 功率 | 800W |
| 运行温度 | 0~40℃ |
| 存储温度 | -40℃~55℃ |
| 相对湿度 | 10%~90%RH 非凝结 |

# 6 产品资质

| 证书名称 | 认证机构 |
|---|---|
| 计算机信息系统安全专用产品销售许可证 | 中华人民共和国公安部 |
| 计算机软件著作权登记证书 | 中华人民共和国国家版权局 |
| ISCCC | 中国网络安全审查技术与认证中心 |
| IPV6 Enabled Phase-2 logo 认证 | IPV6 Ready Logo 委员会 |

| | |
|---|---|
| 军用信息安全产品认证 | 中国人民解放军信息安全测评认证中心 |
| EAL 产品信息安全认证证书 | 中国网络安全审查技术与认证中心 |

# 7 关键技术

## 7.1 数据格式识别技术

对已知的数据类型能够准确识别，同时可以方便友好的提供扩展，用以识别未知的数据格式；数据内容抽取标注引擎和工具，完成对文档类数据内容的抽取。采用机器学习及神经网络完成对光学字符的有效识别，采用神经网络完成对图片内容的有效标注。

## 7.2 数据指纹的生成技术

相似数据生成的指纹具有相似性，相似指纹则表示指纹标识的数据具有一定的相似度。伴随着数据的不断演进，其指纹也会发生变化。依据不同的数据，采用不同的特征抽取方法，抽取数据的关键特征，采用基于局部敏感哈希技术和感知哈希技术来生成数据的指纹。对文本类数据，使用滑动窗口，抽取最小值的方法，抽取出文本的特征；对于图片类数据则使用小波变换来抽取图片特征。采用海明距离、欧式距离、jaccard 距离和编辑距离来度量指纹的相似度。

## 7.3 数据防泄漏机器学习

机器学习可以对大量的无特定格式文件样本进行快速学习和分类，分类产生的模型（Model）可用来对数据进行分析并计算该数据是否属于某一个分类。机器学习的优势在于

其生成的模的大小基本恒定，所以很适合处理大量的样本，另外机器学习技术可以对新的、并未出现在样本中的数据进行较为准确的预测。

## 7.4 数据防泄漏防护技术

依据 DLP 策略来完成数据的合规性检查，并且能够执行相应的响应动作；结合安全元数据和数据安全策略提供数据的安全建议；能够依据给定的数据，提供出数据的指纹、相似性数据等信息。

# 声明

1．本文档所提到的产品功能规格及资讯仅供参考，有关内容可能会随时更新，天融信不另行通知。

2．本文档中提到的产品功能或性能可能因产品具体型号、配备环境、配置方法不同而有所差异，此种情况产生的差异为正常现象，产品功能或性能请以产品用户手册等资料为准。

3．本文档中提到的信息为正常公开的信息，若因本文档或其所提到的任何信息造成或可能造成他人直接或间接的资料流失、利益损失，天融信及其员工不承担任何责任。

RELIABLE NETWORK
SAFE WORLD
TOPSEC