Elemental数据科学平台
产品白皮书

孔明科技

2021年3月

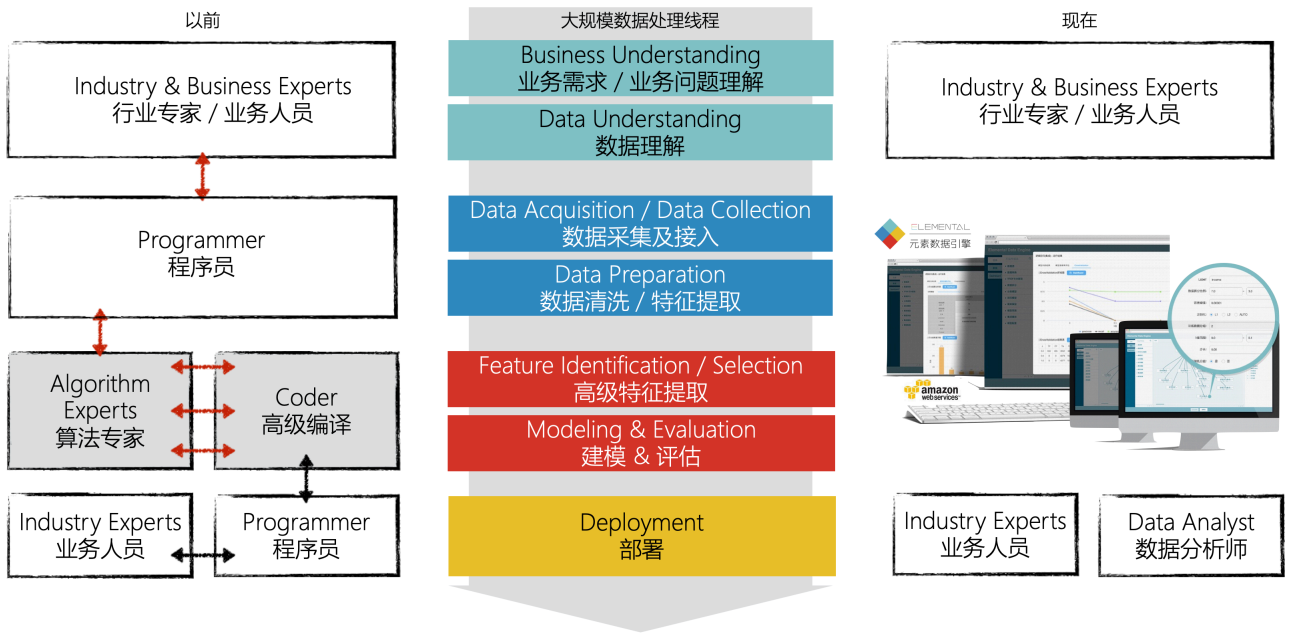
产品概述

随着数据技术发展，各行业信息化水平和数据积累程度越来越高，数据治理、数据挖掘、建模分析、人工智能算法等技术的使用场景越来越广泛，企业对数据资产的重视程度随之大幅提升，对存量数据的应用落地和价值挖掘的诉求愈发强烈。

与此同时，大数据技术在企业的应用过程中暴露了以下问题，严重制约着企业开展数据资产价值挖掘、扩展业务和优化成本的进程：

- 数据来源混乱，采集困难，数据质量参差不齐，难以对数据资产进行有效盘点和统一集中管理；
- 传统数据挖掘技术和机器学习应用的门槛很高，需要数据科学家、大数据工程师非常熟悉企业业务和数据结构，掌握机器学习算法，通过代码编程等方式对算法进行调用和参数调优建立模型；
- 数据价值开发过程与实际业务耦合程度较高，定制化产品灵活性不足，系统扩展和迁移成本高；

孔明科技的通用型数据科学平台Elemental系统，专为解决全行业的各类数据分析问题而设计研发，是进行数据采集、清洗、挖掘和建模分析的工作流程可视化的系统级平台。



通过数据实时采集、集中接入、统一管理，基于底层的分布式存储和索引，Elemental系统最大程度的保证了企业数据的可用性、优质性、一致性、使用便捷性和安全性，具备优秀的性能基础，为数据治理、分析查询、运算建模提供有力支撑。

Elemental系统包括：基于Akka开发的轻量级实时数据采集输出工具、基于分布式运算框架的数据挖掘建模分析平台、以及轻量级数据可视化套件。其良好的图形化交互和特有的“拖拽式”工作流编辑模式，使数据分析人员脱离编程和调试过程，实现数据的在线计算和离线计算，可快速复用流程逻辑，并帮助用户快速实现可视化图表展示和将数据结果集成到业务系统，促成数据应用。

Elemental系统提供以下价值：

- 多源异构数据实时无侵入采集与输出，无需代码开发；
- TB级数据集中治理，多层安全管控，发挥元数据价值，实现数据资产梳理；
- 数据分析建模等过程无需编程，降低研发与运维成本；
- 提供大量ML算法模块与参数自动框架，提供AI应用能力；
- 业务模型可快速复制调整，避免重复工作与二次开发；
- 支持边缘设备的轻量数据管道与流式计算，与云端中央计算平台实现云边协作；
- 快速实现数据展现和业务可视化图表，快速理解数据，提升业务质量；
- 跨部门多人协作，完善的用户鉴权，提升数据人员与业务人员工作效率；
- 强大的Elemental咨询团队，提供行业解决方案与专业服务；

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

产品简介

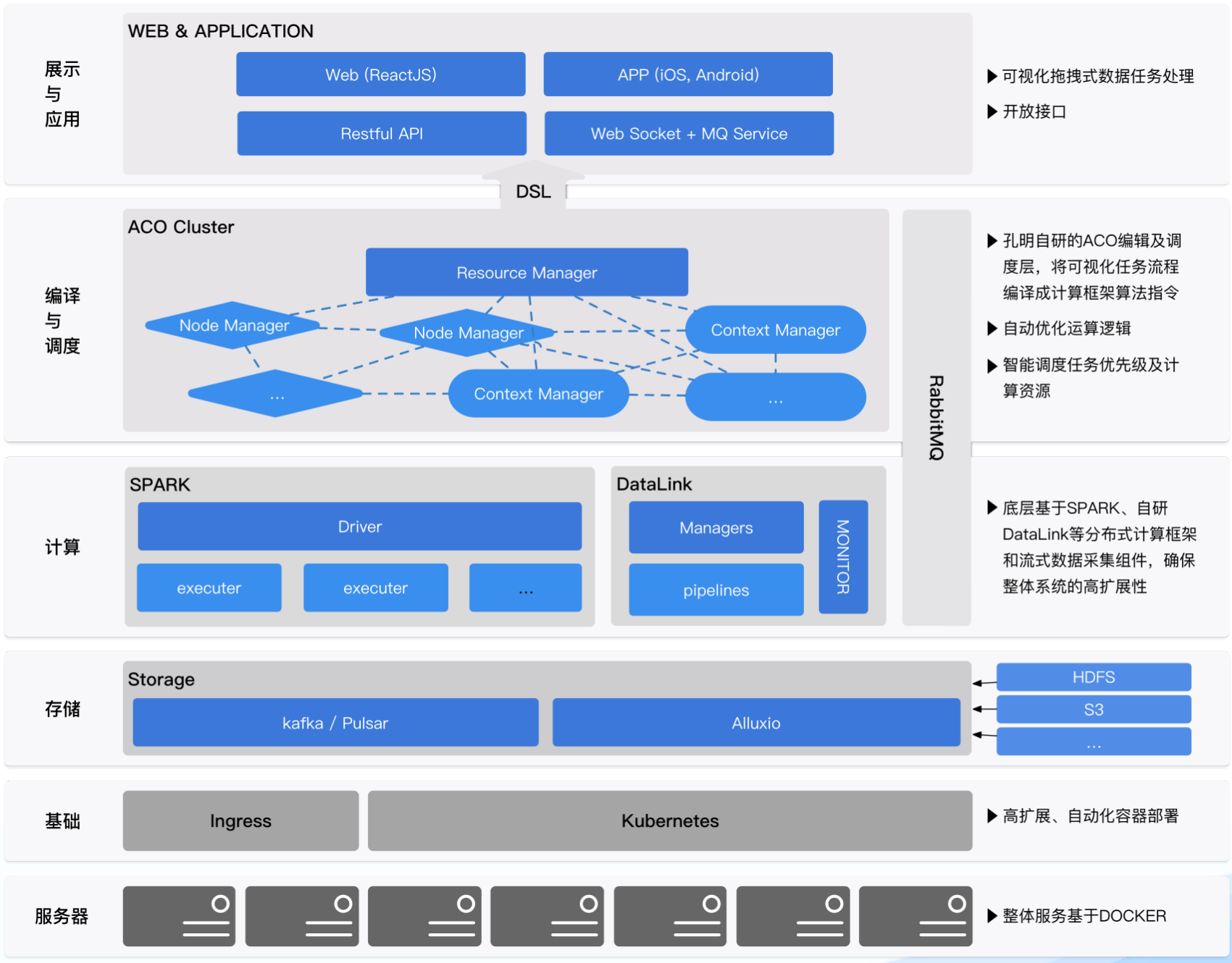
Elemental系统提供一站式大数据解决方案，基于分布式存储和分布式计算框架自主研发。系统集成数据采集、数据治理、分析挖掘及机器学习算法，包括可视化大数据挖掘工具及配套组件，提供高效的数据洞察和业务应用能力。

系统所有功能均提供图形交互界面，使用流程图来组织数据处理过程，使企业客户无需研发人员编写代码即可快速开展数据工作，低成本高效率地解决业务需求到数据模型之间的逻辑转化。系统功能从数据接入、清洗、特征工程、建模、评估、预测，到可视化图表、集成应用，全面覆盖数据业务需求。由于进行了高度可复用的设计，企业可节省大量预算和时间成本。

架构设计

Elemental系统采用微服务架构，具有全异步、高可用、弹性伸缩的技术栈，集成了批量、流式、深度学习等不同风格的分布式计算引擎。服务架构如下：

- 支持HDFS、S3、OSS等分布式存储，使用Alluxio分布式缓存，确保整体系统高扩展性；
- 轻量级数据管道，提供异构数据一对多实时分发与流式计算，可实现边缘计算设备部署；
- 基于SPARK、MXNET、自研流式计算框架等分布式计算底层，大大提高计算效率；
- 基于Docker的底层系统架构，便于轻松部署和扩展；
- 自主研发的ACO调度层，将任务流程动作编译成计算框架语言，确保不同任务在处理过程中使用最优计算逻辑及资源；



info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

服务端技术优势

Elemental系统针对不同的数据来源和不同的计算引擎分别开发了不同的驱动，将计算引擎和计算资源的管理作为服务调用，无缝融合到产品整体架构中，具有多层次的分隔和抽象。用户通过页面交互自定义逻辑计算过程，后台引擎进行流程优化和代码自动生成，并将任务分解为不同规模的Job，智能分发到集群所托管的计算资源节点上，选取合适的资源（CPU、Mem）进行运算，最终将执行结果实时反馈推送给用户。

- 使用Scala语言作为主要开发语言，相对于Java语言表达能力优秀，其强类型检查和函数式的特性，使得服务稳定性更好；
- 自研流程管理引擎，支持本地算子并无缝接入多种开源计算引擎，自动优化运算流程，通过全异步非阻塞的方式实现Reactive风格的复杂拓扑运算管理；
- 自研轻量级数据管道，支持多源异构数据高并发采集、一对多实时数据分发；
- 自研集群资源管理框架，动态调整计算资源并进行运行环境的分发；根据算法和数据特征匹配最合适的运算节点，并支持手动调整等高级功能；
- 基于Websocket的通知服务为用户提供实时消息和运行结果反馈，支持不同客户端之间的协同通信；
- 支持多种底层数据存储（HDFS/S3/OSS），采用Alluxio提高数据读写效率；
- 支持多种底层分布式计算引擎如Spark、自研流式计算框架，在开源产品的基础上进行了大量算法和架构的优化；
- 基于MXNet、TensorFlow等深度学习框架，实现多种算法，借助GPU提供强大的建模能力；
- 基于ServiceMesh的微服务架构，结合Docker和Kubernetes实现服务的自动发现、注册、部署；

用户端技术优势

前端采用最新的MVVM架构，数据、界面、交互行为相分离，React+Redux配合WebSocket最大程度地保证了数据更新在界面上的实时体现，贯穿始终的数据可视化技术让复杂的数据处理流程得以直观体现，无需一行代码、鼠标拖拽即可进行灵活地修改，极大降低了大数据处理操作的难度。

- 基于Facebook开源的React，深度定制Ant.design；
- 借助Echarts、D3.js使系统运算过程和运算结果实时可视化；
- 前后端彻底分离的架构模式，使用下一代API查询语言GraphQL让前后端的配合更灵活；
- 基于NodeJS+MongoDB提供轻量级数据应用；
- 基于ES6规范开发，使用Webpack+Babel增强用户端兼容性和执行效率；

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

产品功能

用户和团队

团队/租户可对内部成员进行角色定义。用户登录Elemental系统后，可根据角色所设权限访问系统中不同产品功能。

系统支持细颗粒度的权限项配置，支持到单个数据的超细颗粒度数据读写权限控制。

以团队/租户形式进行管理功能协同，用户操作和系统消息会实时通知给团队成员。

数据接入及输出

Elemental系统数据采集组件支持多源异构数据的一键接入输出，支持实时数据融合，提供灵活的更新策略。数据采集状态实时可查看，接入系统的数据保存在团队的数据中心内。对接数据源包括以下类型：

- 主流数据库：MySQL、PostgreSQL、Oracle、SQL Server、DB2、MongoDB等；
- 实时及流式数据：Kafka、Pulsar、OPC等；
- 分布式文件系统及数据库：HBASE、HIVE、HDFS等；
- 文本文件：csv、txt等；
- API数据：微博API、微信公众号API、小程序API、用户自定义API；
- 埋点数据：网页埋点、小程序埋点；
- 定制化对接；

数据治理

Elemental系统提供完善的数据管理功能，可帮助企业进行数据资产梳理，提高数据可用性，管理数据生命周期，为数据中台或数据应用提供基础服务。

- 提供元数据管理功能，系统内自动生成数据血缘与传播关系，提升数据价值；
- 所有数据在数据中心集中管理，可创建业务包分配至不同项目；
- 数据保存多个快照版本，增强安全与可用性；
- 数据结果可进行图表展示、导出、下载、通过结果访问；
- 原始文件加密存储，提供冗余备份支持，服务间的通信加密处理，保证传输安全；
- 分布式存储保证了数据安全性和读写效率；
- 任务分布在不同容器中运行，通过容器隔离不同用户运行环境；
- 严格的权限控制，使得用户只能访问授权的数据和任务；

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

产品功能

计算模块

计算模块是将流式计算、分析预处理、数据挖掘、机器学习等领域中常见的算法或计算流程抽象出来，通过用户参数定义来实现特定数据运算的功能单元。计算模块就像一块块积木，互相衔接搭配，可以建立各种数据处理流程。以机器学习过程为例，构建模型需要经过数据清洗、特征工程、建模、评估、实施等几个步骤，每个步骤都可以在Elemental系统中找到对应的计算模块：

模块分类	模块名称
数据认知	统计分布、相关性分析、交叉分布
假设检验	T-test、F-test、卡方检验、KS正态检验
数据操作	文件版本聚合、Save动态文件、Update动态文件、生成数据组、拆解数据组
变量操作	变量赋值、变量传递
数据预处理	缺失值处理、数据类型转换、列名修改、行列交叉统计、地址转换、行筛选、数据值替换、列增加、数据值去重、TopN、列合并、数据关联匹配、行追加、无效列指定、列运算、列拆分、列筛选、数据排序、字符处理、数据抽样、数据离散化、列加密、列转行、行转列、大小写转换、压缩矩阵解析、关系聚合
特征提取	数据离散化、归一化、标准化、应用标准化规则、SVD、PCA、降维映射、IG特征选择、卡方特征选择、t-SNE、样本相似度
文本分析	中文分词、词频统计、TF-IDF、Word2Vec、LDA
时间序列	ARIMA、EWMA、GARCH、LSTM-多变量时间序列
二元分类	逻辑回归、线性SVM、神经网络、决策树、随机森林、GBDT、XGBOOST、LightGBM、朴素贝叶斯、KNN
多元分类	多元逻辑回归、多元决策树、多元随机森林、多元朴素贝叶斯、1vsRest
回归分析	线性回归、决策树回归、随机森林回归、GBDT回归
聚类模型	Kmeans、二分Kmeans
推荐模型	Model-based CF、Item-based CF
关联规则	FP-Growth
模型预测	模型预测
自动调参	交叉验证
集成模块	批量运算、用户自定义模块

除上述系统提供的独立模块外，用户还可根据使用场景创建运算流程并将一组模块打包保存为自定义模块，从而实现数据运算逻辑的复用。

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

产品功能

流程图

Elemental系统提供可视化流程式数据分析和挖掘工具，由流程图的交互方式实现。

每个流程图包含一个或多个数据，以及对这些数据进行处理的计算模块。在Elemental系统中，编写代码的过程被拖拉拽的交互方式代替。模块的参数定义了详细运算规则；模块之间的连线定义了计算流程的逻辑顺序。

流程图中的计算模块包括待连线、待配置、配置完成、配置错误、运行中、运行成功、运行失败等不同状态。拖拽式的流程图有以下优点：

- 数据处理与挖掘方法模块化，处理流程可视化，大大缩短清洗和组织数据的时间；
- 数据人员可轻松进行逻辑追溯校正和流程复用，随时查看中间数据结果；
- 集成大量主流机器学习算法，提供自动调参，模型评估与预测结果可实时展示；
- 可自定义集成模块，直接拖拽于类似场景应用，企业可产出自主IP并共建行业数据挖掘知识产权；

自动运行

Elemental系统支持数据的自动化处理，用户对流程图页配置定时规则后，在指定的激活时间段内，如果文件更新引起数据状态发生变化，则触发系统使用最新的数据版本重新运行流程图，从而实现对数据的自动更新运算。

对于实时数据或在线服务，通过将测试过的流程图上线部署，可实时将运算结果进行输出。

资源管理

用户使用私有化部署或付费服务时，可对系统的计算资源进行管理和监控，包括：

- 计算资源管理：创建、重启、删除计算节点，查看任务运行队列等；
- 存储用量和流量管理；
- 管理用户的模块使用权限；

边缘计算

Elemental系统的云端计算平台以OLAP为主，为超大数据量的离线分析建模所设计研发；而数据采集与分发组件则采用了自研的高效轻量级数据管道，支持实时流式运算，对资源占用进行深度优化，可部署运行在物联网网关等边缘计算设备上。通过云端计算组件与边缘数据管道的结合，可实现：

- 边缘数据实时采集、分析、上传至中央云端计算平台；
- 云端计算平台的数据处理分析流程、ML模型可远程下发至边缘端，实现一键部署更新；
- 边缘计算设备统一管理，资源状态实时查看与控制；

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

产品功能

商业智能

用户可在系统内将流程图运算结果在展示为图表，并将图表组织为页面后一键分享。图表数据支持自动更新，流程图的运行结果会随时更新到图表中。系统支持图标的主题定义，并提供大量配置项帮助用户完成数据到图表的合理转化。

目前支持的图表类型包括：

表格、列表、柱状图、折线图、面积图、饼图、环形图、双轴图、箱线图、漏斗图、词云、中国地图、热力图、雷达图、散点图、气泡图、关系图等；

业务集成

Elemental系统提供业务集成套件，可将数据分析结果、模型预测服务集成至用户的业务系统中。其中数据的使用以开放API、数据库写入、实时消息或对接FTP、文件目录等形式作为输出，模型的在线使用通过API方式提供服务。

- 数据分析结果以多种图表形式配置展示，数据动态更新;
- 将数据分析结果以API或系统集成工具形式，直接应用到业务场景中；
- 开放API将Elemental系统的模块逻辑和数据输出深入集成到客户业务系统中，由Elemental系统提供建模和数据运算服务;

场景应用

Elemental科学数据平台已在制造业、能源、零售等多个行业提供成功应用案例。

以零售行业为例，通过对ERP、会员、营销等数据进行统一采集、整合、清洗、打通、分析、建模等操作，Elemental可提供：

- 零售数字场景应用，如用户画像、会员线上数据采集与行为分析、会员精准营销、会员生命周期管理、库存优化等；
- 营销、会员应用数据回流，实现业务数据闭环；
- 高效业务中台支撑，如OLAP数据统一治理、CRM、业务支持等；

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248

部署与服务

产品服务方式

Elemental系统提供公有化部署服务和私有化部署服务两种方式：

- 公有化部署：用户使用Elemental系统的统一SaaS平台，免费用户使用系统提供的公有计算资源和存储资源，付费用户可根据需求购买存储和计算资源。
- 私有化部署：用户使用私有存储和计算资源进行Elemental系统的部署，包括私有机房，以及在公有云上开通的云服务器等。

Elemental系统的使用对硬件和软件具有一定要求，除客户端访问建议使用最新版的chrome以外，服务端的最低部署要求如下。

私有化部署条件如下：

	自有服务器部署需求列表：
*	服务器(推荐配置): 服务器数量最少3台（不包括高可用） a–CPU：Intel(R) Xeon(R) CPU E5–2620 0 @ 2.00GHz *2 12核; b–MEM: 128G c–DISK:RAID10，RAID5 600G硬盘 d–千兆网卡*2
*	系统: centos7 or redhat7(内核:4.1以上)
*	配置服务器之间密钥认证，hosts配置主机名互联; selinux暂时关闭
*	网络：部署系统时，服务器可连接互联网
*	权限：root权限
*	远程部署：ssh拨入，root权限。如果有访问限制，建议vpn拨入

info@kongming–inc.com
上海市长宁区愚园路753号E栋六层 021–63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010–52434248

关于孔明

孔明科技是一家总部位于北京的大数据解决方案提供商，成立于2010年，人员规模超过200人，在上海设有分公司，业务遍及华北、华东、西南等区域。公司拥有自主研发的数据科学平台Elemental系列产品，和一支经验丰富的数据分析团队。

作为一家面向数据时代的系统级软件供应商，孔明科技自成立伊始就在企业级大数据治理、数据挖掘、建模应用等方面深耕，不断积累行业化经验和实操案例。通过多年锲而不舍的努力，公司对异构数据融合、数据清洗、数据资产管理、大数据挖掘和建模分析、数据可视化、大屏等方方面面获得了丰富的经验，在大数据平台开发、数据建模、数字化营销运营等领域拥有独特的核心竞争力和商业案例。2017年，孔明科技登陆新三板，证券代码 872061。

长期以来，孔明科技秉承“客户的数据需求就是我们可持续性发展的号角”的执业理念，在服务上追求精益求精，在技术上追求实用性领先，在项目执行中提倡务实有效、高可用性，通过数据方式有效改善传统决策，提升企业的洞察力、竞争力和决策力。孔明科技的客户群体横跨母婴快消、食品饮料、休闲娱乐、电力能源、医疗健康等不同行业，累积服务大中型企业数十家。

公司愿景是：赋能企业数据科学能力，让数据更简单，商业更智慧！

info@kongming-inc.com
上海市长宁区愚园路753号E栋六层 021-63669001
北京市朝阳区惠新东街甲2号住总地产大厦10层 010-52434248