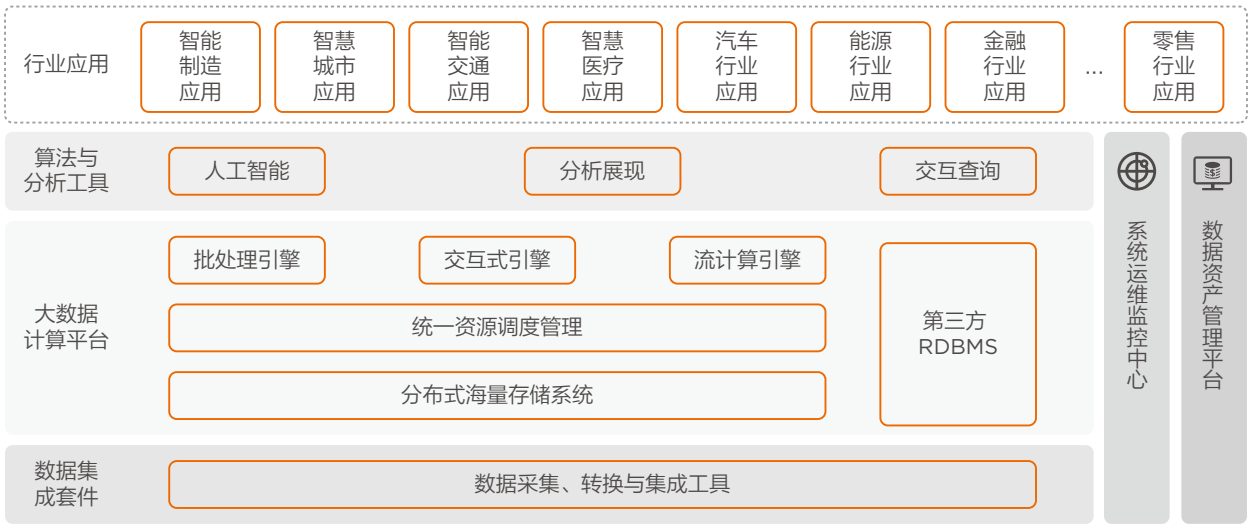
Lenovo 联想

联想企业级大数据分析平台

Lenovo Enterprise Analytics Platform

产品概述 PRODUCT DESCRIPTION

在大数据时代，构建面向海量数据的存储与计算能力、挖掘数据的深层价值正逐渐成为提高企业竞争能力的核心要素之一。联想企业级大数据分析平台（Lenovo Enterprise Analytics Platform，简称：LEAP）是业界领先的处理企业级大数据场景的高性能一站式分析平台。它可以帮助企业快速建立一个统一的数据和计算平台，快速支持企业内部 / 外部数据的采集与集成、实现海量数据的存储、并提供极佳的数据计算与深度分析挖掘能力。在大数据平台之上，用户可以构建相应分析挖掘应用，从而辅助企业及时洞察新的商机和潜在的风险，提升企业竞争力。



满足新一代数据管理需求的企业级一站式大数据平台

- **业界超高性能**

  深度优化 Spark 性能和系统整体能力，使得评测性能比最新开源标准发布产品提升 20% ~ 30%，性能领先主要友商 3—6 个月。

- **全源数据整合能力**

  自主研发，突破传统架构和性能瓶颈，提供五十多种数据接入适配能力；通过全图形化的灵活配置，可以实现多源异构数据的快速采集与集成。

- **完整 SQL 兼容能力**

  扩展计算引擎，率先通过 TPC-DS 99 条语句的全部测试；全面支持 SQL 99/2003 语法和存储过程，支持数据的增删改查处理能力，保证传统企业数据处理流程的无缝迁移。

- **海量数据实时处理能力**

  深度优化实时计算引擎，可支持物联网百万传感器的实时采集需求，实现 500MB/s 的传感器数据的实时预警分析能力，支持分钟粒度的突发事件预警 。

- **一站式图形化的数据开发套件**

  提供强大的开发组件环境，提供丰富的图形化管理和开发界面，支持运行、调试、日志跟踪、结果预览等功能，极大地方便业务人员的使用。

- **最完整的并行数据挖掘算法库**

  提供目前最全的（50 多种）并行数据挖掘算法，同时整合超过 5000 个 R 语言算法包。对原生算法提供深度优化，保证易用性和准确性。

- **最高标准数据安全保障**

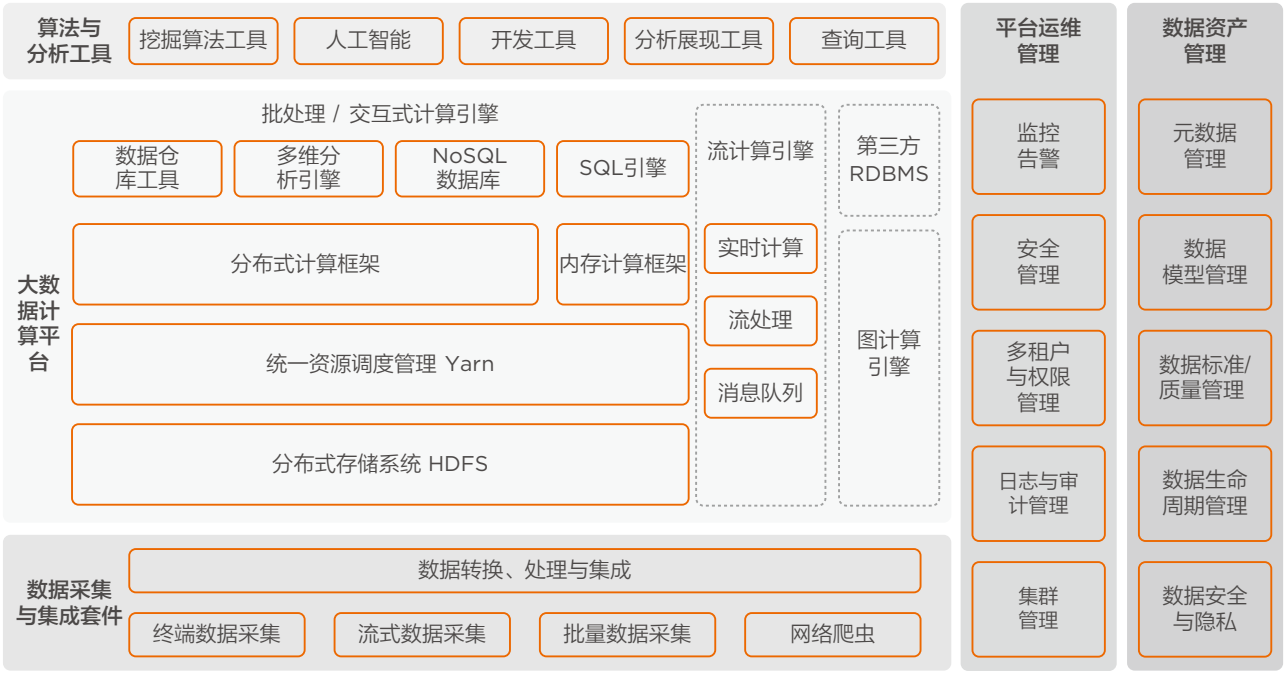
  平台支持细粒度的数据访问控制，并扩展了多租户管理及资源隔离；支持基于 Kerberos&Sen-try&LDAP 实现用户访问和服务间的强认证；支持 TCM 硬件级密码计算和密钥保护；提供全面和高标准的数据安全保障机制。

- **全球实践验证的一流可靠性**

  联想已在全球部署了规模达到 2000+ 台服务器，3000+ 名操作用户的超大规模实证集群；是国内最大的制造企业数据集群，系统经受了 9PB 级复杂业务的实战锤炼，实现 99.9% 的全球高可用性。

能力介绍 ABILITY TO INTRODUCE

联想企业级大数据分析平台涵盖数据采集与集成套件、大数据计算平台、算法与分析工具、平台运维管理、数据资产管理等五大部分。



数据采集与集成套件

LEAP 平台支持多种结构化和非结构化数据的灵活集成。提供与众多系统和设备对接的集成套件，

能够为企业快速收集其内部运营销售数据以及信息系统之外的设备、用户和社交数据等，并且能够根据企业的需求方便地快速扩展。

大数据计算平台

LEAP 平台基于 Hadoop/Spark 生态系统，引入了多种核心功能和组件，对复杂开源技术进行高

度集成和性能优化，面向基础设施层进行深度调优。在分布式存储系统的基础上，建立了统一资源调度管理，高效地支持大规模批处理、交互式查询计算、流式计算等多种计算引擎。LEAP 平台为企业级分析提供最佳性能和高稳定性的大数据计算环境。

算法与分析展现工具

LEAP 平台预集成了各类数据挖掘算法，结合多年的实践优化，显著提升算法的运算效率，为数据

深度挖掘提供能力支撑。同时结合内嵌的应用工具集，能够为用户提供一站式、可视化、低门槛、高价值服务。

系统运维监控中心

LEAP 平台的运维管理工具，快速完成产品套件的安装部署、节点监控、访问权限管理、资源配额

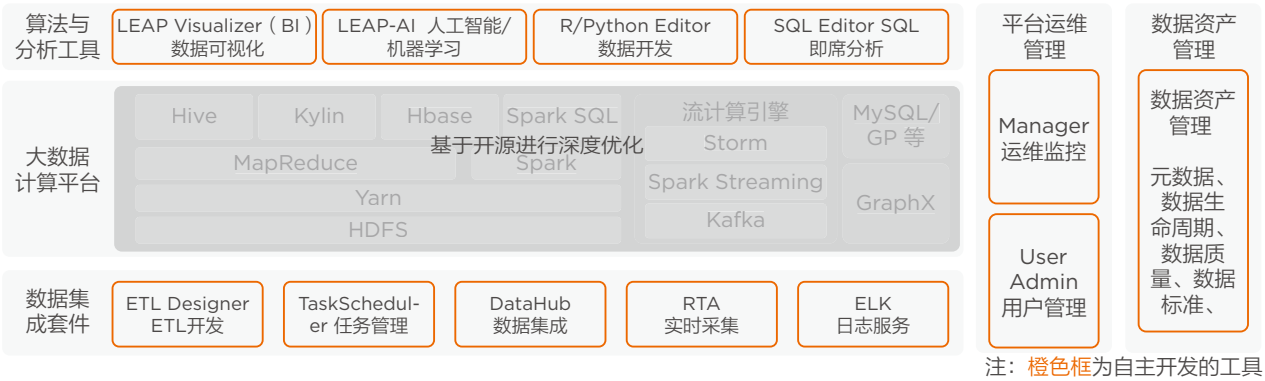
管理、系统告警分析、升级扩容等计算平台维护工作，通过统一界面实现对 LEAP 分析平台及运行状况的易用、易管理。

数据资产管理

将数据对象作为一种全新的资产形态，围绕数据资产本身建立一个可靠可信的管理机制，提供数据标准管理、数据资产管理、元数据管理、数据质量管理、数据安全等，以实现数据的可管、可控、可视，为实现数据价值增值奠定良好基础。

LEAP 产品功能 LEAP PRODUCT FEATURES

联想基于企业内部多年的大数据建设实践经验，针对开源 Apache Hadoop/Spark 框架进行了大量的修复完善及深度优化工作，并自主创新的众多功能和实用工具，易于使用者开发和管理。



注：橙色框为自主开发的工具

功能	内容描述
数据集成 DataHub	
数据库导入	支持 MySql、Oracle、DB2 等多种数据库到 Hive、HDFS 的数据导入；支持常见数据库互导以及导入到 HDFS 和 Hive；
本地文件导入	支持本地文件、Excel、CSV 到 Hive、HDFS 的导入；客户端导本地路径文件源、客户端导数据库数据源；Dump 文件上传到 Hive、HDFS、DB；
公有云数据导入	提供阿里云、亚马逊云 RDS 到 HIVE、HDFS、常见 DB 数据导入；
大数据类数据库导入	支持 Redis、HBase、Impala、MongoDB 等数据导入
Http 流式上传	提供 http 流式上传方式，开放上传接口，上传到 kafka
消息队列服务	提供 Kafka 消息队列服务
其他类型导入	支持 SAP、网络数据爬取等功能；提供 FTP 上传到 HDFS
迁移任务	展示所创建的任务信息及任务运行情况，并可对任务进行管理
	查看所有数据迁移任务的执行历史和日志（比如某条任务是每小时执行一次，就会产生多条执行历史）
资源库	创建资源库连接信息并保存，在之后的上传过程中，可以选择已经保存过的数据库，自动进行连接选择，不用再输入信息，方便操作；
	对保存过的资源库连接进行列表展示，并可以对资源库信息进行管理操作；
流程管理	流程的新建、复制、删除、修改、启用、停止、查询，定义任务调度策略；
大数据平台组件	
分布式文件存储 - HDFS	分布式文件存储、多副本备份与同步机制，提供容错机制，可修改副本策略，支持跨机房备份；
	大文件写入、流式数据访问、高吞吐量数据访问；
	支持数据存储分布策略，支持机架感知与负载均衡，支持高可用；
NoSQL 数据库	分布式、列存储、多维结构存储，支持结构化和非结构化大数据量的高速读写操作；
	面向列表（簇）的存储和权限控制，列（簇）独立检索，以及二级索引，支持数据多版本；
	面向列的数据压缩，高压缩比，有效降低磁盘 I/O；
数据仓库工具 - Hive	海量结构数据批量离线分析；
	提供基于 HQL 的数据查询机制，支持 UDF，自定义存储格式，扩展数据类型，函数和脚本；
批量计算框架 - MapReduce	数据划分和计算任务调度；
内存计算引框架 - Spark	分布式内存计算引擎；
流数据计算引擎	基于 Storm 与 Spark Streaming 的流式计算引擎；
分布式数据库 MPP	支持基于 Spark 的 MPP 架构数据库，基于 Spark 扩展 CRUD 操作；
多维分析引擎 - Kylin	提供 OLAP 分析能力，支持 SQL 查询

分布式消息队列服务 - Kafka	支持消息队列的负载均衡、分区存储、数据压缩等
分布式协作服务 - Zookeeper	配置管理、配置更新通知、节点主备容灾、节点心跳管理等；
统一资源调度 - Yarn	支持资源封装、调度、隔离以及配额管理；
	支持 Capacity(静态)、FIFO(先进先出)、Fair(公平、动态) 等调度模式；
交互式分析引擎 - Impala	支持基于 SQL 的查询分析；支持基于 JDBC/ODBC 的数据库连接，支持 BI 可视化工具连接
数据导入导出 - Sqoop	支持传统数据库到 Hadoop；支持 Hadoop 到传统数据库；
全文搜索引擎 - Solr	基于 Lucene 的全文搜索服务器；
日志采集服务 - Flume	分布式、可靠的日志采集服务；
日志分析服务 - ELK	提供一个分布式多用户能力的全文搜索引擎；支持日志搜集处理框架、快速的日志综合处理能力；支持日志搜索、可视化、分析能力
缓存服务 - Redis	基于 Key-value 的数据缓存库，支持数据同步；
安全保障	支持 Kerberos 认证和 LDAP 集成；
任务调度 Task Scheduler	
实例管理	快速检索查询当前平台的所有流程实例、流程执行实例依赖关系图形化、实例执行流程图查看；
	重跑、补跑、任务重试、终止；
	实例相关流程调度历史时长图形化与列表两种方式展示；
配置管理	数据库等资源的连接配置；
	依赖的 hadoop/hdfs 等相关的配置；
	支持短信 / 邮箱告警服务，如邮箱配置、SMS 配置，告警短信配置；
数据质量 Data Quality	
基础检查	按用户选定模式统计 " 空白 " 数量和占比；
	证字段的唯一性，统计不唯一 id 的占比，计算“孤值”；
类型检查	统计 true/false(/null) 各自占比
	统计字符集统计各自数量
	按用户勾选项统计结果
日期检查	统计日期缺失数量和占比；
	统计各种时间关键数据；
	统计各部分时间分布；
	查找出当中包含的工作日；

其它检查	按用户输入统计各部分数量和占比
	提取顶部 (底部)top N 的值
	统计用户指定参数不匹配的值及数量比例
	模式搜索
数据分析 SQL/R/Python Editor	
数据源管理	获取数据库元数据信息，可以展开并快速检索表信息；
数据查询	提供 SQL 编辑器，支持语法补全、关键字补全、数据库表提示、SQL 格式化；
	支持 SQL 2003 标准，兼容 SqlServer/Oracle 语法，支持存储过程、支持 TPC-DS 测试集 99 个 SQL 语句；
	查询结果可以通过交叉表进行进过展示，默认显示前 100 行；
脚本开发	支持 R、Python 脚本运行；
定时任务	展示当前定时任务列表及执行历史；定时任务添加、修改、删除、禁止，可以配置任务的调度周期
多租户管理 User Admin	
用户管理 - 操作用户	用户及账户的添加、修改、删除，用户启停用；
项目管理 - 多租户管理	提供项目管理、人员分配、权限管理等功能
费用管理	当前计算、存储资源消耗费用计算，并可查看详情；
资源管理	当前租户下各项目资源使用情况，对项目进行资源池分配；
个人中心	密码修改、用户注销；
集群管理 Manager	
安装部署	安装文件拷贝、环境检测与主机环境配置、组件自动化部署；
集群监控	指标监控、监控热图、历史配置信息、版本信息；
服务管理	添加与删除服务、服务启停、部署与移动；
	参数配置、配置组、历史版本；
	HA 配置，支持全组件的 HA 配置，包括 LEAP Manager 管理节点；
主机管理	添加与删除主机节点、主机监控指标、主机及相关组件的告警信息；
告警管理	报警历史记录；
	告警组、告警通知；
版本管理	平台及各个组件版本管理、版本升级；
	平台授权信息注册；
用户管理 - 运维用户	用户添加、修改、删除；
	角色添加、修改、删除；
日志管理	根据检索内容做简单的信息统计，统计不同类型输出信息数量；
	根据组件、关键字信息做信息搜索；

核心技术优势 CORE TECHNOLOGY ADVANTAGE

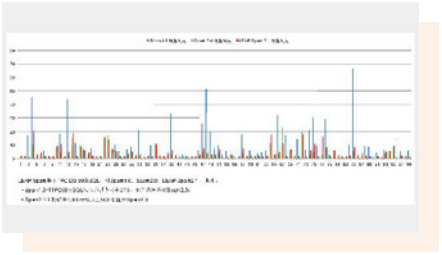
联想紧密研究和实践大数据先进技术，基于企业内部多年的 PB 级大数据建设实践经验，形成了适应企业级应用的稳定可靠、性能优异、易用易管理的大数据平台，更好地满足企业级客户的需求。

LEAP 平台核心技术优势包括：

完全支持 SQL 标准，增强分布式事务处理能力，全面支持 MPP 场景

联想 LEAP 对 Spark 引擎进行了大量优化，全面支持 SQL99/2003 和存储过程方式（兼容 Oracle、SQL Server、MySQL 等）访问大数据基础数据库，全面高标准通过 TPC-DS 标准测试集的全部 99 个测试项；性能优异；LEAP 还支持对数据的增删改操作，并通过分布式事务处理保证数据增删改查过程中的一致性，使得用户更好的迁移历史应用，并降低使用人员的专业要求门槛。

右图是 LEAP Spark 与开源标准 Apache Spark 的性能对比。



突破实时处理计算框架，支持物联网实时业务分析

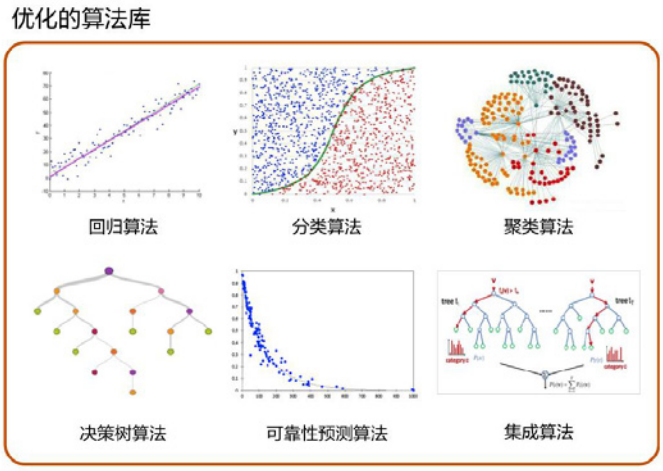
LEAP 采用深度优化的 Kafka 加 Storm 计算引擎，实现 500MB/S 的传感器数据的实时预警分析能力，通过服务器的线性无衰减叠加，可支持百万传感器的实时采集需求。实现了基于时间序列的传感器数据优化，可以支持分钟粒度的突发事件预警。通过高效的物联网数据存储压缩能力，成本最优的解决海量数据的存储问题。

实时物联网数据分析框架

最完整的并行数据挖掘算法库，并原创前沿机器学习组件

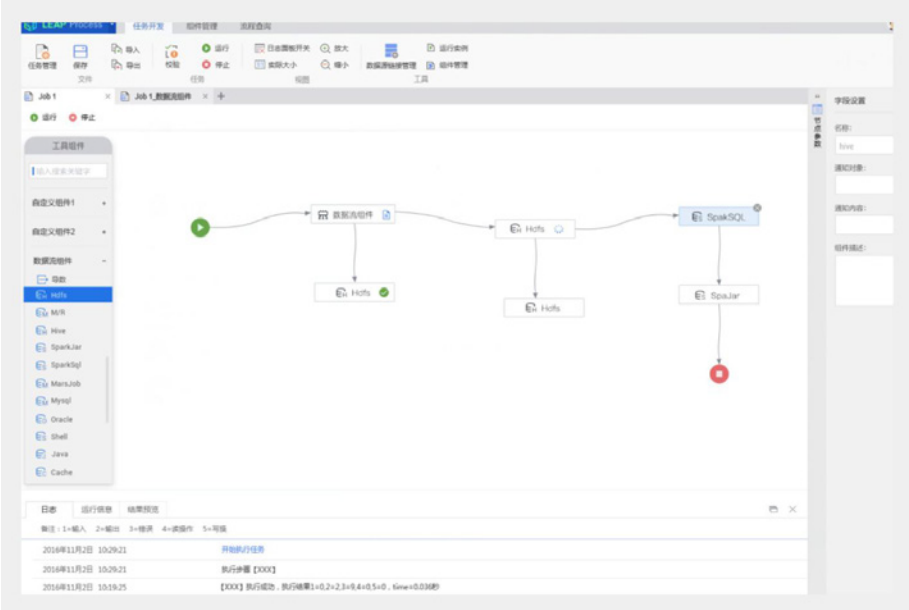
LEAP 的数据挖掘模块内置了大量常见的机器学习算法，涵盖目前最全的 50 多种分布式算法，并且对部分

算法进行了优化，精度平均高于原生算法 10%。LEAP 平台还提供了自然语言处理、文本分析、水军识别、信息传播等原创前沿机器学习组件。联想专为企业级用户打造了业界创新的机器学习 + 流处理引擎，满足大规模模型训练的高并发计算和实时流式计算处理的需要。



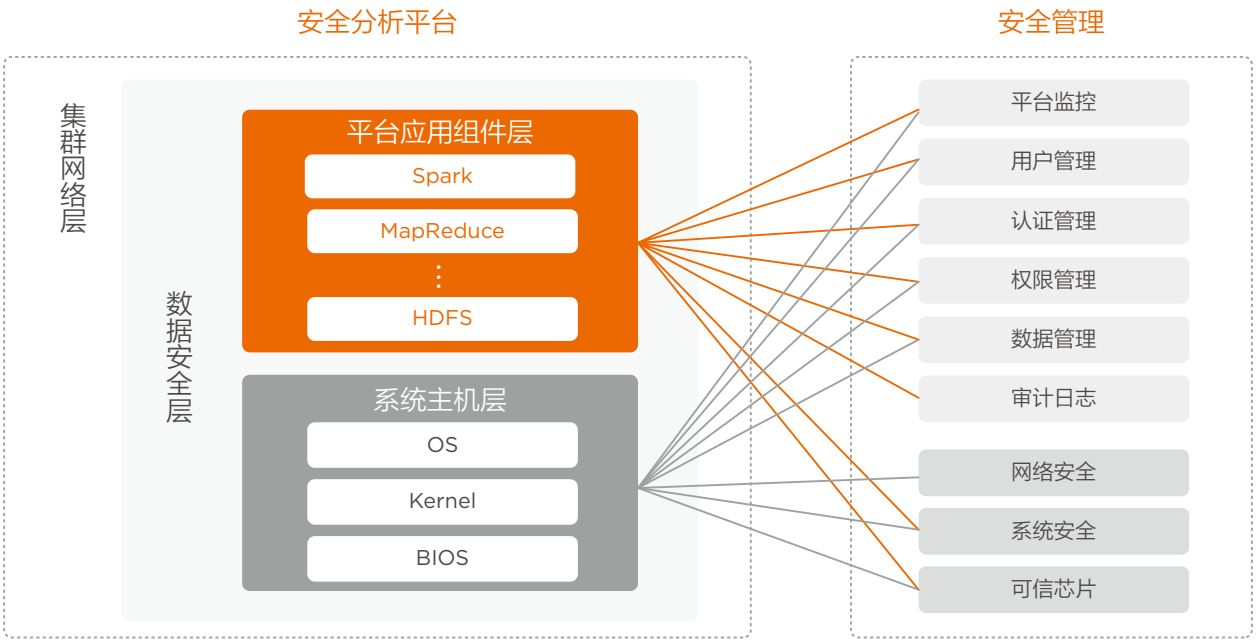
一站式图形化的数据开发套件，快速分析应用

创建各种大数据任务和查询操作全部图形化完成，无需命令行入口，上手简单。支持丰富的开源扩展，图形化添加近百个高质量组件，全面支持和管理各种大数据业务场景。联想 LEAP 采用全图形化的开发和管理界面，方便客户完成大数据集群的安装、升级及监控工作，使得管理非常便捷。



最高标准数据安全保障，构建从平台到硬件芯片级端到端的数据安全框架

联想 LEAP 非常重视信息安全的构建，平台基于 Kerberos & Sentry & LDAP，实现对大数据中心节点用户访问和服务间的强认证。支持 TCM 硬件级密码计算和密钥保护。平台自带集中式安全管理框架，解决用户授权、数据权限难题。平台支持细粒度的数据访问控制，并扩展了多租户管理及资源隔离。平台具备完善的数据生命周期保护：覆盖数据生产、采集、传输、存储、处理、共享、销毁的全生命周期管理。最终实现从硬件到系统平台到大数据服务的一体化安全可信解决方案。



全球实践，一流的可靠性

5 年不断发展，联想已在全球部署了 9 大数据中心的超大规模集群，规模达到 2000+ 台服务器，3000+ 名操作用户；存储总容量规模 12PB, 数据实际总量达到 9PB 以上；日新增数据约 30TB，日处理涉及数据达到 4.3 PB；是国内最大的制造企业数据集群，是仅次于 BAT 互联网公司的实证集群规模；同时系统具有一流可靠性，实现 99.9% 的全球高可用性。

- 总容量12PB，数据总量9PB
- 日处理数据4.3PB
- 日新增数据30TB
- 99.9%的全球高可用性

- 全球化多中心部署，2000多台服务器，3000多名操作用户
- 国内仅次于BAT的实证集群规模
- 国内最大的制造企业数据集群
- 在实践中充分验证系统的高可靠性

• 为何选择联想大数据?

- 技术开放与融合实现高可靠、高性能大数据分析平台
- PB 级数据全球化运营支撑，经过实战证明的成熟产品
- 先进完整的大数据体系方法论、以及产品与服务的交付能力
- 丰富的大数据高端人才、数据科学家、以及行业专家
- 广阔的合作生态圈，为客户提供端到端的整体解决方案

• 关于我们:

联想大数据是业界领先的少数掌握大数据核心技术的高科技团队，专注于企业级大数据分析平台的研发和实践，服务企业级客户。自 2011 年起，联想就启动了大数据分析平台的建设。多年来紧密跟踪研究和不断实践大数据相关技术，优化完善技术，自主创新的众多功能和实用工具，形成了具有自主知识产权的、稳定可靠、性能优异、易用易管理的企业级大数据平台，产品的功能和性能在业界处于领先水平。

联想大数据拥有强大的大数据研发、分析和实施团队，拥有北京、成都和香港三个研发中心，共计 200+ 名大数据研发工程师、60+ 名大数据平台运维工程师、40+ 名应用系统开发工程师；拥有 50+ 名数据科学家，他们是来自中科院、清华、北大、牛津、港大、港科大、以及美国、澳洲等著名学府的博士和硕士人才；拥有 30+ 名大数据领域平台和业务专家。

联想大数据具备全球部署超大规模集群的运维管理能力、PB 级数据与复杂业务实践的丰富经验，能够为客户提供从底层平台到上层应用的端到端全面解决方案。

• 核心技术:

联想大数据产品 (LEAP) 具有业界最完整的 SQL on Hadoop 支持，实现 100% 支持 SQL 标准，增强分布式事务处理能力，全面支持 MPP 场景；突破实时处理计算框架，支持物联网实时业务分析；最完整的并行数据挖掘算法库，并原创前沿机器学习组件；一站式图形化的数据开发套件，可快速分析应用；它是企业级大数据场景的高性能一站式分析平台的优秀选择。

• 应用案例:

已成功部署多个关键行业领域，包括制造业、零售业、能源与公共事业、金融业、医疗、智能交通、环保、教育等。

Lenovo 联想

地址：北京市海淀区上地西路6号
邮编：10085
网址：http://b2b.lenovo.com.cn