Big Data Global Practice and Arch evolution in Lenovo

联想全球制造大数据平台的架构演进及技术创新

Yu Chentao / Chief Architect, ED of Lenovo， 2018/6

关于我

于辰涛，联想集团执行总监，首席研究员

- 国家教授级高级工程师，中关村领军人才
- 北京理工大学自动化学院兼职研究员
- 科技部国家科技重点研发计划专家，网信办大数据和云计算安全专家
- 中国云计算和大数据青年科学家联盟，工信部信通院数据中心联盟大数据发展促进委员会副主任委员，CCF TF大数据专家委员会委员

研究领域：工业物联网，工业智能，云计算，大数据，工业安全

项目经验：

- 主持建设联想智能制造全球云化平台，上百个内部数字化转型项目
- 联想工业大数据产品总体架构师
- 数十家骨干企业提供工业大数据和工业智能优化解决方案

重要奖项：

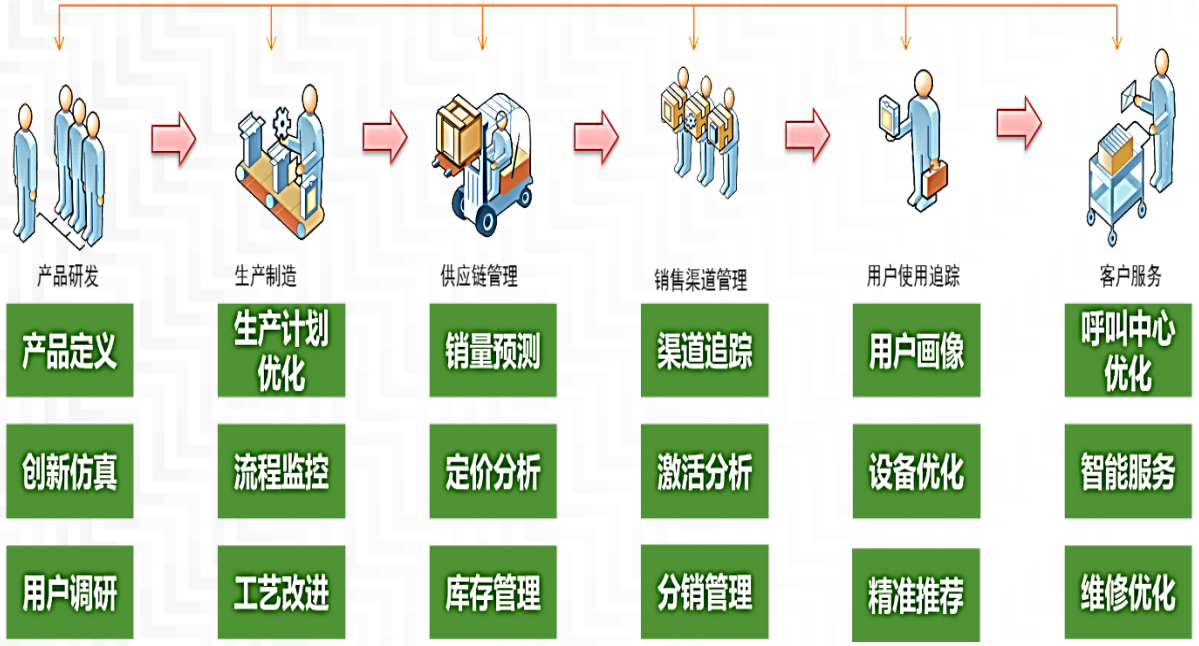
- 领导联想工业智能团队获得了2017 Kaggle数据科学竞赛全球金奖，中国大数据50强，工信部年度大数据优秀案例，中国软件技术博览会金提名奖等荣誉

专利：

- 主导建立了联想工业智能和云计算领域专利壁垒，个人申请发明专利超过110件

六年时间，构建联想工业大数据平台，支撑全球2亿多台设备的全价值链优化能力，国内最大的企业支撑集群

全价值链的产品和业务优化



- *用户需求驱动的产品研发闭环，构建了面向产品全流程的敏捷化和精细化优化能力*
- *用户价值驱动的新型供应链，支持联想千万产品的按全球消费者需求个性化柔性生产*
- *产品质量实时追踪，关键环节预测和优化*

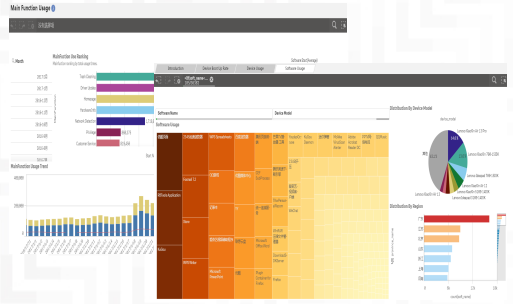
覆盖全球的大规模云化部署

- *全球10个数据中心，日增30TB，日分析数据10PB*
- *实时管理全球2亿台联想设备，31家智能工厂，6亿应用用户，1600亿条数据，已接入内部数百个业务系统*
- ***数据处理完全合规**，帮助联想构建全球化数据整合能力*

支撑500多个大数据场景优化

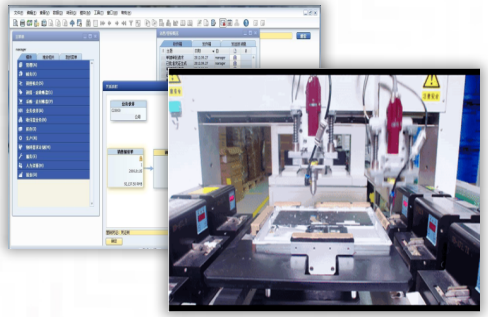
-全面提升产品研发，生产，供应链，客服等关键环节的运营效率

产品研发管理



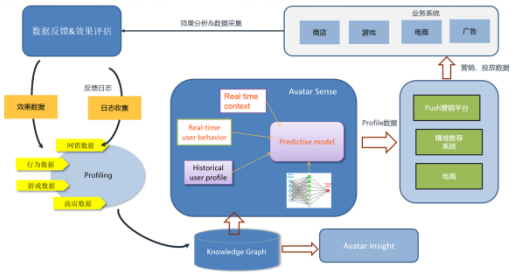
10s内全球亿级设备的产品追踪和量化分析能力

生产制造优化



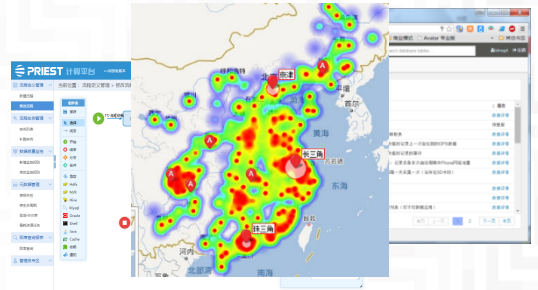
70%产品个性化定制生产，数万个配置组合

供应链管理



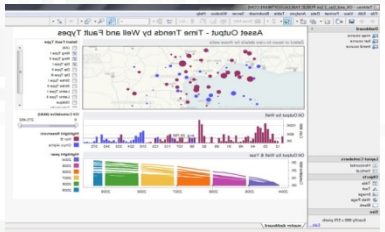
供应链预测准确性提升10～20%

销售渠道管理



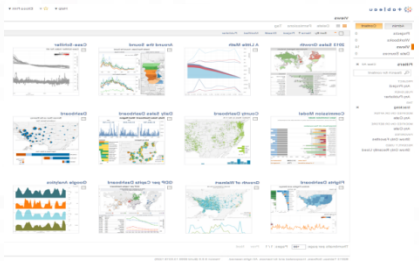
超过200万家，全球渠道和经销商潜在商机挖掘，提升商用业务效率

用户使用追踪



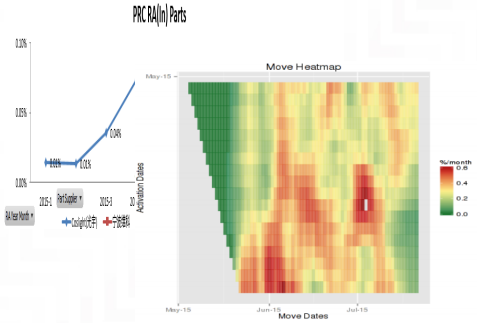
用户使用全程闭环，通过ID打通全球亿级设备

客户服务优化



20分钟内，全球全网用户舆情和用户反馈监测，并做出响应

设备质量和备件优化



2000+部件的全面备件优化，实时监控产线，优化制程

用户洞察

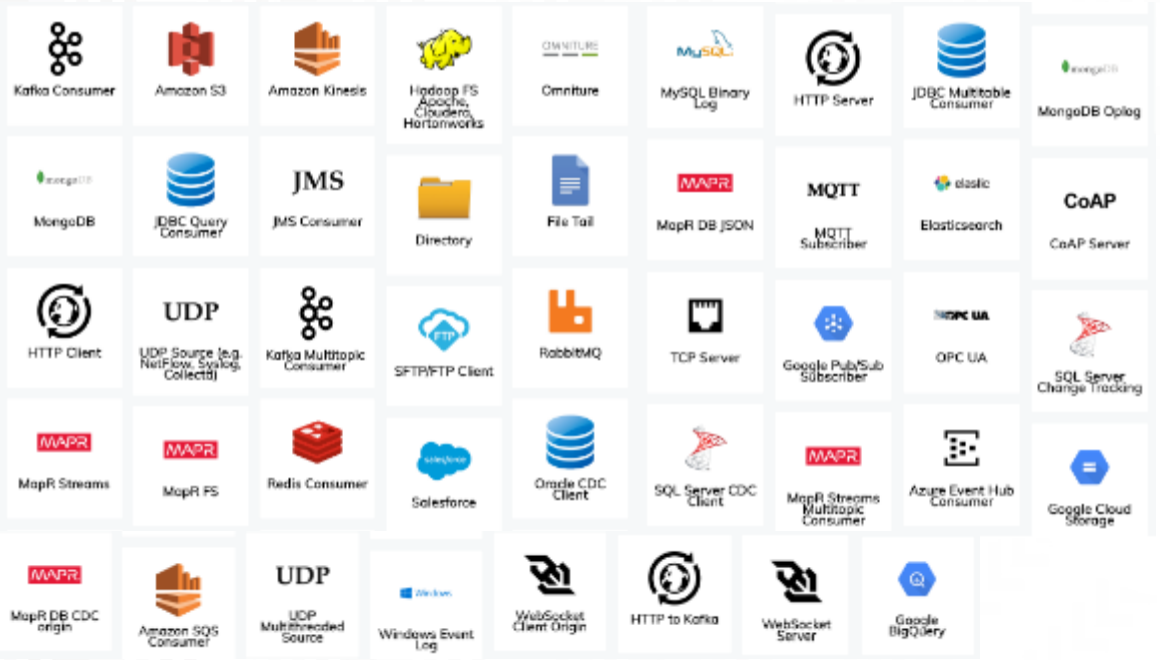
亿级用户的画像，千种不同用户标签，细分目标用户

不仅在联想内部深入实践，而且也广泛服务于大型骨干企业转型升级

01 智能化研发	02 智能化采购	03 智能化生产	04 智能化销售	05 智能化服务
产品持续改进	采购提前期预测	动态智能排产	需求／销量预测	产品预防性维护&服务
新品立项模拟	来料质量预测	现场设备预防性维护&服务	精准营销	舆情分析
文档搜索优化	供应商健康评级	持续质量管理	分销商行为分析	备品备件预防性服务
人员资质能力图谱	采购行为健康评级	成品仓优化	客户画像	服务评价
	原材料价格预测	能耗、环境分析	货物调拨优化	

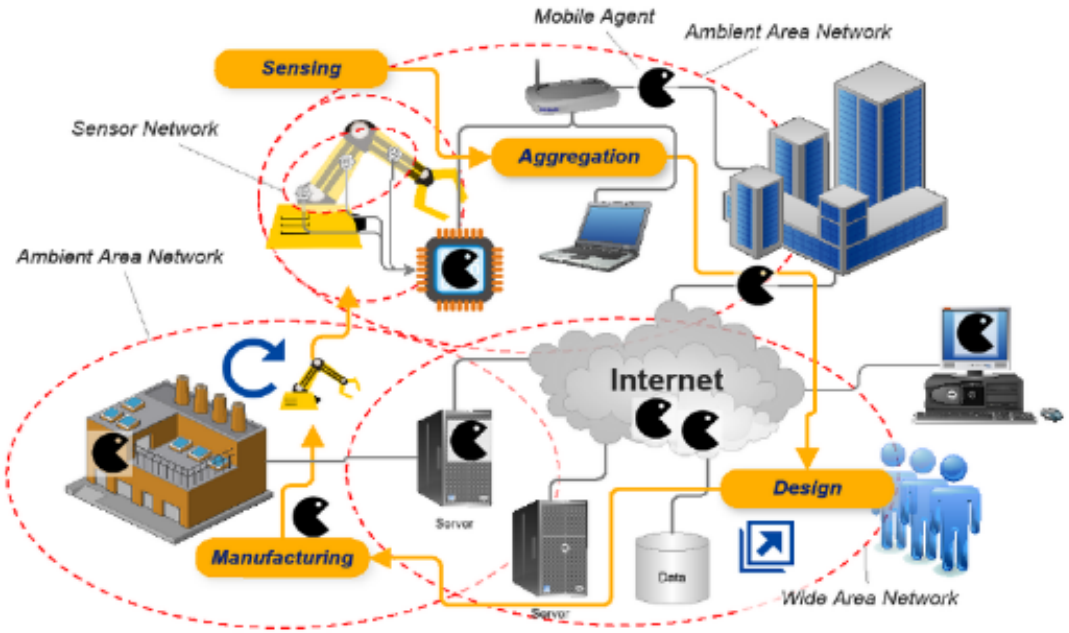
挑战之一，如何实时处理产线设备及终端设备的反馈，实现生产过程智能化

复杂工业连接



- 超过150中不同的工业控制协议
- 支持主流CoAP、MQTT、OPC-UA、TCP、UDP、WebSocket等多种数据源的接入
- 专有工业协议，获取数据，需要额外付费

实时可靠生产协作



- 海量数据量，必须实时处理完成
- 生产协作需要对接MES，PLM，TSDB等工业系统环境
- 可靠性要求高，单机需要持续数月的稳定运行能力

挑战之二，如何处理企业管理系统的海量异构数据，实现管理流程数字化

•43 供应商, 500+ 应用
•170 套商务套件, 185 套自开发应用

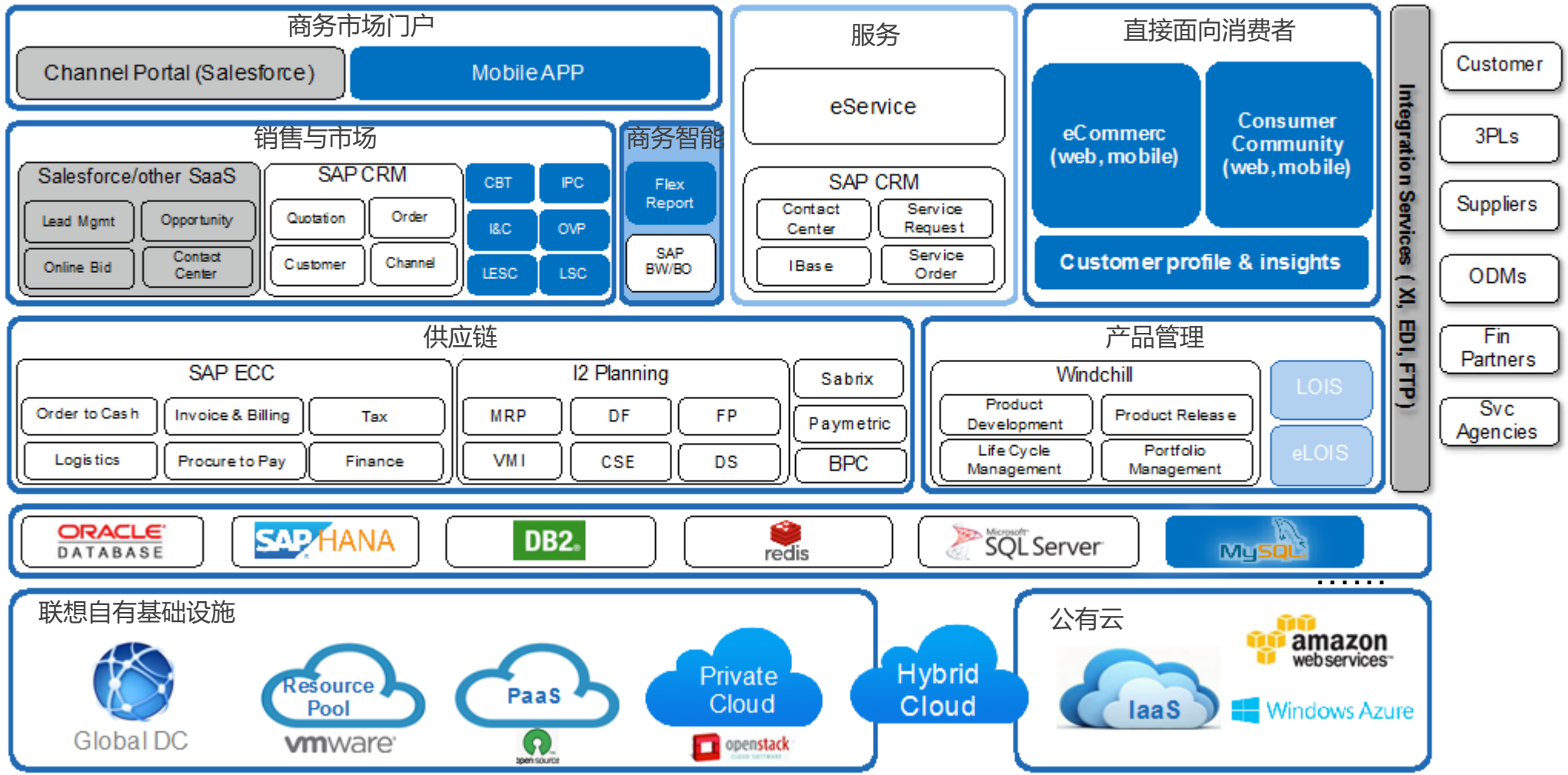
本地部署商务套件　云解决方案　开源或自开发方案

面向客户的应用

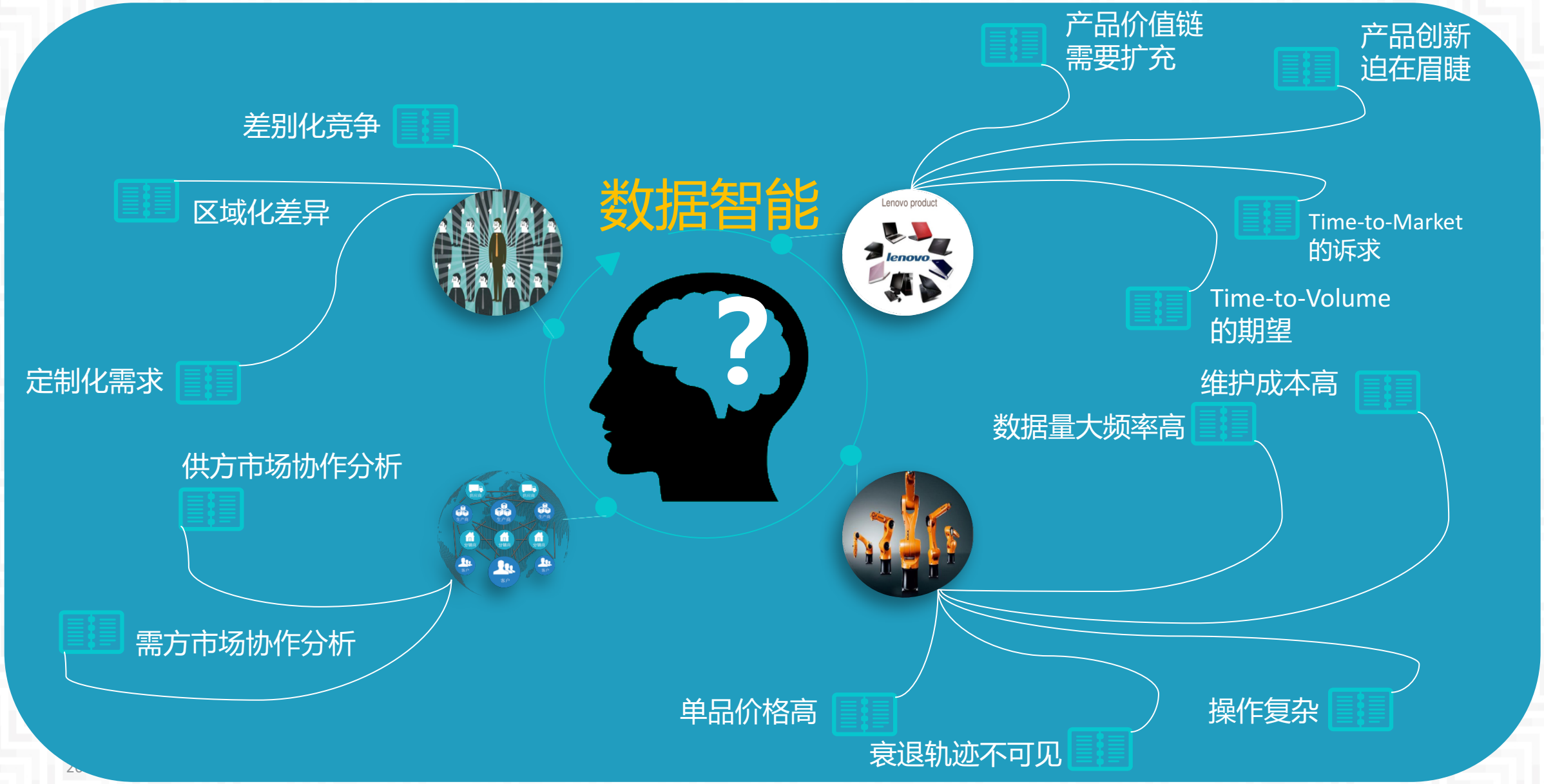
敏捷的前端

可靠的后端

技术平台与基础设施



挑战之三：如何用人工智能技术发现数据的潜在价值，实现决策过程的自动化

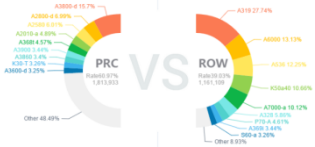


数据平台 1.0（ 2010～2014 ）
大数据技术应用的启蒙与拓荒

主要解决的问题

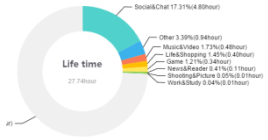
- 应用的细粒度分析，帮助业务建立量化分析能力
- 定制化的统一Dashboard，分析软件和设备的日活，留存，关键路径，软件异常等

产品管理



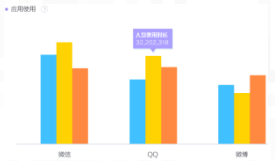
实时激活分析

销售地域分析



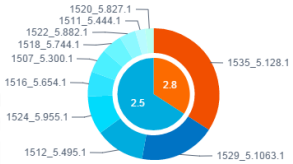
用户舆情分析

设备用户分析



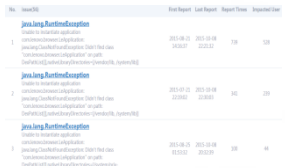
应用使用分析

设备开发和优化



软件版本管理

无线质量分析

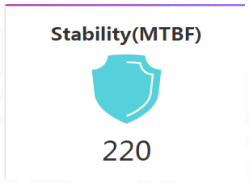


系统错误分析

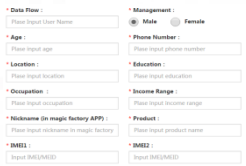
内存可用性分析

开关机分析

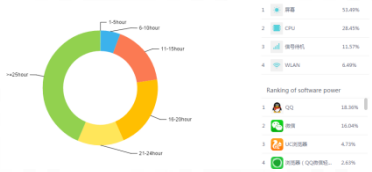
产品质量提升



平均无故障时间



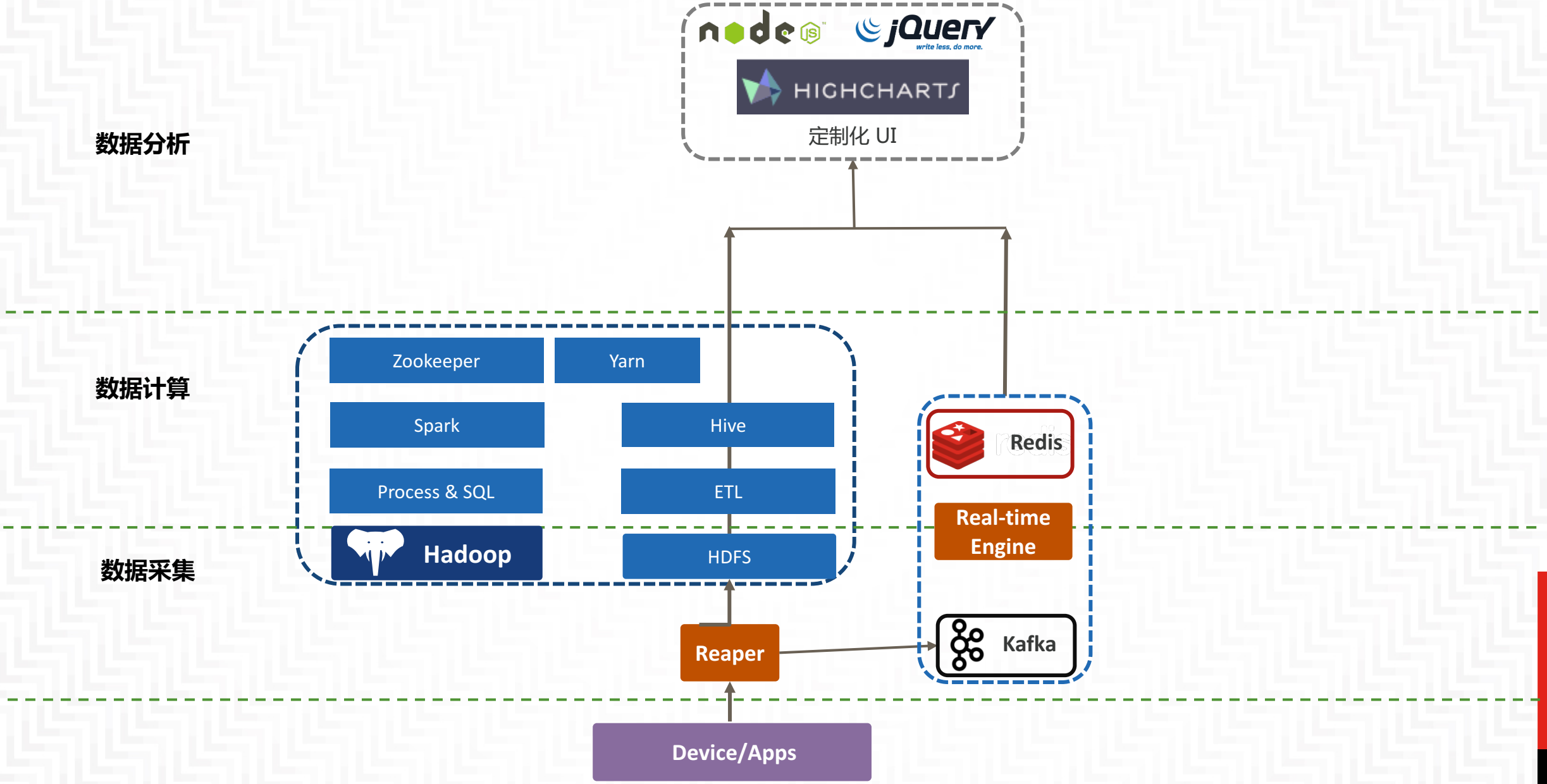
用户试用产品深度分析



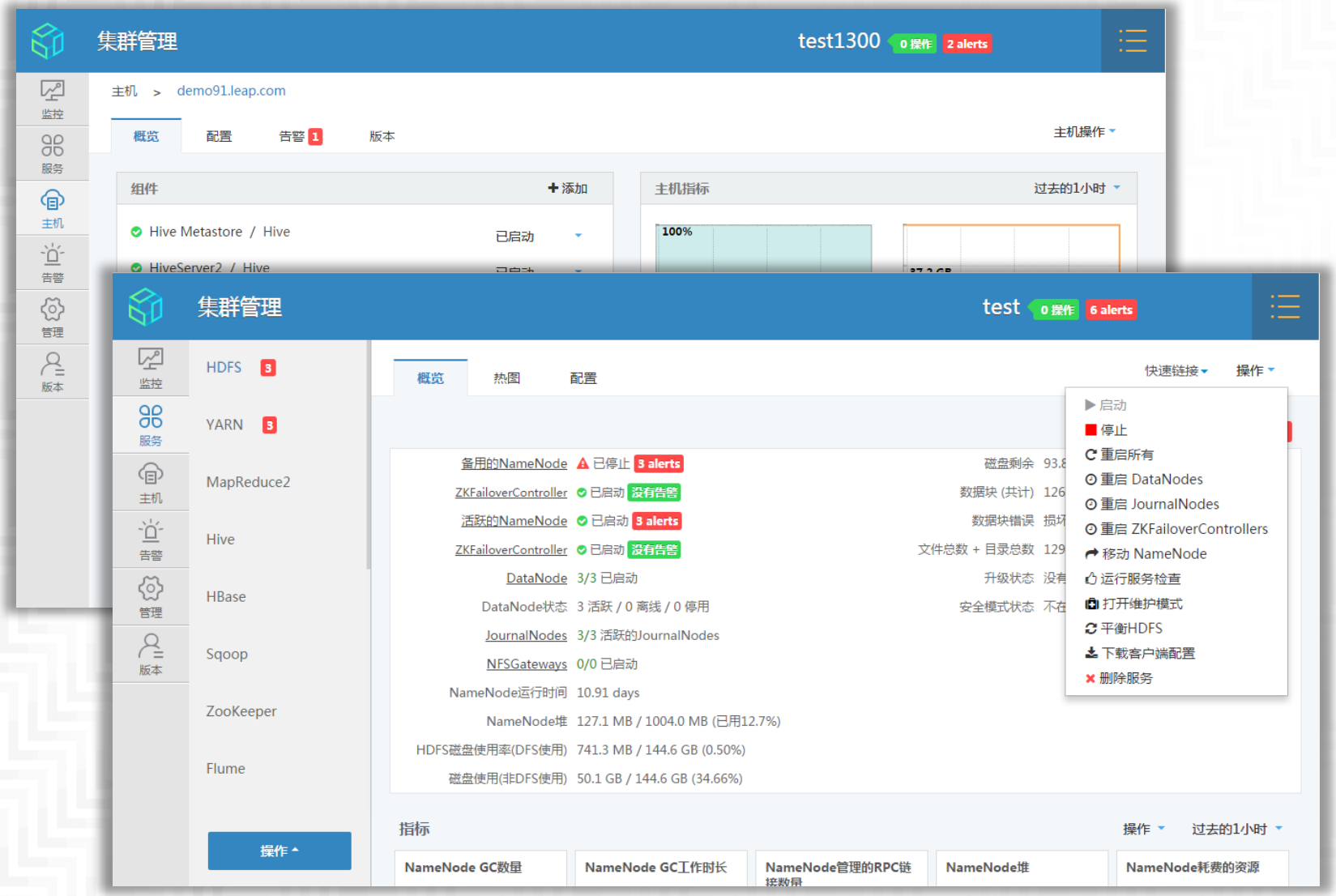
关键部件质量分析

产品质量预测

1.0 架构图（2010～2014），以整合好开源技术为主要目标



全图形化集群配置管理，实现了千台集群的可视化运维

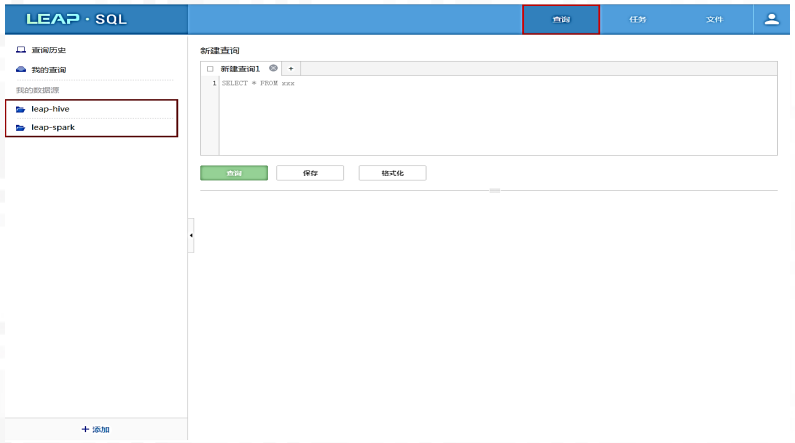


全图形化平台运维

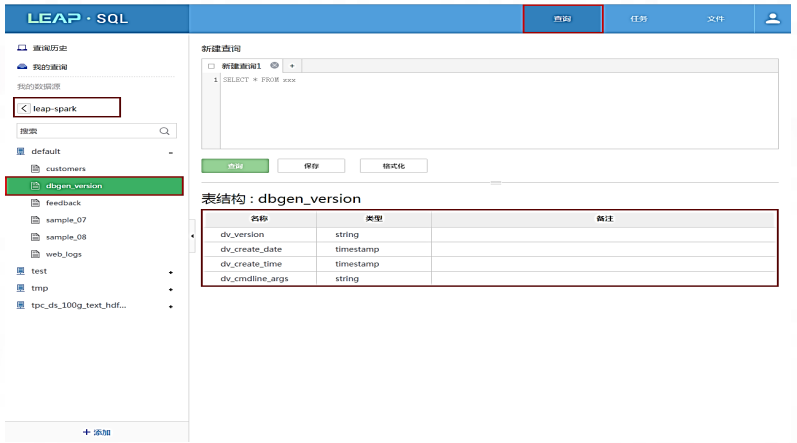
- ✓ **主机管理：**提供快速配置大数据集群主机脚本，批量初始化服务器环境、拷贝安装平台介质；
- ✓ **服务管理：**支持向导式平台配置，对主机、服务等快速配置；
- ✓ **服务配置：**根据选取的服务，自动检测资源匹配度，如CPU、内存资源是否符合服务启动需要；
- ✓ **HA配置：**提供全组件的HA服务与配置管理，包括Manager节点在内

SQL的一站式图形化编辑和执行工具，实现了计算引擎的透明化

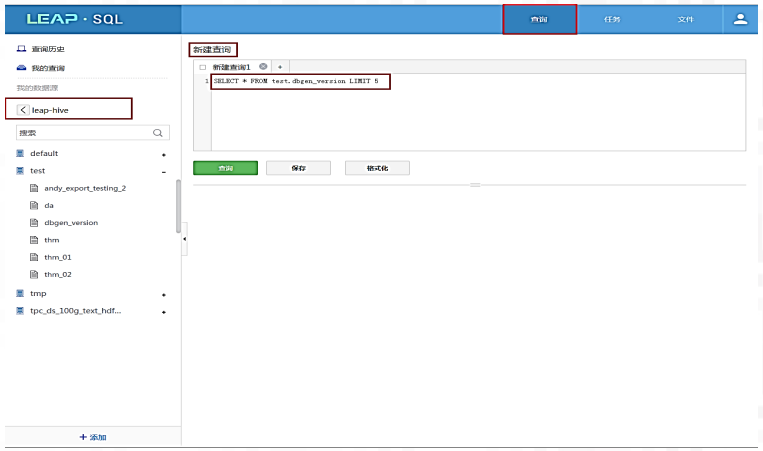
- Supports multiple data sources: MySQL, Impala, Spark, Hive



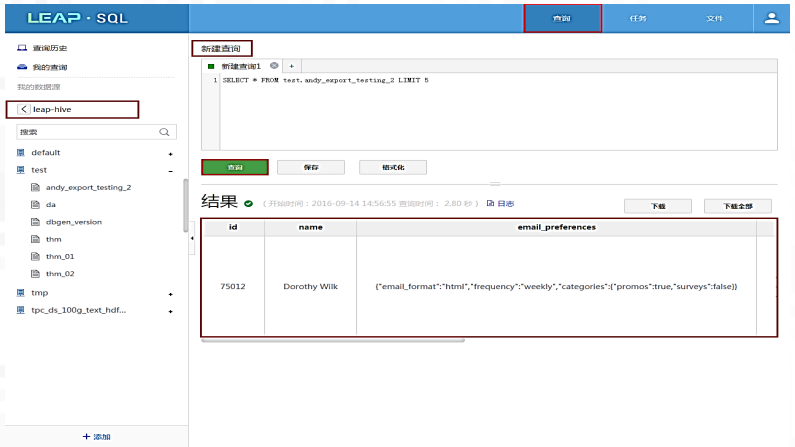
select data sources



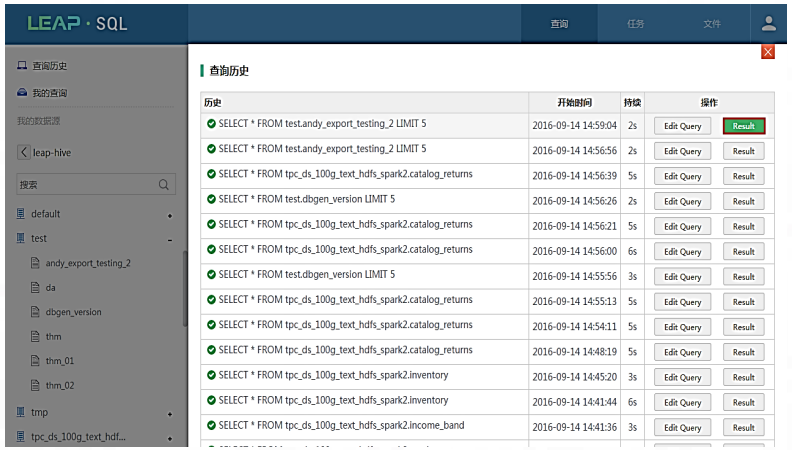
view data tables



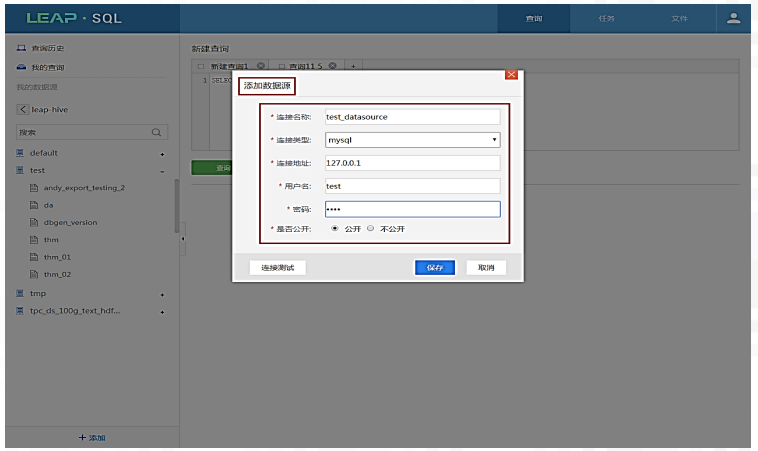
edit SQL statement



execute SQL statement



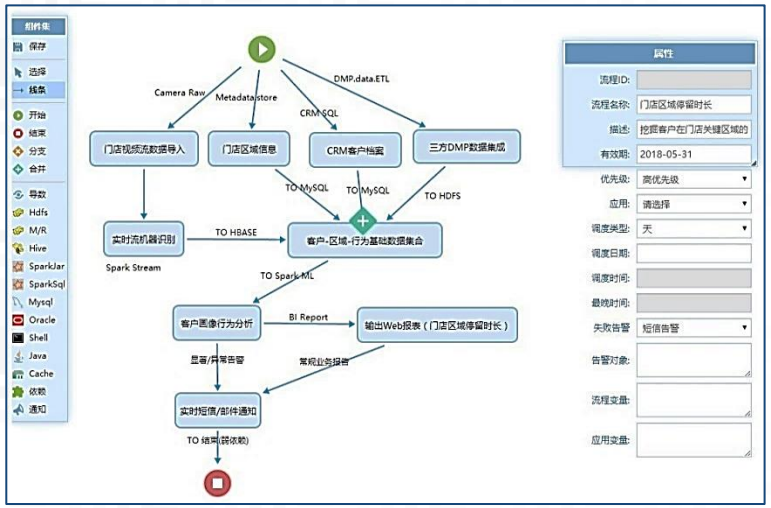
view query history



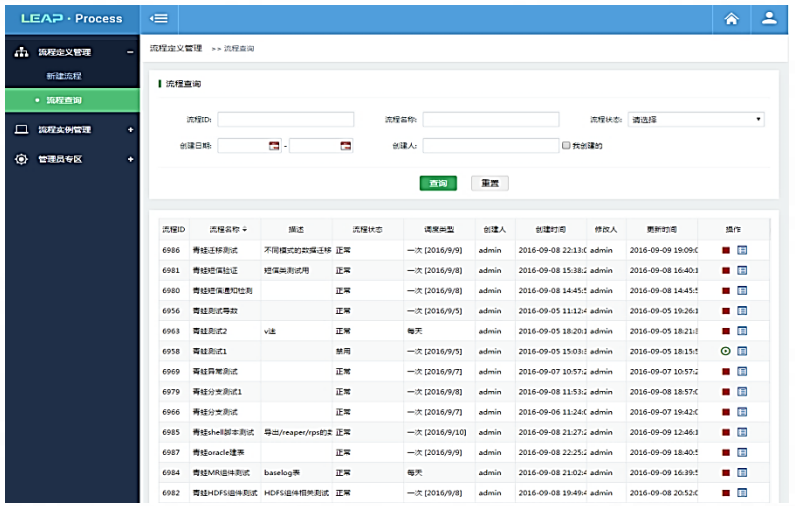
add data source

新的流程调度引擎，替换Oozie，实现万条计算任务的图形化调度和管理

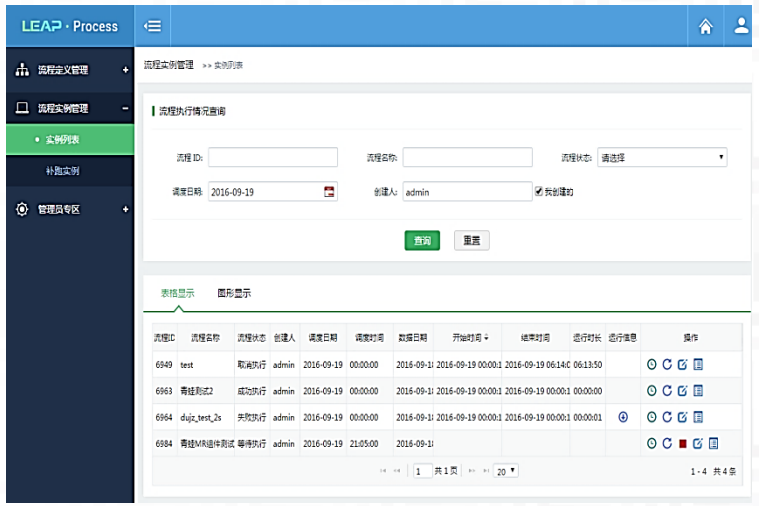
- Conveniently define the process, view process status, re-execute process.



create new process



process query



view process instance

平台1.0架构的主要问题

- 几颗老鼠屎，坏了一锅汤
  - 单层的集群架构，上千台各种配置的服务器放在一起，难于性能优化和定位问题
  - Yarn配置参数极其复杂，后续难于维护
  - 开源软件更新太快，互相之间的版本关联难于管理
- 裸奔的元数据
  - 数据缺少安全保护，这时的Hadoop没有满足要求的安全框架
- 只有设备/应用数据，没有企业数据
- 实时性差
  - 大多是批量计算任务，无法为实时业务决策做出支持
- 业务人员不会编程
  - 数据应用需要开发基础，无法帮助没有任何编程能力的业务分析人员

平台2.0（2014～2016）
大数据全球化部署，全面整合企业数据之路

Lenovo™

平台2.0 — 企业数据智能需求的爆发

业务需求

平台能力

联想统一数据平台

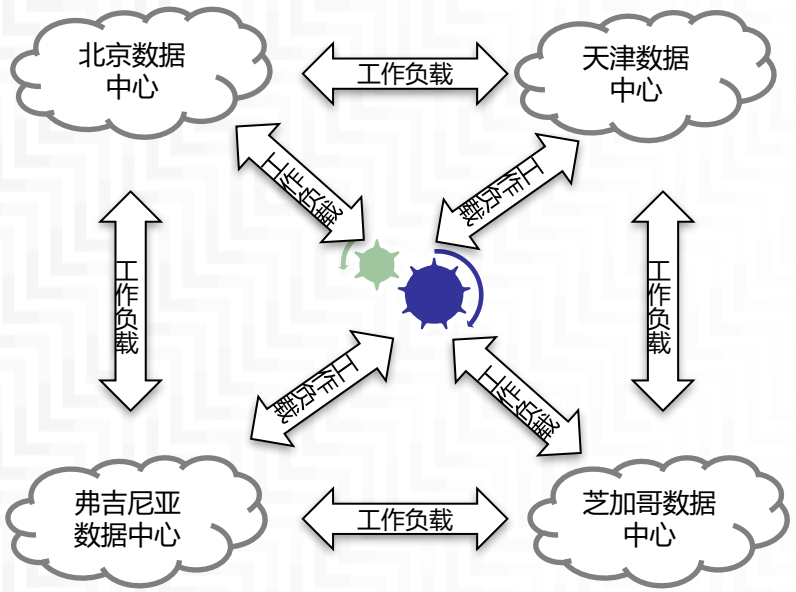
1. **全球数据整合:** 存储企业的全球运维、设备+云、客户数据（属性，行为数据）、外部数据

2. **全球数据治理:** 数据本地采集，本地计算，遵守当地国家有关用户数据的法律法规，清洗后的脱敏数据统一汇总分析，保证数据一致性，完整性

3. **企业内数据分析能力:** 数据仓库，数据集市的建设，提供智能报表工具、算法包、多维分析支持等

全球化部署挑战：需要构建满足企业不同场景应用需求的混合云架构，并达到安全可管理的业务目标

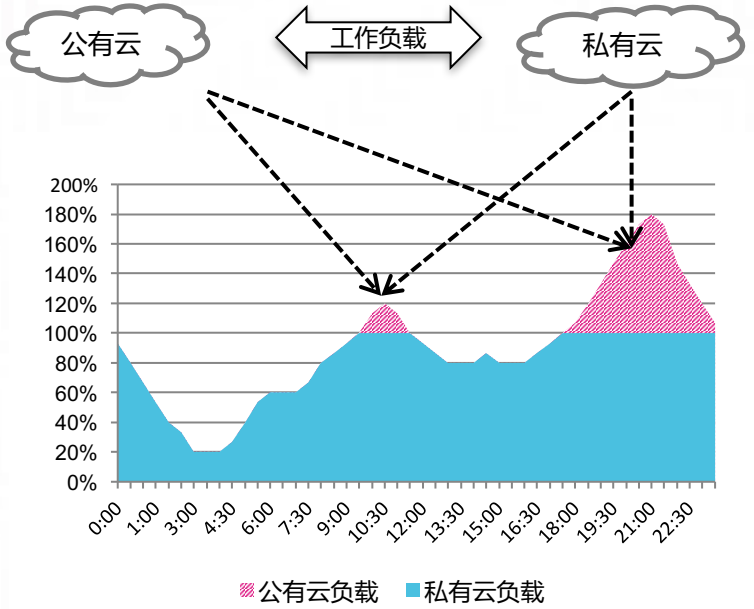
多数据中心工作负载的调度和迁移



多数据中心协同运作

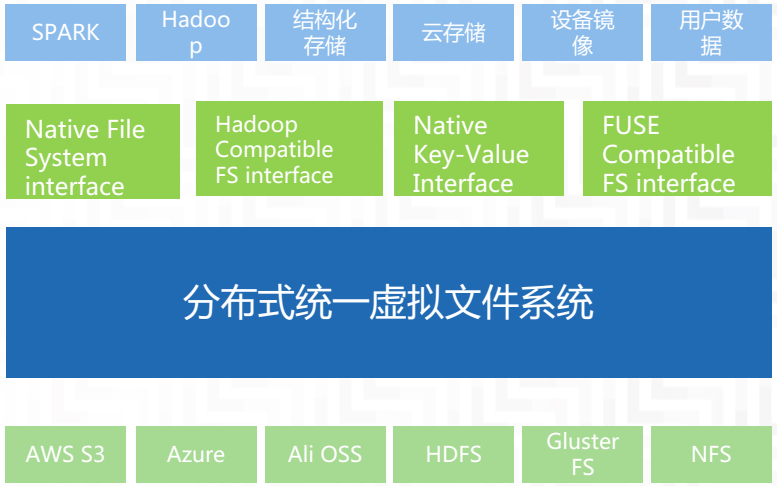
- 易于整合多个数据中心资源，提高扩展性
- 资源协同调度，消除数据中心孤岛
- 统一管理，提高运维效率
- 全局资源优化配置，降低运营成本
- 解决异地的灾备问题
- 数据中心的优势互补

公有云与私有云负载迁移



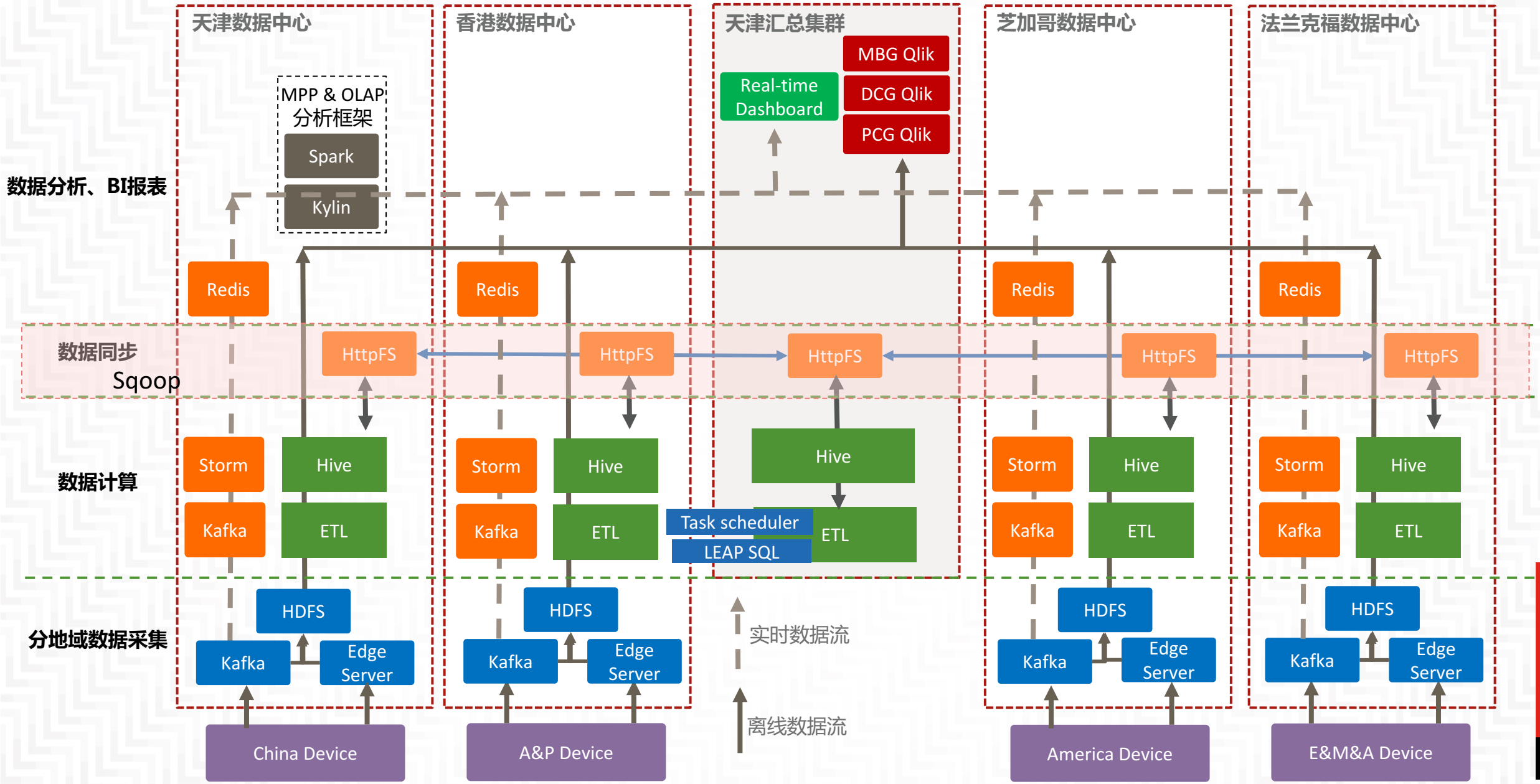
- 在业务高峰期，自动把高峰期工作负载或流量，转移到公有云，利用共有云更高的计算能力，可以缓解对内部私有云的压力和需求。
- 有效的利用公有云API完成工作负载的迁移、智能DNS技术，完善应用程序架构，以适应新型的混合云的模式

统一虚拟文件系统

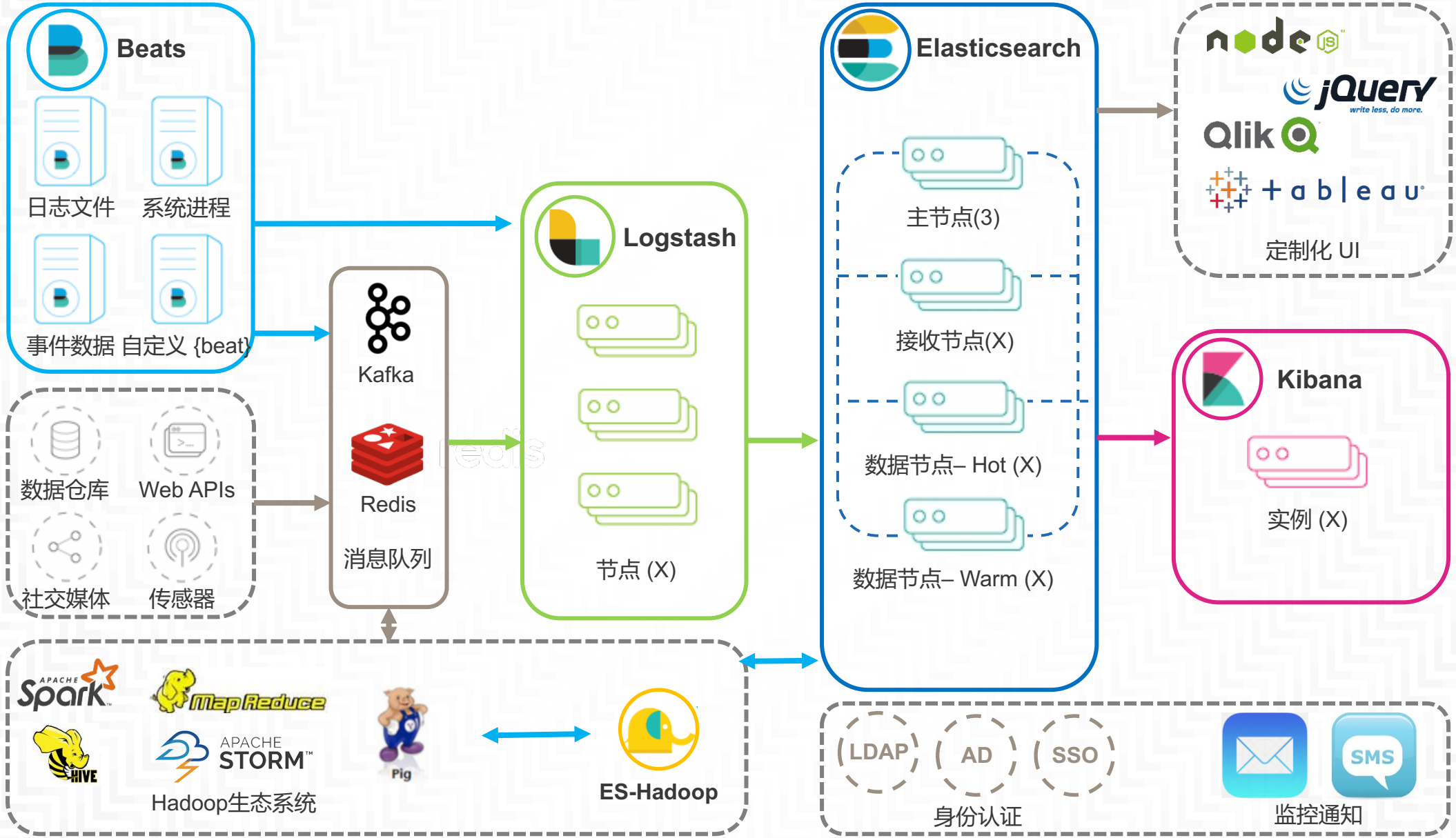


- 可以有效整合底层HDFS/S3/CEPH等异构文件系统，并对上层应用一共统一的文件接口；
- 可以整合异地的文件系统，支持跨数据中心文件系统的建立
- 支持分层读取，预读取，利用缓存技术大幅提高文件系统的性能。

平台2.0 架构，多个分立的小规模集群，每个承担不同的计算任务



引入ELK，构建全面的日志数据采集和实时分析的能力

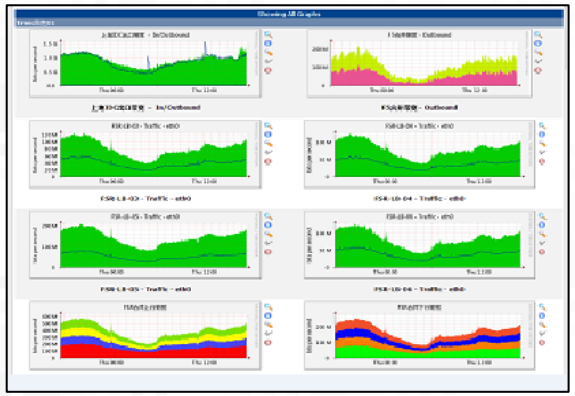


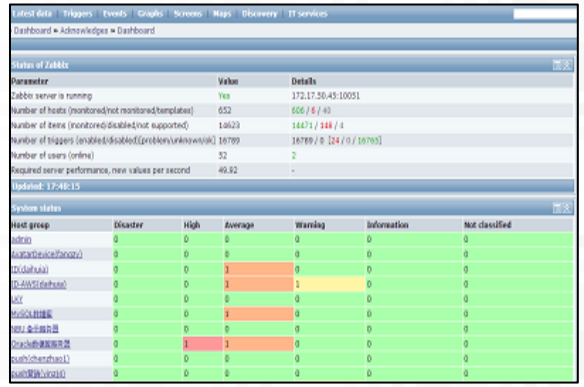
整合各种开源方案，实现全面的系统资源业务监控能力

- 采用多种开源和定制化的监控工具，从基础架构层到应用层进行端到端的监控，可以在5分钟内发现故障，并通过短信、邮件等多种方式报警，并由7×24小时服务的运维团队在第一时间响应。

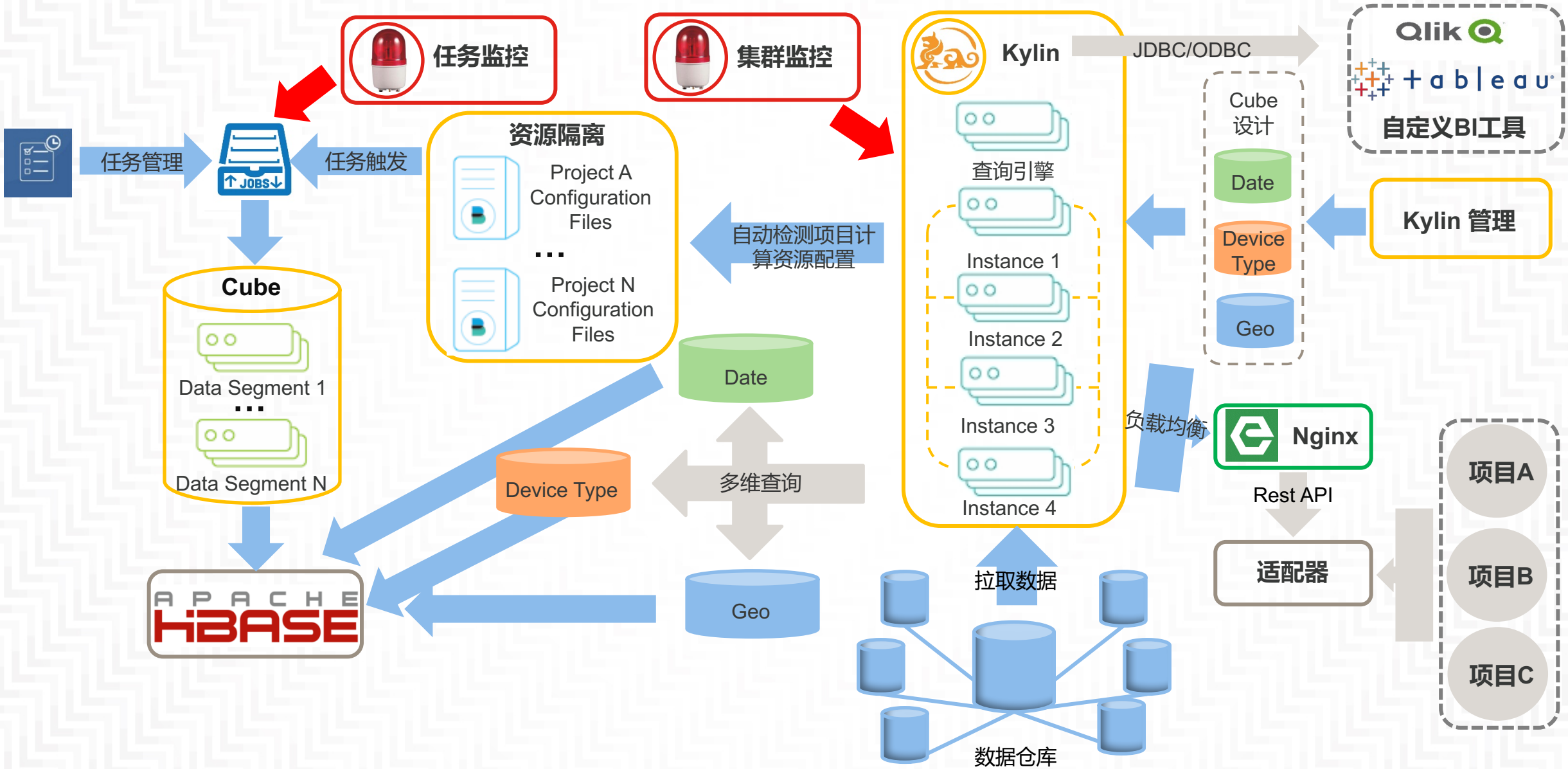
监控层	监控目标	监控参数
基础架构层	服务器	CPU, Disk, Memory, I/O
	数据库	SQL Performance , DB Usage , Running parameters
	网络	Bandwidth, Ping Delay , F5 , Firewall , Switch statistics
应用层	标准服务	Http , SSH , Download Etc.
	定制化	Login , Register, Pay Etc.

监控工具	监控点	报警手段
Zabbix/Falcon	服务器/存储/数据库/应用等	邮件 网页 短信 电话
Cacti	带宽监控	
Smokeping	网络质量监控	
Capacity Watch	容量监控	
NetworkBench	全球网络性能监控	



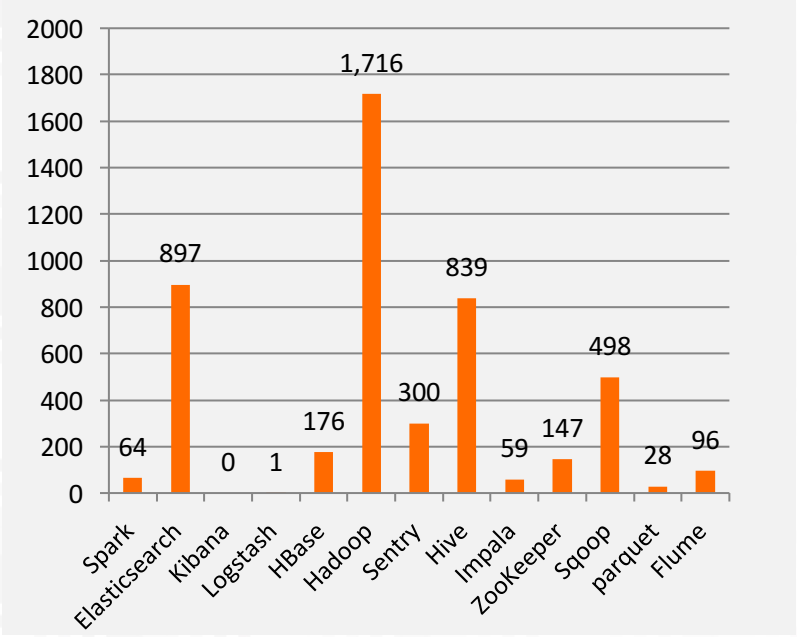


通过开源Cube计算增强，弥补三方报表工具短板，实现业务报表快速构建能力



扩展Kerberos和Sentry，实现全面的数据安全保障能力

漏洞修复

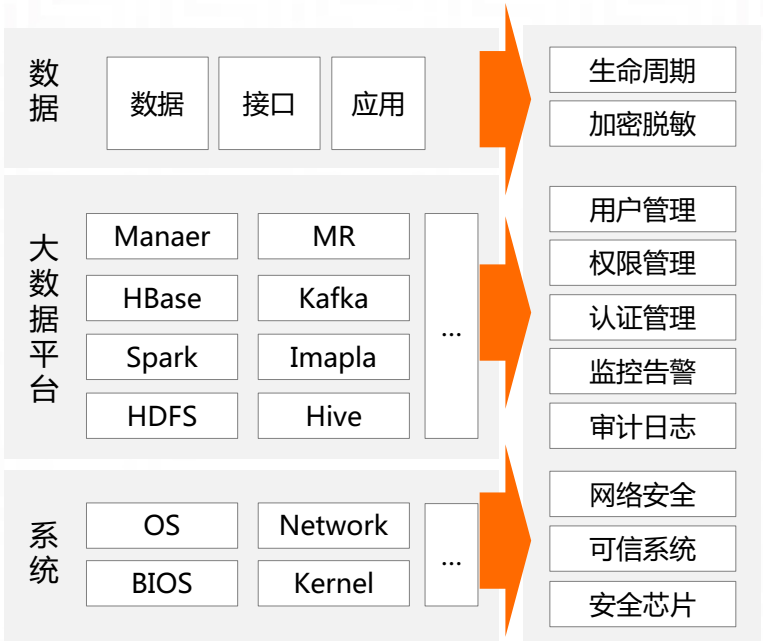


漏洞发现及修复

危险程序、弱安全配置、系统缺陷及渗透等多全方面漏洞检测与安全增强

依照ISO27001标准的构建安全体系，建立了完备的安全扫描和加固能力

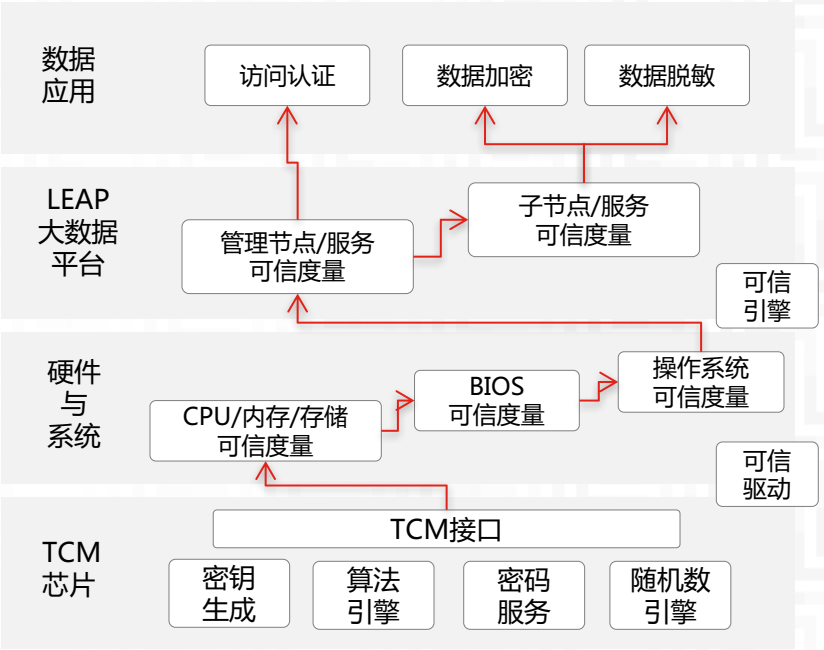
分层保护



安全框架

不是简单的权限管理与认证，而是在系统、平台、数据三层进行全方位安全保障，无安全短板。

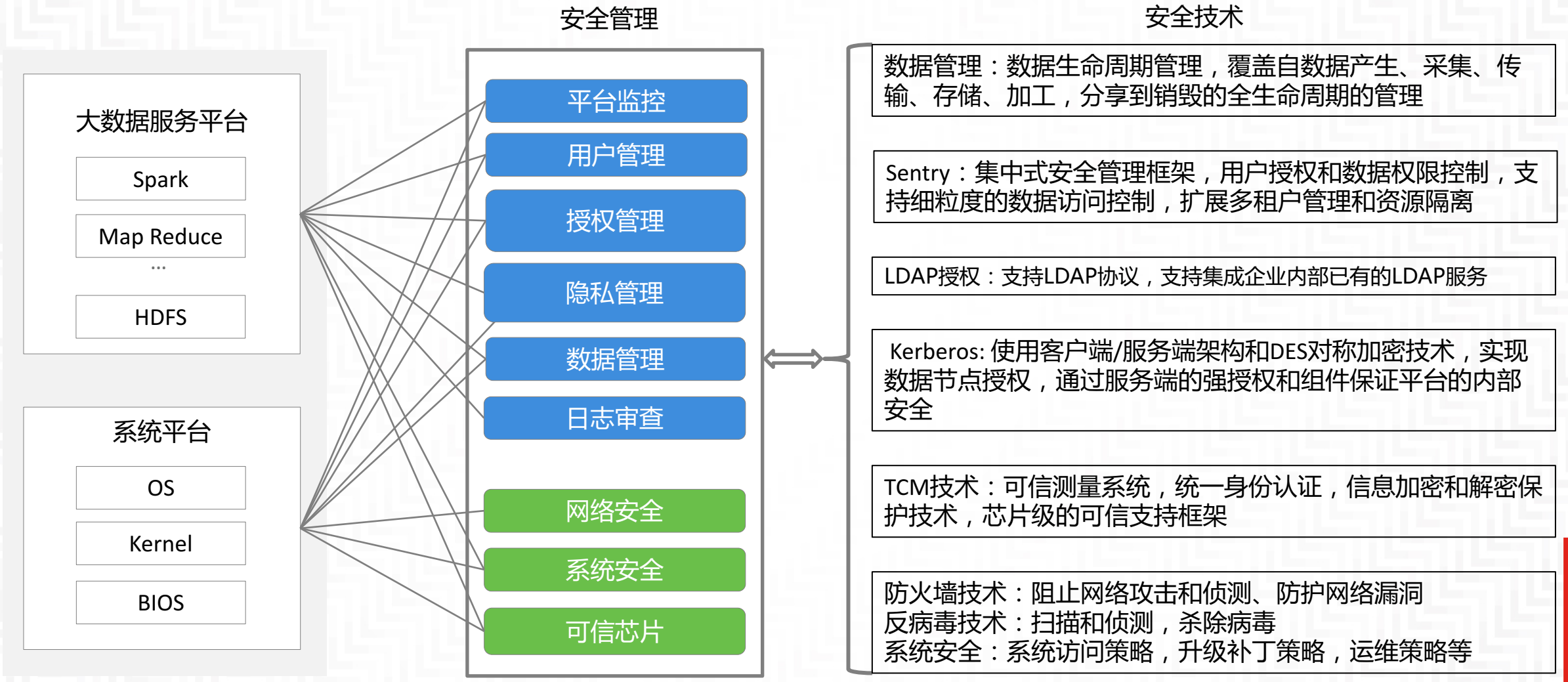
可信计算



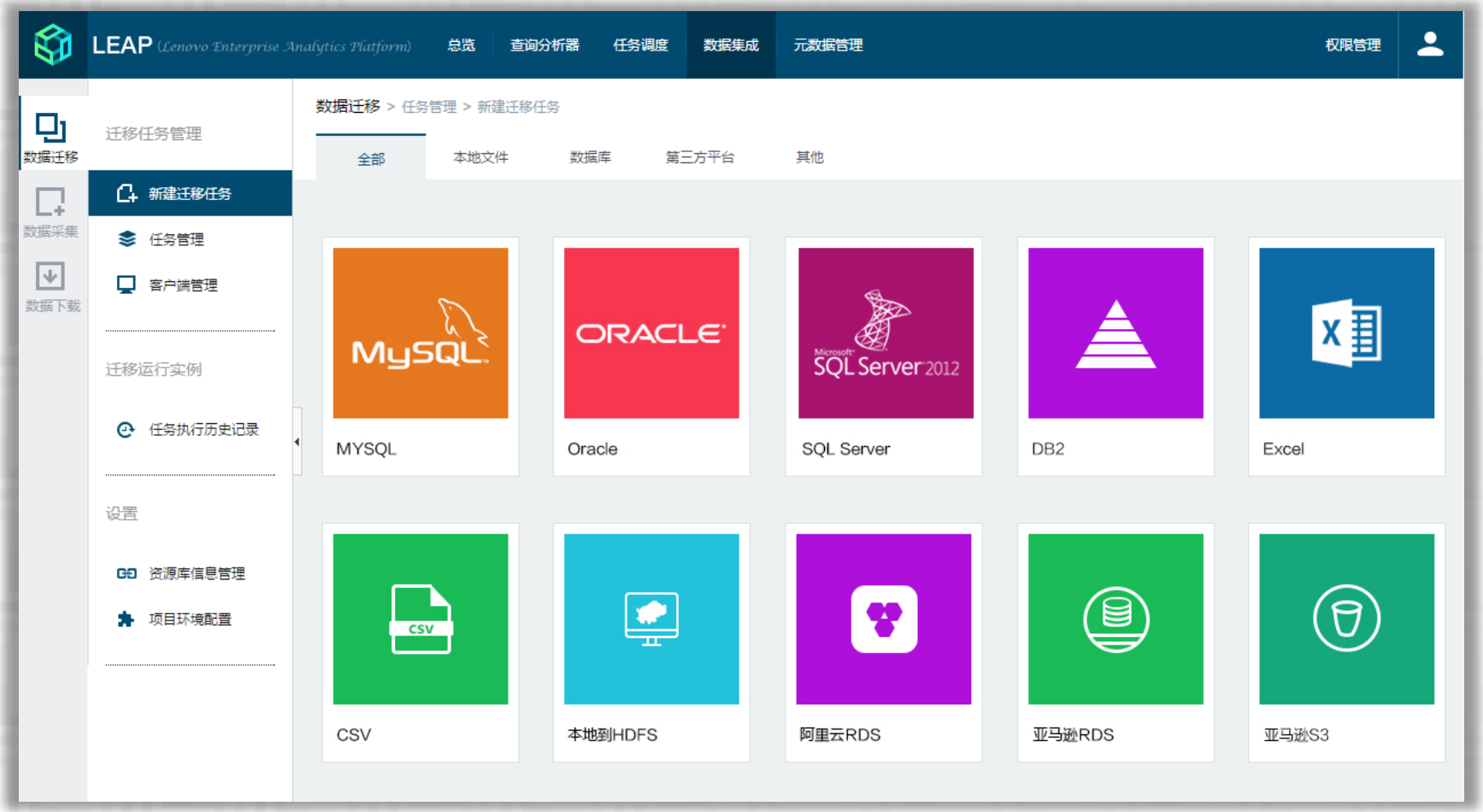
可信计算框架

第一个建立基于TCM/TPM安全芯片的大数据安全解决方案，形成行业壁垒和标准化组织深入合作，推动建立国家的大数据可信计算标准及联盟

覆盖硬件、系统平台到大数据服务的一体化的安全管理方案



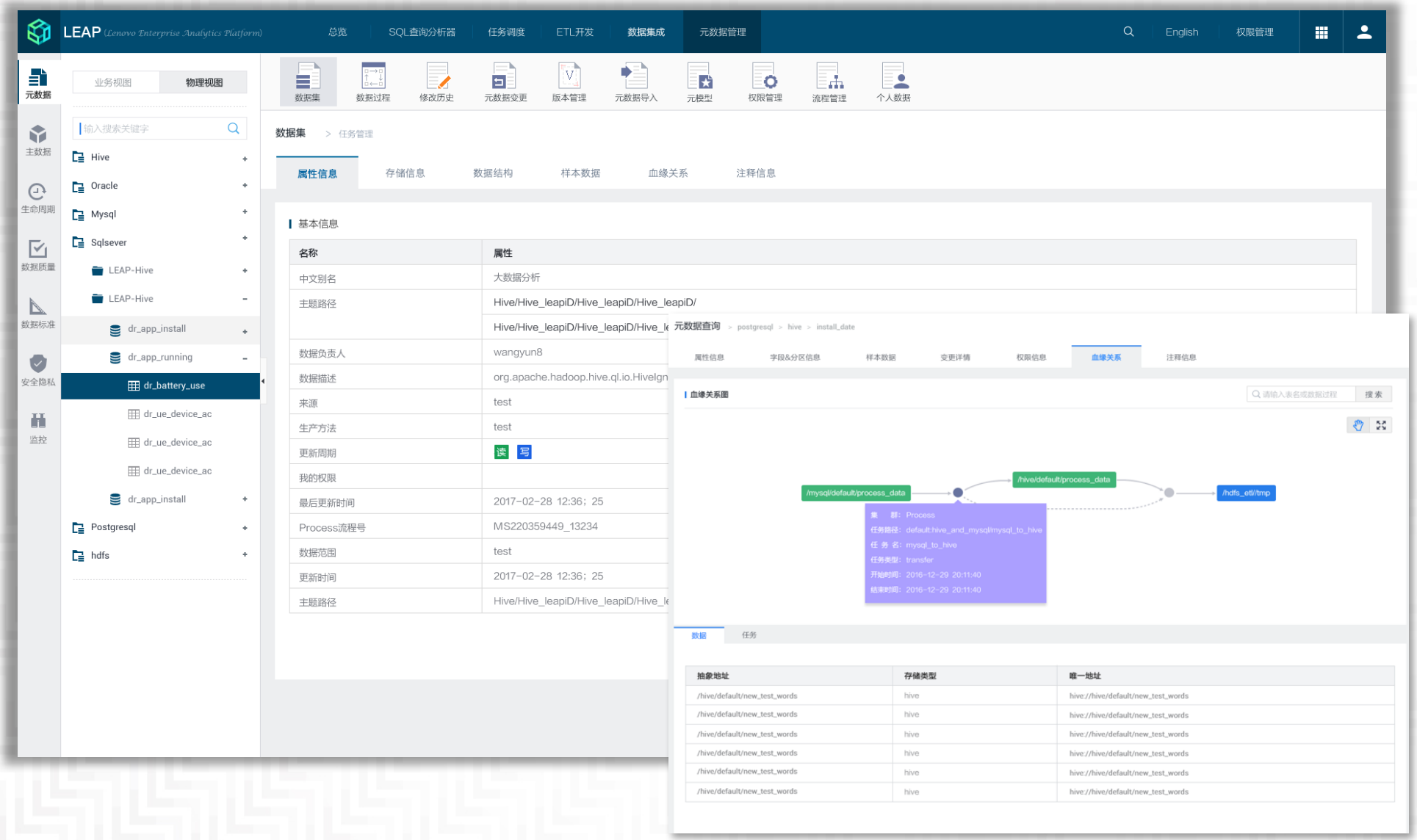
新的数据ETL工具，兼容各种企业信息系统，实现异构数据资产整合



数据集成平台：

- ✓ 全图形化化开发与管理界面；
- ✓ 丰富的ETL数据处理组件；
- ✓ 丰富的数据介入适配器，可对接各类商业产品，如SAP、Oracle等管理软件，及各类数据库；
- ✓ 丰富的接口支持，支持JDBC/ODBC、http、ftp、消息队列等多种数据传输方式；
- ✓ 支持Oracle、SqlServer、DB2等商用数据库，也支持MySQL、MongoDB、PostgreSQL等开源数据库，支持结构化和非结构化数据获取；支持XML、TXT等文本格式数据的处理与解析；
- ✓ 支持Hive，Spark，Impala，Habse等Hadoop生态技术及组件；
- ✓ 强大的开发环境，支持运行、调试、日志跟踪、结果预览，支持工程的导入、导出等；

元数据管理工具，实现数据资产的字典化管理，支持数据接口发布/分享

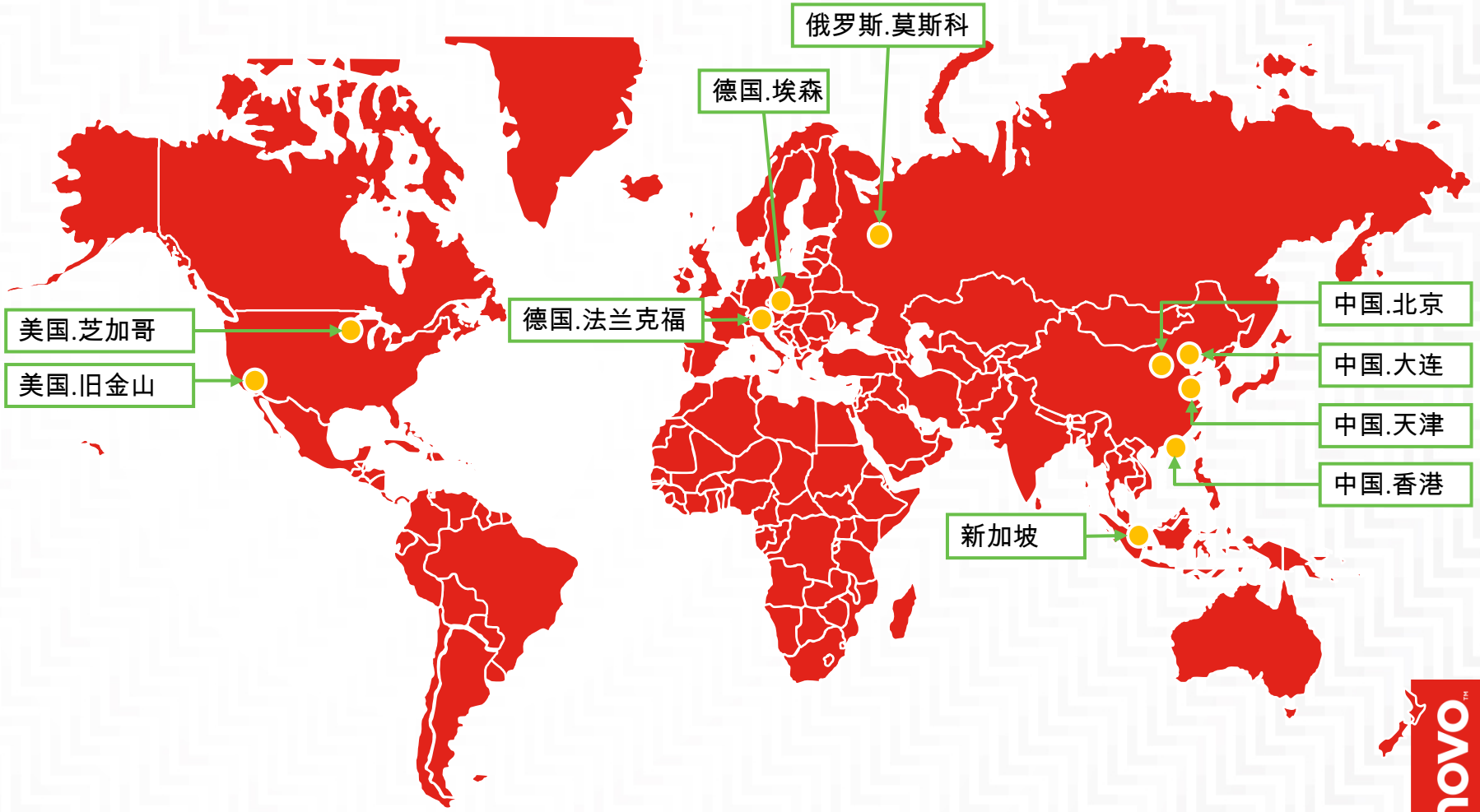


元数据平台

- ✓ **元数据管理：**提供LEAP平台内全部元数据信息的集中、可视化管理，实现对元数据信息的快速定位、查询与检索；
- ✓ **数据质量管理：**构建数据标准、数据质量校验规则及质量分析报告；
- ✓ **数据生命周期管理：**实现对数据的分级定位，从采集到销毁的全生命周期跟踪及管理；
- ✓ 血缘分析与影响分析

实现了5s内，对全球设备和用户进行实时追踪和系统重算的能力，构建了联想统一的全球数据湖

- 数据处理能力
  - 数据存储能力**突破1552亿**条记录，每日新增30TB数据
  - 追踪联想设备**突破2亿**，每月以1000万的数量增加
  - 总追踪**全球用户数突破6亿**，其中全球注册用户突破1.6亿，每月新增接近400万
- 硬件规模
  - **物理服务器突破2000台**，虚拟机实例突破7000个
  - **10个数据中心**在全球5个不同区域：中国，北美，亚太、欧洲、俄罗斯，覆盖全球160多个国家和地区
- 数据隐私保护
  - 所有数据本地保存，**遵循当地政府隐私保护法规**，数据加密并脱敏存储



平台2.0架构的主要问题

- 工厂里面的数据也需要整合，如何处理来自生产设备和工控系统的实时时序数据，并改进生产工艺
- 如何支持广泛的OLAP场景和爆炸式的自助分析需求，使得企业的IT资产得以复用
- 如何提供业务弹性，为突发任务调动足量的计算资源

平台3.0（2016~2018 ）
大数据平台突破，支持广泛的智能化场景

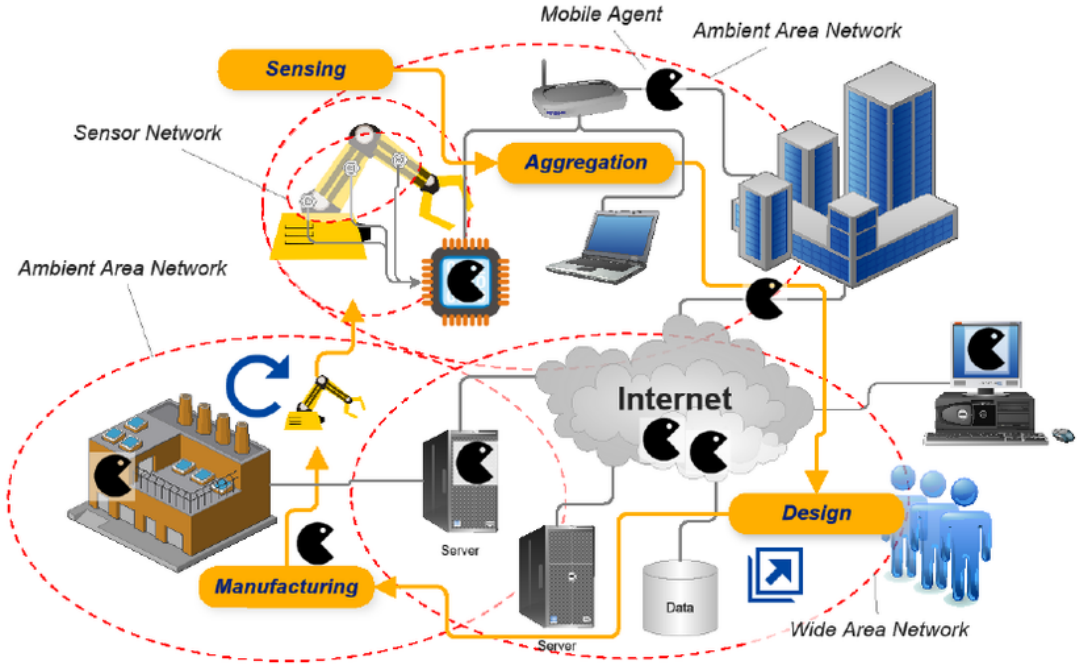
Lenovo

挑战：全方位的数据融合，海量的企业原有数据资产，实现从IT到OT的数据即时分析

•超过20000张数据表，6000名报表用户，2000+数据分析用户，来自数十个部门

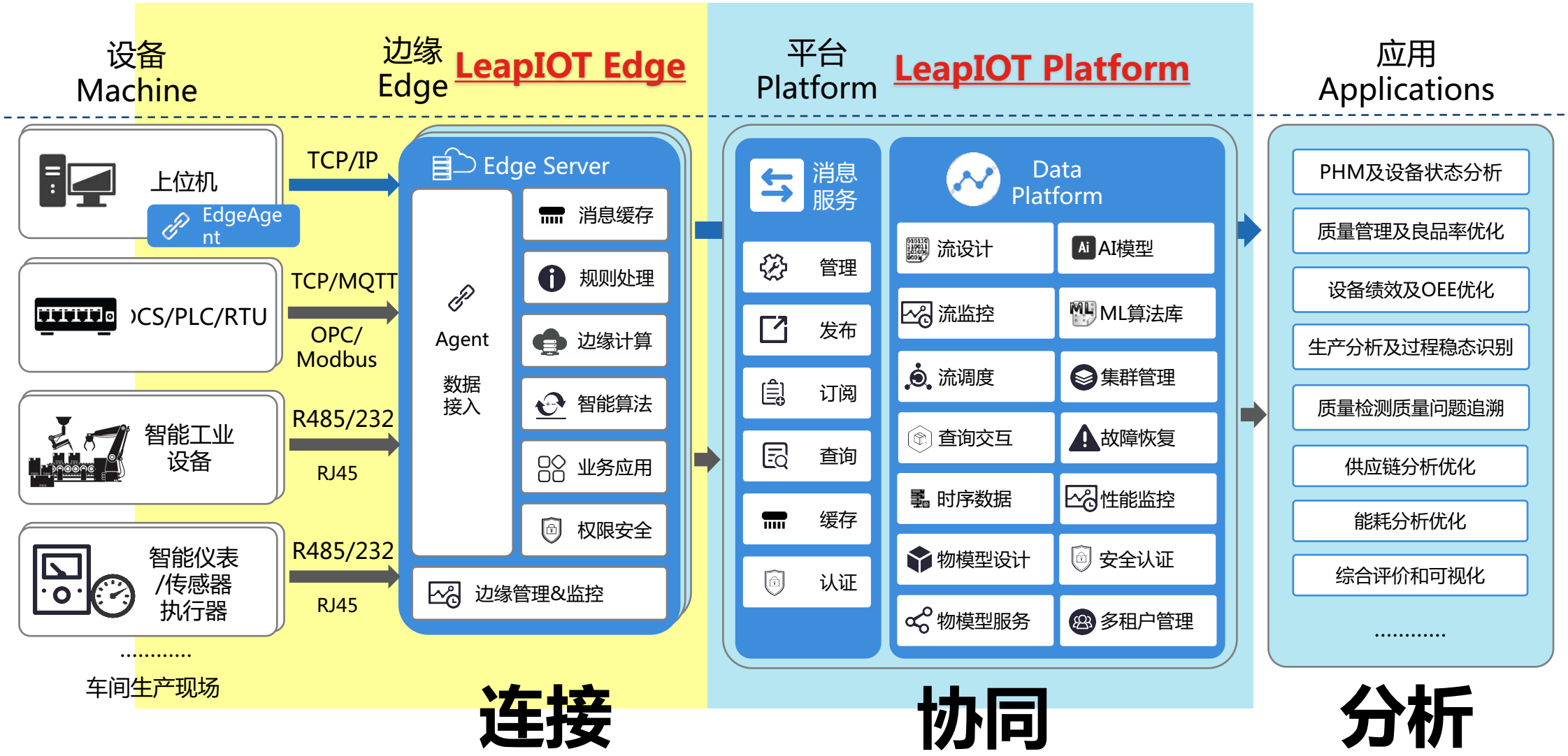
Data category	Data Objects				
	Master Data	M&S	Supply Chain	Service	Finance
Enterprise Data (IT system)	**Product Data** Component,BOM,CV,FG(SN ,IMEI),PH,FRU	**Pre-Order** 4P campaign, leads, activities, quotation, contract, Rebate	**Order execution** Fulfillment,shipment, logistics	**Readiness** Warranty(ibase), service offering	**Balance sheet** AR. AP,inventory Fixed asset,
	BP Supplier,B2B customer, B2C, UID(Lenovo id)	**Order loading** Offering, Sales order.	**Planning** S/D planning, MRP, FP, inventory	**Delivery:** Parts, field service (service order, records), online service(CC/IVR, social),complaint.	**P&L** Revenue,cost, expense,PTI,
	Analytic hierarchy BG, GEO, Region, Country, model(T/R)	**5S mgt.** Sell in, sell out, sell through, sell stock,SO.	**Procurement** PO,ODM/OEM/VMI/SRM		**Transaction** Billing, tax
			Manufacture MO,MES(Delivery,Quality,Effic iency)	**Cost:** Parts Inventory, services fees.	
Device Data (Device collection)	**Activation**		**Activities**	**Equipment & sensor (absence)**	
	Device info IMEI/IMSI/SN ， Device model, SW version...	**Activation info** Time, Geography(GPS/ base station,IP)	**Device Activities** Call time/duration, Power on/off, SW Installation, Touch Panel...	Robots, Factory Power, Factory ESD ， logistics GPS....	
APP Activities (APP collection)	**Lenovo Web Sites**		**Lenovo APP platform**(store/game/PC manager)		
	Web Activities Record Log on/off, Web ID cookies, Steps/times, favorite, A.d., conversion rate		**User Activities** Payment, Log on/off, Download, Stay time, key word... Launch/ Close, Feature Usage, Crash, Step , conversion rate		
3Rd party data (Scrawl or buy)	**Survey&Promo** Survey, Promotion(e.g., MIDAS email address)	**Social Media** **Social Media Record** Customer Comments, Product Evaluate, MI/ Governents (In Weibo, Forum, Blog...)		**3rd E-commerce(absence)** **JD, Tmall...Record** Customer, Payments, Orders, UE...	

•全球31个工厂，上百条产线
•年生产设备1.5亿台，每天40～100万台设备的生产能力

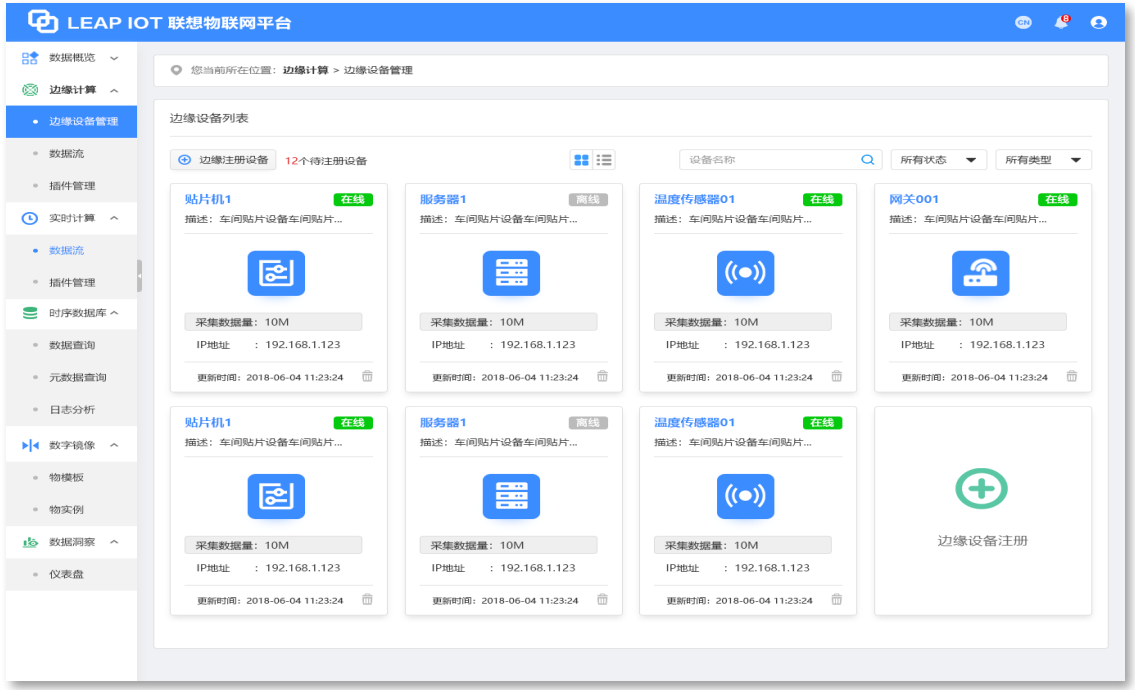


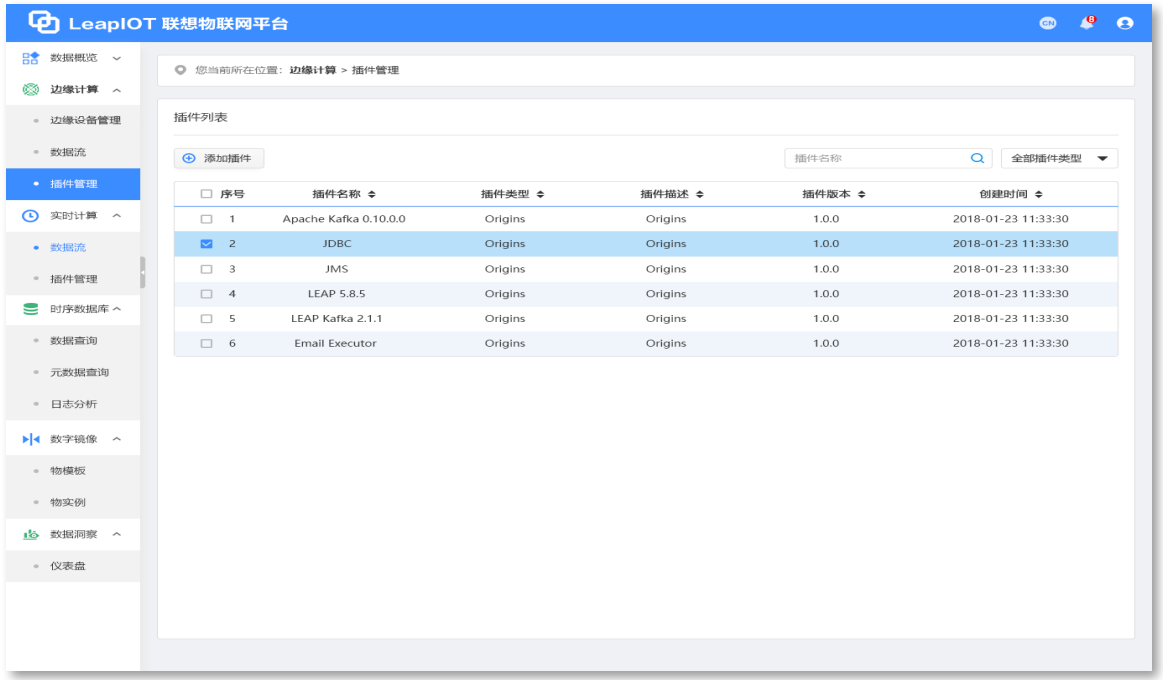
融合复杂的数据逻辑与存储、提供弹性的计算资源

LeapIOT 提供的边缘计算和平台两部分，分别解决连接和协同、分析的需求



LeapIOT Edge – 强大的边缘接入与灵活的计算框架





- **边缘接入**：适配OPC、Modbus、CAN、CoAP等十余种常见工业协议，并支持在线扩展；毫秒/秒级 传输及计算延迟；
- **设备管理**：百万级网关及千万级设备接入能力，涵盖智能仪表或传感器、边缘网关、边缘服务器；
- **设备监控**：实时监控边缘物理设备接入、数据吞吐、设备状态等；

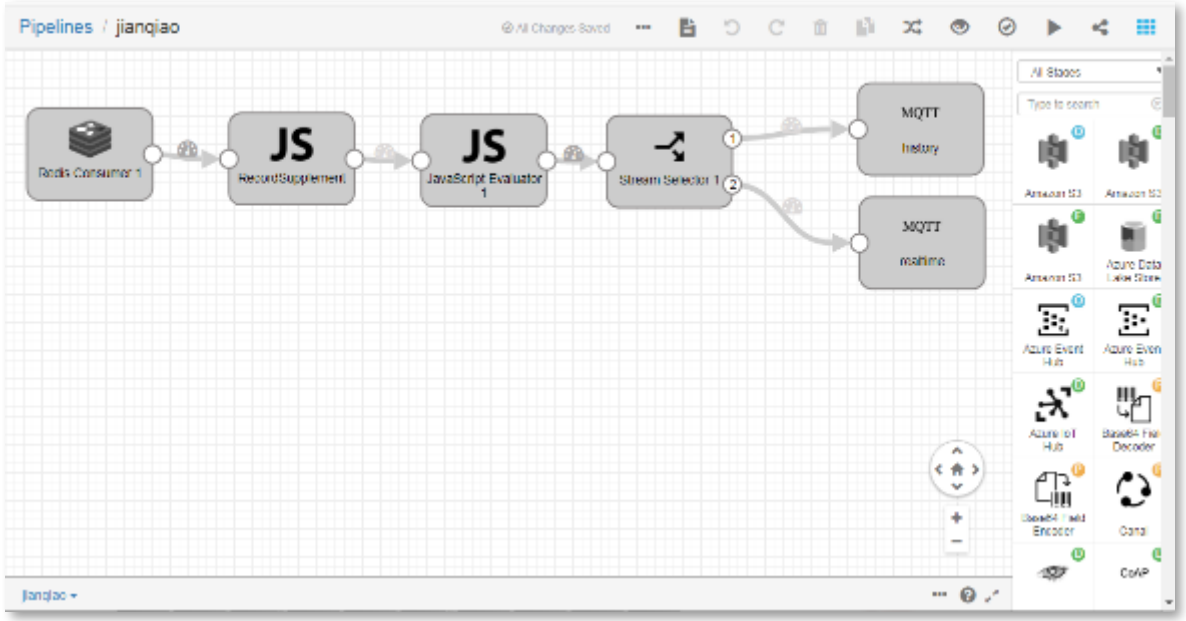
- **终端适配**：基于GO语言编程，最小10MB，可运行在ARM/X86环境，适配上位机、工业网关及各类型服务器；兼容主流ADLink、树莓派等硬件，可按需求定制软硬件；
- **灵活扩展**：独特的插件模式，在边缘端灵活扩展工业协议转换、数据预处理、算法脚本甚至是独立的存储与应用服务；

实时计算 – 可视化设计与高性能

数据流可以运行在上位机、工业网关、边缘服务器、大数据平台、公（私）有云平台

边缘上**单一边缘服务器**支持**100万条/s**数据处理能力，延迟5s以内，平台整体支持亿级设备数据同步与实时处理

在线设计每一条数据流及其处理逻辑



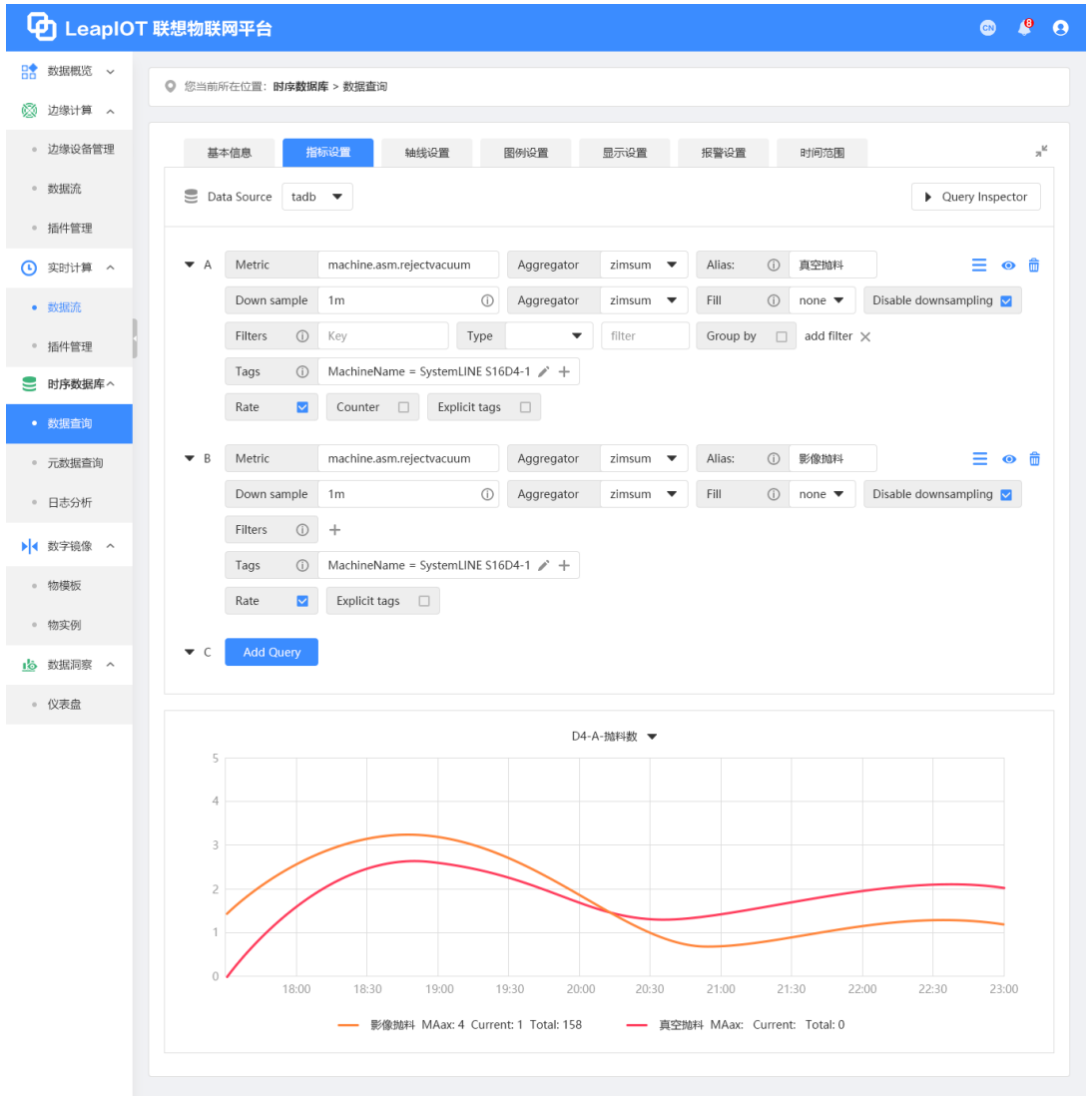
提供百种数据采集、存储、处理、转换、计算的调用组件，可视化设计、在线调试、发布并即时生效

在线监控每条数据流的性能与数据质量

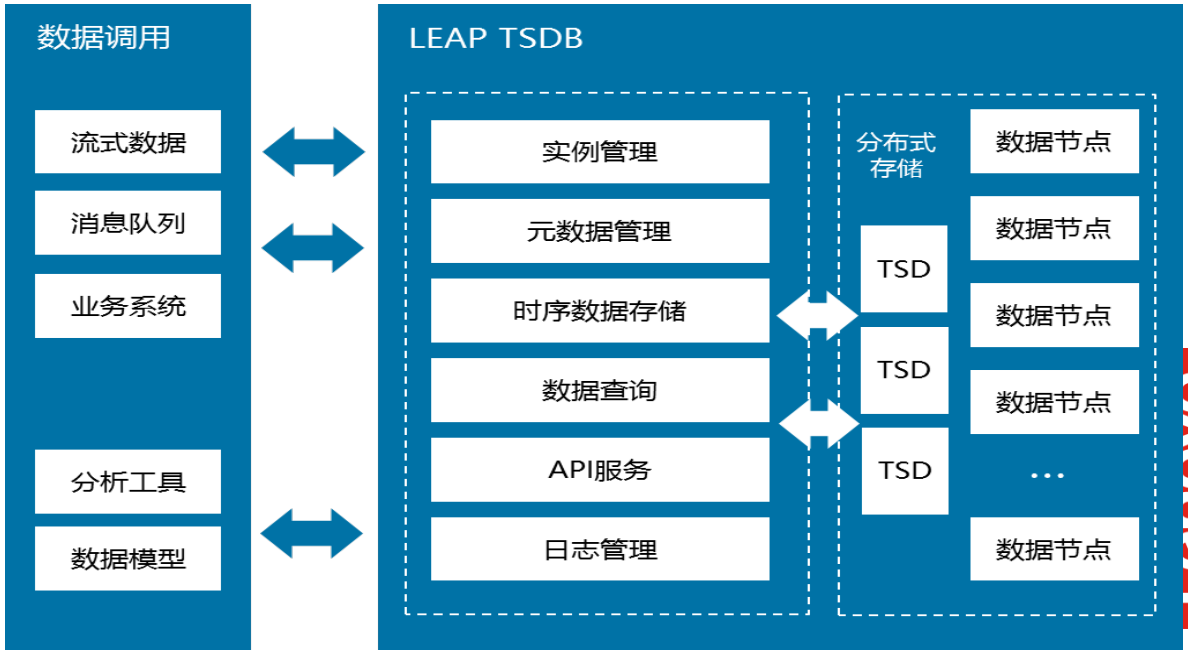
支持CPU、内存、网络、JVM等多种资源的使用监控，对数据处理过程、条数、错误数量等进行统计。

时序存储 – 建立海量分析基础

作为工业应用的数据核心系统，提供分布式、海量数据存储和分析能力

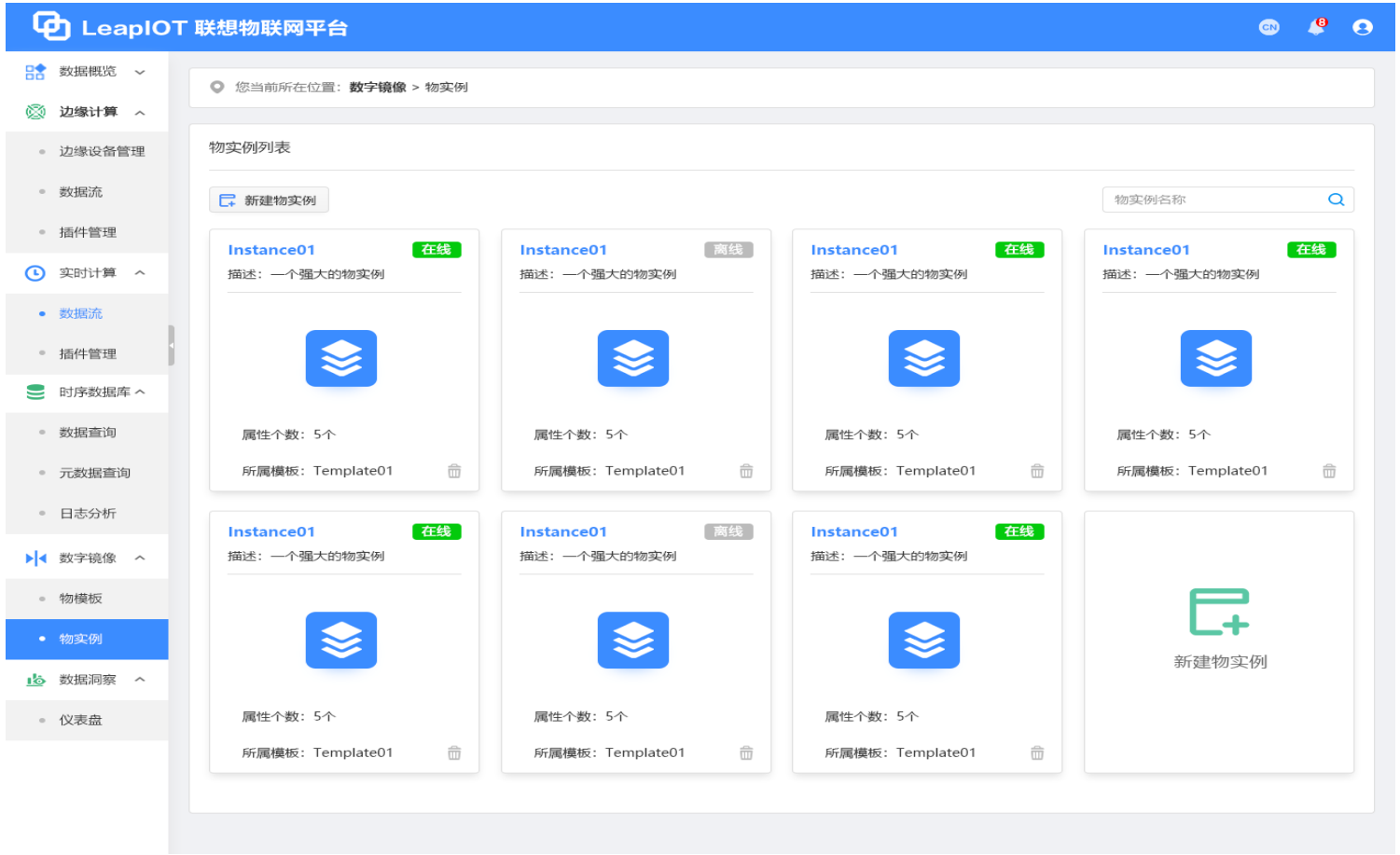


- ✓ **数据接入**：基于REST API提供高频、高速、大规模的采集点数据接入，低延时、高可靠；
- ✓ **数据存储**：基于分布式的、列式的底层存储机制，提供高压缩比数据存储方案与高性能保障机制；
- ✓ **数据查询**：提供基于标签的数据趋势分析与可视化查询工具，一站式展示时序数据变化；



数字镜像 – 建模，让系统再现现场

建立用户认识、构建和管理系统的基础



- ✓ **数字建模**：建立数字化物模型体系，构建物模板、物对象、物属性、将物理设备抽象为数字对象；
- ✓ **设备服务**：数字镜像成为物理设备的服务的载体，构建设备即服务的模式，任意对数字镜像的访问与操作都转化为对物理设备的访问与状态变更，通过抽象与服务代理方式完成工业应用与物理设备的解偶与业务重构；
- ✓ **设备智能**：与人工智能、机器学习工具与服务进行整合，在数字世界为数字镜像添加智能算法，在物理世界做设备智能改造；
- ✓ **设备可视**：对工业设备数字建模后的内容，通过数据分析工具或可视化工具，展现物理设备的2D、3D、AR/VR监视内容；

使用Spark，进行即时分析的性能优化挑战，如何达到10x以上的性能提升

挑战一： 很多在TPC-DS上有效的优化，在实际数据上效果不明显

挑战二：不同配置的服务器构成集群，CPU/内存/硬盘/SSD速度各异，导致整体集群利用率不高

挑战三： Spark 2.3引入了一些新特性，比如CBO打开后，对于很多SQL任务的执行计划无影响，无法带来宣称的性能提升

挑战四：Spark不理解元数据的含义和用途，无法有效的管理业务数据加载

挑战五：Spark多个短小任务连续执行时，大量系统开销用于重新加载数据和资源重新分配

挑战六：Spark SQL的SQL解析器对于很多在关系数据库（Oracle）上性能下降不多的脚本，执行效率低下

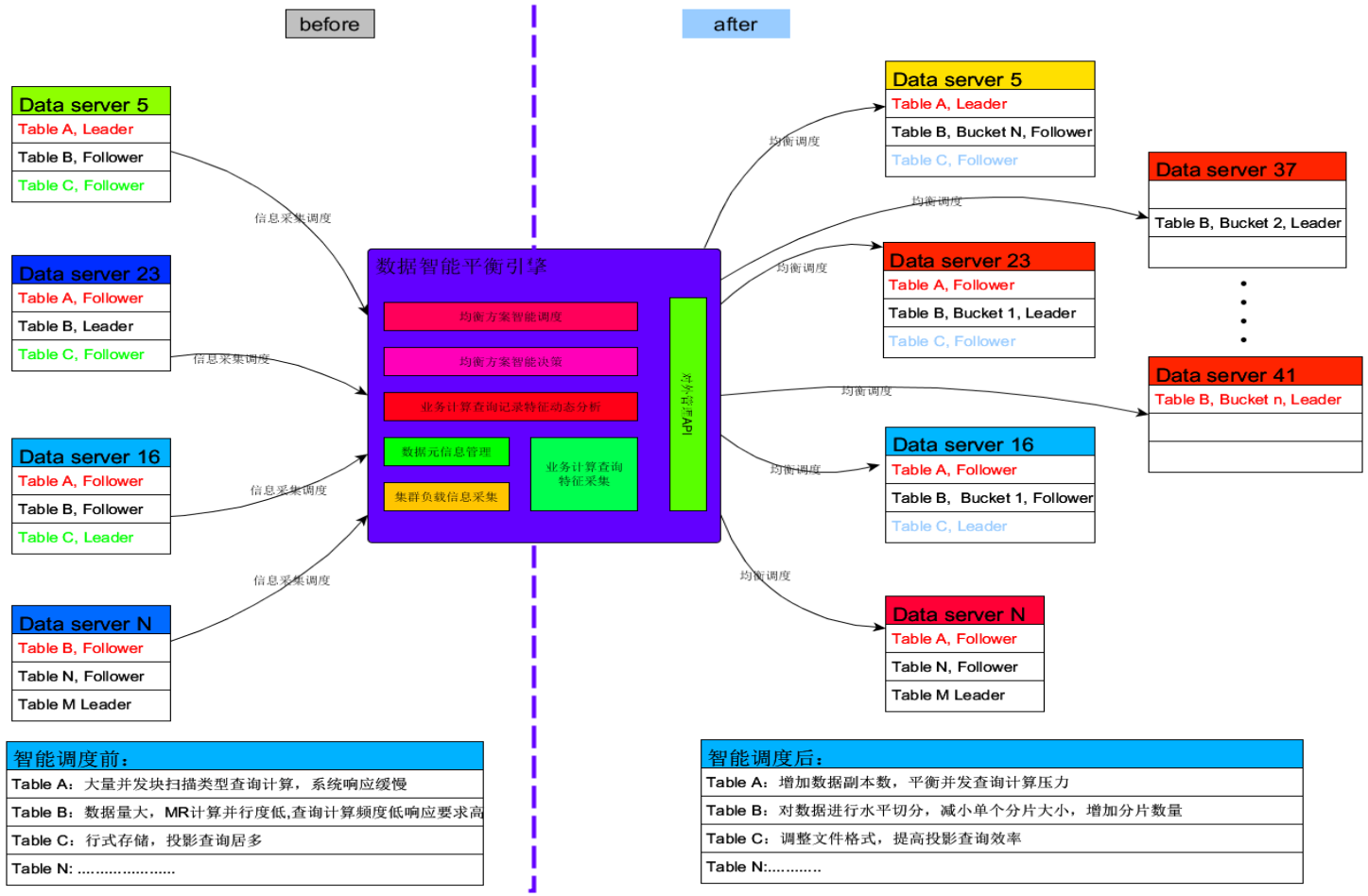
挑战七：如何充分发挥新硬件的优势，比如NVMe，RDMA，FPGA

数据不合理的分布和低效的内存管理，是限制Spark集群性能的最大挑战

问题一：在单表查询或多表Join时，需要频繁I/O，CPU利用率不高
问题二：不同配置的服务器，承担的计算任务不合理，集群等待某几台机器的计算结果返回
问题三： 热数据在新查询开始时被重新从硬盘加载，Spark并不知道什么是业务最常用数据
问题四：多并发的快速查询任务，把大量的时间消耗在计算准备和资源调度上

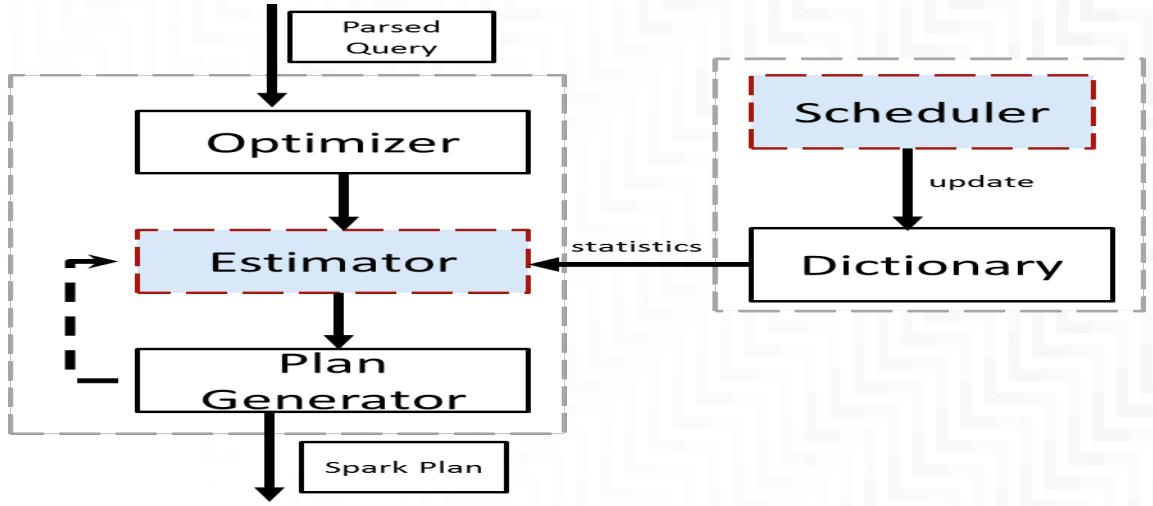
数据智能平衡引擎，配合元数据管理引擎，理解业务，最优化数据Balancing策略

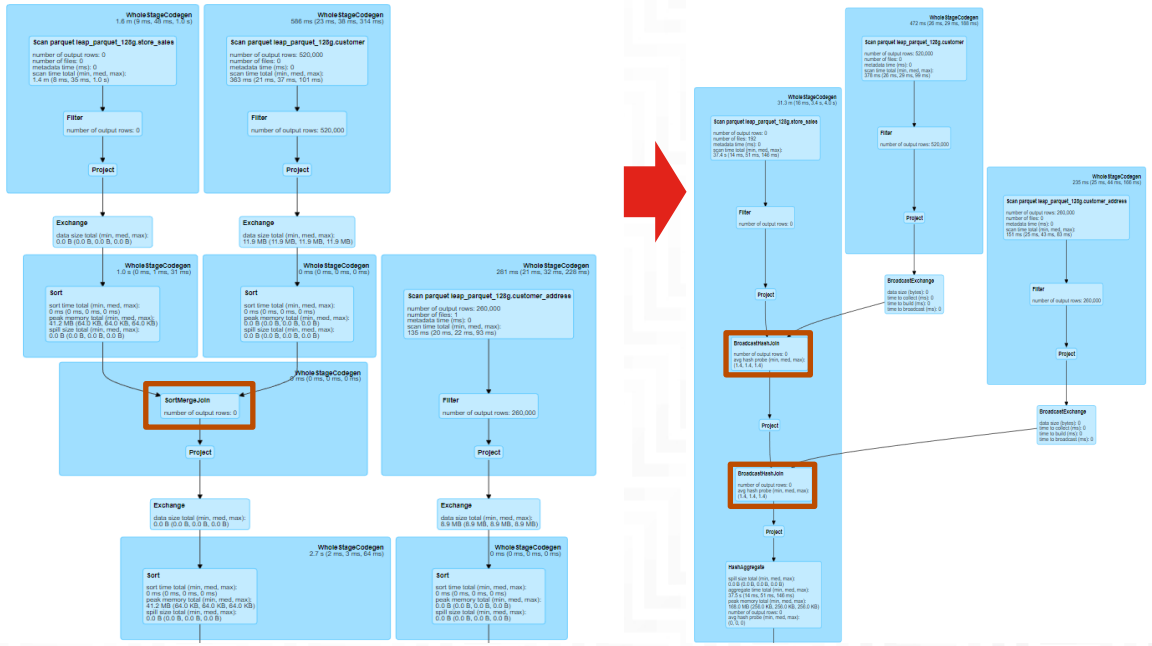
- 结合业务数据特点，通过对数据的查询特征（查询频次、随机读写/批量扫描等）及元数据特点，智能的制定出数据再均衡（分布、负载）和内存加载方案
- 适当增加关键维度表数据的冗余，降低Shuffle比率
- 对于事实表数据按照服务器的性能，进行均匀分片，提升并行度和ShuffleHashJoin
- 通过自动化任务调度平台，对数据分布、副本、存储格式等方面进行主动调整，使系统的负载均衡及性能响应达到最优



根据业务数据和分析自动调整执行计划，提升CBO的效率

- 优化后的Cost Based Optimizer：
  - 引入调度器，根据用户SQL访问特征，按需/定时刷新统计信息
  - 根据集群资源与任务执行Stage实时统计信息，动态调整执行计划
  - 实时计算关键信息，规避数据更新统计信息丢失问题
  - 根据业务数据和场景特性引入更多因素，更好支持符合业务场景的场景的代价评估





构建全新的SQL语言查询生成工具，根据元数据信息生成最适当的SQL脚本

COUNT(DISTINCT)时，若有数据倾斜，非常影响性能

```
604-2370-hive

CREATE TABLE IF NOT EXISTS
d_lenovopcmanager.lenovidcnt
(
   p_event_date  string, lid_cnt bigint
);

INSERT OVERWRITE TABLE d_lenovopcmanager.lenovidcnt
SELECT
    date_sub(from_unixtime(unix_timestamp()),1) AS p_event_date,
    COUNT(DISTINCT(lid))
FROM
    d_lenovopcmanager.lenovoid
WHERE
    event_action='F0000' and
    p_event_date<=date_sub(from_unixtime(unix_timestamp()),1);
```

23’21“ → 1’43“

两步操作，先进行GROUP BY再进行COUNT()

```
SELECT
    '2018-06-06' as  p_event_date,
    count(1)
FROM
    (SELECT
        lid
    FROM
        d_lenovopcmanager.lenovoid
    WHERE
        event_action='F0000' and
        p_event_date<="2018-06-06"
    GROUP by
        lid
    )t0
```

列转行再行转列查询效率低

```
    )lid_mapping
JOIN
    (SELECT
        collect_set(history_domains) AS history_domains,
        mac
    FROM
        (SELECT
            history_domains ,
            mac
        FROM
            d_upe.gdm_pcmsdk_browser_history_modify_di
        WHERE
            p_event_date='${date}'  --2003095601(20亿条数据)
        )history_names
    GROUP BY
        mac
    )his_names_list
        (his_names_list.mac=lid_mapping.lps_did)
    GROUP BY
        his_names_list.history_domains,
        lid_mapping.lid
    )lid_history
    lateral view explode(history_domains) lid_domains AS history_domains_name
GROUP BY
    lid_history.lid,
    lid_domains.history_domains_name
```

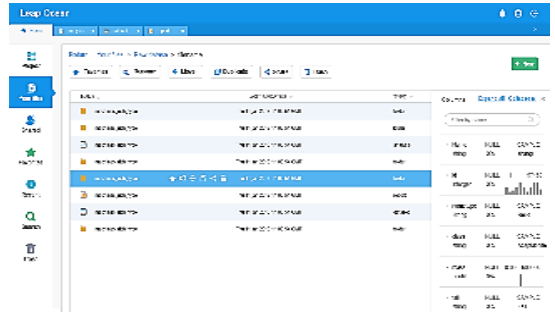
分别进行两次行列转换

40’ → 6’

先做表连接，再转换行列

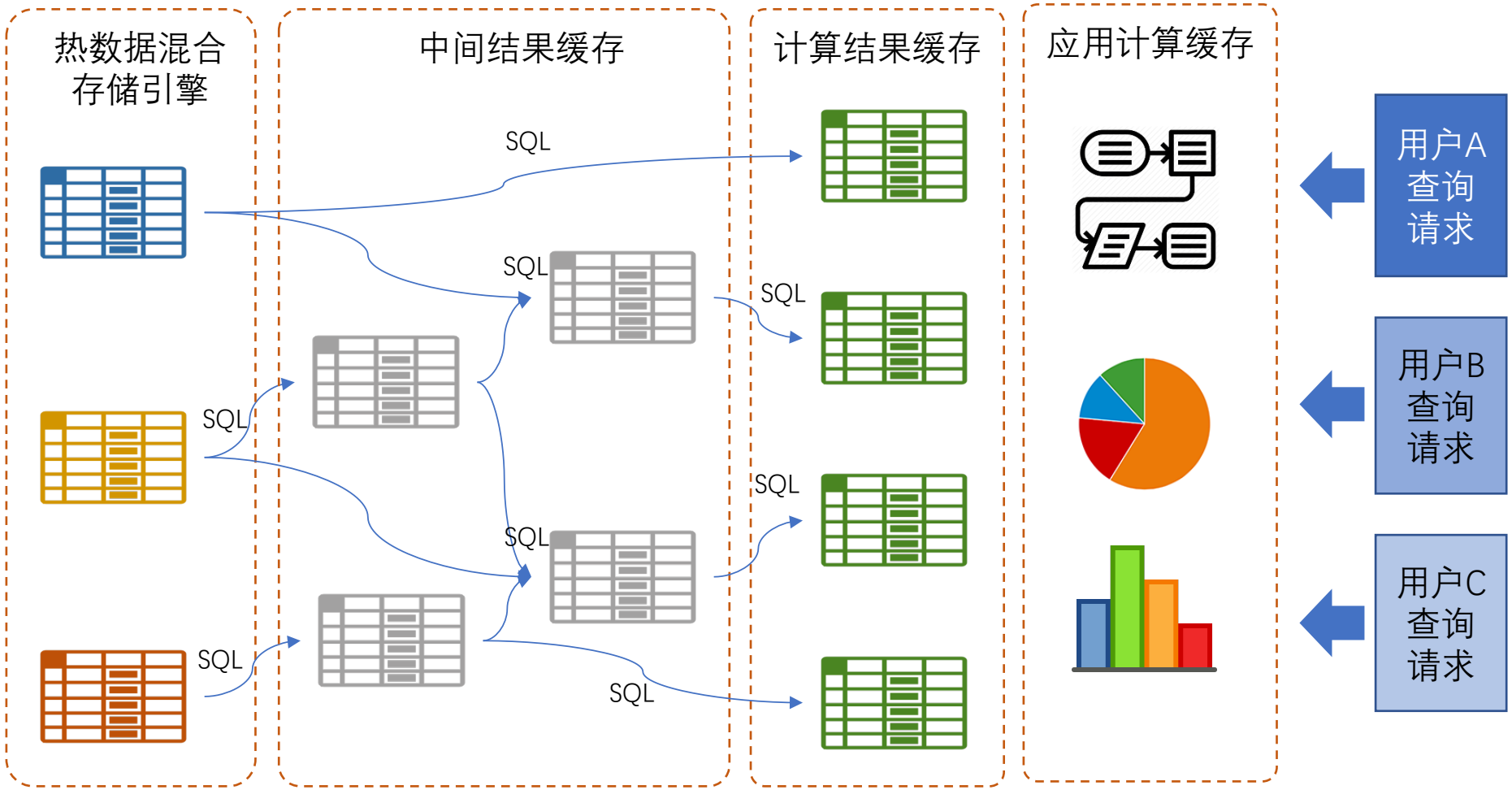
```
INSERT OVE                                                    id_di
PARTITION(
SELECT
    lid_ma
    his_na
FROM
    (SELECT
        lps_did,
        regexp_replace(lid,'"','') AS lid
    FROM
        d_lenovopcmanager.idmap_structure_total
    WHERE
        p_event_date='${date}' AND
        length(lps_did)=12 AND
        lid IS NOT NULL
    GROUP BY
        lps_did,
        lid
    )lid_mapping
JOIN
    (SELECT
        history_domains ,
        mac
    FROM
        d_upe.gdm_pcmsdk_browser_history_modify_di
    WHERE
        p_event_date='${date}'
    GROUP BY
        mac,
        history_domains
    )his_names_list
ON
    (his_names_list.mac=lid_mapping.lps_did)
GROUP BY
    his_names_list.history_domains,
    lid_mapping.lid
```

- 高质量的代码可遇不可求，任意一行低质量的代码将会让一切的优化成为性能消耗的饲料
- 不依赖于“高阶”DBA的数据系统才有普适的优化可言



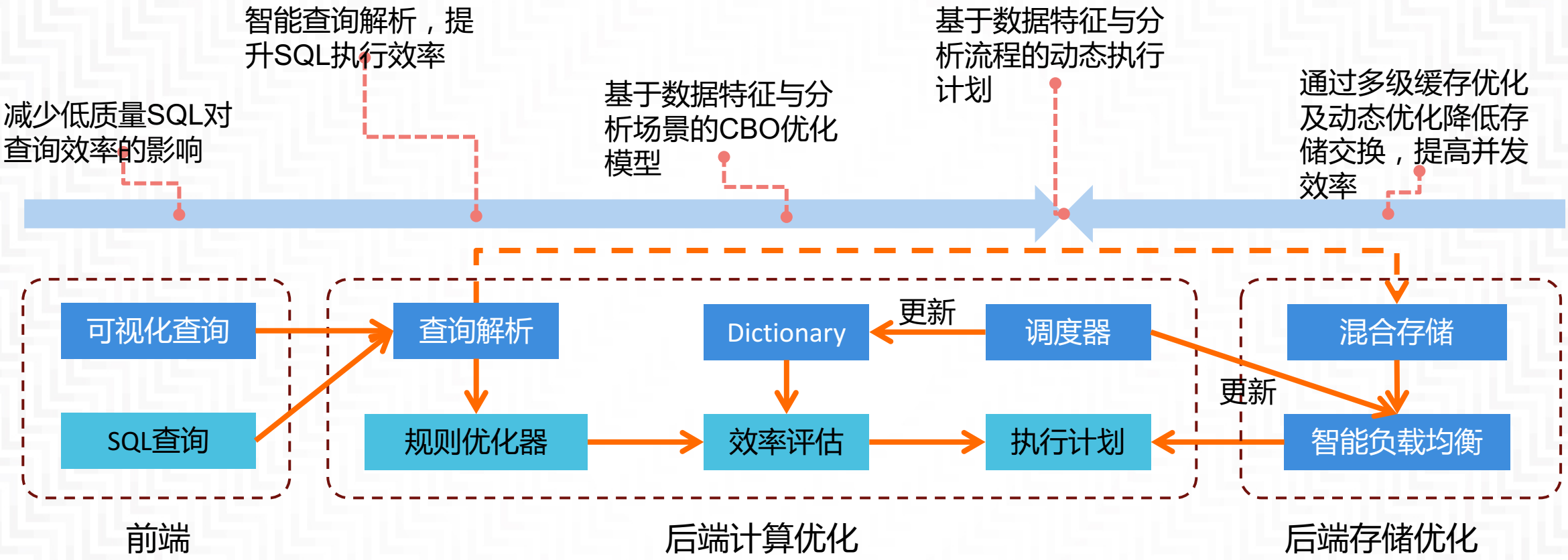
多级缓存管理与调度，提升组用户的SQL查询并行度

统一分析计算调度和缓存管理引擎负责从应用层到计算引擎和存储引擎的多级缓存管理与调度，以业务分析价值和计算资源消耗价值策略来动态地评估和调度管理多级缓存，维护缓存数据的生命周期。高度复用计算产生的结果既提升了同类分析和计算的效率，同时释放更多计算资源供其他计算使用。统一的缓存调度架构保证了数据的一致性。



- 多级缓存，动态评估，统一调度。
- 基于数据价值评估的计算结果缓存与中间结果缓存，减少计算，加速查询响应

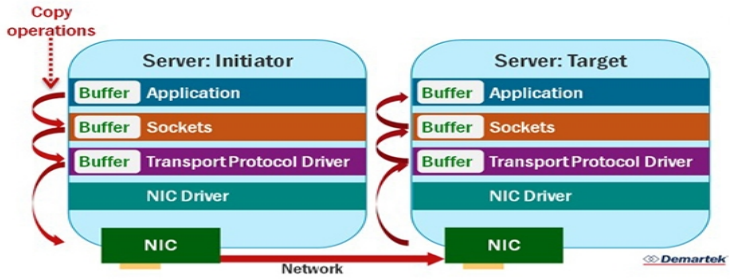
即时分析时，CPU/IO/Networking在分布式集群上满负荷运行，比缺省Spark实现效能提升10x~50x



- 总结
  - 从查询分析场景与数据特征两个方向，提升即时分析效率
  - 通过构建新的元数据管理引擎，提升Balancing和CBO的效能

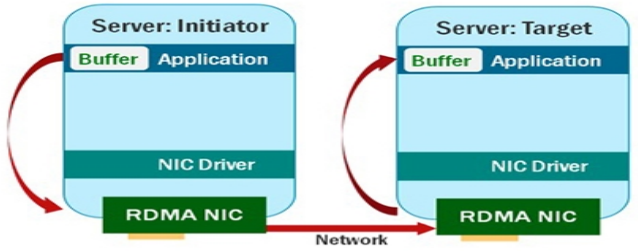
下一步：CPU是分析的瓶颈，提升CPU的利用率，进一步提升性能的关键

通过引入RDMA降低网络数据交换对CPU的消耗与性能损失



传统TCP Socket数据传输

- 跨节点数据数据传输需经过6次Copy
- 需操作系统介入
- 延迟波动较大



RDMA数据传输

- Zero-Copy
- 不占用系统内核时间及IO
- 亚秒延迟，可硬件流控

RDMA Shuffle Server端

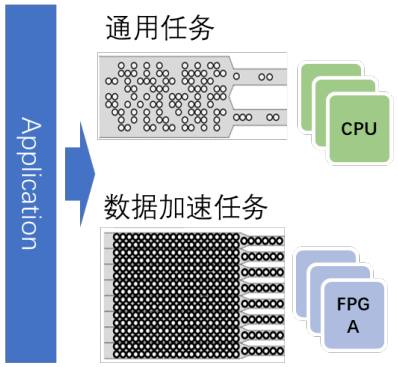
- 0 CPU
- Executor端GC不再影响Shuffle传输
- No Buffering

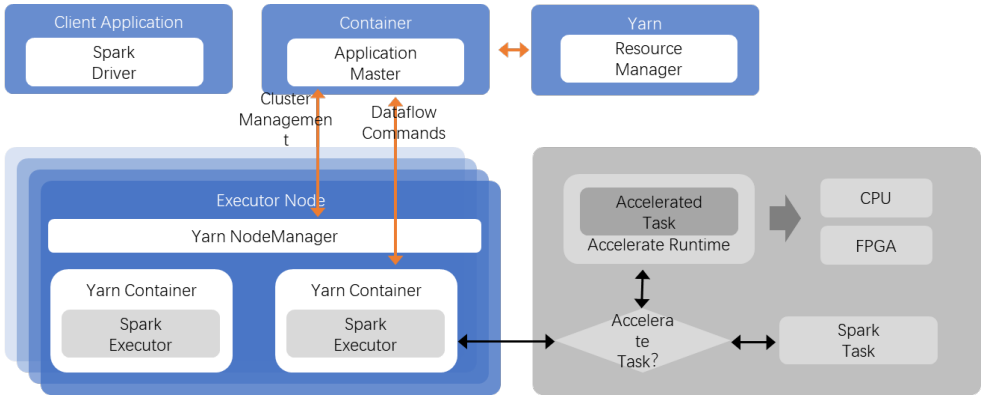
RDMA Shuffle Client端

- 即时传输，无等待
- 减少不必要的消息传输量
- 远端Block的直接无阻塞访问

实际业务测试中，有10%～20%的性能提升

通过FPGA降低特定计算任务（数据压缩/解压）对CPU的消耗

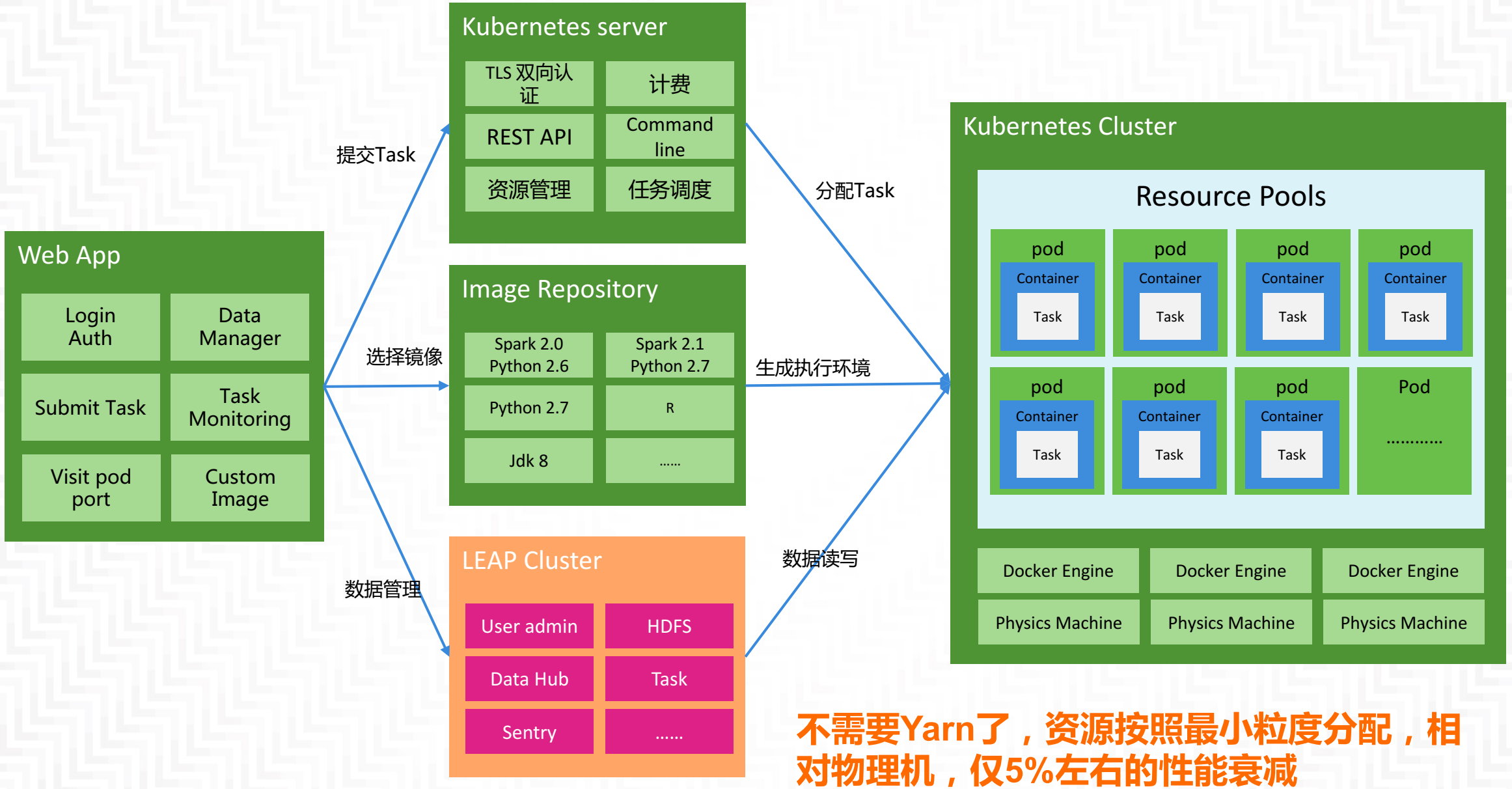




利用FPGA吞吐量和并行处理的优势提升数据加速能力。

采用CPU+FPGA混合计算模式，基于Spark架构进行扩展，对特定任务利用FPGA硬件进行加速处理。

使用Docker技术，轻量化Hadoop，使其支持业务弹性计算



不需要Yarn了，资源按照最小粒度分配，相对物理机，仅5%左右的性能衰减

说到最后

- 联想大数据平台的演进来源于一个传统企业的摸索实践，走了很多弯路
- 大数据开源技术处在快速发展的过程中，仍有很多不足，需要大家一起贡献力量
- 大数据和人工智能技术可以很好帮助传统企业转型升级，这一切刚刚起步，这将是我们中国人的机会

THANK YOU

DAKUJEM DANK BEDANKT MERCI TAKK 谢谢

ありがとう СПАСИБО GRACIAS DZIĘKUJĘ DANKE

OBRIGADO БЛАГОДАРЯ GRAZIE תודה GRACIAS

Lenovo™