

Tech World '18

联想大数据产品的优势和技术创新

张成松 | 2018-09-27

联想

Lenovo

完整的大数据产品体系

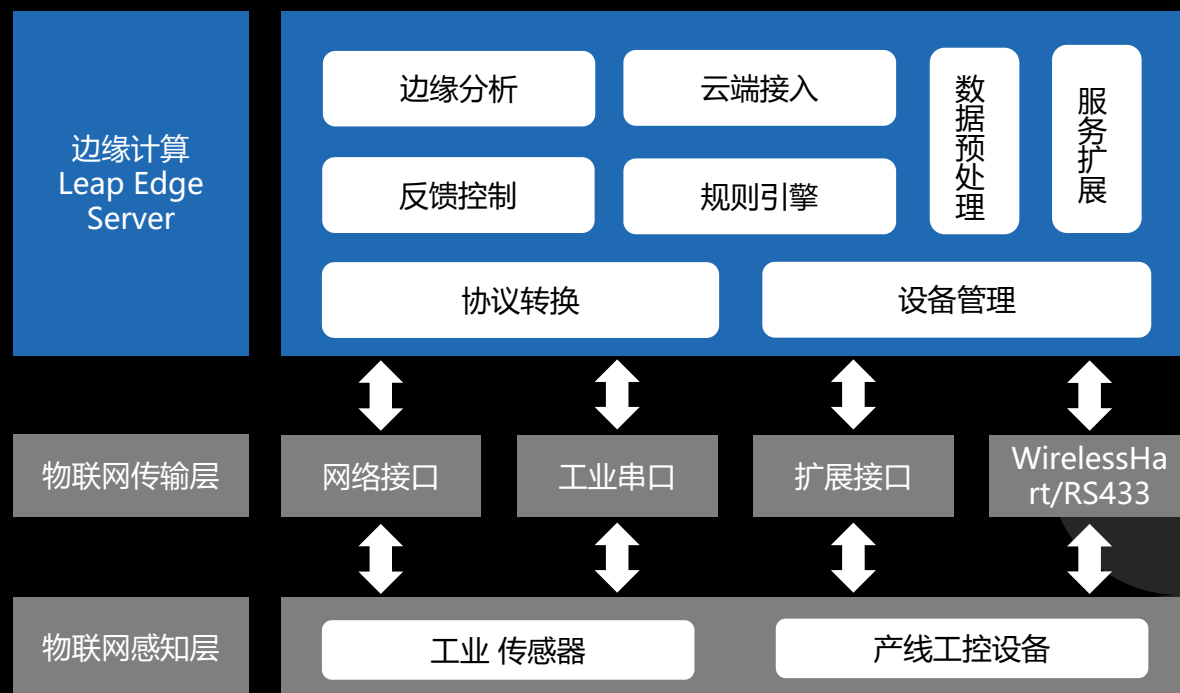
联想工业大数据技术体系



LEAP EDGE: 基于物联网的边缘计算

- ✓ 支持**近20种**物联网数据协议/常用接口，快速解决工业设备数据接口对接、协议兼容解析等难点
- ✓ 高性能传输处理，提供**500MB/秒**级别的工业消息处理能力
- ✓ 支持数据传输的多协议，复杂网络环境的要求，支持**断点续传、带宽限制、数据压缩加密**等多种传输要求
- ✓ 提供**工业数据流设计工具**，可视化配置从边缘设备到时序数据处理、存储，以及工业智能计算的全过程
- ✓ **边缘分析算法集成**：稳态、正态、趋势、周期等
- ✓ 边缘计算可在工业网关、工业计算机以及服务器等**多种环境中运行**，贴近现场设备端，在复杂工况下提供实时智能分析能力

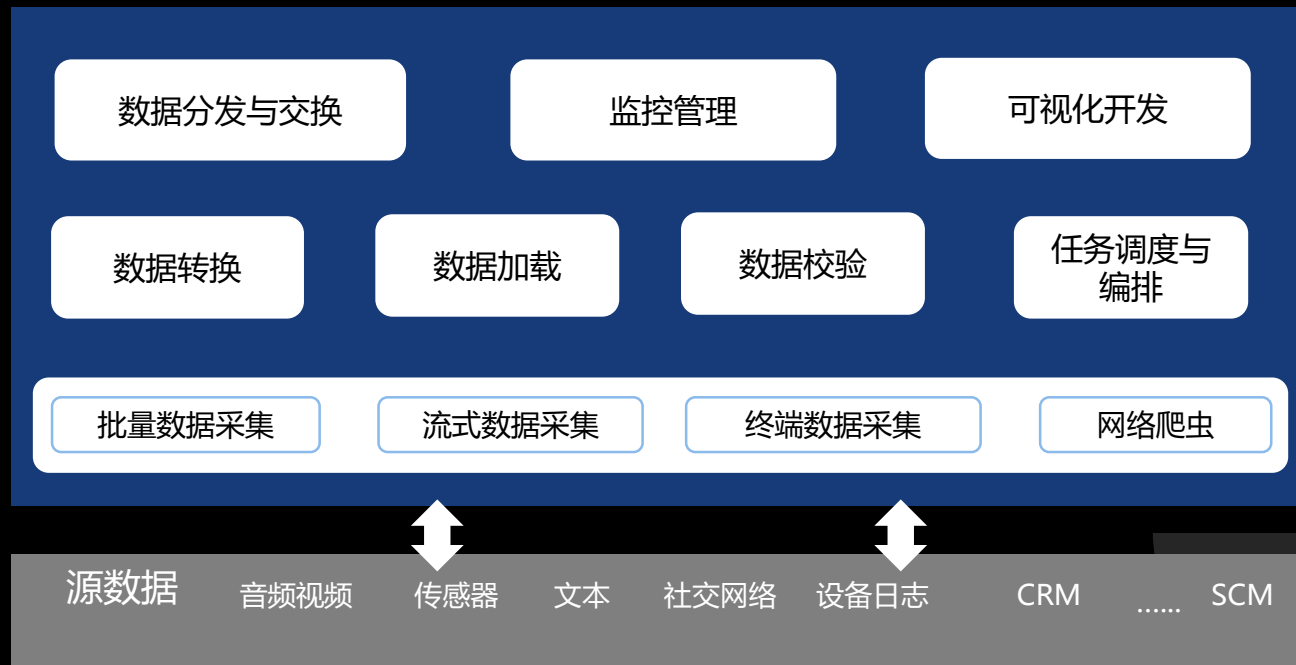
远端设备的感知、计算、分析、管理、控制



LEAP DATAHUB：多源异构数据的高效集成

- ✓ 提供**50种**数据采集接口，包括：各种主流的关系型数据库如mysql、oracle等、第三方云平台、文本文件、HDFS文件等。
- ✓ 支持文件批量采集，数据库迁移单个任务最大支持**500张表导入**
- ✓ 灵活的流式数据处理引擎，支持**20多种**的流式数据源，自定义触发任务条件，在实时传输过程中进行数据处理
- ✓ 丰富的**图形化界面设计**和监控数据采集过程执行状态
- ✓ 灵活的数据清洗规则配置，支持记录级的多字段的**混合运算、过滤、去重、清洗、复杂条件过滤**等，实现数据的标准化
- ✓ **统一的作业调度及监控功能**，前台提供多种任务逻辑的组合处理及配置

多元、异构、实时，集成企业内部各种数据源



LEAP HD：高可靠高性能的大数据计算平台

- ✓ 深度集成与优化近**20种**组件，自研**10余种**实用工具，功能丰富，满足多样化场景要求
- ✓ 支持**最全SQL语法、存储过程**，易于使用和快速实现应用分析
- ✓ 支持**分布式事务处理**，是业界少数拥有此能力的厂家之一
- ✓ 基于联想实战经验的混合云架构优化，满足不同规模客户的需求
- ✓ 基于Docker容器技术的**弹性计算资源分配**和**多租户管理能力**，满足大规模弹性使用
- ✓ 支持**2000多台服务器**、10个数据中心的**12PB**数据稳定运行实践，支持业务每天20000+任务处理、4.3PB/天的计算量，**99.99%的一流可靠性**

获得奖项：

1、完全通过**数据中心联盟(DCA)能力测试**，在功能、运维、多租户、可用性、安全性、兼容性、扩展性等7项指标完全满足评测要求



2、荣获第**21届中国国际软件博览会金提名奖**



大数据计算平台
LEAP HD

数据仓库



MPP

分布式
计算框架

No SQL



一站式运维

分布式存储
系统



HDFS

统一资源
调度管理

内存数据库



图数据库



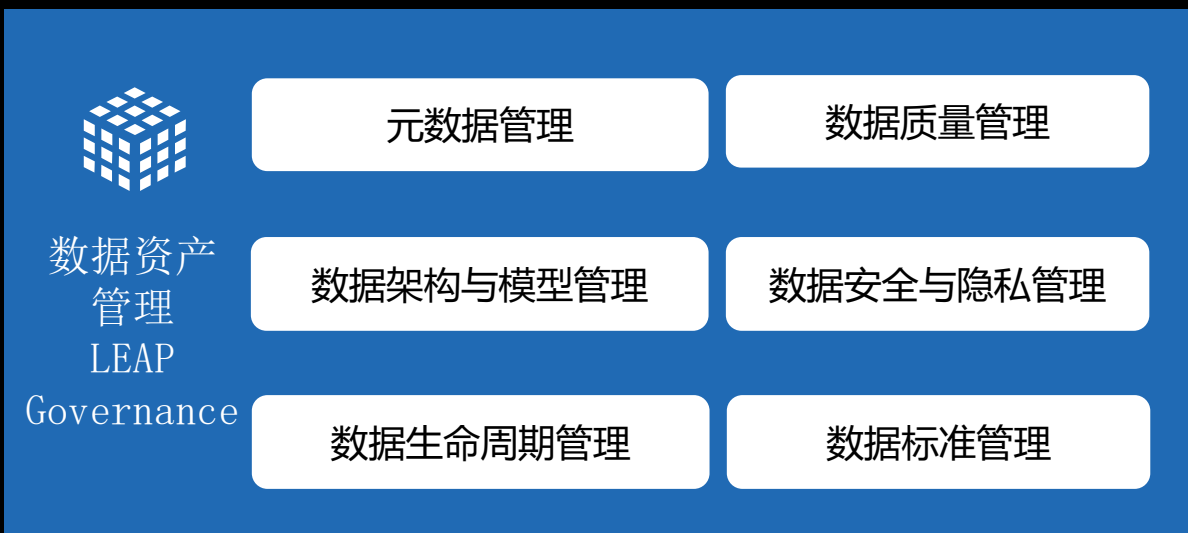
时序数据库



LEAP DATA GOVERNANCE: 智能化的数据资产管理

- ✓ 支持HIVE, HDFS的**元数据自动获取**及全量更新和增量更新的方式
- ✓ 10分钟可同步系统单数据源支持**3万+张表**元数据的初始化获取; 支持**单次10万行**文本数据的导入和导出
- ✓ 支持**数据质量智能评估及自定义检查规则**, 并提供相应的信息统计及自定义告警;
- ✓ 支持全链路的数据处理流程**血缘追踪及影响分析**
- ✓ 支持多层多级多角色的数据权限模型管理元数据, 以及**敏感数据过滤**及主动探测隐藏;
- ✓ 支持**数据操作安全审计及日志分析**, 主动拦截非法操作, 及**自动数据生命周期管理**

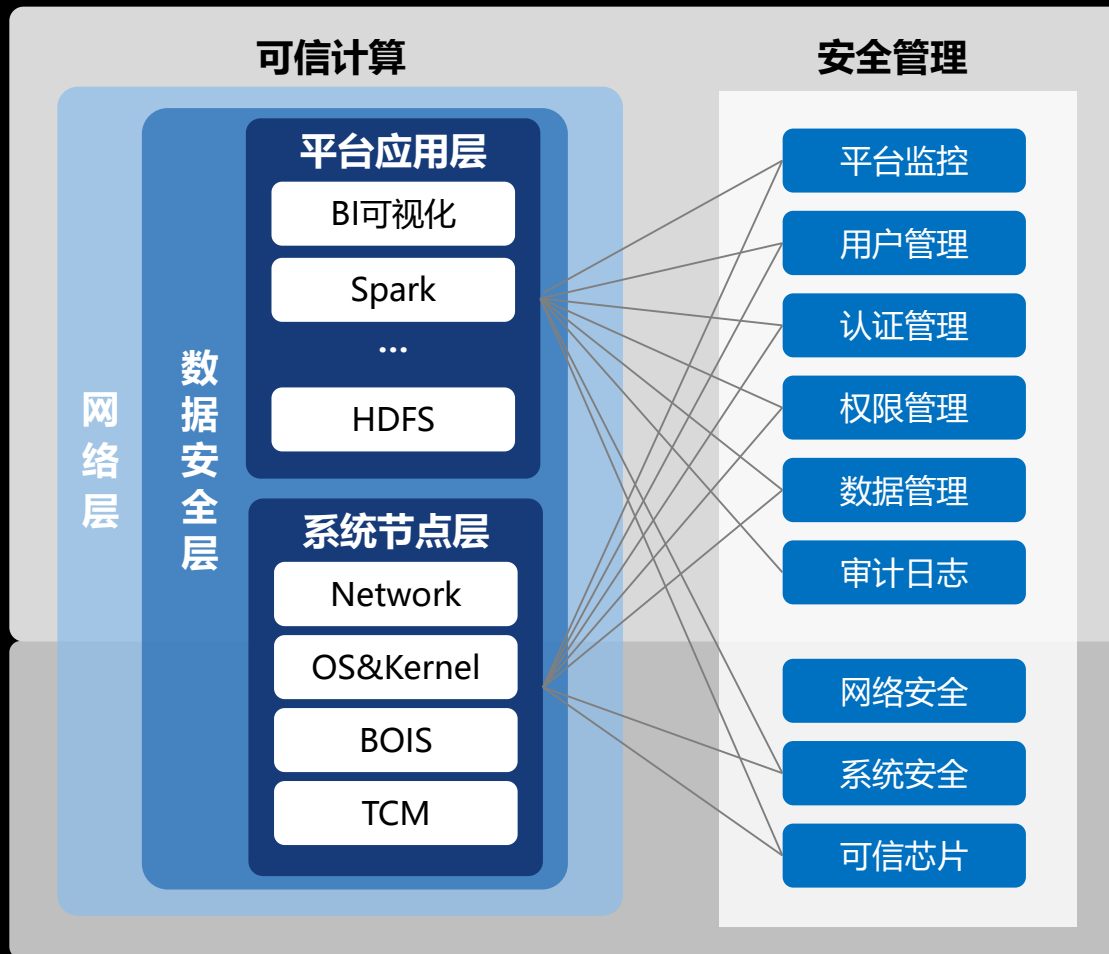
数据全生命周期可管、可控、可视



LEAP TRUSTED: 全链路的数据安全可信

多层次工业互联网安全保障体系

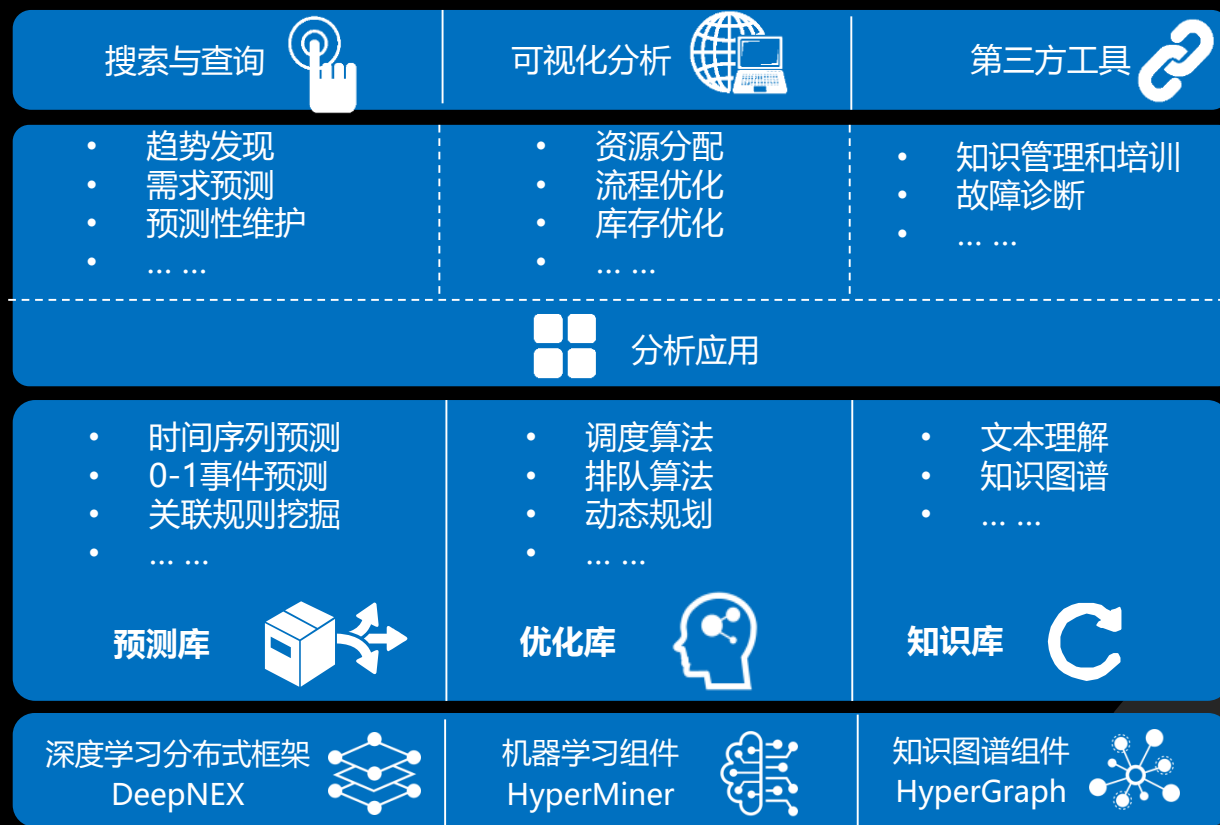
- ✓ **全链路可信：**
 - 支持**硬件级的可信平台**（基于TCM可信技术）
 - 支持安全操作系统
- ✓ **全周期安全：**
 - 支持Https的安全传输
 - 支持**SM3、SM4等国密算法**
 - 支持硬件级的加解密
 - 支持**三权分立**机制的权限管理
- ✓ **全体系统监控：**
 - 支持系统节点层的日志管理审计
 - 支持平台应用层的日志管理审计



LEAP INTELLIGENCE: 专业团队助力构建数据智能

- ✓ 500+名研发工程师, 100+名实施与运维工程师, 50+名数据科学家, 30+名大数据领域专家,来自国内外超一流著名学府的人才, 博士与海归比例超过80%.
- ✓ 集成**深度学习、机器学习、知识图谱**机器智能独立组件
- ✓ 支持**自动化建模、自动化调参**, 让机器学习用户可非常简单地相关应用开发
- ✓ 支持**数亿节点**图数据挖掘, 容易扩展; 包含**20多种**常用图算法, 涵盖分类, 聚类, 链接预测, 网络特征提取, 推荐, 查询等功能
- ✓ 提供**10余个**原创前沿机器学习行业优化方案, 涉及时间序列预测、自然语言处理、文本分析等领域
- ✓ 框架、工具、算法与应用具有模块化、弱耦合关系

结合企业场景需求, 加速从数据到业务的价值实现



LEA产品六大核心优势和突破

1. 支持最全SQL，支持存储过程

- 支持SQL 99和2003语法(99%)
- 业界唯一支持SQL Server存储过程(98%)
- 支持Oracle PL/SQL(80%)
- 支持DB2 SQL/PL(70%)
- 帮助用户低成本迁移传统应用

2. 超强的计算能力

- 率先通过50TB TPC-DS的99个复杂业务场景测试。
- 基于SSD 列式存储支持对数十亿条记录的秒级交互式探索分析能力

3. 支持分布式事务处理

- 保证分布式事务处理的ACID特性
- 支持批量增删改的分布式事务处理
- 多版本(快照)隔离可以保证只读事务的高并发性
- 业界极少能在分布式系统中保证事务一致性

4. 完整的图形化数据处理工具

- 提供全图形化的数据处理工具
- 图形化的数据实时同步工具(Leap DataHub)
- 便捷的SQL,PL/SQL开发调试辅助工具(Leap IDE)
- 图形化的数据治理工具(Leap Governor)
- 极大的提升开发及运维人员的效率和用户体验

5. 丰富的数据挖掘、机器学习算法和深度学习平台

- 支持最全的分布式统计和机器学习算法，支持5000多种标准R语言算法
- 业界最全的多个行业挖掘模型，可以更有效地帮助客户构建行业应用
- 整合TensorFlow和MxNet的深度学习框架
- 帮助数据科学家方便的构建DNN(深度神经网络) 或者CNN(卷积神经网络)，使用更大的数据做模型训练，提高算法的精准度。

6. 多租户管理/细粒度访问控制

- 资源共享和隔离、配额管理
- 支持一键部署Leap、扩容、缩容
- 支持基于优先级的抢占式资源调度和细粒度资源分配

紧跟社区技术发展，持续实现LEAP能力的优化和增强

组件	Leap3.4.4.0发行版本	优化或改进内容
Hadoop	2.6.0	存储优化、性能优化
Hbase	1.2.0	增强索引使用，安全扩展
Hive	1.1.0	数据安全扩展、元数据管理扩展
ZooKeeper	3.4.5	分布式应用程序协调服务
Spark	2.1.0	更强的计算性能 扩展MPP CRUD 存储过程支持 扩展SQL语法及解析支持
Impala	1.2.0	优化及调优
Kafka	0.9.0	增强数据实时处理能力
Flume	1.6.0	实时处理解决方案
Elasticsearch	5.4.1	增强日志收集、处理、搜索、存储及分析能力
Kibana		
Logstash		
Kylin	1.6.0	增量OLAP分析能力
Oozie	4.1.0	兼容，并提供自研的流程调度引擎
Sentry	1.5.1	增强多租户，扩展ID管理、认证、LDAP及权限管理
Solr	4.10.3	
Sqoop	1.4.6	
Storm	1.1.0	实时处理解决方案
Kerberos	1.10.3	网络认证协议
Phoenix	4.8.2	HBase数据查询引擎

组件更新策略：

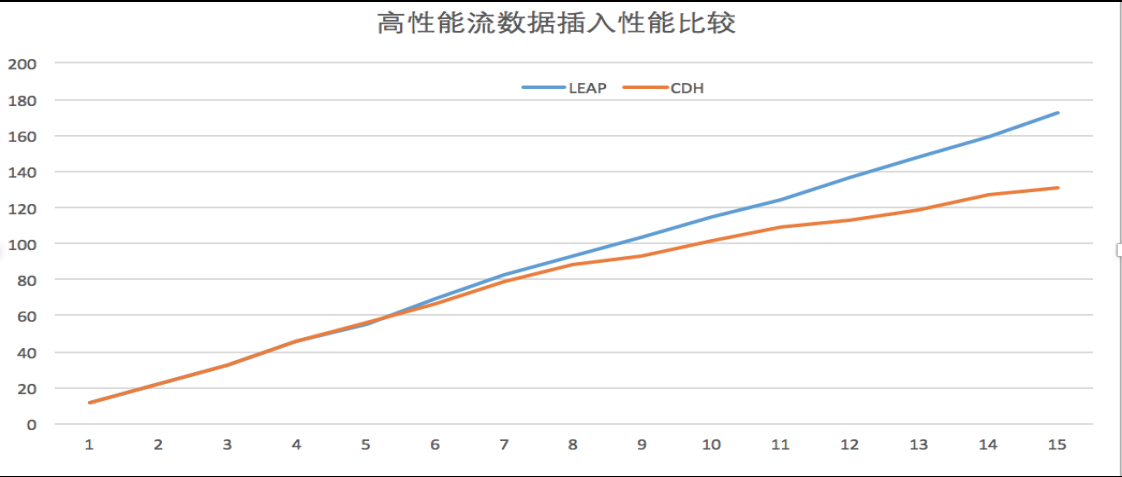
- ✓ **版本跟踪：**跟踪社区组件变化，不断更新组件版本
- ✓ **版本选择：**自动化校验，比对组件不同版本**功能、性能、稳定性及兼容性**差异，择优选择集成
- ✓ **集成优化：**不断集成组件版本至Leap平台，优化功能及配置参数；
- ✓ **功能增强：**在**安全、性能、计算框架**上不断增强与扩展；

自主研发，构建自动化压力测试工具，使得开源整个Spark生态在Apache正式发布后1个月内正式导入Leap产品

深度优化SPARK计算引擎，领先主要国内竞争对手3~6个月

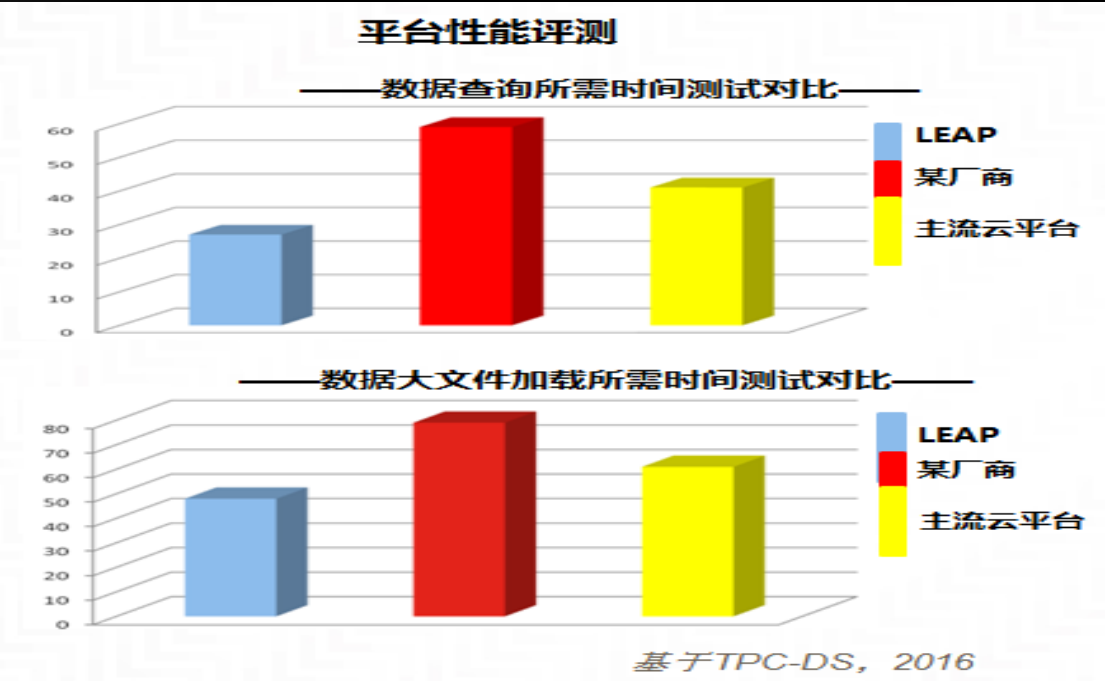
- ✓ 深度优化Spark性能，使得评测性能比最新开源标准发布提升20%~30%
- ✓ 性能领先主要竞争对手3—6个月
- ✓ 新技术采纳和融合迅速

IOT传感器数据吞吐量测试，Leap优于Cloudera 20%



某实际项目测试，比客户要求提升1.6—2.45倍

项目内容	客户要求	实测值	提升
1、NoSQL数据库单台机器1秒内扫描记录数	1亿条	3.37亿	2.37倍
2、从文本文件中导入数据单机性能	30MB/秒	103.6M/秒	2.45倍
3、Spark计算任务从平台内置的NoSQL数据库中导入数据的单机性能	20000条/秒	52083条/秒	1.6倍



LEAP产品得到了业界和客户认可!

2018, 工信部国家“大数据优秀产业、服务和应用解决方案”, 排名第一

2017, 牵头成立工业大数据产业应用联盟

2017, 中国国际软博会金提名奖

2017, 中国大数据企业50强

2017, 中国数字化转型企业TOP10

2017, 中国国际大数据大会明星平台奖

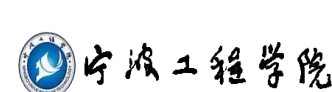
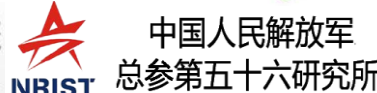
2017, Kaggle数据科学竞赛全球金奖

2017, 中国数据英雄

2016, 中国大数据企业50强

2016, 中国大数据产业十大领军人物

2016, 工信部物联网优秀解决方案



thanks.

Lenovo 联想

Tech
World '18

