CAICT 中国信息通信研究院
China Academy of Information and Communications Technology

大数据技术与产品标准实践

姜春宇 张诚

中国信息通信研究院 移动与大数据研究部
数据中心联盟 大数据技术与产品工作组组长

2016年10月25日

提纲

1 大数据标准发展情况

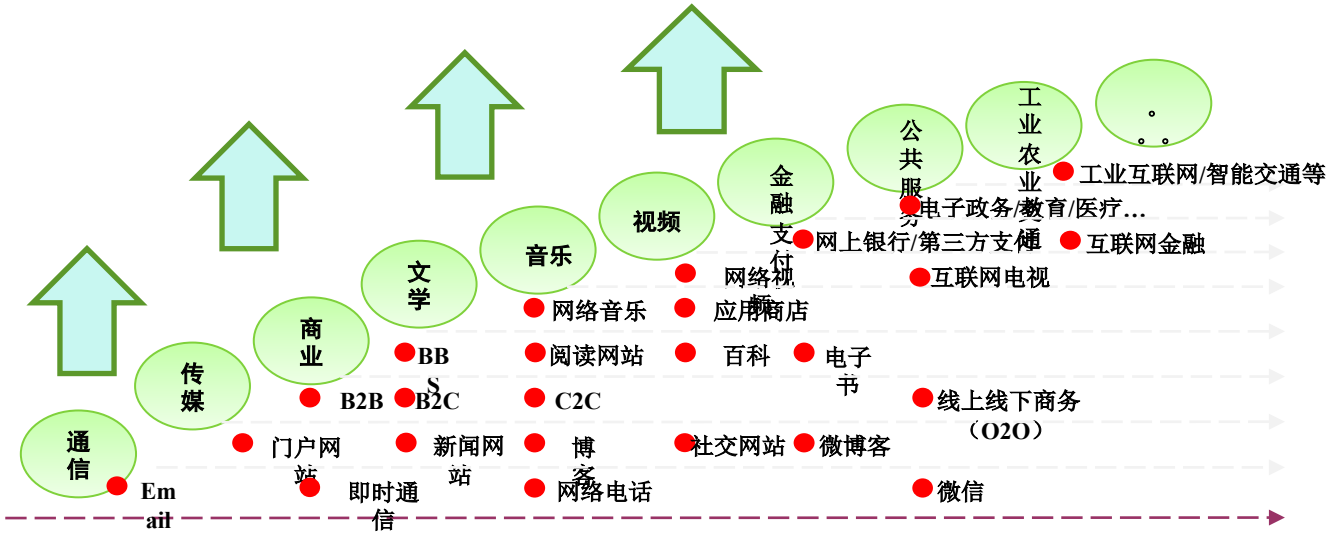
2 测试实践

3 总结和下一步建议

大数据产业发展背景

数据流通

大数据应用



大数据技术和平台

开源主导的大数据技术

- 以Hadooop、Spark、NoSQL数据库为主的大数据技术从开始就走上了开源的道路，通过开源走向壮大
- 开源技术 ≠ 成熟的产品
  - 技术种类多
  - 技术门槛较高
  - 不够稳定
  - 安全漏洞
  - 易用性差
  - 缺少服务支持

大数据技术与产品的标准化需求

供应商

- 技术和产品众多，缺少标准来规范市场
- 重复的POC测试
- 缺少产品和技术间横向比较
- 需要引入用户需求来引导产品研发

共性的评估体系和标准

将技术与应用场景相对应
将复杂的产品转化为容易理解的指标

用户

- 需求模糊
- 技术选型困难
- 产品评估过程成本很高
- 使用难
- 日常的运维很繁杂

供应商和用户之间存在明显的信息鸿沟

重要意义

厂商

- 保障大数据技术和系统的健康发展
  - 客观的指标体系保证厂商之间有序竞争
  - 对技术难点集中进行攻关，专注于性能改进
  - 定义合理的用户场景需求引导产品研发方向

公众

- 大数据系统度量标准
  - 指标易于理解
  - 测量方法公开公平

用户

- 帮助客户进行数据库系统的选型
- 较少POC测试的花费

大数据产品标准发展历程

第一阶段标准制定

联合20多家企业，四次会议讨论

《大数据平台基准测试 第一部分：技术要求》——方法论、负载和数据需求、指标

《大数据平台基准测试 第二部分：测试方法》——10个测试用例，条件、流程、方法

2014 6月-2015年1月底

2015 10月-2016年3月

第二阶段标准制定

联合20多家企业，3次工作组会议，5次电话会议讨论形成

《 Hadoop基础能力能力测试方法》--

7大指标项，38个测试用例

第三阶段标准制定

20多家企业，5次电话会议讨论

Hadoop/Spark大数据性能测试方法

面向4种任务类型，12个用例

2016 4-6月底

2016 7-10月底

第四阶段标准制定

联合国内10家数据库厂商，5次工作组会议，2次电话会议

完成《MPP数据库基础能力测试方法》

--6大指标，50个测试用例

《MPP数据库性能测试方法》 继续中

基础能力认证指标体系

运维管理	可用性	功能	兼容性	安全	多租户	扩展性
自动化部署	Namenode主节点失效恢复	数据导入	ODBC兼容性	认证	租户管理	集群动态扩展
资源监控	Namenode备节点失效恢复	SQL任务能力	JDBC兼容性	授权	资源管理	集群动态收缩
作业监控	Datenode节点失效恢复	NoSQL数据库	SQL支持度	加密	资源隔离	
集群操作	HMaster节点失效恢复	机器学习	传统数据库同步	审计	权限管理	
故障管理	RegionServer节点失效恢复	流处理能力	跨不同数据库表关联操作			
日志管理	HDFS备份恢复					
配置管理	HBase备份恢复					
权限管理	运维管理节点失效及恢复					
用户管理						
无宕机升级						

大数据产品基础能力认证包括七大项：功能、运维、多租户、可用性、安全、兼容性、扩展性，总共38项测试用例

性能专项认证用例分布

SQL任务	NoSQL任务	机器学习	批处理
I/O密集型任务	数据并发导入	Kmeans 无监督聚类	Terasort MR I/O密集
CPU密集型	95%的读，5%的写	贝叶斯 有监督分类	
报表任务	50%的读和50%的写		
分析型任务	读、更改、写		
交互式查询			

大数据产品性能专项认证包括SQL任务、NoSQL任务、机器学习和批处理四类任务，总共12个测试用例

大数据产品标准化

标准化框架

基础能力

指标导向

- 功能
- 运维管理
- 可用性
- 安全
- 兼容性
- 扩展性
- 其他（可扩展）

性能

场景导向

- 批处理
- SQL任务
- Nosql 任务
- 机器学习任务
- 图计算任务
- 流处理任务
- 其他（可扩展）

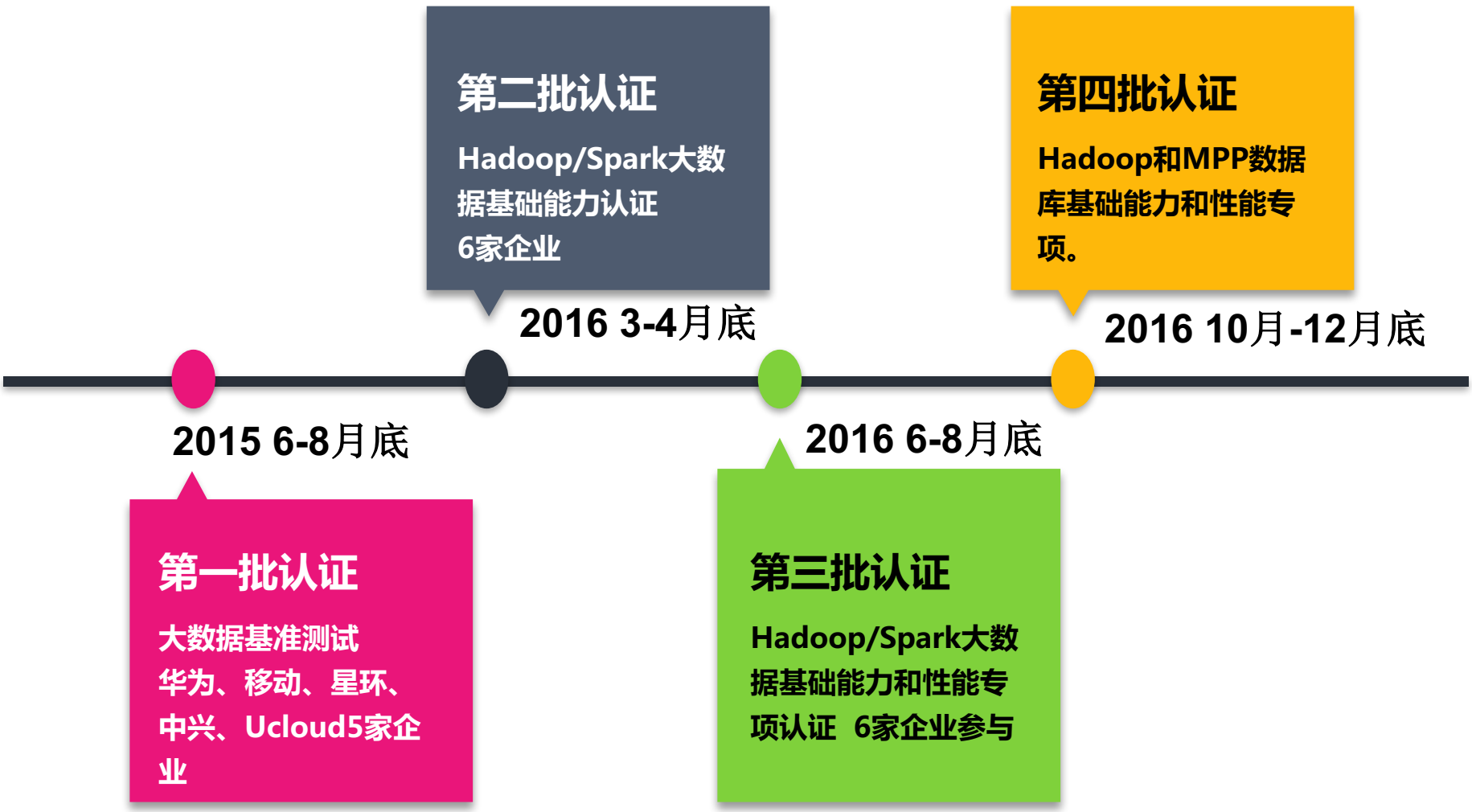
提纲

1 标准发展情况

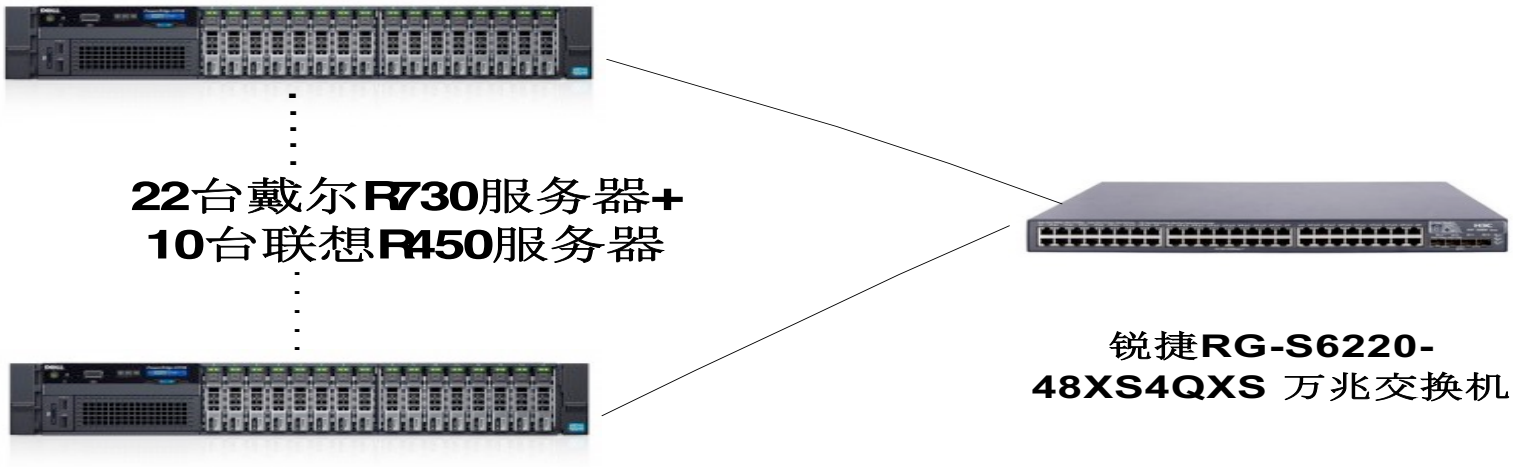
2 测评实践

3 总结和下一步建议

大数据产品能力认证发展历程



测试实验室环境



组件	配置	台数
CPU	2*英特尔至强 E5-2620 v3 2.4GHz,15M 缓存	32
内存	4*16GB RDIMM, 2133 MT/s	31
	8*16GB RDIMM, 2133 MT/s	1
硬盘	10*1.2TB 10K RPM SAS 6Gbps 2.5英寸 热插拔硬盘	22
	10*1.2TB 10K RPM SAS 12Gbps 2.5英寸 热插拔硬盘	10
网卡	单口万兆网卡	32
交换机	锐捷RG-S6220-48XS4QXS 万兆交换机	1

参与认证的厂商

第一批认证

中国移动通信 CHINA MOBILE

HUAWEI

TRANSWARP

ZTE

UCLOUD

第二批认证

东方金信 Eastern Jin Technology

明略数据 MININGLAMP

byit 博易智软

H3C

TRANSWARP

腾讯云

第三批认证

百分点 BAIFENDIAN.COM

GRIDSUM 国双

H3C

腾讯云

东方金信

TRANSWARP 星环科技

通过认证的产品

第一批	第二批	第三批	
Hadoop基准测试	**Hadoop基础能力**	**Hadoop基础能力**	**Hadoop 性能专项**
华为FusionInsight	东方金信 SeaBox Big Data platform	百分点 BD-OS	新华三 H3C DataEngine
中移动苏研 BC-Hadoop	明略 Minglamp Data Platform	国双 国双大数据平台	腾讯云 Data Intelligence
星环Transwarp Data Hub	博易智软 Super Center Big Data Platform		东方金信 SeaBox Big Data Platform
中兴 Golden Data	新华三 H3C DataEngine		星环 Transwarp Data Hub
Ucloud	星环 Transwarp Data Hub		百分点 BD-OS
	腾讯云 Data Intelligence		

提纲

1 标准发展情况

2 测评实践

3 总结和下一步建议

衡量大数据产品的维度

- 功能完备性
  - 38个认证项覆盖功能、运维、可用性、安全、扩展、兼容性、多租户等7大维度
- 性能
  - 产品本身的易部署性稳定性、易运维性、性能
  - 考察参测团队综合使用大数据平台的能力
    - 环境部署与集群规划
    - 测试工具的使用
    - 多任务调优能力
    - 时间进度安排
    - 集群的故障处理与运行维护
- 稳定性

性能评测总结

- 性能测试是在统一平台（32台集群）、统一测试数据、统一测试工具、统一测试周期、统一测试规则下进行的测试认证
- 任务重，周期紧
  - 7天的测试周期内，包括操作系统安装、大数据平台安装、数据生成，以及完成12项不同种类的测试用例，其中Terasort中位执行时间为3小时19分，HBase任务中位执行时间分别是29、54、50、79分钟
  - 模拟SQL复杂业务查询、NoSQL高并发任务、机器学习和批量处理等多项典型任务
- 领先性
  - 业界领先的测试集群规模和集群配置（32台集群规模）
  - 10项以上用例使用TB级别的数据规模

视频大数据平台标准规范建议

- 梳理视频领域大数据基本应用场景
  - 视频流调度
  - 存储
  - 其他
- 面向视频应用场景的性能测试方法和测试工具
  - 业务描述
  - 数据描述
  - 数据生成工具

请各位专家批评指正

谢谢！

中国信息通信研究院
标准所移动与大数据研究部