拓尔思

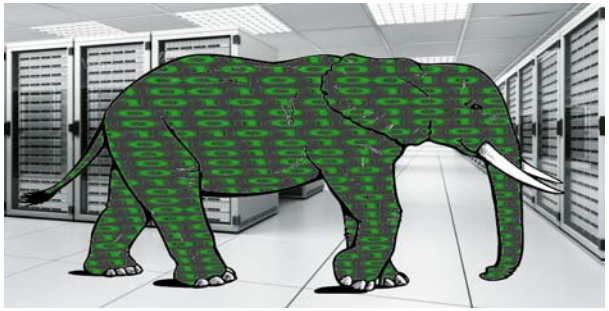
中文信息处理技术和应用发展的新推动力

施水才

北京拓尔思信息技术股份有限公司

2012年4月13日 江西婺源

这些大趋势深刻影响中文信息处理的未来



大数据

云计算

移动互联网

社会化计算

- 大数据时代使中文信息处理技术成为***必须***
- 云计算可以实现智能中文信息处理技术的***落地***
- 移动互联网的小产品大市场特性释放了中文信息处理应用的***潜力***
- 社会化计算需要中文信息处理技术的***创新***

大数据时代已经来临

- 大数据既是一项破坏力，也是一个业已影响到传统认识和业务模式的紧迫问题……它打乱了现行趋势，同时亦代表了公共部门、业务和IT领导者们无法忽略的巨大机会，*- Gartner Stephen Prentice*

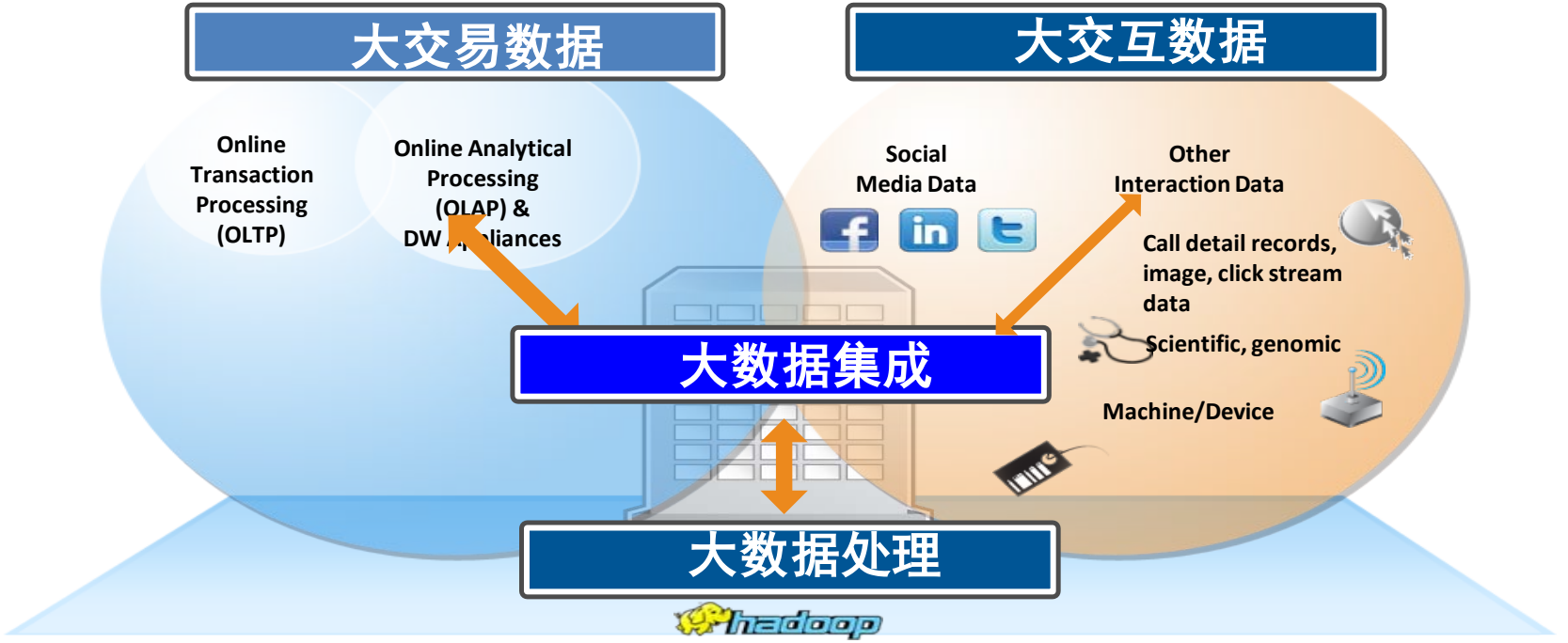
数据的爆发式增长和社会化趋势（一）

- 企业信息和数据：在美国15-17%的企业存储的数据,超过了国家图书馆的信息量。信息和数据每增长40%，IT投入将增长5%
- 社会化媒体：在Facebook上每个月分享的信息达百亿条。国内最活跃的新浪微博、腾讯微博、搜狐微博等月发布微博信息达数十亿条，而且在快速增长中
- 电子商务：淘宝目前每天的活跃数据量已经超过50TB，共有4亿条产品讯息和2亿多名注册用户在上面活动，每天超过 4000万人次访问

数据的爆发式增长和社会化趋势（二）

- 智能手机和移动互联网：中国超越美国，成为世界上最大的智能手机市场。智能手机的社交功能、娱乐功能以及商业方面的用途已经超过了我们的想象，智能手机将成为最重要的数据承载终端
- 智慧城市、物联网、传感器：愈来愈大量的数据不断地产生出来，但这些数据并不一定出自人类之手。随着物联网的发展，从交通信号，到汽车、医疗设备等都会不断地产生大量的数据

数据的爆发式增长和社会化趋势（三）



大数据对传统IT设施带来的颠覆和挑战

- 大数据已经成为企业或者组织的新型资产，大数据通常指PB的数据规模。1PB=1024TB 1TB=1000GB
- 大数据的4V特性：海量（volume）、多样（variety）、实时（velocity）、价值（value）。数据多格式、数据处理速度和大容量是大数据时代必须解决的问题
- Gartner认为到2015年，超过85%的财富500强企业将在大数据竞争中失去优势

大数据是信息社会化的必然结果

- 信息社会化是推动大数据产生的根本原因
- 大数据中的"行为和关系"是大数据的核心价值
- 大数据时代IT产业的三大趋势:数据成为资产 行业垂直整合(服务化) 终端应用(移动互联网)

大数据: 创新 竞争 破坏力

- 互联网的本质是对用户的精确理解
- 对行为和关系的挖掘将推动业务模式创新
- 大数据将颠覆、重构现有产业链

传统企业天然缺失“用户行为”数据，难以获取消费者的真实需求！

用户行为暴露其真实的需求，保存与散失，差别犹如云泥。

互联网公司完整追踪“用户行为”数据，用大数据技术形成“用户档案”，甚至可以洞悉消费者潜在需求。

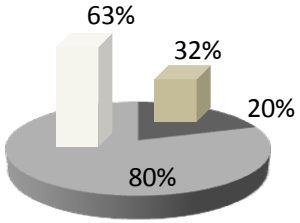
“新一代基于互联网DNA企业的核心能力在于利用新模式和新技术更加贴近消费者、深刻理解需求、高效分析信息并作出预判，所有传统的产品公司都只能沦为这种新型用户平台级公司的附庸，其衰落不是管理能扭转的。互联网的魅力就是“the power of low end””

——索尼前总裁出井伸之

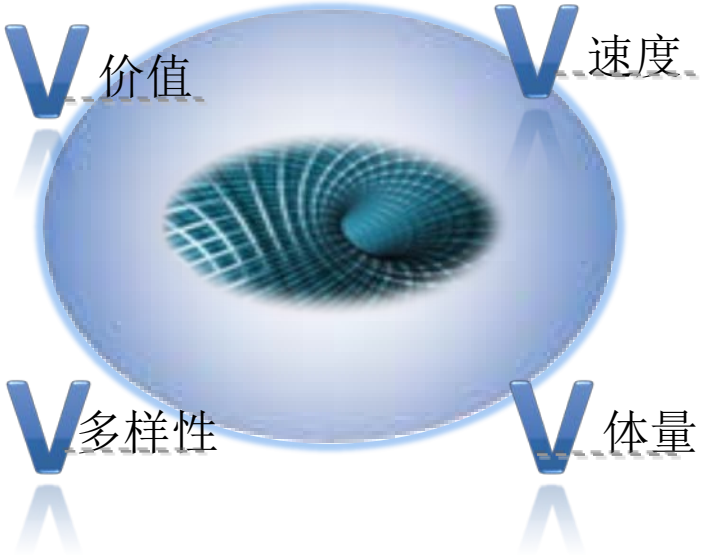
大数据技术

- 数据多格式、数据处理速度（实时性）、大容量（PB级别）
- 大数据中80%是海量非结构化信息，挖掘和分析
- Hadoop开放框架本身已经成为一种大数据技术标准

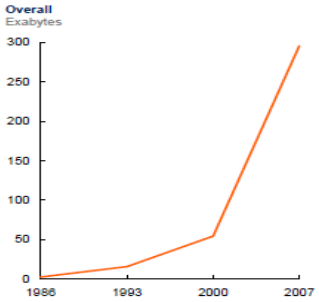
沙里淘金



大数据包括大量非结构化数据、邮件、Word、图片、聊天记录、视频、音频等信息



1秒定律
实时获取需要的信息



DC最新数字宇宙研究报告表明，到2020年，全球数据使用量预计暴增44倍，达到35.2ZB

大数据管理的新挑战
-从管理数据到理解和分析内容

- 虽然大数据是一个重大问题，Gartner分析师表示，真正的问题是让大数据更有意义
- 目前海量数据管理多从架构和并行等方面考虑，解决高并发数据存取的性能要求及数据存储的横向扩展，但对非结构化数据的内容理解仍缺乏实质性的突破和进展，这是实现大数据资源化、知识化、普适化的核心
- 非结构化海量信息的智能化处理：自然语言理解、多媒体内容理解、机器学习等

目前所有大数据管理的解决方案
没有解决语义计算的基本问题：理解内容

云计算可以实现高级中文信息处理技术的落地

- 云端计算能力
- 自学习和自我改进能力
- 人工干预的机会

移动互联网的潜力

- 移动互联网的小产品大市场特性释放了中文信息处理应用的*潜力*
- *Siri*

16岁程序员获得亿万富翁李嘉诚的种子投资来“简化互联网”

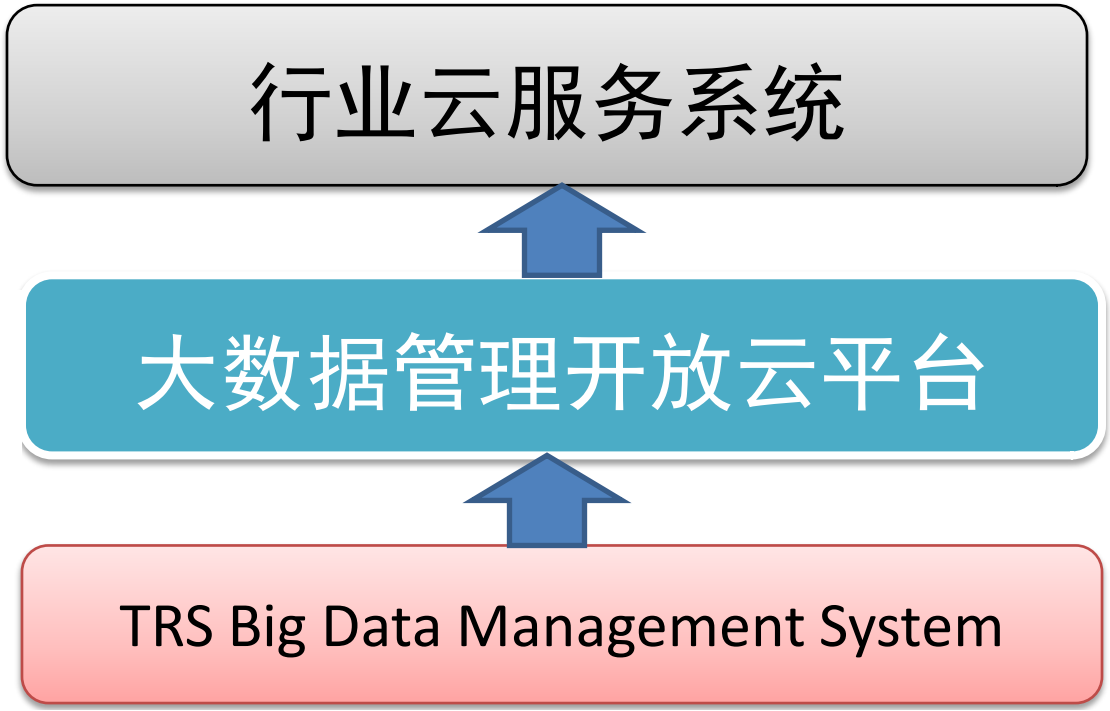
今天，一个名为“Summly”的应用程序登陆App Store，这款应用的初衷是为了“简化互联网”，它的前身就是著名的Trimit，其创始人－16岁的程序员Nick D’ Aloisio已经从香港亿万富翁李嘉诚旗下的Horizons Ventures获得了种子投资。

到Trimit允许用户在浏览网页时将“网页内容缩减到1000、500或者140字的概括版，Trimit的本质是将文本自动转换成摘要，使它得以适应不同的设备，并且可以通过短信，电子邮件，Facebook或者Twitter等进行分享－你只需要简单点几下鼠标。”

社会化计算需要中文信息处理技术的创新

- 数据的社会化-行为和关系
  - Facebook, Twitter, 微博 Splunk
- 数据的实时性
- 信息可视化

我们想干什么，在干什么？



行业应用，如舆情服务，行为分析

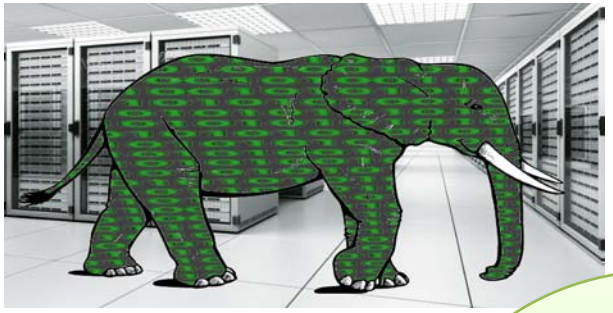
数据、知识、关系
第三方应用接入

核心技术创新

100亿记录
PB级数据
200台服务器
每天增加1亿记录

分布式和并行计算构架, 支持结构化、非结构化、半结构化、多媒体、实时及用户行为数据的高效管理，
兼容Hadoop标准，支持PB/EB级的海量数据管理
创新的多引擎机制（全文检索引擎Lucene和TRS、关系数据库引擎、多媒体检索引擎、第三方引擎）

串起来



大数据

云计算

中文信息处理

移动互联网

社会化计算

总结

- 大数据时代使中文信息处理技术成为***必须***
- 云计算可以实现智能中文信息处理技术的***落地***
- 移动互联网的小产品大市场特性释放了中文信息处理应用的***潜力***
- 社会化计算需要中文信息处理技术的***创新***

谢谢！请批评指正

联系方式：shi.shuicai@trs.com.cn

微博：weibo.com/shuicai