数据技术时代的未来

张茂森
阿里巴巴网络有限公司 北京 100022

摘要
数据应用是数据技术时代的价值承载，数据技术时代和已有的信息技术时代的区别在于是否将数据作为生产资料。信息技术时代解决的是“过程”智慧的问题，从而提升效率，而数据技术时代解决的是“感知”和“协同”智慧的问题，使效率大幅提升并能完成业务的创新。归纳了数据应用类产品的分类，给出了一个通用的数据应用实现架构，同时对大数据的数据共享和交换的本质和难点做了简要分析。

关键词
大数据应用；数据共享与交换；信息技术；数据技术

Future of Data Technology Era

Zhang Maosen
Alibaba.com Limited, Beijing 100022, China

Abstract
Data application is the key element of data technology (DT) era, the difference between DT and information technology (IT) is whether data is the key product element. IT suppose to solve the “process” problem to make business more effective, but DT suppose to solve the “cognitive” and “collaborative” intelligence to make business to renewable and creative. The catalog of data application product was given. A general architecture of data application platform and organization was proposed. The key and difficult point of data sharing and exchange were analyzed.

Key words
big data application, data sharing and exchange, information technology, data technology

1 引言

前段时间，杰克·马的CEBIT欧洲巡回演唱会[1]非常成功。面对从总理到各大商业巨头，杰克坚定无比地讲述了一个数据技术时代的梦想，不得不说，杰克真的是神奇的外星人。杰克说过，他负责吹牛，然后他的团队负责把他吹过的牛实现，从而打造了一个如此强大的阿里巴巴，让美国人甘心叫BABA的公司。

这次杰克在欧洲巡回演唱会又吹了一个什么牛呢？这头牛不再是电子商务，而是DT（data technology，数据技术）。DT和IT 的区别是什么？为什么DT就是利他，而IT（information technology，信息技术）就是利己？笔者一直很困惑，IT不是也让人们的生活更美好了吗？从经济学的角度来讲，反而是人人利己创造了人类的进步。

2 IT与DT

笔者查了好多资料，也跳出互联网圈子接触了传统行业的朋友，似乎有些理解了。

IT这个词诞生于何时，笔者暂时没有查到，但是它的大规模商业化发端应该是20世纪70年代，具体是指利用电脑和网络让企业的内外业务与流程更加高效。换句话说，没有IT系统，业务也能运行，只是“慢、卡、丑、挫”而已，当然当大家都用上IT系统后，就再也回不去了。

IT的引入让企业拥有了更强大的业务能力，使全球化、大规模、深层次的协作变成可能，让大象也能跳舞，所以说IBM是通过输出IT能力，让别的大象跳舞，从而让自己在资本市场也风姿绰约。

企业级IT市场的原有商业模型如图1所示。

1 http://money.163.com/15/0317/10/AKTF1GD200253B0H.html

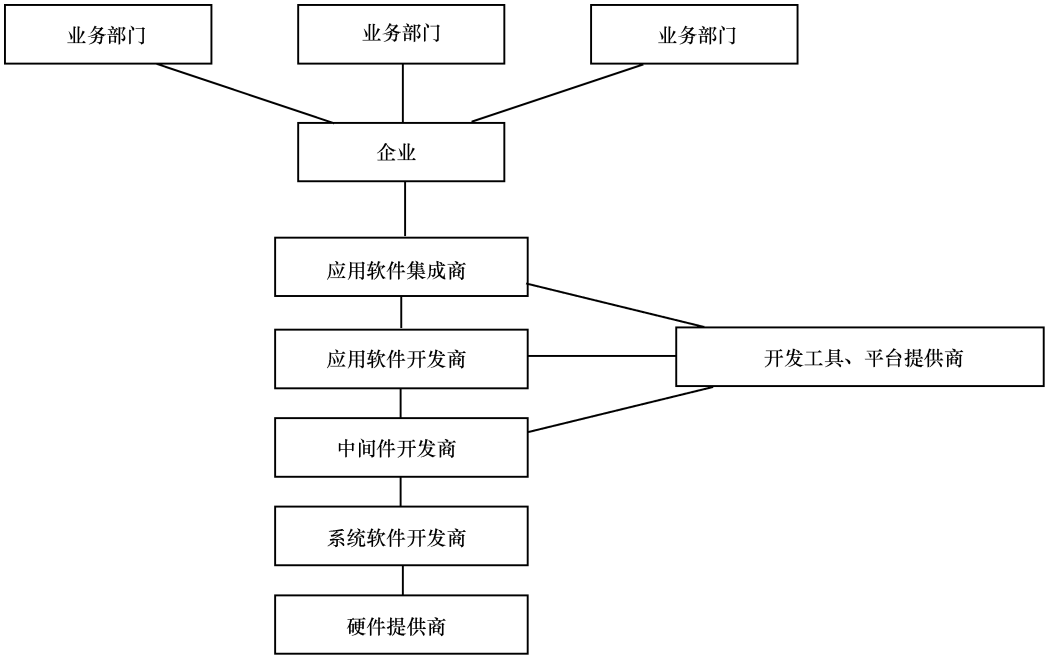

图1 企业级IT市场的原有商业模型

在旧有模型下，IT解决的是企业自有业务流程的信息化问题，比如OA系统、CRM系统、ERP系统、绩效管理系统、BOSS等，这些系统依赖的架构就是应用代码（BS或者CS）+数据库+PC服务器。

互联网企业的出现，成为了一个IT产业的异类，一家家什么都不生产却又市值奇高的公司，当然是泡沫。只是互联网企业经历一个个大起大落后，越发青春焕发，甚至都开始秀健硕的肱二头肌，开始讲颠覆了。背后的原因是什么？互联网企业正因为没有实体生产，经历的过程正好契合了DIKW的知识金字塔[2]，如图2所示。要知道人类历史的推进就是知识的累积和进化，近100年人类社会的高速发展，也是知识的超常规累积的结果。

2 http://en.wikipedia.org/wiki/DIKW_Pyramid

第一代的互联网企业完成的是实体到数据（data），把线下的东西数字化后搬到线上来，比如以Yahoo公司为代表的Portal。

第二代互联网企业完成的是从数据到信息（information），Google公司把全类目的数据聚合、整理、结构化后索引起来，变成了可供大家快速检索的信息。

第三代互联网企业完成的是从信息到知识（knowledge），目前还在快速推进，一种企业是通过人和人的连接，从而利用人机结合来填补信息到知识的鸿沟，比如Twitter、Facebook；一种企业是通过大数据+机器学习+人工智能来填补信息到知识的鸿沟，比如Google。然而，这两个方向随着后续的推进和大数据的介入，正在融合为一。

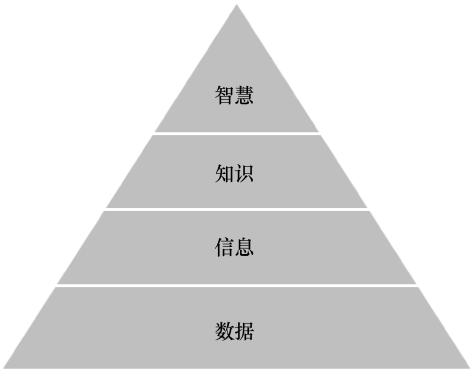

图2 DIKW金字塔

可以看到这些互联网企业在“吹泡沫”的同时，也构建起了从实体到数据、从数据到信息、从信息到知识的基础架构和设施，比如非结构化数据的处理、分布式数据处理、人工智能与机器学习以及在专业领域的方法论（如精准营销、搜索引擎、社交关系等）。

目前，传统的IT企业帮助传统企业仅仅完成了业务流程到部分数据化、数据到部分信息化的过程。换句话说，传统IT与自动化解决的是“流程”智慧的问题，“感知”与“协同”智慧是由人来完成的。比如传统汽车制造，流水线就是“流程”智慧，大幅提升运行效率，而流水线上的熟练工人依靠他们的“感知”与“协同”智慧保证了高品质汽车的生产。

传统企业中的CRM、ERP解决的也是过程智慧的问题，大幅提升客户管理和生产管理的效率，而使用软件的业务人员依靠他们的“感知”与“协同”智慧（领域经验与知识等），保证了业务的顺畅运行和优化。

互联网企业更是如此，所有的业务天生就是信息化的，处处是IT也就没有IT了，互联网企业的价值由于轻资产的模式反而落在人上面，如它的技术人员、运营人员和产品人员。互联网企业要应对快速变化的市场，必须依靠这些人的“感知”与“协同”智慧来推进公司的创新与变革，从而不被时代抛弃。“流程”智慧在其中的附加价值已经不是很大了，云计算等基础设施的出现，更加剧了这一点。

而最近的10年，情况在发生改变，一个是工业智能机器人的出现，它们具有了

对周围环境的感知能力，并且拥有了更强大的学习能力，如果进入宝马最先进的工厂，基本上看不到多少工人了。

CRM、ERP等传统IT软件中也越来越多地引入了智能推荐、智能搜索、决策辅助、运营辅助的功能，试图大幅提升人在“感知”与“协同”方面的效率。而互联网企业则更不用说，从Google大脑到深度学习、无人驾驶，无不在把“感知”与“协同”能力推向极致。对于高阶智能的追求，让工业4.0和互联网+走到一起，而孕育智能的基础就是“大数据”，换句话说，大数据将工业4.0和互联网+粘合起来了。

3 DT时代的应用架构

如同多次工业革命的进程一样，先是基础原理技术的突破，如蒸汽动力技术、电力技术、信息技术，然后真正让社会福祉得到巨大提升的是，构建于这些原理平台型技术上的广阔而多姿多彩的应用型技术。笔者相信，大数据技术在经历最近10年的技术原理探索与构建后，大数据应用将真正地把人类引入“数据技术时代”。

说实话，大数据最成熟的应用目前看来还是在互联网领域，从搜索到营销再到智能手机，处处都有数据作为生产资料的影子，通过对数据价值的挖掘来提升业务的能力。最近笔者走访了很多传统的大型企业和政府部门，和大家聊需求的同时，也让笔者从应用架构的视角来思考相关的问题，找寻其中的共性。要做到真正的大数据应用，有两点缺一不可。

（1）对业务的理解、对数据的剖析和大数据分析的方法论

没有对业务的理解就没有应用场景和商业未来规划，就根本不会有应用的诞生，往往这一步是最难的。而大数据应用还需要对数据进行深入理解，如自己有哪些数据、数据的分布如何、数据质量情况如何等。最后是大数据分析的方法论，要把数据当作生产资料而非报表资料，对数据中蕴藏的旧有现象，通过多维度的拼接和长历史的对比，就能够构建起关联关系，从而进行推演和预测，进而构建因果机理。

（2）大数据开发平台与数据科学团队

现在市面上有很多的开发平台或者PaaS平台，都在标榜自己能做大数据，然后像传统IT时代一样把软件卖出去。笔者认为这样是不对的，大数据平台除了能够进行数据开发、建模、集成等工作之外，还需要大量真正的非传统数据技术能力的支撑，如数据安全、数据可信交换或共享、数据探索与协作等，这就需要能够使用这个平台的人，即数据科学团队。数据科学团队不是科学的老学究，而是一群不同侧重的角色组合，如偏业务与创新、业务数据模型与算法、基础数据处理。而现实中往往需要一个人具备以上多种角色，这可能也是他们被称为数据科学家的原因。构建数据应用的后端结构如图3所示。

4 大数据的交换与共享

任何一次工业革命里面都会有基础的、可被标准化交换与共享的载体，如蒸汽、电力、公知信息。在大数据时代，数据的交换与共享也是必然的，如果数据的能力仅仅是锁在政府里，锁在几个互联网巨头、几个工业巨头手里，是不能构建起多姿多彩的应用世界的[3]。

而数据交换与分享的形态是由数据应用产品的形态决定的。对数据应用产品的分类如图4所示，越到顶层的数据发挥的

3 http://www.dataguru.cn/article-5726-1.html

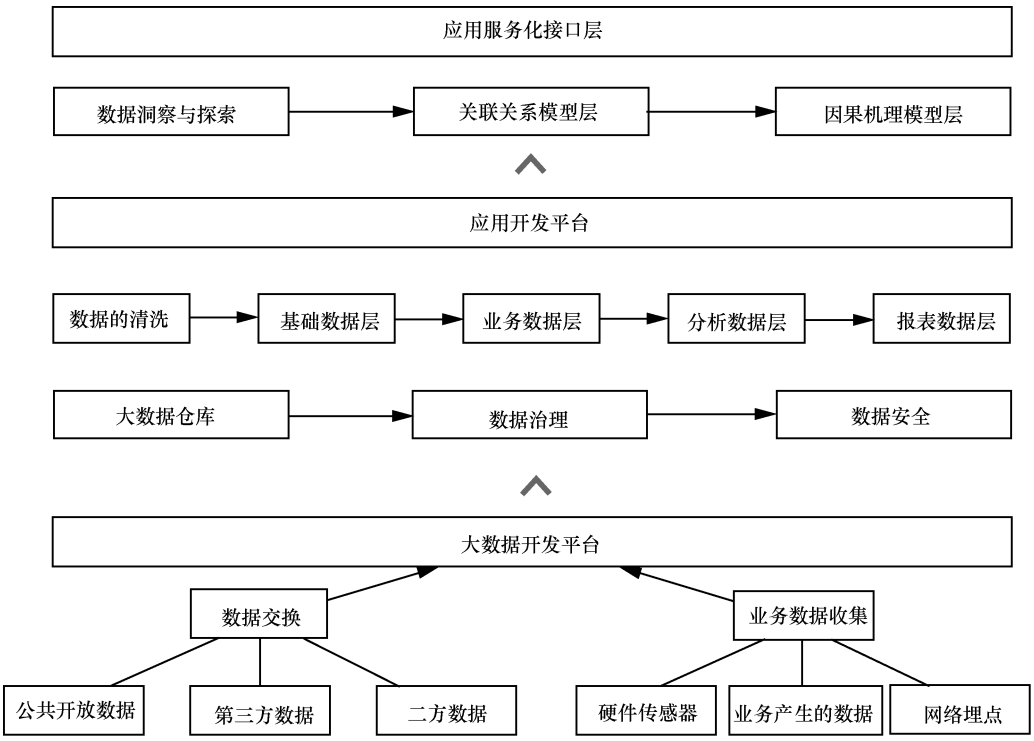

图3 构建数据应用的后端结构

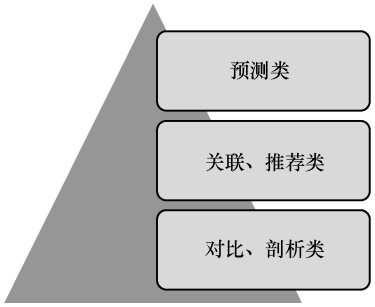

图4 数据应用产品的分类

价值越大，对大数据的需求也越大（如体量、多样性与全面性、稳定性和质量）。但是，从数据交换的角度来讲，越往上层越容易。其实RTB就是一个关联、推荐类的数据产品，完成了个人属性的数据交换，银行的征信也是一样的。之所以比较容易，是因为交换的数据是业务链条中的最终结果数据，它代表的是业务形态，而如果这个业务形态能够被公众和社会接受，交换的数据也是没有问题的，换句话说，此时交换的不是数据而是业务价值。但是对比、剖析类的数据产品要进行交换就非常困难了，Facebook 开放平台、淘宝开放平台面临的一些困境就是基于此，开放出去的是拥有无数未知和可能性的信息。这也是前几年很多Data Marketplace模式的公司衰亡的本质原因，数据一旦被真正地“交换”和“分享”，将造成巨大的未知，如信息的泄露、价值的低估和市场的撕裂。

而随着互联网和工业4.0的快速发展，越来越多的领域需要关联/推荐、预测类的数据产品帮助他们进行业务创新和优化，快速获取价值，也就是说，数据交换与共享的大幕开始启动。如何才能迎接这个趋势而又不陷入已有的错误中呢？答案可能是数据可“用”不可见。因为从应用的角度来看，大家不是需要数据，而是需要数据在杂交、关联、分析、预测后，在对应应用的业务领域的价值，也就是说要的是业务结果。如图5所示，数据可被使用，但是数据生

图5 数据可用不可见

产资料不能被拿走，而是被锁在一个可信平台中，平台输出的是业务结果。

数据可“用”不可见，还有两个最关键的事情尚未解决：平等可信的交换/交易模型与机制；定价和市场管理模型与机制。

最近几年经济学的成就也集中在了博弈论以及衍生出的市场机制设计方向。笔者相信，随着整个社会对大数据应用的认同和需要，工业企业+互联网企业+经济学模型+合理的监管一定能找到问题的解法。

整个社会把数据作为生产资料来看待才刚刚开始，这也正是大量应用蓬勃发展的契机。很幸运能生活在这个技术、业务剧烈变革，同时社会也在变革的时代，充满了梦想实现的机会，数据人加油！

作者简介

张茂森，男，阿里巴巴高级技术专家，2006年加入阿里巴巴，主要从事大数据仓库与开发平台、数据应用产品架构和数据PaaS平台等数据相关工作。

收稿日期：2015-05-11；**修回日期**：2015-05-13
论文引用格式：张茂森. 数据技术时代的未来. 大数据, 2015011
Zhang M S. Future of data technology era. Big Data Research, 2015011