

1 概要

全世界每天都有几十亿人使用计算机、平板电脑、手机和其它数字设备产生海量数据。据 Forrester 的最新研究显示：在线或移动金融交易、社交媒体、GPS 坐标等数据源每天要产生超过 2.5 X10¹⁸ 的所谓“大数据”。今后几年，数据的增长速度将超越摩尔定律。

Big Data Phenomenon



这些数据或“大数据”最近被誉为新的“金矿”，对于很多行业而言，如何利用这些大规模数据正成为赢得竞争的关键。基于以下原因，在所有行业中，电信运营商拥有明显的优势，而且能从这一演变中获得最大收益：

- 数量

电信市场的渗透率通常很高（近 100%）。作为一个垄断行业，每个国家通常只有三或四家电信运营商（有些运营商拥有超过 1 亿用户），这意味着，运营商通常能够接触到大量客户的数据。

- 数据量

客户打电话、使用互联网、发送消息或导航时，他们每一秒钟都在产生数据。即便客户只是将手机连接到运营商的网络中，也会产生位置、移动速度、计费甚至生物计量等数据。而只有运营商才能采集到如此之多地与用户行为有关的信息。

- 多样性

最后，具有潜在价值的大量承接关系数据每天以客户位置、设备交互、购买行为、在线状态、社交地图和人口统计数据的形式从运营商这里大量流走。因此，运营商具备了解客户的潜力。

我们相信，由于在大数据领域拥有上述优势，电信运营商正处于一个他们从未能够充分利用和赚取收入的“富矿”上。传统而言，运营商数据中心中的大型业务支撑系统只是为了确保运营商能够对其客户所使用的服务计费。但是，随着电信运营商的竞争格局不断变化，谷歌、Skype 等 OTT 服务提供商正在蚕食他们的收入。从他们的现有资产中获益并提供良好的客户体验正成为一个关键的成功要素。被 Ovum 誉为“增长燃料”的数据是运营商最宝贵的资产之一，而且他们也越来越热衷于更加充分地利用用户数据。

2 什么是大数据

大数据指的是超出传统数据库系统处理能力的数据。这些数据量太大，移动速度太快，或者与您的数据库结构不匹配。为了能从这些数据获益，你必须选择另外一种方式来处理它们。

大数据通常使用 3 个“V”来定义，Gartner 对其的定义如下：

“大数据是大数据量、高速度、种类繁多的信息资产，它们需要经济有效和创新型处理方式来提升洞察力和决策水平。”

- **数据量 (Volume)：** 企业系统内数据量的增加是由交易量以及其它传统数据类型和新数据类型引发的。太大的数据量不仅在存储方面，在大规模分析方面都会出现问题。
- **速度 (Velocity)：** 这涉及数据流、结构化记录的创建以及数据的可访问性和可交付性。速度既包括数据产生的速度，也包括满足需求所需的数据处理速度。
- **种类 (Variety)：** IT 主管在将大量交易数据转化为决策时总是遇到问题，而我们现在有更多类型的信息需要分析，这些信息主要来自社交媒体和移动领域（承接关系感知）。数据种类包括：表格数据（数据库）、分层数据、文档、电子邮件、计量数据、视频、静止图像、音频、股票报价机数据、金融交易数据等等。

现实中的大数据应用通常涵盖上述一或两个“V”，但也有很多企业的大数据项目涵盖所有三个“V”，这些项目通常涉及来自多个数据源、大量的流数据。

传统数据和大数据之间的差异



	Small Data era	Big data era
Data Source	Internal	Social + External
Data Type	Structured	Un-Structured
Data Integration	Data-in-rest	Event captured
Analysis method	Table + graph + Analysis in back office	Dynamic data visualization + analytics in war room
Analysis environment	Dw + Server	Distributed process + cloud

3 电信网络中的数据

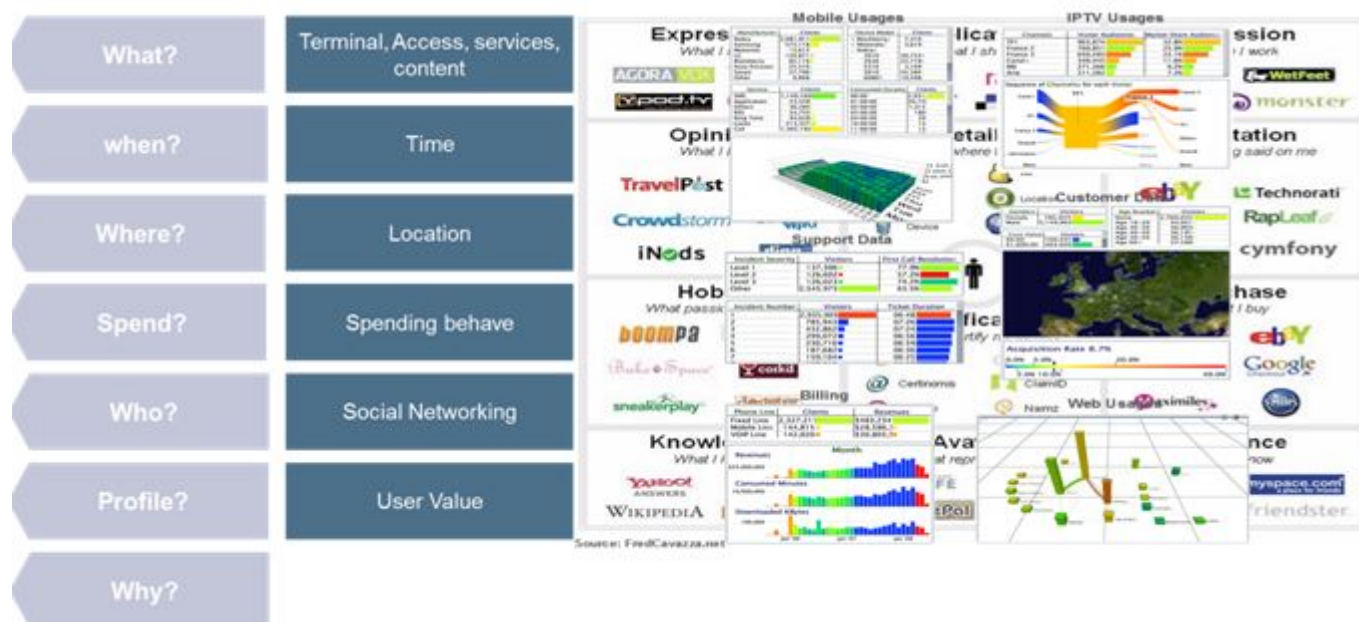
如前所述，全世界每天都有几十亿人使用计算机、平板电脑、手机和其它数字设备产生海量数据。这些数据通过运营商的网络传输，但却没有得到充分利用和货币化。在我们讲述数据货币化之前，让我们首先看下运营商的数据：

存在于运营商网络中的数据（举例）：

		
终端信息 节点：DM DB	位置信息 节点：LBS 系统等	互联网行为 节点：GGSN 等
终端品牌； 终端价格； 发布时间； OS 版本； 网络支持； 屏幕尺寸； GPS 功能； Wifi 支持； 摄像头； 等等	用户位置（工作时间）； 用户位置（工作时间之外；娱乐区中的用户频率位置； 不同地点的用户行为（与其它数据源合作）； 用户的社会身份（通过社区分析）	网上冲浪时间 网上冲浪技术（TD、GPRS、EDGE 等） 最喜爱的网站 网络行为（下载、视频、游戏） 社交网络的使用（QQ、微信、博客等） 数据量和时间分配 消费行为（淘宝、JD、VIP 等）

			
用户身份 节点：BSS、CRM	网络信息节点： HLR、HSS 等	CRM 信息 节点：CRM 系统	消费行为 节点：BI、BSS
姓名，年龄，性别； 预付费或后付费； 职业； 家庭状况； 业务订购历史； 开支； 奖励积分； 积分使用行为； 等等	SIM 卡信息； 漫游信息； 状态； 每英寸点数（DPI）信息； 业务信息； 等等	客户问题； 咨询历史； 反馈； 申告； 客户关系管理（CRM）频率； 等等	业务订购； 消费水平； 增值业务的使用情况； 数据和语音业务的使用情况； 每用户平均收入（ARPU）； 消费历史； 等等

通常，运营商数据中心中的大型业务支撑系统只是为了确保运营商能够对其客户所使用的服务计费。但是，在整合所有数据以及某些外部信息后，运营商确实将拥有每个用户的详细信息。



随着电信运营商的竞争格局不断变化，谷歌、Skype 等 OTT 服务提供商正在蚕食他们的收入。从他们的现有资产中获益并提供良好的客户体验正成为一个关键的成功要素。被 Ovum 誉为“增长燃料”的数据是运营商最宝贵的资产之一，而且他们也越来越热衷于更加充分地利用用户数据。

4 电信数据的货币化

电信行业在过去十年经历了多次更新换代，但其主要战略从未改变：

- 优化资本支出
- 降低运营支出
- 开辟新的收入来源

采用正确的大数据技术并实施一个有效的数据管理战略，可帮助运营商获得上述所有成功要素。

大数据将促进整个电信价值链的增长，并提升其效率和盈利能力。以下各图显示了大数据相对于传统数据仓库技术的优势。它们包括：

- 开辟新的收入来源

大数据项目与其说是 IT 项目，不如说是为了提升企业的盈利能力。运营商目前正处于一个过渡时期，他们既要提供高质量的传统业务，也要投资开发诸如机器对机器、移动商务和企业云等将能提升盈利能力的新业务。大数据是实现这一过渡的重要前提之一。以下举例列出了大数据所能开辟的一些新的收入来源：

“驮运”（Piggy back）业务：

运营商可以采用打包销售数据的方式为银行、零售商和 OTT 服务提供商提供增值服务。

定向广告与营销：

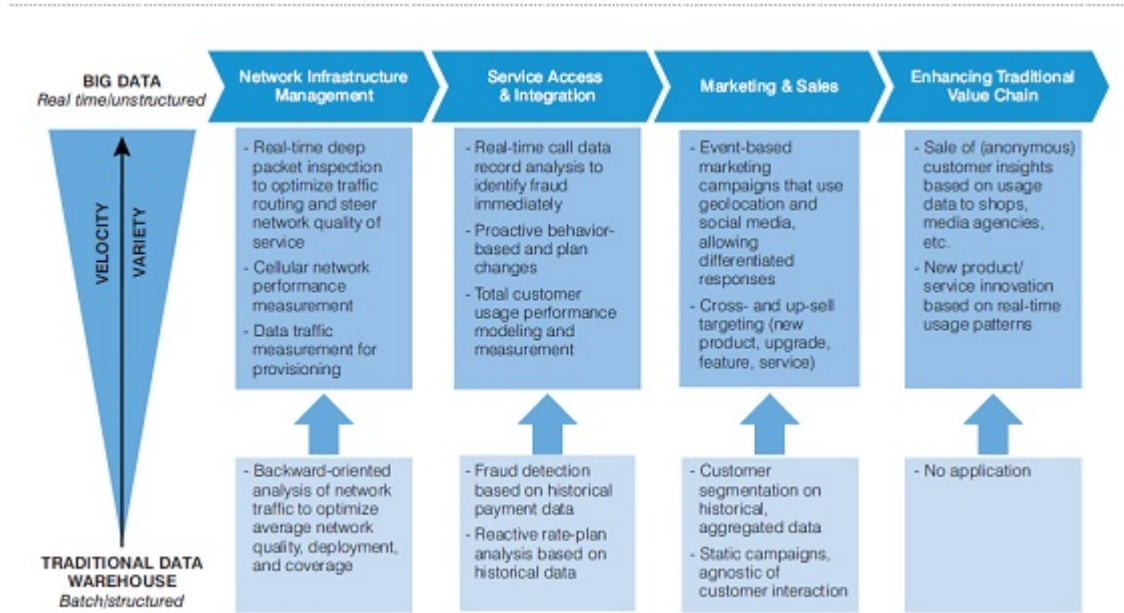
通过提供综合数据包，运营商可以帮助广告客户根据目标客户的资料、位置和消费习惯更加有效地投放相关广告。

- 改善客户体验

提高客户忠诚度和降低客户流失率是当今电信市场中的两个关键问题。通过从数据资产中获益，运营商能够更好地了解客户，并改善内部流程，例如，了解客户的行为、所喜爱的内容、设备类型等等。同样，人口统计和位置数据可帮助运营商做出有关部署网络和销售渠道的正确决定。此外，客户关怀部门也可以利用这些数据预测某个客户何时有可能流失，并采取相应措施。

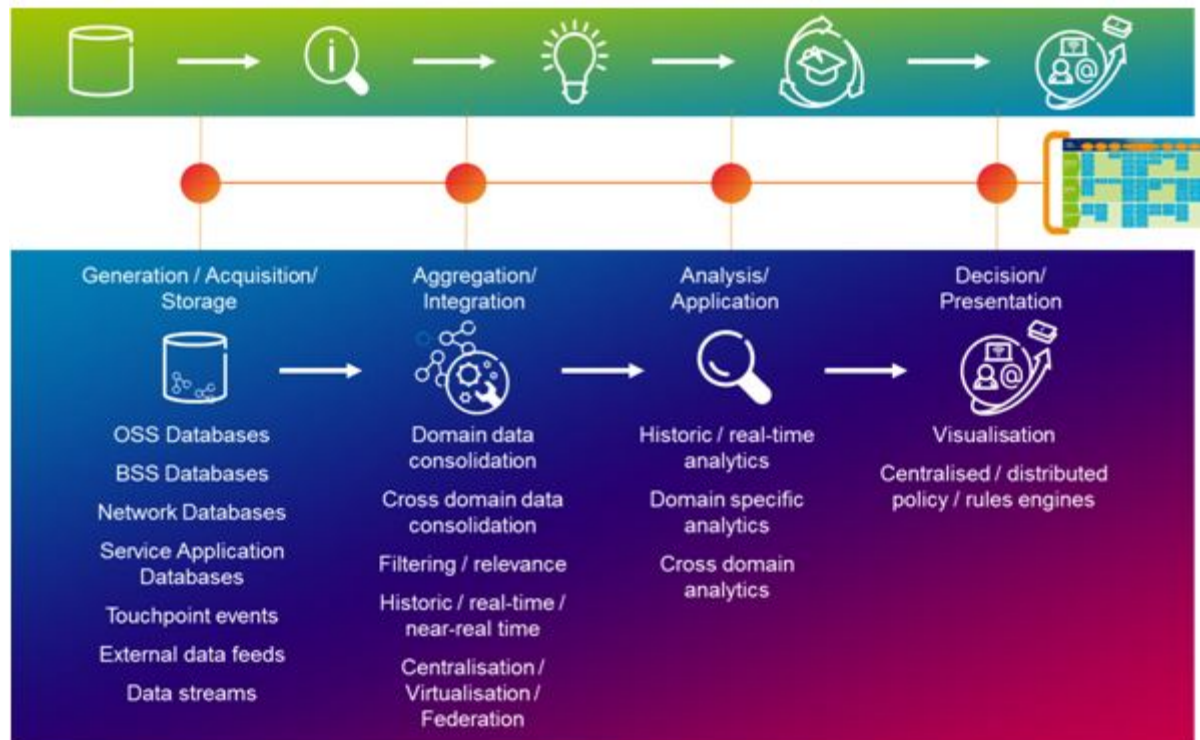
近年来，Orange、Telefonica、Vodafone 等欧洲大型运营商已开始利用数据分析技术来改进他们的管理决策。

Big Data Offers Benefits across the Entire Telecom Value Chain



Source: Booz & Company analysis

如上所述，“大数据”的重要性并非数据本身。事实上，我们已经拥有了大量数据。“大数据”是一个迅速增长的市场，包括捕获、存储、处理和分析运营商所拥有的海量数据，并从中获益。下图显示了运营商大数据流程的价值链。



大数据的最终目的是整合和关联所有信息来源，以便生成一个完整、透明、全面的视图，描述每个客户或家庭与运营商之间的所有交互。

但是，为了真正利用大数据，运营商必须彻底改变他们采集、验证、了解和利用他们所拥有信息的方式。

此外，运营商还需要学习谷歌、Facebook 等公司；在这些公司中，数据为王，几乎每一项产品决策都源于现有数据所透露的有关客户的信息以及如何使用这些数据的方式。大数据战略应涉及所有部门，包括网络运营、IT、产品开发、营销、财务等部门，甚至包括用户，这是因为他们可以利用自身的专业知识，采用各种新方法分析数据。

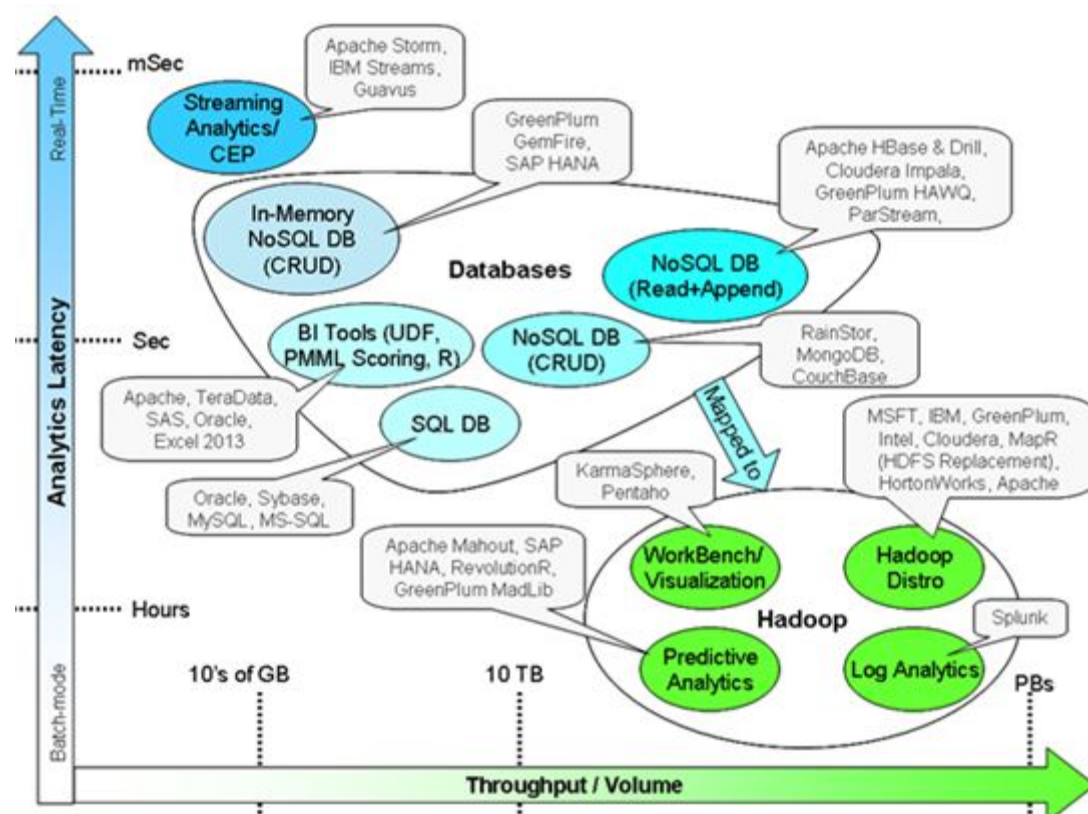
5 大数据中的软件技术

大数据技术描述了新一代技术和架构，目的是通过高速捕获、发现和/或分析，经济高效地从种类繁多的大量数据中获益。

目前共有两种顶层分析方法：

1. 分析过去，不预测未来（关联分析）
2. 分析过去，预测未来（具备监督式学习功能的预测分析）。

下图描述了大数据中的主要软件技术分类。这些分类中有很多正在开发的工具（既包括 Apache 旗下的开源工具，也包括各厂商开发的工具），这些工具可用于支持数据摄取、数据准备、数据库编程、文本处理、分析可视化等工作。



上图描述了大数据中的软件技术分类，并使用插图提示显示了每个分类中的厂商/产品。目前共有三种顶层的软件技术分类：

1. 流数据分析和复杂事件处理

结构化数据从多个来源持续流出，以便对它们进行“线速”分析和关联，而不是首先将它们存储在某个数据库中。Apache Storm 和 IBM InfoSphere Streams 等某些解决方案提供“表述编程设计”（declarative programming）框架，让数据经历转换、加入、分割、开窗等一系列处理步骤。这种模式通常被称为“复杂事件处理”（Complex Event Processing）。

流数据分析的结果通常被存储在一个数据库（SQL 或 NoSQL）中，并能触发其它事件。单位时间（例如 1 小时）内所处理的数据量通常以吉字节为单位，处理时延以毫秒为单位。关联分析和预测分析均能以线速运行，但预测分析中通常仅评分部分以线速运行。流数据分析的范例包括股票预测、自动交易引擎、M2M/传感器分析等。

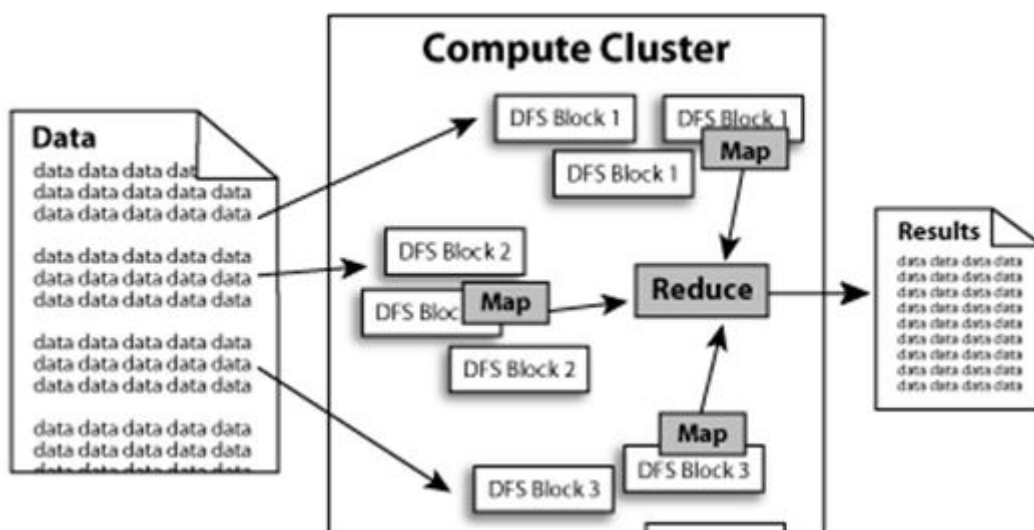
2. 数据库

目前主要有两类数据库，即 SQL 和 NoSQL。SQL 数据库向 RDBMS 确认（即提供事务处理和引用完整性）。NoSQL 数据库提供各种表、schema 和接口，但不提供事务处理和引用完整性，与 SQL 数据库相比，它们的时延更小，吞吐量更大，数据存储容量也更大。两种数据库均用于管理结构化数据。NoSQL 数据库既可以基于磁盘，也可以基于内存。内存数据库以牺牲较大的磁盘容量来换取更小的内存时延。

3. Hadoop

流数据分析和数据库用于处理结构化数据，而 Hadoop 用于分析 Web 浏览日志、IT 系统日志等半结构化数据以及社交网络、Twitter feed、图像、音频文件等非结构化数据。

为了分析数据，Hadoop 软件在计算机集群上运行一系列“MapReduce”任务。给定计算机上的每一个 Map 任务负责在给定时间处理某个数据子集；给定计算机上的每一个 Reduce 任务负责编译在预定义的计算机集群子集上运行的一组预定义的 Map 任务所产生的处理后的数据。下图显示了 MapReduce 任务的迭代过程。



数据和结果与分布式计算机集群中的每台计算机上的 MapReduce 任务共址。这些集体

数据构成了一个 Hadoop 分布式文件系统（HDFS）。任何一个 Hadoop Distribution 必须至少包含 Hadoop 集群软件和 HDFS。

大数据分析技术的未来发展

SQL 的重新兴起:

很多传媒大肆宣扬 NoSQL 是兼容 RDBMS 的 SQL 数据库的“终结者”。但是，RDBMS 可确保数据完整性，而这对于很多应用至关重要。因此，业内将来有可能搭建性能媲美当今 NoSQL 数据库的 RDBMS 数据库。

Hadoop 2.0:

Hadoop 将来有可能更多地被视为支持大型 NoSQL 数据库的一个平台，而不仅仅是一个批量分析引擎。在增添了流处理能力后（始于 Apache Storm），Hadoop 也有可能用于实时分析。大多数大数据厂商依赖于 Hadoop 的未来成功，因此，我们可能会看到在此方面的投入（如与 Hadoop 的 RESTful 接口，集成 Node.js 等）。

6 结语

大数据为电信运营商提供了一个更加全面了解其业务和客户、进一步加大创新力度的真正机遇。以研发投入占销售收入的比例计算，整个电信行业的研发投入远低于任何一个技术型行业，而其改变运营方式的努力也尚未取得广泛的成功。大数据要求各个行业采用一种完全不同的非传统方法来拓展业务。如果运营商能够以最快速度将全新灵活的战略整合到企业核心业务之中，就将获得真正的竞争优势，从而战胜行动较为迟缓的对手。

2013-09-04