中国数据分析行业核心刊物

数据分析

CHINA DATA ANALYSIS 用数据说话·做理性决策

中国商业联合会数据分析专业委员会 主办

《中国数据分析》会员特刊
2016年第04期 总第28期（季刊）
咨询热线：010 59000991 / 59000339
http://www.chinacpda.org/
投稿请至 xiehui@chinacpda.org

China Data Analysis

/ 预见2017 /

数据分析
CHINA DATA ANALYSIS 用数据说话 · 管理用决策

《中国数据分析》会员特刊 2016年第四期

本期目录 CONTENTS

主办单位
中国商业联合会数据分析专业委员会

编委成员
袁硕 / 李缘

出版时间
2016年第四期 12月出版

美工 / 设计
崔峻珩

联系我们
中国商业联合会数据分析专业委员会
地址：北京市朝阳区朝外soho C座9层
电话：+86-10-59000991 / 59000339
传真：+86-10-59000991转 607

投稿

欢迎广大读者踊跃投稿，内容包括学术观点、教学体验、教学活动、学习感悟、实战经验、随笔文章等。 稿件附图格式为JPG或TIFF格式，大于1M，分辨率在300dpi以上。

感谢您对《中国数据分析》的支持！

投稿邮箱：xiehui@chinacpda.org

/ 预见 2017 /

2016年，大数据开启了一次重大的时代转型，就像望远镜让我们能感受宇宙，显微镜让我们能够观测微生物一样，大数据正在改变我们的生活方式以及理解世界的方式，成为新发明和新服务的源泉。

在数字社会时代，智能手机、智能家电、智能工业仪表等嵌有传感芯片的机器无处不在，人类的一切活动都有了永远无法抹去的数字痕迹。每一次鼠标点击、每一次手指点按触屏，就会留下一个数字，这个数字与其它无穷无尽的数字一起，组成了一个数字化的社会。

用爆炸来形容今天的数据，实在是再恰当不过。这种数据爆炸是非常真实的，这种爆炸是前所未有的。我们今天面临的生活事实上在全面数据化，数据不是一切，但一切都将成为数据。一切业务数据化对任何企业来说都是一个终极状态，业务过程数据化就可以记录下来，就可以再现事实，可用于分析和预测未来。

只有分析才能真正让数据创造价值，这是从数据大国走向数据强国的基础。只有抓住数据分析这个根本，唯“分析”说话，才能真正从做大走向做强。所以物联网应该首先是一个数据分析工具的联网，对所有产生的数据进行价值挖掘，把数据转化为真正的数据资产。没有预置数据分析工具的物联网，并不能真正创造价值。

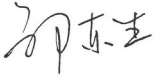
中国正从制造大国走向制造强国，这其实也是一个从数据大国走向数据强国的过程。在这个过程中，中国与制造业的关系不仅没有减弱，反而越来越强。尽管中国的劳动人口规模正在减少，工厂里面的工人数量不断减少，但这同时也是一个机遇，因为中国仍然是一个转型中、成长中的经济体。如果能够在生产中提升科技投资，这将会进一步推动中国的创新经济，让中国的工厂不仅仅依赖于工人的人数。

我们目前正在进行一个巨大的转型，转型到一个更依赖于网络、更网络化的经济模式。

作为“大数据”布局十年多的行业协会，我们看到了巨大的市场需求，同时也清醒的看到了大数据1.0时代带来的问题。

我们看到部分早期成立的事务所，在这样一个日趋繁华、需求日益旺盛的大市场中，仍然迷茫、缺少方向，是时候让我们真正沉下心来，努力积淀自己的实力了，否则再多的机遇也留不给“投机者”们。我们希望无论是协会还是数据分析行业的从业者们，能从各种热闹的场景下回归到自身能力的研发上，在人才体系培养以及深度咨询应用方面形成自主研发、不断更新的产品体系。让我们“以接地气”的方式为在这个时代下生存的企业和个人做点实在事儿。喧嚣结束后，专注经营决策数据分析业务的深化、把握“咨询+技术”的先进经验，才是事务所成就辉煌的“王道”！

中国商业联合会数据分析专业委员会



/ 第七届全国优秀事务所评选结果揭晓 /

文 / 协会会员处 袁硕　编辑 / 协会会员处 李缘　日期 / 2016-11

经过一个月紧张忙碌的评审工作，2016年数据分析行业优秀事务所评选活动圆满结束了。优秀事务所评选活动从2010年开始至今已连续举办七年。今年不仅保留了往年的评选方式，而且结合了事务所在大数据背景下的综合业务开发能力进行全面综合的评估。

大数据已经不简简单单是数据大的事实了，而最重要的现实是对大数据进行分析，只有通过分析才能获取很多智能的，深入的，有价值的信息。现在，这个理念被越来越多的人认同。通过评选活动，我们希望有针对性的进行行业规范和指导，促进事务所的发展，帮助事务所在不同领域中进行项目对接，推动大数据行业的发展，带动事务所的积极性，让事务所在大数据行业中更好的发挥优势。

2016年有三家优秀事务所从众多参选事务所中脱颖而出，这三家事务所不但保持着稳定的发展，同时在规范化经营等方面进行了严格的把控。长期以来积极参与协会组织的各类活动，并为协会提供了诸多宝贵的意见。

他们分别是:

湖南翰林数据分析师事务所

湖南翰林数据分析师事务所是湖南省较早成立,规模较大，附和资质较齐的专业从事各类数据分析及管理咨询的专业机构。事务所以专业项目数据分析服务为核心，同时与其他职业联盟单位紧密联合，构筑数据分析大平台，拓宽数据分析宽度，增强数据分析的深度。

承接了邵阳新奥奥迪4S店数据分析、年产3000吨无纺布生产线数据分析、京都世纪城涉税数据分析等16个规模达到百万元以上的项目，取得了丰硕的成果，为数据分析行业的发展添砖加瓦。

联 系 人：卿启伟
联系电话：13187299268
联系ＱＱ：835033160

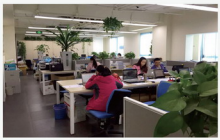
上海天元数据分析师事务所

上海天元数据分析师事务所是一家专业从事项目数据分析的服务性机构，事务所由多位高级数据分析师发起成立，并拥有一支集经济研究、金融投资、财务分析、工商管理等多领域的复合型团队。

上海天元在今年前三个季度中，承接了苏州凯联石英玻璃有限公司新产品销售预测分析、上海祁樱新材料发展有限公司扩大生产线项目、上海喷波新材料科技有限公司商业计划书等几十个项目，涉及行业广泛，深得用户好评。

联 系 人：王会会
联系电话：13917778657
联系ＱＱ：1412813247

北京鼎盛恒信数据分析师事务所



北京鼎盛恒信数据分析师事务所以全新的管理模式，完善的技术，卓越的品质为生存根本，始终坚持用户至上、用心服务于客户的理念，用自己周到的服务去打动客户。

鼎盛恒信数据分析师事务所在今年1月、4月、6月~7月间，承接了北京市西城经济科学大学绩效指标相关数据整理与分析、北京市西城区少年宫绩效指标相关数据整理与分析等8个项目，事务所通过更多的大数据服务，来提升不同客户的满意度。

联 系 人：唐丽丽
联系电话：010-58362048
联系QQ：794997164

/ 大数据落地，带来了权威的转移 /

文 / 网络大数据　编辑 / 协会会员处 李缘　日期 / 2016-11

翻看科技发展史，当某些跨时代的技术转折点来临之时，两种状况似乎同时出现：大众观念里的鸡同鸭讲，以及从业者对于概念追逐的狂热。近几年一个好例子即是：大数据。

公众视角

到了2015年，当马云数次宣称阿里正从一家IT企业转型为DT企业，舆论普遍关切的依旧是最新出炉的商业鸡汤。从农业社会，工业社会，到信息社会，中国两三步并作一步的路径，让数据概念几乎远离大众语境。

产业视角

大概三年多前，创业公司十之八九说自己能做大数据，与之相关的产业链也被切分为采集，处理，分析，可视化四个节点，一时间暗潮汹涌——泡沫也随之而来，从2014年进入2015年，行业开始小步洗牌，大数据从一个纯技术名词，或者说一个虚妄的概念，转变为应用范畴，以更为务实的姿态蔓延到一个又一个行当。时至今日，对大数据概念的热炒已有些无力，各种峰会与论坛所谈话题几乎都是如何让大数据“落地”的声音更为揭地有声。

当然，让数据落地并非新鲜论调，真正令人着迷的是落地的“程度”——数据即权力——大数据。

早在几年前关于大数据的哲学意涵呼之欲出的时候，数据是一种“宗教”就被不少学者大肆探讨。最近，关于数据未来最具象的描述，来自《人类简史》作者尤瓦尔·赫拉利，他为我们勾勒了一个“数据主义”的未来时代。

在他看来，那些极端的持数据主义世界观的“信徒”将整个世界视作一个数据流，任何事物的价值判断都由它对数据处理的贡献决定。“正如自由市场资本主义者相信市场无形的手，数据主义者相信数据流无形的手，当全球数据处理体系变得全知全能，接入这个系统就成为了一切意义的来源。”

从几年前从业者对大数据概念的热炒便知，大数据是个无远弗届的概念。赫拉利就曾写道：“数据主义允诺了人类在过去几个世纪里求而不得的科学‘圣杯’：一项将从音乐学，经济学一直到生物学的科学学科统一起来的无所不包的理论。根据数据主义，贝多芬第五交响曲，股票交易泡沫和流感病毒不过是三种数据流形式，能用相同基本概念和工具进行分析。”

数据主义的未来对数据从业者而言非常诱人——倘若数据是这个世界的主体，侍奉它的人无疑将收益颇丰。

搁置在上述宏大叙事框架中，如果你忘了未来由现实铺就，以下数字似乎顿时显得渺小。

2016年上半年，共有18家大数据相关创业公司获得上千万融资，新三板与大数据相关企业有50家左右。在寒冬中，资本市场对大数据项目怀有巨大热情。其中一个原因是，越来越多投资者厌倦甚至惶恐于B2C疯狂烧钱的迷途，开始转向那些由技术驱动，商业模式清晰健康的领域，大数据就非常符合他们的胃口，在包括赫拉利在内的一众预言家眼中，它看起来就是未来本身。

现在看来，无论数据采集，传输，建模存储，统计分析挖掘还是可视化，都存在创业者的短兵相接。但从他们的服务对象一端分析，必须承认，诸多企业在试图驾驭数据的过程中，都面临着如何快速有效地处理海量数据，以及密集的多源异构数据的现实境遇，对数据关键节点有效整合的方案缺失，也让决策者丧失了对自身业务的最终判断。

而站在数据运营商的立场，将项目充分产品化，脱离难以复制且交付难度

相对较大的项目制，无疑是将生命周期延续下去的最佳商业模式——在大数据行业，这并非易事，困扰大数据产品化至少有三个痛点：1、数据量太大，这对软硬件系统都会带来巨大冲击；2、作为决策依据，数据处理的效率必须非常之高；3、多样性，行业和业务场景的不同，会对数据的呈现方式有不同要求。

所以买卖双方因素相加，做数据的生意，理想之举无疑是提供一整套完善的解决方案——用户的需求加快了大数据行业从软件到硬件到一站式应用的产品化进程。毕竟，面对行业的多元化需求，人们对过去单一的数据分析产品似乎愈加不满，在理解数据的方式上，他们希望最好能有从数据发现，存储，到可视分析，再到交互模式的一站式产品。

举例来说，最近被投资人颇为看好的海云数据就发布了通用性综合解决方案“图易大数据决策产品生态平台”，提供一站式整体解决方案，协助企业，园区，城市，政府，重新通过数据认知自己的业务。

举这个例子是因为，在面对任何项目和节点里，他们80%的工作都已经完全实现标准化和模块化，另20%则可根据不同行业属性和应用节点定制。

总之，这个时代，驾驭数据的能力是所有决策者“技能清单”里最重要的一章，因为任何行业，无论竞争，合作，还是管理，其本质都将趋向于“信息战”。而技术变迁史又同时告诉我们，无论哪个行业，谁能率先在行业中拾起新工具是多么重要。

譬如对于数据的掌握，谁都知道数据可视分析能最大化做到价值变现，但当不少决策者还将思维停留在用饼状图和百分比了解业务，或者用守旧的IT系统装备自己时，他们已经落伍了，而那些对诸如图易这种“正在发生”的新工具敏感的决策者，无疑拥有了赢得信息战的利器。

因为，历史早已证明，当一项跨时代的技术转折点来临之时，除了开头所说的“大众观念里的鸡同鸭讲”以及“从业者对于概念追逐的狂热”，还有另一件更重要的事同时发生：新技术的诞生本身即是一个人群分类过程，它将人群划分为“会用它的”和“不会用它的”——率先拿起新工具的人总会走在竞争者前面。

越是在科技发展深处，这种分野的力量就将越加强大。

TIN

/ 全球深度学习系统市场报告：TOP6深度学习企业 /

文 / 大数据实验室　编辑 / 协会会员处 袁硕　日期 / 2016-10

Technavio 是一家全球技术研究与咨询公司，其最近的《全球深度学习系统市场报告》选出了全球Top 6的深度学习机构，分别是谷歌的母公司Alphabet、伯克利视觉学习中心(BVLC)、Facebook、蒙特利尔大学的LISA lab以及微软公司。

报告称，到2020年，全球深度学习系统市场规模将超13亿美元，2016~2020年期间的复合年增长率达到38.73%。

深度学习具有在现实生活中应用的巨大潜力，这也使得它越来越受到关注。在实际应用中，社交媒体、软件服务协议、硬件、网站Cookies以及应用程序权限等为训练神经网络提供了大量数据。深度学习网络在从这些数据中提取有价值的信息方面很有优势，因为它们擅长无监督学习。

上图中的：1）美洲市场得到BFSI行业快速增长的数据的驱动；2）北美有超过1500家AI公司；3）2010年以来，金额最大的四项AI收购案例均发生在英国；4）亚太地区将是AI增长最快的区域，到2020年预计年复合增长率将达到41.58%。

全球范围里的深度学习系统是一个高度分散的市场，既有许多大型公司，也有无数的初创企业。

Technavio的分析师预计，随着不同类型的企业进入这个市场，市场竞争将加剧。拥有差异化产品的企业具有更大的竞争力，为了在这个市场上站稳地位，企业将持续创新。此外，随着先进技术的广泛普及，产品更加多样，产品的选择将变得更加复杂。Technavio评出的全球Top 6深度学习企业如下：

1、Alphabet

2015年11月，Alphabet的子公司谷歌宣布在开源Apache 2.0许可下开放TensorFlow深度学习框架。TensorFlow

是由谷歌机器智能研究团队开发的深度学习框架，可用于使用数据流图进行数值计算。该框架由Python API组成，使用数据流图执行数值计算。TensorFlow拥有快速增长的用户和贡献者社区，使它成为非常受欢迎的深度学习框架。

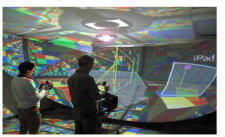
2、BVLC

BVLC(伯克利视觉学习中心)及其社区贡献者在 BSD 2-Clause许可下开发了名为Caffe的深度学习平台。Caffe是使用C++开发的，具有表达简洁、快速和模块化的优势。Caffe拥有许多特征，如富于表达力的结构、可扩展的代码以及不错的速度。

3、Facebook

Facebook 开发了深度学习框架Torch，能用于训练大规模卷积神经网络，适用于图像识别等AI应用。Torch 是一个系统的计算框架，广泛支持各种机器学习算法。它同时提供的Tensor库具有非常高效的CUDA后端，且神经网络库可用于构建具有自动微分功能的随机非循环计算图。

4、LISA Lab



蒙特利尔大学的LISA Lab开发了名为Theano的深度学习框架。Theano是一个软件包，或说是一个数学表达式的编译器，能够有效地定义、评估和优化包含多维数组的数学表达式。Theano允许用户在不同的架构(如CPU或 GPU)上编程。Theano不仅可用于CPU密集型的机器学习，也可用于大规模神经网络或深度学习。

5、微软

CNTK(微软认知工具包)是微软的一个开源深度学习工具包，可用于加速 AI开发，使在多个GPU和服务器上组合训练多种深度学习模型变得容易。CNTK适用于许多应用，例如：语音识别、机器翻译、图像说明、图像识别、语言建模、自然语言理解、文本处理等。

6、Nervana Systems

今年8月英特尔正式收购深度学习创业公司Nervana Systems，其目的是加强英特尔内部AI解决方案的作用。Nervana 开发了基于Python的深度学习框架 Neon，并且最近在Apache 2.0许可下开源了。Neon 具有定制化的CPU和GPU后端，分别名为Nervana CPU后端和Nervana GPU后端。

TJS

/ 解析AI大数据，行业创业与投资趋势 /

文 / 中国大数据 编辑 / 协会会员处 李缘 日期 / 2016-12

大家都知道人工智能和大数据这块现在特别火，从数据来看都获得了非常多的投资，更多的可以看到的是机器视觉，语音识别，虚拟助理，智能语音机器人这块，从整个趋势来看国内目前比较火的是机器学习和场景应用这两个方向。

大数据其实从很多年前就开始在讲述这个概念，随着计算量的增加和计算能力的增强，最近才得到长足的进展，用机器学习来找出数据之间的关联和因果关系变得越来越可能，所以这也是最近人工智能和大数据比较火的一个原因。

从整个产业发展来看人工智能已经几起几落好几次了，和互联网刚刚起来那个阶段差不多，移动互联网从IPHONE开始出现那个阶段开始起来，目前人工智能也刚好处于这个阶段。

未来的10年20年人工智能会成为一个主流，未来我们可能不会叫互联网+，而是叫人工智能+。

AI产业链分布

我把它划分成：基础层，基础层又分为数据层和计算能力，数据层像自身能够不断产生数据的Alphago一样，它会自己不断地产生数据和进化它的模型，计算能力就像之前提到的硬件的加速，神经网络芯片这一类的计算能力提供商，也有做像GPU云计算的开放给创业公司。

技术层：在基础层上面是技术层，它又分为框架层，像IOS这些框架和操作系统。另外一个是算法层，比如机器学习，深度学习这些增强的这一类学习的算法。

应用层：应用层有两类，一类是通用这一类，就是我们所看到语音识别，图像识别、NLP、FLAME、传感器融合、路径规划等这一类的；另外一类就是特定的一类应用层，就是我们所说的行业应用，就是我们所说的自动驾驶，医疗，教育，大数据征信等这一类的应用层，整体我会把整个产业链分为这三层吧。

从创业和投资角度来看AI大数据的机会。

从整个发展来看我们可以分为三类：一类是底层的基础构架；二是通用场景应用；三是专用应用。底层基础构架比如BAT、华为这些公司会利用自身服务器的优势，数据的优势，为各类创业公司或者传统的公司提供基础资源的同时，他会把自身的优势转化为通用和专业领域的一些研究，让自己形成闭环。这个是初创的公司比较难去touch的在基础的构建上面。

还有一种我们在看计算机视觉、语音视别这一类通用场景的应用方面，像FACE++,商汤包括我们投的山世光老师的中科时拓这些都是用来提供某一类的通用场景的一些应用。但这一块从现在看来，我们的初创企业没有特别的技术上的突破和优势的话，在这方面的机会也不会太多，因为已经不少明星的创业者已经拿到钱。

专用应用领域里面目前看来是初创公司机会最多的一个领域，一定要想好如何把技术和应用相结合，完善应用场景。

另外提一点，在基础算法这一方面，虽然我们看到的很多的基础算法都是二三十年前甚至是六十年前人工智能刚刚被提出这个概念的时候就已经有了一定基础。但最近大家也知道中国人在算法这个领域其实是非常强的，大家有志于算法研究的，觉得也可以进行一些研究。

中美创业者在AI领域的差距

在关于整个人工智能领域的研究和探索上，中国还是落后于美国的，从之前的技术角度讲差距还是比较大，在人工智能这块的应用层面我认为差距没有那么大。我们可以看到很多的论文发表之后很快就会被消化吸收，就像Google前几个月做的图像识别的竞赛当中我们的结果也都非常的好。所以我认为在应用领域其实在接近，但是在基础研究方面差距还是非常大的。整体来看美国的人工智能方面的创业这块其实还跟其他领域的创业一样，会更加倾向于更加基础的深度研发，而且是市场在驱动。

中国的创业者在实际应用方面探索得更多，在商业化的这条路上从一开始就比较他们小更多和走的更具体，因为毕竟我们有这么大的一个市场。具体来讲中国会在商业化方面走的比较快，美国会在技术方面走的比较强，很多时候是从应用层面来推动技术方面的研究。

美国这些创业者，一开始起步比较慢，融资也比较慢，主要是从整个市场来找一些需求，是让市场来推动他们的产品。但是中国的创业者融资都非常快，有一些TO VC的倾向，反而走到后端的时候就会比较慢，甚至有些停滞不前或者找不到方向。反而在这块中国的创业者是需要向美国的创业者多学习的。

AI大数据哪些赛道有投资机会及原因

未来的世界会是一个AI+的世界，所以广泛来说是每个赛道都有投资机会。但是如果我们把它拆分起来看，其实大数据是整个人工智能的一个基础，单从大数据上面来看，会有数据源的问题。

那如何去获取这些数据，我们看到像八爪鱼这一类企业他们去爬取公网的一些数据，去建立一些数据基础，另外一方面你只看大数据的时候你会发现这些数据需要做数据清理，清洗，甚至有些数据都是一些孤岛，那可能需要做一些ETL的整个数据的融合。数据的融合之后，你需要做数据的分析，在整个链条上其实都有机会，反而我认为这些机会比人工智能来得要更为实在。

如果我们再看人工智能这块，人工智能如果你看它的基础层，其实是他的基础的数据和他的运算能力和他的算法，如果你有什么特殊的渠道可以抓住数据源的话，一定是一个非常高的壁垒。另外的话是运算能力，运算能力更多的是体现在硬件上面，所以你会看到最近有非常多的GPU，FPGA这些创业公司出来，他们都是在解决运算能力这一块。解决完运算能力就要去看算法，算法这一块你可以看但是会比较难，因为基础算法这二三十年来没有发生太大的变化，而且对算法能真正做出有价值的贡献从全球来看这方面的人才也不多。

同时我会说到从应用层面，我们觉得现在这些底层技术要想看哪些解决得比较好，一定要去找应用场景，我们目前会觉得比如自动驾驶，智能语音助，语音客服这些会是一些比较好的方向，当然还有一些跟人工智能跟医疗结合的辅助诊断，人工智能跟IOT结合的，我们来预测设备发生故障的概率，用来节省人工成本和提升效率。

从应用层面来说，整个赛道会非常非常多，因为人工智能就像当年的互联网一样他就是这样的一个底层技术，它用来提升各行各业的效率，甚至我们会认为它会比互联网更加深入到我们的各行各业中，比妨说用在第一产业用来预测气候状况，用来提升农产品的产能，也可以放在第二产业用在工业4.0上面，用很多AI的方法帮他们提升效率。

对人工智能大数据领域的思考和给

创业者的建议

这一块我其实在前面几个问题里有提到，我还是再强调下我个人的看法，大数据和AI一定要和实际场景结合才能产生价值的，就像我们过去几年的创业，如果我问你是做哪方面创业的，你说是做互联网创业的，同样你说你是做大数据或者AI的，投资人都会打一个大大的问号。

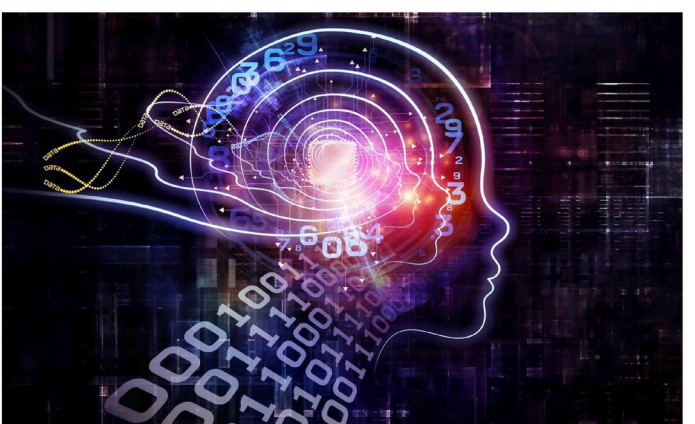
因为其实大数据和AI目前的形势下没有一个直接变现的模式，从过去互联网的创业最赚钱的两个行业来看，一个游戏，一个电商这个其实是非常清晰和直接的，大家都知道商业模式越直接，商业路径越短越好。但大数据和AI一定是要绕个湾的，所以一定要找实际应用场景结合的一些领域来进行创业。

另外我们大数据和AI的创业可能和我们以往的创业很不一样，人会非常非常关键，现在所有的公司都在抢AI和大数据方面的人才，我们可以看这块的初创公司一上来就会有巨额的融资，因为这块人才成本比较高，起步资金比较高，而且因为没有非常明确的变现的模式，所以对投资人来说我们宁可给到创业者一大笔钱，让你慢慢的跑，慢慢的去尝试。

市面上所有技术本身是不赚钱的，必须要跟实际业务结合才能获利，所以必须要想清楚谁是你的用户，需求是什么，痛点是什么。必须要找到靠谱的小伙伴，切入点一定要小，尽量深耕一个领域去试错去积累，不要上来就做一个大的平台，因为做平台是非常难的。不论你在哪个领域里面去创业，但是在AI里面去创业你一定要找到你的稳定的独特的数据源。我相信所有的投资人都会问你，你能不能拿到属于你的稳定的数据源。这个我想信是每个有志于在AI大数据领域创业的人需要去思考的。

曾提问过曾颖哲：未来一两年内AI会给智能硬件带来哪些重大改变？

“我们基金智能硬件投的不少，黄明明之前投了极路由，小牛电动，我们也投了车和家，但其实说句实话我并不会把他们叫做智能硬件。因为我觉得很多人就是联个网做个APP就是智能硬件，但其实我认为很多需求是想象出来的一些需求，比如智能开关，其实睡觉了就是按一下开关就好，你还要打开手机，打开APP再按一下，你觉得其实这真的叫智能吗？没有带来本质上的变化，只是一个噱头而已。所以我认为做智能硬件一定要去想究竟用在一个什么样的场景下给用户解决了什么问题。真正和人工智能结合起来的硬件我们才会真正把它叫智能硬件，就是一个无缝的连结，不需要过多的人工干预的就像Google收购的NEXT，他会自动学习你的偏好，你的喜好，他也会学习你到家的匹配时间，去把你温度匹配到你最适宜的温度，真正带有人工智能的硬件才叫智能硬件，去和你的生活无缝的衔接，真正给你的生活带来方便”。



/ 大数据会怎样改变一个传统行业的运营模式？ /

文 / 大数据观察　编辑 / 协会会员处 李缘　日期 / 2016-12

大数据会怎样改变一个传统行业的运营模式呢？

在前不久举行的2016英特尔中国行业峰会上，英特尔公司销售市场部副总裁兼行业解决方案集团全球总经理乔农·波林分享了一个传统农业是如何与大数据结合的故事。

他以美国一家非常大的农机公司为例。这家生产拖拉机等农业机械设备的传统行业公司也要拥抱新趋势：他们要把依赖手工和人工劳动力投入的行业转变成依靠技术发展的行业。

他们是怎么做呢？他们在拖拉机的维护上就作了一些新的技术尝试，比如说安装芯片收集GPS信息，除了采集拖拉机本身的数据，还通过拖拉机在农田中的耕作收集农田的情况，比如土壤的湿度、温度、构成等。然后，他们告诉农民，根据收集到的土壤信息，应该做什么样的农业生产。

现在这家农机公司不仅卖给农民农业机械，卖的是更多、更高端的服务，而这种服务是基于技术收集的数据基础。

"这就是创新和变革的力量。"波林说。

实际上，整个行业环境的竞争格局都在发生变化。行业内生性变革和跨界竞争交织在一起，不断冲击着传统行业的固有业务模式和竞争格局。在2016英特尔中国行业峰会上，来自金融、能源、医疗、交通、零售、教育等行业的专家、企业高管、英特尔业务负责人以及合作伙伴们，一起分享了数字经济如何挑战传统行业并推动新兴行业的发展。

在零售业，从云到端的技术和应用为消费者打造了完美的购物体验，打通零售全渠道营销，助用户更加精准、快速地响应市场变化，捕获无限商机；在医疗行业，创新技术正在帮助用户提升运营效率、完善就医体验，让高质量、个性化、便捷的医疗服务触手可及；而在传统金融业，银行正在从古老的限定位置、时间的线下网点服务，转向即时可用、随时随地的个性化金融服务，这使得用户可以更加灵活地应对不断涌现的金融业务创新和技术挑战。

各行各业都面临着相同的答卷：在数字时代经济大潮涌动的时候，巨大的市场需求和科技创新驱动着商业变革。

前段时间，波林与一些金融行业，特别是银行的客户进行交流。他们告诉波

林，现在银行业处在巨大变化的过程中：银行业过去的对手比较传统，在未来的几年里，一些新的业务模式将给银行业带来更强劲的挑战，比如说互联网金融。如果放在20年甚至10前，这些新兴的公司不可能成为传统公司的竞争对手。

据波林了解，银行业正在努力实现自身的变革，让自己能更好地成为颠覆的力量，而不是成为被颠覆的对象。而这一切才刚刚开始。

在过去10~15年，全世界都经历着计算带来商业模式的变革。以云计算、大数据和万物智能互联等为代表的新兴技术正在为全球范围内的商业和企业带来变革。以Airbnb为例，它用6年时间发展到200万间客房，而著名的酒店连锁企业美国万豪酒店集团用了90年的时间才做到100万间客房的规模。

“云在颠覆我们的传统业务。”英特尔公司数据中心事业部副总裁及云服务平台事业部总经理芮洁安妮·斯基尔伦女士说。在她看来，云实际上是现在业界最大的颠覆者，它是一个总值达到2000亿美元的行业，其中一半是软件及服务，以网络为基础面向普通消费者的服务，同时也会对传统业务在向数字化进行转型的过程中起到巨大的推动作用。

云意味着很多全新的可能。斯基尔伦每一次来到中国都会和很多新公司交流，它们是在新趋势下应运而生的新企业，他们每天做的工作是改变消费者的生活方式，社交、娱乐、即时沟通工具，都是基于公有云计算的基础上。

以美国的UPS公司为例，每天运输的包裹数量达到2000万个，这些邮包如何准确送到用户手中？UPS用人工智能进行他们的业务。他们打造了一个系统，使用非常先进的远程信息系统GPS路由以及卡车司机过去的驾驶数据分析，综合在一起，这个系统能够帮助司机实现最优的路线规划。

实际上，一天少开一英里可能不算什么，关键这个数据以年来计算，每年就是8500万英里，换而言之，意味着相应量的汽油的节约。

云会不断扩展，每秒钟都会有新的数据产生，每12个月产生的数据量要翻番。斯基尔伦说，我们在创造数据时，也要存储数据，但是在使用数据方面做得并不是很好。斯基尔伦和同事们做了一些研究，在现有捕捉的数据中，只有5%的数据实际上真正得到了使用，95%的数据被存储了，然后就再也不去碰了。

必须承认的一个事实是：未来的世界是属于新型创新的公司。

波林认为，很多企业在创新中被颠覆，在未来几年，40%的企业将会以某种方式受冲击：他们可能被收购，也可能因为商业模式不再奏效，完全没有业务。也许还有一些企业因为能够创新，带来整个行业同类公司的变化。

在未来的竞争中，除了技术，还必须能够打造一个吸引创新人才的职场。“如果做不到这一点，5-10年后这家公司就会死去。”波林说。

ITX

/ 中科院秘书长邓麦村：发展大数据人才是关键 /

文 / 网络大数据　编辑 / 协会会员处 李缘　日期 / 2016-11

11月17日，第三届世界互联网大会大数据论坛在浙江乌镇举行。该论坛由中国科学院主办，中国科学院信息工程研究所承办，数据中心联盟协办，是第三届世界互联网大会“互联网创新”专题的重要组成部分，也是首次由中国科学院主办的大会专题活动。

中国科学院党组成员、秘书长邓麦村在代表主办单位致辞时指出，当前，以信息技术为代表的新一轮科技革命方兴未艾，信息技术与经济社会发展深d度交汇融合，数据迅猛增长成为重要的基础性战略资源。

邓麦村提出三点倡议：

第一，推进大数据基础研究和技术攻关。

第二，加强大数据专业人才培养。

创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系，重点培养专业化数据工程师等大数据专业人才，大力培养具有统计分析、计算机技术、经济管理等多学科知识的跨界复合型人才，积极培育大数据技术和应用创新型人才，注重培养网络信息安全专业人才。

第三，深化大数据国际合作交流。

下面是邓麦村先生的现场致辞：

当前，以信息技术为代表的新一轮科技革命方兴未艾，信息技术与经济社会发展深度交汇融合，数据迅猛增长成为重

要的基础性战略资源。大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要而深远的影响。

如何突破大数据关键技术，如何运用大数据推动经济发展、完善社会治理，如何在推动大数据发展的同时确保信息安全，已成为世界各国和各行各业普遍关注的热点问题。

本次论坛以“大数据的发展与安全”为主题开展研讨，就是希望能对相关领域的技术和产业发展有所促进。

中国科学院作为中国自然科学最高学术机构、科学技术最高咨询机构和自然科学与高技术综合研究发展中心，按照国家经济社会发展和科技创新总体部署要求，秉承“三个面向、四个率先”的办院方针，一直重视大数据基础研究和技术攻关，以及科学大数据的开发利用。

在大数据前沿技术研发与应用方面，为应对终端接入规模、海量数据处理性能、能耗和安全等四大挑战，中国科学院于2012年启动了“面向感知中国的新一代信息技术研究”战略性先导科技专项，组织二十多个研究所的科研力量协同攻关，现已形成以人工智能芯片“寒武纪”、代数处理器芯片、深度可编程网络、三元融合安全技术、海云大数据系统等为代表的一系列原创性成果，在智慧城市、社会治理等领域得到了成功应用。

在可信大数据技术、大数据安全通信、大数据访问控制、身份认证授权等大数据关键技术上，以及量子通信技术上已取得重要突破，成为国家大数据安全领域的中坚力量。同时，在数字地球、全球变化、高能物理、基因组计划、深空探测等领域，利用大数据技术驱动科技创新，也取得了一些重要成果。

在科学大数据积累与应用方面，上世纪70年代，中国科学院就开始建设专业数据库。经过几十年的持续部署和推动，中国科学院现已建成服务全国科技界的“中国科学院数据云”，整合了各学科领域的1340个数据库资源，共享数据达

655TB，年均在线访问超过千万人次。此外，中国科学院的网络安全工作也得到了行业主管部门和业界同仁的肯定。

按照《中国科学院率先行动计划》和《“十三五”发展规划纲要》，未来一段时间，中国科学院将在大数据领域加强相关基础科学问题研究和软硬件关键技术开发，继续引领国家科学大数据建设，为国家大数据发展和大数据安全保障体系建设作出应有贡献。

借此机会，我愿向大家提三点倡议：

第一，推进大数据基础研究和技术攻关。大数据的快速发展提出了许多新的科学问题，仍有很多关键技术亟待突破。我们应深入开展数据科学研究，在大数据理论、方法及关键应用技术等方面进行探索，不断提升数据分析处理能力、知识发现能力和辅助决策能力，形成安全可靠的大数据体系。

第二，加强大数据专业人才培养。发展大数据，人才是关键。大数据的快速发展对专业人才提出了非常急迫的需求。我们应创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系，重点培养专业化数据工程师等大数据专业人才，大力培养具有统计分析、计算机技术、经济管理等多学科知识的跨界复合型人才，积极培育大数据技术和应用创新型人才，注重培养网络信息安全专业人才。同时，还应依托社会化教育资源，广泛开展大数据知识普及和教育培训，不断提高社会整体的认知和应用水平。

第三，深化大数据国际合作交流。大数据的快速发展给世界各国都带来了共同的机遇和挑战。我们应坚持平等合作、互利共赢的原则，建立完善国际合作机制，积极推进大数据技术的交流与合作，充分利用国际创新资源，共同促进大数据相关技术和产业发展。

中国科学院愿意同国内外同行和社会各界朋友一起，在新一代信息技术和服务业态蓬勃发展的浪潮下，共同为大数据发展与安全贡献力量！最后，再一次对国内外同行和社会各界对中国科学院科技创新工作的支持表示衷心的感谢！

END

/ 2016第四季度全国大数据业务成交汇总 /

编辑 / 协会会员处 李缘　日期 / 2016-12

采购单位	中标时间	中标金额	中标单位	项目名称	项目编号
国家信息中心	2016 年 10 月 8 日	¥3498824.8 元	厦门市美亚柏科信息股份有限公司	国家信息中心互联网大数据经济主体行为识别服务采购项目	0714-EMTC02-ZC5696
交通运输部海事局	2016 年 10 月 12 日	¥1810000.00 元	山东中创软件工程 股份有限公司	海事局船员大数据应用项目	0745-1640CCIEC050
国家质量监督检验检疫总局	2016 年 10 月 13 日	¥1700000.00 元	北京中百信软件技术有限公司	统一社会信用代码（组织机构代码）大数据应用服务平台系统开发项目	0733-166212226601
广州市地方税务局	2016 年 10 月 13 日	¥2420000.00 元	北京华胜天成科技 股份有限公司	广州市地方税务局大数据分析二期	0692-169C04420304
西南大学	2016 年 11 月 2 日	¥4895000.00 元	北京天同诚业科技 有限公司	大数据创新应用平台采购	2016-02-85
中日友好环境保护中心	2016 年 11 月 4 日	¥590000.00 元	北京人民在线网络 有限公司	中日友好环境保护中心生态环境大数据2016 年度建设项目—环境舆情与公众热点数据集成分析项目	0702-1641CITC2139
上海出版印刷高等专科学校	2016 年 11 月 18 日	¥1885692.00 元	上海睿泰信息科技 有限公司	上海出版印刷高等专科学校基于大数据的移动信息化课程建设	SQ169174
吉林市科技信息研究所	2016 年 12 月 1 日	¥570000.00 元	吉林市东杰科技开发有限公司	吉林市科技信息研究所购置吉林市科创云计算大数据综合平台及吉林市大型仪器共享服务平台开发建设项目	JSZFCG—J2016ZX265
河北地质大学	2016 年 12 月 2 日	¥2188000.00 元	石家庄硕威实验室 设备销售有限公司	大数据与地学空间信息技术实验平台（电磁法遥探系统）	HB2016103602080009
央视国际网络有限公司	2016 年 12 月 8 日	¥5099000.00 元	三江天地（北京）信息技术有限公司	央视云和大数据 项目-IPTV 基础设施扩容项目设备项目	0701-164110080440

从以上2016年第四季度大数据业务成交量中可以看出：

（1）大数据在各行各业尤其是政府、高校、企业应用广泛，并且与我们的生活联系紧密；

（2）大数据项目的中标金额平均不低于百万元，可见大数据项目关注度在持续升温；

（3）政府、企事业单位以及高等院校投入大量资金主要用在大数据平台的开发建设以及基础设施扩容；

（4）从招投标的总体情况来看，大数据项目依旧朝着良好、稳定的方向快速迈进。

/ 基层社会治理要善用大数据 /

文 / 人民网　编辑 / 协会会员处 袁硕　图 / 崔峻珩　日期 / 2016-11

社会生活中常有一些揪心现象，引民众抱怨。比如：办理准生证，“16天跑了8个单位，开了5份证明材料，涉及8个盖章环节”；假期旅游，景点有的人满为患，有的门可罗雀；路口红绿灯发挥治堵效用不高，绿灯无车和红灯长龙并存……诸如此类，不胜枚举。

如何解决？

其实，基层政府部门掌握着人口普查、经济普查以及大量与社会经济生活息息相关的数据，但由于一些部门存在数据“小农意识”、缺乏大数据思维、习惯对政务数据秘而不宣等原因，部分数据被锁在柜中、束之高阁，各职能部门之间形成了“数据壁垒”。如果职能部门能做到数据共享，打造诸如“一站式办理”“城市大脑”“互联网+信号灯”“实时旅游热力图”等项目，即可高效快捷破解“公章四面围城，审批长途旅行”困局，解决城市发展与人民生活难题。

党的十八届五中全会提出，“实施国家大数据战略”。习近平总书记强调，要深刻认识互联网在国家管理和社会治理中的作用，加快用网络信息技术推进社会治理。创新基层社会治理，必须重视大数据的价值。

不管承认与否，我们已快速进入大数据时代。可以说，谁掌握了大数据，谁就掌握了主动权。在基层治理中，大数据正日益成为社会管理的“强力推手”、政府治理的“幕僚高参”。这里的关键是，广大基层干部的格局视野、思维理念、紧迫感和敏锐度，要与大数据时代相适应，善于运用大数据优化政府服务和监管，提高行政效能。通过日益完善的电子政务系统，在推进更加方便群众的在线服务同时，还能做到权力运作有序、有效、“留痕”、“规矩”，极大促进政府与民众的沟通互联，构建新型政民关系和畅通民主协商渠道。

如何做好大数据模式下社会治理创新？首先，就要将其作为一项紧迫且重要的一把手工程，由基层党委政府主管亲力亲为，自上而下做好顶层设计并一抓到底。其次，是将其作为一场基层执政理念提升的思想革命，积极引导基层干部涵养“互联网+”思维，把“节约居民每一秒钟”等理念根植于心。第三，是将其作为一项发动群众参与的系统工程，紧紧把握群众需求“痛点”进行整个政务流程、服务标准等机制再造，不断提升群众对政府信任和依赖的“黏度”。第四，是将其作为一项政能政府主导下的民心工程，充分依托大数据技术优势，探索群众“不上门”的行政审批和“送入户”的政务服务，彻底消除“条块不畅”“信息孤岛”等壁垒，解决服务群众“最后一公里”甚至“最后一米”问题。

基层社会治理创新，离不开这场席卷全球的大数据革命。惟有尽快同步升级“互联网+”思维下服务群众理念，积极推动高效精准政务服务，努力让过去诸如群众办事“跑断腿”、社会管理“粗线条”、部门信息“不并联”、政府决策“样本少”等现象彻底消失，才能乘上大数据这列时代快车。

END

“十三五”规划年内有望发布，大数据产业迎重要发展机遇

编辑 / 协会会员处 李缘　图 / 崔峻珩　日期 / 2016-11

“工信部信息化和软件服务业司牵头组织编制的《大数据产业发展规划(2016-2020年)》将在年内出台。”工信部信息化和软件服务业司司长谢少锋日前出席大数据与信息安全企业家峰会时透露，新一代信息技术产业加速变革，市场应用需求处于爆发期，大数据产业迎来重要的发展机遇。

随着大数据产业“十三五”发展规划的发布，大数据产业将迎来新的发展机遇。业内人士认为，未来五年大数据产业市场仍将保持高速增长。去年发布的《促进大数据发展行动纲要》指出，到2020年，培育10家国际领先的大数据核心龙头企业，500家大数据应用、服务和产品制造企业。

政策大力支持

据了解，上述规划将作为引领DT(数据处理技术)发展的指导性文件，内容包括推动大数据在工业研发、制造、产业链全流程各环节的应用，支持服务业利用大数据建立品牌、精准营销和定制服务。

目前多地大数据“十三五”规划开始启动。《兰州市大数据产业发展“十三五”规划》提出，到2020年，大数据产业将成为新的产业增长极，相关企业达到5000家以上，部分关键技术研发及特色应用达到国内领先水平，相关产业规模达1000亿元以上。

今年5月，贵阳大数据交易所发布《2016年中国大数据交易产业白皮书》，预计2016年末中国大数据产业市场规模将达到2485亿元。随着各项政策的落实，到2020年，中国大数据产业规模或达13626亿元的高点。

工信部信息化和软件服务业司人士介绍，2015年8月，《国务院关于印发促进大数据发展行动纲要的通知》明确了大数据发展的指导思想、发展目标和发展任务，标志着大数据已成为重要战略资源，大数据发展将充分享受政策红利。

国民经济和社会发展“十三五”规划纲要提出，“实施国家大数据战略”，将大数据提升至国家战略层面，明确要把大数据作为基础性战略资源，全面实施促进大数据发展行动，深化大数据在各行业的创新应用，探索与传统产业协同发展新业态、新模式，加快完善大数据产业链。

业界对大数据“十三五”规划充满期待。华泰证券分析师认为，作为引领DT时代发展的顶层设计文件，规划主要涉及到个人信息采集应用的范围和方式的界定，以及大数据在工业研发、制造、产业链全流程各环节应用的推动等，数据资产将加快整合，推动全产业链创新发展。

存储领域率先受益

目前A股市场涉足数据产业的公司不在少数。数据显示，大数据概念板块目前有32家公司，主要集中在计算机行业。此外，不少传统行业公司通过并购涉足该领域。

“这个统计并不完整。”参与大数据产业“十三五”规划制定的中国电子信息发展研究院电子信息产业研究所所长安晖指出，通信、计算等行业的公司大多与大数据有关。从数据生命周期看，可以将大数据企业分为数据采集、整理、存储、分析挖掘和数据应用这几个部分。各个环节都会涉及相应的软件、硬件开发和服务。

从相关公司的具体情况看，可分为大数据资源类、大数据存储和运行维护、大数据分析应用、大数据安全等。“数据采集”处于产业链上游，占据开发价值较大的流量入口。目前除互联网巨头、电信运营商外，行业信息化龙头企业也积极卡位资源入口。

上市公司方面，易联众基于海量社保卡资源聚焦民生大数据，恒生电子发力金融大数据。此外，还有语音输入公司科大讯飞以及地图大数据公司四维图新。

不过，在大数据淘金热中，最先

受益的可能是大数据产业链中游的大数据存储和运维。中信证券分析师指出，受益智能时代数据量爆炸式增长，存储行业有望持续保持15%以上行业复合增速，国内厂商将数倍于行业平均增速实现市场扩张。

业绩体现方面，同有科技前三季度实现营收2.75亿元，同比增长44.27%；净利润6576.1万元，同比增长156.93%。对于业绩增长的主要原因，公司表示，耕耘国防军工行业多年，是军队信息化领域最大的国产存储系统供应商，受益于国防信息化和军方大订单，业绩支撑动力充足。

同时，政务和大型企事业单位数据中心以及IT基础设施建设提供商，将受益于政府资源平台开放和共享建设加快。天玑科技2014年启动超融合一体机研发，一体机2014年度实现收入约1062万元，2015年实现收入约2090万元，2016年前三季度已销售将近6000万元。

分析挖掘服务是核心

“数据的分析挖掘服务是产业核心，也是最具有商业价值的一部分。”安晖指出，作为大数据产业链的下游，对大数据进行分析应用是实现价值的终点。把软硬件的研发生产、数据收集等行业都加入到大数据产业中，这使得大数据产业的概念更为广泛。

中银国际研究员吴友文表示，2015年是大数据应用启动元年。虽然大数据产业链并没有迎来预期的爆发，但在调研中发现大数据产业链依然在高速成长，未来一到两年内实现大数据应用的全面爆发确定度较高。

吴友文经过对整个大数据产业链的调研发现，大数据底层软件、数据强化产业以及相应的数据分析产业已逐渐成形并加速走向成熟。海量数据经过有效的分析处理从而可以支撑大数据应用，并完成“数据→信息→知识→决策”的整套数据到应用变现链条。

“目前存在一批专业从事大数据分析挖掘的服务型企业。”安晖介绍，大数据行业还处在发展初期，短时间内还没有形成规模特别大的企业，但不少企业的成长性很高。

在A股市场，开展大数据应用的公司主要涉足医疗健康和互联网金融领域。“医疗和金融行业离开IT技术基本上是寸步难行。”有IT企业负责人表示，随着“互联网+”在各个行业加速融合，大数据资源的积累逐渐丰富，大数据技术和应用产业有望保持快速发展。

网络安全应运而生

伴随着大数据产业的快速发展，信息安全需求日趋提高，优质的信息安全龙头企业将率先受益。

谢少锋在大数据与信息安全企业家峰会上指出，加强信息安全保护、构建强有力的大数据安全保障体系至关重要。作为政府主管部门，将为大数据信息安全发展营造良好的政策环境，全力支持相关企业、科研院所开展大数据生命周期安全研究，研发数据来源可信、多源融合安全数据分析等新型安全技术，推动数据安全态势感知、安全事件预警预测等新型安全产品研发和应用。

谢少锋提出，支持建设一批大数据安全实验室。组织研究建立软硬一体化的模拟环境，支持工业、能源、金融、电信、互联网等重点行业开展数据入侵、反入侵和网络攻防演练，提升数据安全防护水平和应急处置能力。

中国互联网协会理事长邬贺铨表示，2015年全球连接到互联网上的设备达到49亿台，2020年将超过260亿台。随着“互联网+”计划的推进，传统行业逐步数据化、在线化、移动化、远程化；人、企业、社会服务甚至整个世界都与网络深度绑定，将产生巨量连接和巨量数据。

“在云计算、大数据、移动互联网等新兴技术广泛应用的万物互联时代，过去相对独立、分散的网络已经融合为深度关联、相互依赖的整体。”中央网络安全和信息化领导小组办公室副主任王秀军强调。

奇虎360总裁齐向东指出，奇虎将大数据的方法与现代网络安全技术相结合，通过对各类网络行为数据的记录、存储和分析，可以发现异常、捕获威胁，实现快速监测、发现和响应，通过“数据驱动安全。”

在此背景下，上市公司积极完善业务布局。卫士通完成了对三零盛安、三零瑞通和三零嘉微三家信息安全企业的重大资产重组，形成从芯片到模块、从单机到系统的信息安全完整产业链；启明信息通过并购快速扩大版图；2016年以来，绿盟科技持续投资及并购安华金和、深之度、剑鱼科技、敏讯科技、亿赛通等。安华金和是国内专业的数据库安全产品和服务提供商，剑鱼科技在移动终端相关产品的研发及推广方面具有优势，敏讯科技是国内领先的专业反垃圾邮件安全厂商，亿赛通主营数据安全和网络内容安全管理产品。

数据显示，网络安全行业上市公司今年前三季度业绩实现大幅增长。券商分析师指出，2016年以来，安全可控国产化替代进程加快，加速接盘国外安全厂商的部分客户，相关公司的业绩呈现强势增长。

/ 国家信息化先破政府信息孤岛 /

文 / 光明网　编辑 / 协会会员处 袁硕　图 / 崔峻珩　日期 / 2016-12

12月7日,国务院常务会议通过了《“十三五”国家信息化规划》。信息化既然是国家规划，政府当作表率。李克强总理的表态和要求，便是给各部门作出的表率。

总理明确说，信息孤岛要坚决打通，起码政府系统不应再有。会议确定的规划重点，首先便是打破信息壁垒和“孤岛”，构建统一高效、互联互通、安全可靠的国家数据资源体系，打通各部门信息系统，推动信息跨部门跨层级共享共用;实施“互联网+政务服务”等信息惠民工程，加快推进公共数据资源向社会开放。

所谓信息化，不仅在于“有”，更在于“通”。尤其是在互联网时代，信息必须实现有效流通和共享，才谈得上“化”。关于这一点，本届政府的思路和举措始终瞄准全球科技革命新浪潮和政府施政新理念。李克强总理力倡“互联网+”，写进了政府工作报告。其中，在政府端，结合简政放权放管结合优化服务的“自我革命”，就是围绕“互联网+政务服务”着重发力的。

客观说，我国信息化已有相当基础。一个显著标志是，各部门建设起为数众多的各类信息系统。然而进一步信息化的梗阻恰恰也在于此。这些系统之间缺少联接和沟通，往往“各自为政”，形成了一个个信息“孤岛”。

正因为此，总理今年以来已多次要求治理政府信息孤岛这一弊端。他指出，目前我国信息数据资源80%以上掌握在各级政府部门手里，“深藏闺中”是极大浪费;一些地方和部门的信息化建设各自为政，形成信息孤岛和数据烟囱，给企业群众办事创业造成很大不便。不久前召开的深化“放管服”改革座谈会上，李克强再次“点题”，各部门的各类信息系统，甚至一个部门内的各个系统，打不通的话，“是多大的资源浪费?”“老百姓办事怎么能方便?”

政府信息孤岛的背后，既有部门利益、揽权的算盘，也有不愿作为、懒政的拖延。事实上，另一个被总理屡屡提及、老百姓也抱怨不止的问题——“奇葩证明”，就与信息孤岛直接相关。因为部门信息不共享，公民的户籍、就业、生育、婚姻等基本信息处于分散的碎片化状态。这导致一些明显不合理的证明，由于不同部门难以掌握相关信息，彼此之间仍要求通过纸质的证明，才能验证事实真伪。这样一来，群众便只能在分散的部门之间“跑断腿”、“寒透心”。

“孤岛”隔绝的不仅是信息本身，它甚至在某种程度上疏离了政府的服务和这种服务首要的对象——老百姓。简政放权，首先是为了百姓的福祉;推进信息化，也首先应体现在群众能享受信息化的便利。国家信息化必先破政府信息“孤岛”。总理在常务会上给“十三五”国家信息化规划定的重点，可谓切中肯綮。

END

如何成为优秀的数据分析师

/ 中美资深人士访谈：如何成为优秀的数据分析师 /

编辑 / 协会会员处 李缘　　图 / 崔峻珩　　日期 / 2016-11

导读：做数据分析师这个行业，从表面上看像是高富帅的行业。很多毕业生趋之若鹜，然而对这个行业有所了解的人都知道做这一行不容易，所需的技能、勤奋、精力时间只多不少。

我们今天面临的生活事实上在全面数据化，数据不是一切，但一切都将成为数据。只有分析才能真正让数据创造价值，这是从数据大国走向数据强国的基础。只有抓住数据分析这个根本，唯“分析”说话，才能真正从做大走向做强。

那么，什么样的人才是大数据人才？如何从“菜鸟”快速成长为优秀的数据分析师？国内外数据分析人才价值体现的不同以及未来几年数据分析师行业会出现“人荒”？我们特别邀请了中国商业联合会数据分析专业委员会，邹东生会长、北京犀数科技有限公司首席数据官，孙雪女士、龙霖国际 董事长，美国资深大数据专家Jimmy先生，三位老师，分别从国内行业角度、专业从业人员角度和美国行业角度来与大家进行分享和探讨。

Jimmy Wu

龙霖国际 董事长

拥有30余年海外从业经验，80年代末创立美国西海岸最大的IT公司，创建并运维美国也是全球最大的电影评论、影迷网站，00年代创建美国少数几家团购网站，与Acer共同成立专注于Linux产业的公司，曾获英国肯特亲王接见。

邹东生

中国商业联合会数据分析专业委员会 会长

北京市青联第十届委员、北京大学光华管理学院MBA、北京大学光华管理学院MBA校友导师。中国数据分析行业发起人、奥基人、丰富的企业经营管理咨询经验，资深数据分析专家，主持编写《投资数据分析》、《经营数据分析》等书。

孙 雪

北京犀数科技有限公司 首席数据官

研究个性化推荐模型、营销模型、消费者行为预测等方向。擅长文本大数据的统计建模及数据挖掘，从事文本内容的深入研究、 机器学习，信息安全方面的研究，曾主持多项数据分析重点科研项目的研究。

主持人：大家好，我是本次访谈的主持人，很开心我们邀请到了龙赛国际董事长美国资深大数据专家Jimmy老师，中国商业联合会数据分析专业委员会会长，邹东生先生以及北京犀数科技有限公司首席数据官，孙雪女士参与此次访谈！

主持人：首先有请几位嘉宾和大家聊聊：想要成为优秀的数据分析师需要具备什么样的职业素养？

邹会长：作为数据分析行业的监管协会，我认为合格的数据分析师首先要遵守行业执业操守，坚持数据客观、真实的引入分析，不对数据进行人为造假。当然，这与在分析过程中进行合理的假设是不同于人为造假的；其次，在大数据时代，优秀的数据分析师应注意知识体系的完备和均衡，具体说就是：知识体系要健全(如：战略、管理学、营销、财务、统计、投资、行业知识、基础计算机知识等)，大数据的分析是一个综合学科、需要复合型人才，只精一门知识，很难成为合格的数据分析师，所以，业内优秀的执业数据分析师，大多有商学院背景；另外，就是要注意先进的分析工具的学习，毕竟现在是大数据时代，对大数据的处理和分析，要依赖专业的大数据工具及技能，只有这样才能使我们的数据分析师能够更快地进入角色，引入大体量的数据研究。

Jimmy老师：我补充说明，要成为好的数据分析师除了培养对数字的敏感度，喜欢这些数字，还要从看似杂乱的数字理念当中找出一些别人看不到的头绪，整理出来，这点很重要，所以说你一定要喜欢这门行业。而且如果分析出别人没有分析出的结果，你会越做越好。

孙雪老师：我很赞同以上两位嘉宾的观点，在我看来一名优秀的数据分析师肯定要具备基本的数据分析能力，比如数据收集、数据清洗、建立数据库，熟练掌握各种算法，数据可视化，对行业大数据场景的深入研究，撰写数据分析报告等等。那么优秀的数据分析师，就是要具备多元化的分析能力及实际操作能力，就是要对你所处行业有深刻的了解，对项目的关联性有全局的把控，在分析的各阶段都具有很强的纠错能力。我觉得只要具备以上能力就可以称之为优秀的数据分析师了。

Jimmy老师：另外，还要注意加强行业间的专业知识，因为做数据分析需要对这门行业足够了解或者说对公司流程很了解，如果对跨界知识有一定的理解能力就更好了，因为你可能会用到不同行业的外部数据。

主持人：大家都知道，IT人员从事一些程序开发的工作，那么IT从业人员算数据分析人才吗？

邹会长：现在越来越多的IT人员开始关注大数据，但我认为即使从事与大数据相关的程序开发工作，也不能完全算是大数据人才，从国际上而言，对大数据人才的认知，更多倾向于从事数据研究，也就是学会分析数据，这很好理解，因为数据需要分析才能展现价值，否则数据越多越无效。当然，现在有很多IT人才开始向分析领域延展，比如统计算法、模型及应用场景等的实现，开始关注企业真正的决策需求，如果是这种模式的话，就应该算是大数据专业人才了。

Jimmy老师：严格意义上来讲，IT开发人员不算数据分析师，有IT开发者本人对数据分析也很有经验，在国外，有的IT开发者就会做些数据分析的工作，把它写到计算机语言中。

主持人：可企业要拥抱“大数据”，毕竟是从数据收集开始，第一步是不是要先由IT人才进行数据搭建，然后才能有数据分析师发挥空间呢？

Jimmy老师：数据收集不一定由IT人才操作，其实很多公司本身已有数据了，有从物联网、社群、论坛里来的，这些都可以用来分析。例如，我们帮戴尔做数据分析的时候，社群就发挥了很大的效果，有5000多万的使用者每天在社群上交互，戴尔一般是不管理它们的，但社群上有好的也有不好的内容，那么数据分析在其中就显得尤为重要。

邹会长：主持人提的这个问题挺好的，很多企业或个人都有这个疑惑，觉得没有数据怎么分析，所以应该让IT或大数据技术人才先搭建大数据平台，将数据存贮下来，再考虑分析如何开展。但我认为恰恰相反，越是这种情况，越要逆向行之，即一定要先从企业的需求出发，先分析企业最需要的决策难点，研究企业需要提升业绩的关键点或风险控制的关键点，再设计分析解决方案，进而引入模型、算法，查找相应的数据由内部的数据引入加上外部的数据融合，并根据数据体量及未来发展的规划，考虑云端大数据分析平台的规模或实施方案。那么在这基础之上，如果企业想在大数据化过程中，少花钱、少走冤枉路，就一定要先从大数据的分析入手，从大数据的场景搭建入手，这是数据分析师的长项，这一两年，大量的企业开始数据化改造，这正是数据分析师未来可以大显身手的时候。

主持人：从数据分析的应用领域来看，数据分析人才是怎么划分的？

邹会长：我觉得不同领域有很多种方法，我说一种分类的方法，我们可以把它分成三类：1、基础数据分析人才——基于本岗位业务开展需要、有一定的数据处理及基础分析能力，比如一定采集和清洗等，这类人才企业会大批需要，因为很多企业都需要有数据分析能力的人才；2、专业数据分析人才——企业的数据分析部门、专注于本企业、本行业的深度数据分析的长期研究；3、执业数据分析师——在事务所、专业大数据公司从事数据分析工作，服务跨行业、跨领域的客户，要求知识更深入、解决能力更强、对数据的把控更准确有效，可以说是越来越高的要求。

主持人：感谢嘉宾们的精彩解答，今天来参加访谈的伙伴们很多是从CPDA毕业正在从事数据分析工作，还有一些是对数据分析感兴趣的新朋友，我们有请嘉宾和大家分享，如何最快摆脱“菜鸟身份”？

孙雪老师：主持人让我谈谈如何最快摆脱“菜鸟”身份，其实可以理解为，如何在最短的时间内高效率地成为数据分析师。我的建议是要针对性的训练以下这3个方面：1、对业务的熟悉

度，2、对数据的熟悉度，3、对各种模型算法有一定的熟悉度。具体来讲，在业务上要多思考，我们在实战中得出经验就是往往找出一个业务上的关键指标会比一个复杂的数学模型更能得到漂亮的结果。在数据上，主要体现在数据处理上，比如，如何清洗，如何变换，如何抽取特征，只有得到一组干净有代表性的数据才能使模型的结果更真实准确。

Jimmy老师：摆脱“菜鸟身份”首先一定要很努力，培养对大数据分析的兴趣，找到好的思路，最后就会是事半功倍，有很好的效果。

孙雪老师：在数据建模上，需要学习大量的背景知识，比如各种模型算法的思想、用处、适用范围、优缺点等，能够在拿到数据后知道选用哪些模型能够得到想要的结果。注意到这3方面后，就需要大家迅速行动起来，快速学习必要知识，多多进行练习和实战。相信大家会越来越有信心的。

主持人：分析工具很重要，那这些工具与分析技能之间的关联是什么？

邹会长：可以这样理解，分析工具是帮助数据分析师实现价值提升的重要助力，大数据分析一定要充分利用IT、相关软件、云计算、云分析平台等先进的工具，否则在随着数据体量的增大、或实时数据时，就会出现困难。

但有一点大家要注意，毕竟大数据要通过分析才能引导企业进行有效决策，所以我们经常说：大数据的核心价值是如何将数据玩起来，所谓的“玩”就是指你的分析能力，否则，你对工具再熟练，没有一定的分析思路和思考深度，那也只是浅浅的介入，无法产生真正的商业价值。

Jimmy老师：我觉得工具与分析技能之间的关联就是：你希望能够达到什么样的目的，做数据分析希望得到什么，才知道选择什么样的工具，现在工具越来越多，如果真的找不到，那你就需要组合几种不同的工具甚至是自己来开发工具。

邹会长：另外，很多没接触过大数据的企业和接触过大数据的企业经常会说大数据没有很多人想像的那么神奇，不就是对数据进行统计、展现，或简单的算法模型来解决实际的问题，其实这种误解，多数是因为数据分析师专业分析能力缺失造成的误解，其实优秀的数据分析师可以帮企业改变的东西是很多的。

孙雪老师：是啊，所有的数据挖掘和分析都和行业的发展特性息息相关。只有对自己所处行业有所了解才具备较强的数据分析预测能力，有目的解决企业的问题。我们常常能看见很多大数据分析平台，在分析效果上或解决问题方面很差，形成的统计结果偏向简单，对实际工作的指导作为不大，这些都是分析能力的不足造成的。所以，建议数据分析师一定要关注在数据分析能力的培养上。

主持人：现阶段，很多人关注大数据的一些工具使用，从简单的Excel、SPSS，到R语言或Python，请问，这么多的工具，是不是要全部学会如何使用呢？

邹会长：其实我觉得不必纠结你想使用哪些工具，这些分析工具都是一个目的：帮助数据分析师迅速从纷繁复杂的数据中，迅速找到数据背后的决策规律，帮助企业提高决策效率。所以，无论哪个工具，只要你熟练运用即可，不必样样都学，最快、最好解决数据问题才是最佳选择。

当然，分析工具有简单或复杂之分，EXCEL可以处理简单的数据分析问题，但R或PYTHON则更灵活，可以自己设计计算脚本，解决具体数据化问题，所以从专业分析师的角度，我建议学习R或PYTHON两种之一。

Jimmy老师：我觉得对于工具的使用可以是条条大路通罗马，只要你专心的把一种工具学好，其它的当作辅助也是可以的，有时间都学是好的，但如果一种工具学精了，再学其它的工具就会很快。

主持人：R语言或Python，如何选择，有没有什么建议呢？

孙雪老师：以我从业经验来说，个人观点，我更偏于后者。虽然R语言作为常用软件包含统计、科学计算都有，但Python还是有独特的优势，例如：爬虫、与平台对接，场景设计等，更有优势。我们强调工具是指分析工具，如EXCEL、SPSS、R、python等，但不要本末倒置，不是要求分析师去学习编程等计算机知识，当然对数据库有一定的了解甚至操作是必要的，毕竟要取得数据，但一定不要认为大数据的编程技术就是大数据的核心。

主持人：接下来请老师解答：从数据的收集、整理、预处理、分析、纠错、可视化、撰写报告，数据分析师必须要掌握每个环节吗？

孙雪老师：这要看你自己的职业规划是什么，或者说你在企业中处于什么角色。但你要成为优秀的数据分析员或数据分析师那肯定是要掌握每个环节的。

Jimmy老师：我建议每个环节都要知道，都要去理解，都要做过，然后选择其中一个环节专一地去学习，在美国、德国、英国等国家分工已经很细了，大家都专心做一部分，但是，全盘了解是有必要的。

主持人：请孙老师谈谈：数据分析师如何帮助企业介入数据分析，整合难么，有外部数据加入吗？你在操作过程中认为最难攻破的点是什么？

孙雪老师：外部数据是很重要，如果没有外部数据融合，数据分析的字段不丰富，那可分析的角度或深度就会受限，而且引入外部数据，也是大数据分析场景设计过程中，数据分析师十分重要的思考和要解决的一个关键问题。

主持人：接下来我们有请Jimmy老师和大家分享美国数据分析行业的现状。

Jimmy老师：关于这个话题我想说现在不仅是美国，欧洲一些工业强国，尤其德国和英国大数据行业发展也是很好的，而且这些国家对数据的分析和应用方面都是蛮重视的。因为我在硅谷待了有20多年，我就着重聊聊美国的情况。大家肯定都知道，2012年3月，美国奥巴马政府宣布投资2亿美元启动“大数据研究和发展计划”，这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署。

此后，美国的数据分析行业形成了

政企联手的局面，以美国科学与技术政策办公室、国土安全部、美国国家科学基金会、国防部、美国国家安全局、能源部等都和民间企业、大学开展很多关于大数据的研究与开发。

据我所知，美国中央情报局通过利用大数据技术，将分析搜集的数据时间由63天缩减到27分钟。大家都形成了共识，如果你不用大数据介入自己的业务，你就注定会失去竞争力。现在美国的数据行业很细分，也很专业，每个部分都有专精的公司或者人才专员，这些细分和专业加起来就形成了完整的产业链。

经过很多年的积淀，AI和BI的部分都应用的很纯熟了，大多数的企业有了数据驱动意识，能够利用专业的数据分析师团队将数据转化为价值，在决策层面已经摒弃了经验决策走向数据决策。

主持人：感谢Jimmy老师带领大家扩宽了视野，那国内的现状是怎样的呢？

邹会长：国际数据分析人才已开始深度细分：商业数据分析师、行业数据分析师、营销数据分析师等等，原因：大数据基础好、开展早、行业研究深入，相对国内，则还刚刚扫盲，很多所谓专业的大数据事务所或公司，在研究深度、解决方案等方面还有很大的差距，但从另一方面也说明了机会所在。

首先，企业的需求在未来是旺盛的：企业在大数据时代必然经历数据化的改造包括存、洗及决策、企业未来要大量运作数据进行各项工作：新产品设计、市场、销售、供应链、财务、投资等等；有相当一部分企业对大数据认知深度不够！企业对大数据还是一知半解，那么解决方法就是提高数据驱动意识，因为，很多企业还没有意识到数据驱动能为企业带来的巨大价值，或者说只有少数超大规模的公司意识到了这一点，这使得创始人的决策、商业知觉远远重于数据驱动。

其次，人才引入是关键：关注技术多于关注分析！分析引导技术，而不是技术引导分析，成功的案例少、落地的大数据企业不多，数据分析师要更务实，从企业的收益提高或成本降低入手，不能只有概念，没有效果。让企业少花钱、多赚钱。再加上先进的分析理念引入不多，要加强数据科学体系的引入，协会加强与国际行业组织的交互，这次课堂就是很好的载体。

主持人：刚才Jimmy老师提到，美国很多政府职能机构都在大数据应用方面取得很好的成效，那么，国外政府的数据资源已经完全开放了么？

Jimmy老师：美国政府自始至终都在推进这项工作的开展，在联邦政府大力推动下，7年来在Data.gov开放的数据集由最初的47个增加至18.3万个。而且在美国有很多公司是依靠加工政府开放数据而实现其商业价值的，例如处理天气数据的Zillow公司，theweatherchannel公司，以及处理GPS数据的Garmin公司，它们的总市值已经超过了一百亿美金。

主持人：中国数据分析行业发展迅猛，数据分析人才供不应求，接下来，我们有请会长和大家聊聊，他是如何看待我国大数据人才“人荒”的？

邹会长：“人荒”是当前一个明显的事实，因为国内大数据人才培养速度明显低于大数据发展和应用的速度。据调查，尽管全国50%的大数据人才集中在北京，但是北京的很多公司仍然普遍遇到了合格的大数据人才“招聘难”和“留人难”的问题。这一现象在短期内不会有明显改善。也正因为如此，我们协会的工作核心之一依旧是加强数据分析人才的培养。

主持人：我们都知道中数委向社会输送了大批数据分析人才，而且CPDA数据分析师和CDA数据分析员的课程都在持续更新，会长能就这方面和大家聊聊吗？

邹会长：近期我们一直在加强CPDA和CDA课程在应用实战层面的升级优化。与国际先进理念接轨的课程改革也在有条不紊的进行。而且我们也在积极建设全国最大、最权威的数据分析师社交体系。让新老学员在协会这一平台上能够得到更有价值，更有实际意义的，更紧密的信息共享。

主持人：大家有人问道，数据分析人才的薪资待遇怎样？

邹会长：这个要看数据分析师是否具有实战经验，一般年薪是20-30万。这个和工作年限成正比，北京某地产公司年薪100万招数据分析人才。数据分析人才缺口180万，好工作来找你。

主持人：数据分析工作，是一个不断循环、不断提升的过程。不是说，大数据时代是我们能赶超欧美国家的拐点么，国内的数据分析师们，我们一起加油吧！感谢各位嘉宾的精彩分享，访谈活动到这里就结束了！

END

/ 想做数据分析师？先弄懂这10种分析思维 /

文 / 51CTO 编辑 / 协会会员处 李缘 日期 / 2016-10

一、逻辑思维

逻辑思维即明白价值链，明白各项数据中的关系；该方法的关键在于明白其中的关系要求你对这项工作要了解、熟悉，要细致和慎密，要清楚充分性和必要性的关系。 实际上也就是指：你需要那些数据？如何获得这些数据？数据之间的关系如何？

二、向上思维

在看完数据之后，要站在更高的角度去看这些数据，站在更高的位置上，从更长远的观点来看，从组织、公司的角度来看，从更长的时间段（年、季度、月、周）来看，从全局来看，你会怎样理解这些意义呢？也许向上思维能让你更明白方向。 该思维方法的关键是：建立长远目标、全局观念、整体概念、完整地分析数据，不做井底之蛙。

三、下切思维

数据是一个过程的结果反映，怎样通过看数据找到更多的原因以及隐藏在现象背后的真相，需要我们下切思维，把事物切细了分析，把过程拆分细了分析。此 时 关键是要知道数据的构成、分解数据的手段、对分解后的数据的重要程度的了解。也就是说那些数据需要分解分析？ 这也如同 显微镜原理

四、求同思维

当一堆数据摆在我们面前时，表现出各异的形态，然而我们却要在种种的表象背后，找出其有共同规律的特点。 关键是找到共性的东西进行分析，还要客观。 实际上就如同：现在的整体数据表现出什么问题？ 是否有规律可行？

五、求异思维

每一个数据都有相似之处，同时，我们也要看到他们不同的地方，特殊的地方 。 这就需要对实际情况的了解，对日常情况的积累，对个体情况的了解，对个体主观因素的分析。 正如：你了解你的下属员工吗？ 如何帮助她们分析问题，从自身找到解决方案。

六、抽离思维

当你从一个旁观者的角度不思考看待数据时，你往往能发现那些经常让我们迷失方向的细枝末节并没有太多的意义，我们迷失方向，忘记了自己的价值，同时 深受情绪困扰。这时，你用用抽离思维更加能够帮助到你。 关键是要用多种分析方法，多角度看问题，不要钻牛角尖，多学习别人的好方法，学会集思广益，发散性思维。 比如说：你的学习能力和方法有效吗？

七、联合思维

很多销售数据，需要我们能站在当事人的角度去思考和分析，这样你才会理解人、事、物。

关键在与多了解当事人的情况，学会换位思考。 比如：你了解你周边的情况吗？你了解你周围的人吗？

八、离开思维

通过数据分析，你发现你处在一个不太有利的地位，那么，此时，你就要有离开思维去替你想办法，离开困境 。关键是学会自我调节，自我放松。 实际情况如：遇到难解的结，你怎么办？

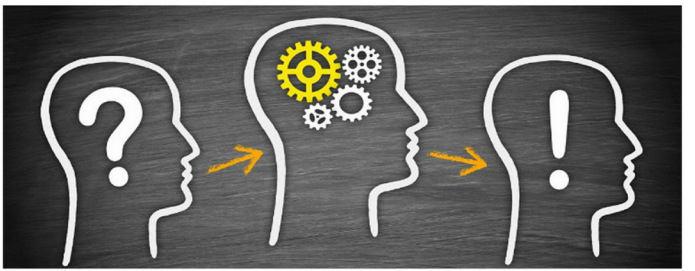
九、接近思维

怎样达成目标，实现销售增长，这时候你需要接近思维来帮助你 。 关键是多接触你要解决的问题，花时间分析，你要的是方案，不是问题。 实际情况如：你在做选择题还是问答题？ 责任点在哪？

十、理解层次

问题发现是第一步，要怎样分析问题，找到真正的原因，那么熟练的运用理解层次 。

关键是：你需要熟悉客观环境，员工的能力、行为的规律、他需要什么？实际情况如：你能够分析到哪一步？



/ 互联网数据分析能力的养成，需一份七周的提纲 /

文 / 人人都是产品经理　编辑 / 协会会员处　宣硕　日期 / 2016-11

不论对大数据分析或大数据运营，我都希望它是一篇足够好的参考。更准确地说，这是一份七周的互联网数据分析能力养成提纲。

比如网站分析，用户行为序列等。比如什么是产品埋点？在获得埋点数据后，怎么利用Python / Pandas的shift（）函数将其清洗为用户行为session，进而计算出用户在各页面的停留时间，后续如何转换成统计宽表，如何以此建立用户标签等。

下面是各周的学习概述：

第一周：Excel学习掌握

如果Excel玩的顺溜，你可以略过这一周。不过介于我入行时也不会vlookup，所以有必要讲下。重点是了解各种函数，包括但不限于sum，count，sumif，countif，find，if，left/right，时间转换等。

Excel函数不需要学全，重要的是学会搜索。即如何将遇到的问题在搜索引擎上描述清楚。我认为掌握vlookup和数据透视表足够，是最具性价比的两个技巧。

学会vlookup，SQL中的join，Python中的merge很容易理解。学会数据透视表，SQL中的group，Python中的pivot_table也是同理。这两个搞定，基本10万条以内的数据统计没啥难度，80%的办公室白领都能秒杀。

Excel是熟能生巧，多找练习题。还有需要养成好习惯，不要合并单元格，不要过于花哨。表格按照原始数据（sheet1）、加工数据（sheet2），图表（sheet3）的类型管理。

下面是为了以后更好的基础而附加的学习任务：

了解单元格格式，后期的数据类型包括各类timestamp，date，string，int，bigint，char，factor，float等；了解数组，以及怎么用（excel的数组挺难用），Python和R也会涉及到 list；了解函数和参数，当进阶为编程型的数据分析师时，会让你更快的掌握；了解中文编码，UTF8和ASCII，包括CSV的delimiter等。

这一周的内容我会拆分成两部分：函数篇和技巧篇。

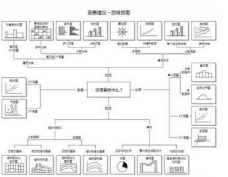
这是一道练习题，我给你1000个身份证号码，告诉我里面有多少男女，各省市人口的分布，这些人的年龄和星座。如果能完成上述过程，那么这一周就直接略过吧。（身份证号码规律可以网上搜索）

第二周：数据可视化

数据分析界有一句经典名言，字不如表，表不如图。数据可视化是数据分析的主要方向之一。除掉数据挖掘这类高级分析，不少数据分析就是监控数据观察数据。

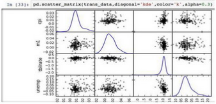
数据分析的最终都是要兜售自己的观点和结论的。兜售的最好方式就是做出观点清晰数据详实的PPT给老板看。如果没人认同分析结果，那么分析也不会被改进和优化，不落地的数据分析价值又在哪里？

首先要了解常用的图表：



Excel的图表可以100%完成上面

的图形要求，但这只是基础。后续的进阶可视化，势必要用到编程绘制。为什么？比如常见的多元分析，你能用Excel很轻松的完成？但是在IPython只需要一行代码。



其次掌握BI，下图是微软的BI。

BI（商业智能）和图表的区别在于BI擅长交互和报表，更擅长解释已经发生和正在发生的数据。将要发生的数据是数据挖掘的方向。

BI的好处在于很大程度解放数据分析师的工作，推动全部门的数据意识，另外降低其他部门的数据需求。

BI市面上的产品很多，基本都是建立仪表盘Dashboard，通过维度的联动和钻取，获得可视化的分析。

最后需要学习可视化和信息图的制作。

这是安身立命之本。这和数据本事没有多大关系，更看重审美、解读、PPT、信息化的能力。但值得花一点时间去学习。

数据可视化的学习就是三个过程：了解数据（图表）、整合数据（BI）、展示数据（信息化）。

第三周：分析思维的训练

这周轻松一下，学学理论知识。好的数据分析首先要有结构化的思维，也就是我们俗称的金字塔思维。思维导图是必备的工具。之后再了解SMART、5W2H、SWOT、4P理论、六顶思考帽等框架。这些框架都是大巧不工的经典。

分析也是有框架和方法论的，主要围绕三个要点展开：

1）一个业务没有指标，则不能增长和分析；

2）好的指标应该是比率或比例；

3）好的分析应该对比或关联。

举一个例子：我告诉你一家超市今天有1000人的客流量，你会怎么分析？

这1000人的数量，和附近其他超市比是多是少？（对比）

这1000人的数量比昨天多还是少？（对比）

1000人有多少产生了实际购买？（转化比例）

路过超市，超市外的人流是多少？（转化比例），这是一个快速搭建分析框架的方法。如果只看1000人，是看不出分析不出任何结果。

优秀的数据分析师会拷问别人的数据，而他本身的分析也是经得起拷问，这就是分析思维能力。需要确切明白的是，一周时间锻炼不出数据思维，只能做到了解。数据思维是不断练习的结果，我只是尽量缩短这个过程。

第四周：数据库学习

Excel对十万条以内的数据处理起来没有问题，但是互联网行业就是不缺数据。但凡产品有一点规模，数据都是百万起。这时候就需要学习数据库。

越来越多的产品和运营岗位，会在招聘条件中，将会SQL作为优先的加分项。

SQL是数据分析的核心技能之一，从Excel到SQL绝对是数据处理效率的一大进步。

学习围绕Select展开。增删改、约束、索引、数据库范式均可以跳过。

主要了解where，group by，order by，having，like，count，sum，min，max，distinct，if，join，left join，limit，and和or的逻辑，时间转换函数等。

如果想要跟进一步，可以学习row_number，substr，convert，contact等。另外不同数据平台的函数会有差异，例如：Presto和phpMyAdmin。

再有点追求，就去了解Explain优化，了解SQL的工作原理，了解数据类型，了解IO。以后就可以和技术研发们谈笑风生，毕竟将“这里有bug”的说话，换成“这块的数据死锁了”，大大的不同。

SQL的学习主要是多练，网上寻找相关的练习题，刷一遍就差不多了。

第五周：统计知识学习

这是数据分析的基础。比如产品的AB测试，如果产品经理并不清楚置信度的含义和概念，那么好的效果并不意味着真正的好。尤其是5%这种非显著的提高；比如运营一次活动，运营若不了解检验相关的概念，那么如何去判别活动在数据上是有效果还是没有效果？别说平均数。

再讨论一下经典的概率问题，如果一个人获流感，实验结果为阳性的概率为90%；如果没有获流感，实验结果为阳性的概率为9%。现在这个人检验结果为阳性，他有多少几率是得了流感？

如果你觉得几率有50%、60%、70%等等，那么都犯了直觉性的错误。它还和得病的基础概率有关。统计知识会教我们以另一个角度看待数据。如果大家了解过《统计数据会撒谎》，那么就知道很多数据分析的决策并不牢靠。

我们需要花一周的时间掌握描述性统计，包括均值、中位数、标准差、方差、概率、假设检验、显著性、总体和抽样等概念。不需要学习更高阶的统计知识，谁让我们是速成呢。只要做到不会被数据欺骗，不犯错误就好。

	A	B	C	D	E	F	G
1	注册日期	注册人数	投放金额				
2	2015/8/1	18021	15188				
3	2015/8/2	20723	37117				
4	2015/8/3	20116	43307		列1		
5	2015/8/4	32545	55040				
6	2015/8/5	26122	69457		平均	23713.4444	
7	2015/8/6	14076	46641		标准误差	2190.63816	
8	2015/8/7	11205	45621		中位数	20419.5	
9	2015/8/8	17767	40023		众数	#N/A	
10	2015/8/9	30759	95427		标准差	9294.0906	
11	2015/8/10	31752	80707		方差	86380120.1	
12	2015/8/11	39569	93210		峰度	-0.8656001	
13	2015/8/12	41661	10921		偏度	0.50925622	
14	2015/8/13	27151	60783		区域	30456	
15	2015/8/14	32233	49040		最小值	11205	
16	2015/8/15	19230	50322		最大值	41661	
17	2015/8/16	16593	47393		求和	426842	
18	2015/8/17	14178	41214		观测数	18	
19	2015/8/18	13141	51405		置信度(95	4621.8423?	
20							

以上图的Excel的分析工具库举例。

在初级的统计学习中，需要了解列1的各名词含义，而不是停留在平均数这个基础上。

第六周：业务知识（用户行为、产品、运营）

这一周需要了解业务。对于数据分析师来说，业务的了解比数据方法论更重要。业务学习没有捷径。

我举一个数据沙龙上的例子，一家O2O配送公司发现在重庆地区，外卖员的送货效率低于其他城市，导致用户的好评率降低。总部的数据分析师建立了各个指标去分析原因，都没有找出来问题。后来在访谈中发觉，因为重庆是山城，路面高低落差比较夸张，很多外卖人员的小电瓶上不了坡…所以导致送货效率慢。

这个案例中，我们只知道送货员的送货水平距离，即POI数据，根本不可能知道垂直距离的数据。这就是数据的局限，也是只会看数据的分析师和接地气分析师的最大差异。

对业务市场的了解是数据分析在工作经验上最大的优势之一。不同行业领域的业务知识都不一样，我就说介绍到这里了。在互联网行业，有几个宽泛的业务数据需要了解。

产品数据分析，以经典的AAARR框架学习，了解活跃留存的指标和概念。并且数据分析师需要知道如何用SQL计算。因为在实际的分析过程中，留存只是一个指标，通过userId 关联和拆分才是常见的分析策略。

网站数据分析，可以抽象吃一个哲学问题：用户从哪里来（SEO／SEM），用户到哪里去（访问路径），用户是谁（用户画像／用户行为路径）。

虽然网站已经不是互联网的主流，但现在有很多APP+Web的复合框架，朋友圈的传播活动肯定需要用到网页的指标去分析。

用户数据分析，这是数据化运营的一种应用。

在产品早期，可以通过埋点计算转化率，利用AB测试达到快速迭代的目的，在积累到用户量的后期，利用埋点去分析用户行为，并且以此建立用户分层用户画像等。

例如：用贝叶斯算法计算用户的性别概率，用K聚类算法划分用户的群体，用行为数据作为特征建立响应模型等。不过快速入门不需要掌握这些，只需要有一个大概的框架概念。

除了业务知识，业务层面的沟通也很重要。在业务线足够长的时候，我不止一次遇到产品和运营没有掌握所有的业务要点，尤其涉及跨部门的分析。良好的业务沟通能力是数据分析的基础能力。

第七周：Python/R学习

这时应该学习编程技巧。是否具备编程能力，是初级数据分析和高级数据分析的风水岭。数据挖掘，爬虫，可视化报表都需要用到编程能力。掌握一门优秀的编程语言，可以让数据分析师事半功倍，升职加薪。

以时下最热门的R语言和Python为学习支线，速成只要学习一条。

R的优点是统计学家编写的，缺点也是统计学家编写。如果是各类统计函数的调用，绘图，分析的前验性论证，R无疑有优势。但是大数据量的处理力有不逮，学习曲线比较陡峭。Python则是万能的胶水语言，适用性强，可以将各类分析的过程脚本化。Pandas，SKLearn等各包也已经追平R。

学习R，需要了解数据结构（matrix，array，data.frame，list等）、数据读取，图形绘制（ggplot2）、数据操作、统计函数(mean,median,sd,var,scale等)。高阶的统计暂时不用去涉及，这是后续的学习任务。

R语言的开发环境建议用RStudio。

学习Python有很多分支，我们专注数据分析这块。需要了解调用包、函数、数据类型(list,tuple,dict)，条件判断，迭代等。高阶的Numpy和Pandas在有精力的情况下涉及。

Python的开发环境建议Anaconda，可以规避掉环境变量、包安装等大部分新手问题。Mac自带Python2.7，但现在Python 3已经比几年前成熟，没有编码问题，就不要抱成守旧了。

对于没有技术基础的运营和产品，第七周最吃力，虽然SQL + Excel足够应付入门级数据分析，但是涉及到循环迭代、多元图表的分析部分，复杂度就呈几何上升。更遑论数据挖掘这种高阶玩法。

我也相信，未来了解数据挖掘的产品和运营会有极强的竞争力。

到这里，刚刚好是七周。如果还需要第八周，则是把上面的巩固和融会贯通，毕竟这只是目的性极强的速成，是开始，而不是数据分析的毕业典礼。

如果希望数据分析能力更近一步，或者成为优秀的数据分析师，每一周的内容都能继续学习至精通。实际上，业务知识、统计知识仅靠两周是非常不牢固的。

再往后的学习，会有许多分支。比如偏策划的数据产品经理，比如偏统计的机器学习，比如偏商业的市场分析师，比如偏工程的大数据工程师。

/ 盘点十大最热门的数据岗位，掌控你的未来 /

文 / 上海大数据联盟　编辑 / 协会会员处 李缘　日期 / 2016-11

随着很多大公司对数据分析需求增多，数据相关岗位的人才需求量也越来越大。数据学作为一门学科，已经受到时代的追捧。数据学，或者更准确来说，大数据，在2000年早期还是个冷门，而现在早已成为人们关注的焦点。早在2014年，高德纳咨询公司就预测，到2016年将有73%的公司企业将在大数据项目中投入重金。

2016年的尾声即将到来，我们是时候回顾一下大数据的发展，盘点十大最热门的数据岗位。

TOP1——首席数据官(CDO)

三军不可无帅也，所有想在大数据项目中取得成功的公司都需要首席数据官坐镇指挥。2014年CDO数量只有400人，2015年增长到了1000人，据此，加德纳预计，到2019年90%的英国大公司都会拥有自己的首席数据官。

首席数据官的工作内容非常多，职责也很复杂，他们负责公司的数据框架搭建、数据管理、数据安全保证、商务智能管理、数据洞察和高级分析。因此，首席数据师必须个人能力出众，同时还需要具备足够的领导力和远见，找准公司发展目标，协调应变管理过程。

TOP2——营销分析师/客户关系管理分析师

客户忠诚度项目、网络分析和物联网技术积攒了大量的用户数据，很多先进公司已经在使用相关策略来支持公司的发展计划。尤其是市场部门能够运用这些数据进行更有针对性的营销。营销分析师能够发挥他们在Excel和SQL等数据分析工具方面的专业特长，对客户进行细分，确保数字化营销能够到达目标客户群体。当与Adobe Campaigns等广告系列管理软件配合使用时，公司企业就可以确保其营销策略达到最佳效果。

TOP3——数据工程师

随着Hadoop和非结构化数据仓库的流行，所有分析功能的第一要务就是要得到正确的数据。商务智能和数据科学都要求有干净的、有序的且可用的数据框架，而这通常是通过SQL服务器、甲骨文(Oracle)和SAP公司数据库来实现的。高水平的工程师需要掌握数据管理技能，熟悉提取转换加载过程，很多公

司都急需这样的人才。事实上，很多首席数据官甚至认为，数据工程师才是大数据相关行业中最重要的职位。

TOP4——商务智能开发工程师

商务智能开发工程师的最基本职能，是管理结构数据从数据库分配至终端用户的过程。商务智能(BI)曾经只是商务金融的基础，现在已经独立出来，成为了单独的部门，很多商务智能团队正在搭建自服务指示板，这样运营经理就能快速且有效地获取高性能数据，评价公司运营情况。

商务智能最重要的技术目前都掌握在主要科技巨头手中，包括微软商务智能软件包，甲骨文，SAP和IBM。

TOP5——数据可视化

可能会奇怪，为什么把可视化摆在商务智能研发工程师前面。但是随着指示板和可视化工具的增多，商务智能“前端”研发工程师需要更熟练掌握Tableau、QlikView/QlikSense、SiSense和Looker。能够使用d3.js在网络浏览器中制作数据可视化的研发工程师也越来越受到公司欢迎。很多大公司开出的年薪已经超过了7万5千英镑，平均日薪500多英镑。

TOP6——软件研发工程师

这个也是大数据相关岗位吗？随着大数据的发展，很多公司都开始打造基于大数据平台的网页应用。除了掌握Javascript、C#、PHP和Diango Python框架等传统软件研发工具，大数据软件研发工程师还需要熟练使用Pyramid或者Flask。

TOP7——大数据工程师

正如上文提到过的，数据工程师的工作是负责管理公司的数据，包括数据的收集，存储、处理和分析。从经验来看，这涉及到使用关系型数据库，来管理以表格方式存储的数据。有很多关于数据怎样才能被定义为大数据的讨论。为了得到这个问题的结论，必须综合考虑结构化和非结构化数据(图像，视频，音频文件等)，它们往往是实时收集的，并且过于复杂，因此不能由传统数据结构处理。

大数据工程师需要能够搭建并维护大型异构数据框架，这些数据通常是在MongoDB等NoSQL数据库中。很多公司采用Hadoop框架和很多Hadoop次级软件包，如Hivel(数据软件)，Pig(数据流语言)和Spark(多编程模型)，当然数据基础设施还远远不止这些。

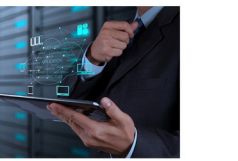
TOP8——洞察分析师

可能每个公司对这个职位的叫法不一样，但不可否认，现在具备执行力且精通技术的分析师炙手可热。通常，他们都会和产品部门、市场部门紧密合作，运用数据编程工具来整合大数据集，得出分析结论，支持发展客户群，制定维持客户关系策略。

从技术的角度来说，洞察分析师需要掌握各种数据编程工具，如SQL、SAS和SPSS等。但是很多公司都希望能够使用R和Python来获得更深度的分析，同时还要与RStudio等软件包配合使用，来生动地表达可视化数据分析结果。

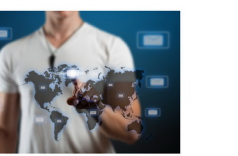
TOP9——数据架构师

在大数据环境中运行程序是一回事，而构建大数据基础设施则是另一回事。一个卓越的数据架构师可为尖端的大数据解决方案提供基础，其职责包括使用AWS，Azure和Google Cloud了解云中的数据存储和使用Hadoop或NoSQL设计基础架构数据库来管理非结构化数据。



TOP10——数据科学家

最近，Glassdoor表示，数据科学家是“美国的最佳工作”，是数据世界的常驻“摇滚明星”。关于谁才是真正的数据科学家，曾引起了世界范围内的讨论，参与这场讨论有许多强大学术背景的博士硕士，他们在统计学，数学，物理学，经济学，数据挖掘和机器学习方面都具备深厚专业知识。



优秀的数据科学家能够使用先进的分析原理和Python，R或Spark等数据编程工具来识别并解决高度复杂的业务问题。他们的分析将在决策中发挥核心作用，提供智力支持，以确保公司能够在日益复杂的商业环境中获得成功。

/ 管理大数据存储的十大技巧 /

文 / 灯塔大数据　编辑 / 协会会员处 李缘　日期 / 2016-12

在1990年，每一台应用服务器都倾向拥有直连式系统(DAS)。SAN的构建则是为了更大的规模和更高的效率提供共享的池存储。Hadoop已经逆转了这一趋势回归DAS。每一个Hadoop集群都拥有自身的——虽然是横向扩展型——直连式存储，这有助于Hadoop管理数据本地化，但也放弃了共享存储的规模和效率。如果你拥有多个实例或Hadoop发行版，那么你就将得到多个横向扩展的存储集群。

而我们所遇到的最大挑战是平衡数据本地化与规模效率，这是一个鱼与熊掌兼得的话题。数据本地化是为了确保大数据集存储在计算节点附近便于分析。对于Hadoop，这意味着管理数据节点，向MapReduce提供存储以便充分执行分析。它实用有效但也出现了大数据存储集群的独立操作问题。以下十项是Hadoop环境中管理大数据存储技巧。

1.分布式存储

传统化集中式存储存在已有一段时间。但大数据并非真的适合集中式存储架构。Hadoop设计用于将计算更接近数据节点，同时采用了HDFS文件系统的大规模横向扩展功能。

虽然，通常解决Hadoop管理自身数据低效性的方案是将Hadoop 数据存储在SAN上。但这也造成了它自身性能与规模的瓶颈。现在，如果你把所有的数据都通过集中式SAN处理器进行处理，与Hadoop的分布式和并行化特性相悖。你要么针对不同的数据节点管理多个SAN，要么将所有的数据节点都集中到一个SAN。但Hadoop是一个分布式应用，就应该运行在分布式存储上，这样存储就保留了与Hadoop本身同样的灵活性，不过它也要求拥抱一个软件定义存储方案，并在商用服务器上运行，这相比瓶颈化的Hadoop自然更为高效。

2.超融合VS分布式

注意，不要混淆超融合与分布式。某些超融合方案是分布式存储，但通常这个术语意味着你的应用和存储都保存在同一计算节点上。这是在试图解决数据本地化的问题，但它会造成太多资源争用。这个Hadoop应用和存储平台会会争用相同的内存和CPU。Hadoop运行在专有应用层，分布式存储运行在专有存储层这样会更好。之后，利用缓存和分层来解决数据本地化并补偿网络性能损失。

3.避免控制器瓶颈(Controller Choke Point)

实现目标的一个重要方面就是——避免通过单个点例如一个传统控制器来处理数据。反之，要确保存储平台并行化，性能可以得到显著提升。此外，这个方案提供了增量扩展性。为数据湖添加功能跟往里面扔x86服务器一样简单。一个分布式存储平台如有需要将自动添加功能并重新调整数据。

4.删重和压缩

掌握大数据的关键是删重和压缩技术。通常大数据集内会有70%到90%的数据简化。以PB容量计，能节约数万美元的磁盘成本。现代平台提供内联(对比后期处理)删重和压缩，大大降低了存储数据所需能力。

5.合并Hadoop发行版

很多大型企业拥有多个Hadoop发行版本。可能是开发者需要或是企业部门已经适应了不同版本。无论如何最终往往要对这些集群的维护与运营。一旦海量数据真正开始影响一家企业时，多个Hadoop发行版存储就会导致低效性。我们可以通过创建一个单一，可删重和压缩的数据湖获取数据效率。

6.虚拟化Hadoop

虚拟化已经席卷企业级市场。很多地区超过80%的物理服务器现在是虚拟化的。但也仍有很多企业因为性能和数据本地化问题对虚拟化Hadoop避而不谈。

7.创建弹性数据湖

创建数据湖并不容易，但大数据存储可能会有需求。我们有很多种方法来做这件事，但哪一种是正确的？这个正确的架构应该是一个动态，弹性的数据湖，可以以多种格式(架构化，非结构化，半结构化)存储所有资源的数据。更重要的是，它必须支持应用不在远程资源上而是在本地数据资源上执行。

不幸的是，传统架构和应用(也就是非分布式)并不尽如人意。随着数据集越来越大，将应用迁移到数据不可避免，而因为延迟太长也无法倒置。理想的数据湖基础架构会实现数据单一副本的存储，而且有应用在单一数据资源上执行，无需迁移数据或制作副本。

8.整合分析

分析并不是一个新功能，它已经在传统RDBMS环境中存在多年。不同的是基于开源应用的出现，以及数据库表单和社交媒体，非结构化数据资源(比如，维基百科)的整合能力。关键在于将多个数据类型和格式整合成一个标准的能力，有利于更轻松和一致地实现可视化与报告制作。合适的工具也对分析/商业智能项目的成功至关重要。

9. 大数据遇见大视频

大数据存储问题已经让人有些焦头烂额了，现在还出现了大视频现象。比如，企业为了安全以及操作和工业效率逐渐趋于使用视频监控，简化流量管理，支持法规遵从性和几个其它的使用案例。很短时间内这些资源将产生大量的内容，大量必须要处理的内容。如果没有专业的存储解决方案很可能会导致视频丢失和质量降低的问题。

10.没有绝对的赢家

Hadoop的确取得了一些进展。那么随着大数据存储遍地开花，它是否会成为赢家，力压其它方案，其实不然。比如，基于SAN的传统架构在短期内不可取代，因为它们拥有OLTP，100%可用性需求的内在优势。所以最理想的办法是将超融合平台与分布式文件系统和分析软件整合在一起。而成功的最主要因素则是存储的可扩展性因素。

ITPUB

/ Apache的Update / Delete功能设计实现 /

文 / 中国数据分析网　编辑 / 协会会员处　李缘　日期 / 2016-10

这里将介绍Apache CarbonData 0.3.0的Update/Delete功能设计实现。CarbonData是由华为开发、开源并支持Apache Hadoop的列式存储文件格式，支持索引、压缩以及解编码等，其目的是为了实现同一份数据达到多种需求，而且能够实现更快的交互查询。目前该项目正处于Apache孵化过程中。

当前，CarbonData暂不支持修改表中已经存在的数据。但是在现实情况下，我们可能很希望这个功能，比如修改维度表，事实表的数据校正以及数据清洗等。很多使用CarbonData的用户很希望其能够提供数据的修改和删除功能。为此，社区已经有人提了Issue(CARBONDATA-440)，其目标就是为CarbonData提供Update/Delete功能，这个功能应该会在CarbonData 0.3.0版本发布。本文将介绍CarbonData的Update/Delete功能设计实现。下面是实现这个功能的高层次设计目标：

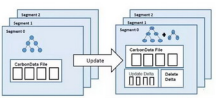
(1)提供标准的SQL接口，以便能够执行更新和删除操作；

(2)对CarbonData表执行更新和删除操作的时候，不需要对已经存在的整个CarbonData块重写，而是将修改写到差异文件中(differential files)；

(3)在更新和删除操作之后，CarbonData readers应该能够跳过删除的记录，并且能够无缝地读取更新的记录，而这些操作不需要用户更新自己的应用程序。下面将详细地介绍CarbonData 的修改和删除实现设计。

更新操作实现

我们都知道，CarbonData的数据是存储在HDFS之上，而HDFS中的文件是不可修改的(immutable)，所以CarbonData的数据块并不能原地进行修改。更新数据的一种方法就是删除和重写整个数据块。然而这种方法的效率很低，会导致性能瓶颈。其实我们可以把更新操作认为是先"删除"，然后"插入"，这也就是CarbonData中更新的实现。下面将详细地介绍CarbonData的更新操作实现：



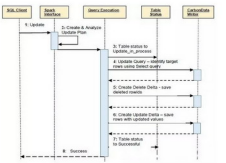
CarbonData的更新操作分为两步——1、第一步包括两个部分：

(1)首先，CarbonData能够通过执行过滤和Join操作识别出需要更新的行。为了能够唯一标识行数据，CarbonData会使用到ROWID属性。一旦需要更新的数据被标识后，这些数据将会在单独的文件中被标识为deleted，而且这些文件是存放在当前表的目录下，这些文件被称为"Delete Delt"。

(2)然后，CarbonData将会从源表中收集需要更新的列值并组成新的一行。新的行数据是由更新后的列值和目标表现有的列值数据组成的。这些更新的行数据将会在Spark处理层组成一个源RDD。

2、第二步：CarbonData将会使用现有的数据加载方法将源RDD中的行数据转换成CarbobData数据格式。这个操作类似于数据的增量加载。这个新创建的CarbonData文件称为"Update Delta"。Update Delta文件将存储在同一个segment中，而且Update Delta本身拥有btree和块级别的统计，正如正常的CabonData文件。这个新的btree应该追加到全局的btree中，并且缓存起来。

下面就是CabonData更新操作的时序图：



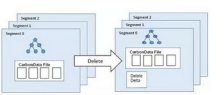
删除操作的实现

在删除数据的情况，CarbonData也是通过过滤和Join操作来识别需要删除的行。为了能够唯一标识行数据，CarbonData会使用到ROWID属性。

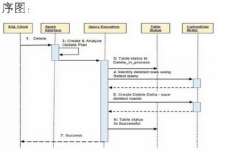
一旦需要删除的数据被标识后，这些数据将会在单独的文件中被标识为deleted，这个文件也称为"Delete Delta"文件。CarbonData记录扫描程序将会把这些删除的文件排除到结果集之外。在删除操作之后，CarbonData不需要更新全局字典表，因为字典表中有些entries对其他的segment还是有效的。

删除操作的原子性

CarbonData的删除操作具有原子性，也就是说，删除的数据要么全部被删除，要么全部都没删除。删除操作产生的Delete delta文件在删除操作仍然进行时，对readers事不可见的;只有删除操作成功进行，新删除的行数据才会对readers可见。删除的操作如下图所示：



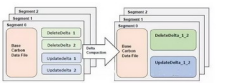
下面就是CabonData删除操作的时序图：



文件合并

对每次更新操作，都会产生update delta 和 delete delta文件，随着频繁地更新和删除操作，会产生越来越多的delta文件。这将会产生许多小文件，这可能会影响scan操作的性能，所以我们需要将这些delta文件合并成单独的delta文件。将许多个delta文件合并成一个delta文件的操作称为compaction或 minor compaction.

操作如下：



而且compaction操作可以通过配置达到多少个delta files来触发。在删除或者更新操作之后，如果delta文件的数量达到了配置的阈值，compaction操作将会触发。

大数据发现："双11"背后的消费新趋势

文 / 网易财经　编辑 / 协会会员处 李缘　日期 / 2016-10

进入第8个年头的"双11"，已经成为洞察全民消费动向的一个窗口。据星图数据发布的《双11网购大数据分析报告》显示，"双11"当天全网销售额高达1770.4亿元，东道主天猫占据"双11"份额将近70%，已经进入交易总额为1207亿元的"破千亿"时代。

梳理"双11"的消费数据，不难发现，消费升级下，一些消费新特点凸显：消费升级带动农村市场成为新的经济增长点，农村居民的消费行为正从"穿"向"吃穿住行"拓展；趋于理性的消费者更愿意选择常用的或者经典的产品，美妆服饰等向品牌化集中；不仅国内消费者热衷海外购，不断进行产品更新的国货品牌也受到海外"剁手党"追捧；零售业经历了互联网化的变革，线上线下的融合已经进入一个全新的阶段，全渠道化的新零售体系已经初现端倪。

趋势一：畅销商品品牌集中度提升

从细分领域来看，个性化和小众化的品类份额在扩大，但在"双11"这样的电商造节时期，消费者更愿意选择常用的产品，经典品类销售份额优势明显，产品向高端化和品牌化集中，这一特点在美妆以及服饰领域尤其突出。

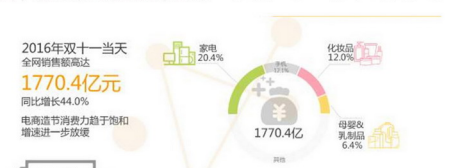
据天猫公布的"双11"数据显示，天猫美妆双11销售额2分钟内迅速突破亿元，百雀羚继续蝉联美妆销量榜榜首，和欧莱雅、自然堂、一叶子以及SK-II组成美妆品牌top5，在"双11"最先破亿的是国外高端热门品牌，预售产品尾款开始支付后，众多高端美妆品牌瞬间超过去年"双11"全天总成交额，其中SK-II仅用5分钟，海蓝之谜、资生堂、芭比波朗仅用10分钟。首次参战双11的多个高端美妆品牌的表现强劲，其中修丽可仅用15分钟，SISLEY用30分钟成交额已突破品牌入驻天猫时全月+"超级品牌日"全天销售额。

据星图数据显示，全网化妆品销售额中，外资品牌销售额占比超本土份额比去年同期增长29.9%，其中，消费者下单最多的是面膜、洗护套装、底妆等女性必选产品。

作为"剁手"男女的最爱，服饰家居类在京东商城成为"双11"下单量最大的品类，下单量占比超过40%。Lee销售额为去年同期43倍，TUMI销售额达去年同期20倍，GUESS销售额是日常销售的70倍，家居家装品类销售额比去年同期增长近100%，当日销售约823万件家居家装用品。消费者对体育品牌持续热衷，运动鞋服热销110万件，著名慢跑鞋品牌圣康尼(SAUCONY)销量是去年同期的6倍，功能性跑鞋品牌亚瑟士(ASICS)销量是去年同期的3倍。

据天猫方面介绍，天猫服饰开场仅5分钟交易额就破20亿，36分钟突破100亿交易额，排名第一的优衣库2分53秒破亿，再次创造纪录，11日下午，优衣库官方旗舰店的双11活动商品已经全部售罄。

"适合自己并且刚好在打折，这样的东西才是我在促销季必买的清单。"陈小姐是"双11"网购大军中的一员，她表示大品牌的经典款相对于一些爆款品牌来说更实用，而且大品牌是实打实地给出五折折扣，不会先涨价再打折。而业内人士分析，今年"双11"大牌之所以被"疯抢"，一方面是由于多家高端大牌在"双11"前入驻天猫，另一方



面还和高端品牌放下“架子”，在预热营销中迎合中国消费者口味有关，比如雅诗兰黛、海蓝之谜、兰蔻、SK-II等高端美妆品牌纷纷推出天猫首发独家定制礼盒，获得中国消费者喜爱。

趋势二：消费升级带动海淘双向流动

从买全球到卖全球，“双11”打通了消费市场的“网上丝绸之路”，消费升级之下海淘出现了新的可能。一方面，出现在天猫跨境购物单前10名的产品，不再只是奶粉、纸尿裤和保健品等简单的几种品类，洗护套装、腕表、洁面仪等多种品类也榜上有名，品牌丰富度有所发展。

天猫国际总经理刘鹏认为，从中深切感受到了中国消费者的消费升级，消费升级不是买贵的，而是消费理念的改变，中国的年轻消费者越来越喜欢尝试新的品类和新的品牌，他们愿意购买更多的有特色的海外商品。

“相信双11之后，到明年，中国年轻消费者的消费观将成为全球重要的流行潮流，而这股潮流将会改变全球的产业结构或者贸易结构，影响跨国公司、跨国品牌的全球定价的策略。”据天猫方面称，天猫国际“双11”只用9个半小时就超过了去年全天的销售额，第二次参加双11的梅西百货，仅用5分钟成交额就突破了去年双11全天成交额；美国第二大零售商Target首战双11就得到骄人成绩，不仅拿下了全球VR购物第一单，其母婴类等多个商品成了天猫国际上领跑的爆款。

另一方面，在买全球之外，卖全球把中国大陆市场和海外市场连在一起。据阿里巴巴集团旗下跨境出口平台全球速卖通数据显示，“双11”当天速卖通平台共产生3578万笔订单，较2015年同日增长68%，创下历史最高纪录。全天交易共覆盖230个国家和地区，共有621万国际“剁手党”参与，无线订单成交占比58%。

据了解，速卖通平台这次增设了香港和台湾的专场，香港和台湾的消费者可以通过专场，用当地的货币进行支付，比如香港的消费者可以使用八达通来支付购买“双11”商品。“卖全球为海外消费者提供了丰富的商品选择。可以看到，在未来，中国一定是世界上最大的消费市场，借助双11，我们有机会把中国大陆市场和海外市场进行联动，开展全球化的销售。”刘鹏表示。

值得注意的是，国产3C产品备受海外消费者欢迎，据了解，小米的电子产品在“双11”单日销量是日常平均水平35倍，而仅在俄罗斯，国产VR眼镜和无人机的订单数就已经接近2万单，有观点认为，供给侧改革和产品创新令国产品牌在全球竞争中的优势有所扩大。

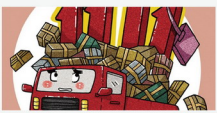
趋势三：品牌打通全渠道运营

“双11”期间，百万门店打通线上线下，全渠道的商品通、服务通、会员通已经成为品牌标配，线下门店变身智慧门店，一方面通过IP、直播等玩法进一步获取新用户，另一方面，则是在产品创新、跨渠道场景体验等方面适应市场环境与消费行为变化，提升全渠道的消费转化，新零售体系已初现端倪。

今年不少品牌加入线下实体店提货、补货的环节，优衣库“双11”半天官方旗舰店销售告罄，迅速引流到线下门店的做法惹人关注，消费者在GAP等品牌的官方旗舰店购买时，品牌可以通过系统对接，实现最近门店配送，最快两小时可以送货上门。冠名天猫“双11”晚会上海家化借助晚会收视率的流量红利，首次派出两位高管以直播的形式实时与天猫网友进行发红包、砸金蛋等形式互动，为上海家化旗舰店和佰草集旗舰店站台，在新零售场景营销下，84家佰草集门店通过深度布置“万庆同庆”，打通“线上—线下—线上”的闭环，实现“进店推送权益—皮肤测试—产品推荐—线上购物”的全场景流程。

此外，上海家化还在机场开设了全品牌体验店E-store、在购物中心开设大型跨品类生活馆，地铁枢纽的化妆品自动贩售机等也将产品融入在生活场景中，实现线下到线上两种体验模式的无缝对接。据统计，截至11日24时，上海家化“双11”全网各渠道总零售额已突破2亿元。

“今年双11，消费者无论在网上、街边还是购物广场中，都能感受到一个立体化的双11，以及线上线下无界限的体验。”天猫商家事业部总经理张阔表示，在全渠道三通大战略中，线上跟线下通过数字化打通的门店已有十万家，后台云端可以通过数字化产品通过线下POS和库存完全打通，可以看线上和线下库存，消费者下单后，可以快速由导购员接单，再把商品发给消费者。



原来“双11”只是线上的节日，现在已经变成全部零售业的节日。“双11”期间，广州各大百货火速参战，天河城百货把促销日期提前了一周，大部分商品跨专柜消费满300元送300元电子促销卡，部分商品5-6折扣，广百利用场地优势打造“情述百老汇”音乐剧主题展，广州友谊则打通线上线下支付渠道，大部分商品6折，持实体VIP卡的顾客凡在广州友谊官方微信新绑定成为e会员，可获赠会员积分2000分。

在“双11”这场消费大戏中，谁也不想缺席，谁也不想被消费者冷落。对于新零售的构想，正如很泰奇业CEO陈晓东所说，当扩展了全渠道之后，有一个词可能会消失，就是所谓的“线下”，实体店还会存在，但是不会以“线下店”的身份存在，将完全与互联网融合，未来，消费者将分不清是处于互联网中还是实体店中，他们可以在很方便的情况下取得想要的商品和服务。

趋势四：农村市场成消费新增长点

今年，有两个数据首次出现在天猫双11全球购物狂欢节的晚会屏幕上，分别是村淘点数量和土特产榜单，从密密麻麻的据点可以看出，农村地区的网购热情已经开始被点燃。据了解，天猫双11全国近2万个农村淘宝服务站都加入到了全球购物狂欢之中，京东方面，覆盖

了44万个行政村的1700多家京东帮服务店在“双11”大促中实现了京东大家电30%的销售额，烟机灶具和热水器等品类的销量也随着农民的生活条件改善而大幅增加，分别比去年双11期间上涨了2倍多和1.8倍以上。

有关农村电商市场发展条件的调查报告显示，我国农村市场消费潜力巨大，2016年网购市场有望突破4600亿元，未来消费规模可能超过城市。目前，农村居民网购接受率高。调查显示，农村居民网购接受率高达84.41%，人均年网购消费金额预测在500至2000元，主要集中于日用品、服装及家电等领域。

但巨大的购买力目前却缺乏渠道释放。里贝恩咨询全球合伙人康雁认为，在传统消费品行业中，最大的一个增长瓶颈就是渠道，多级的分销渠道导致成本飙升，品牌商向下拓展渠道难度很大。对于在3、4、5线城市的消费者而言，想在线下买到一二线城市流行的品牌也很困难。

“电商让品牌商用更低的成本更迅速地渗透到3、4、5线城市。在过去几年的双11中，我们发现欠发达地区甚至包括农村，购买力都是相当强大的，很多快消品牌已经在渠道下沉中尝到甜头。”康雁表示，“网货下行、农产品上行”，像农村淘宝这样的电商平台贡献了品牌与农村之间渠道结构的突破。

阿里巴巴集团资深副总裁，B2B事业群总裁戴珊笼表示，农村淘宝已经升级到3.0阶段，不仅要把货物、第三方的物流、电商消费等各类服务带到农村，还要把线下终端聚集起来，同时要把金融覆盖到农村，通过生态圈的大数据，让农村的消费者和小微企业依托阿里的互联网加供应链的生态中获益。

HX

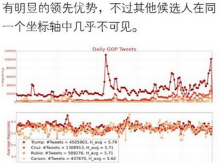
/ 一直在讨论的大数据，这次在美国大选中究竟干了什么 /

文 / 新浪创事记 南七道　编辑 / 协会会员处 李缘　日期 / 2016-11

特朗普入主白宫，美国大选落下帷幕。回首这场混战，大数据技术在其中其实发挥了不小的作用，甚至占据了至关重要的地位。

社交媒体上的大数据分析

从美国大选之初，各候选人各方面的数据统计就已经被统计在各家网站上。比如从这张数据来看，上面折线图表示的是各个候选人的推文提及率。而在同样的时间序列中，特朗普在推文中的提及率占有明显的领先优势，不过其他候选人在同一个坐标轴中几乎不可见。



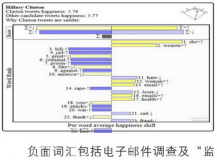
图中下方的折线图则是关于每位共和党候选人的推文的幸福感指数进行的比较，特朗普相对于克鲁兹和卢比奥有微弱的优势，对于卡森有明显的优势。并且，特朗普的平均幸福指数比希拉里略高（5.79:5.70），但仍比桑德斯低（5.79:5.85）。虽说这些数据并不能够直接决定最后的大选结果，但也间接的为特朗普获胜起到了润滑和推动作用。

除了这些幸福指数和提及率，大数据统计还对各个候选人的各个“标签”进行了统计，而正是这些标签决定了幸福指数等相关数据的分值高低。在下图特朗普和希拉里的“标签”对比中，词语的颜色根据感情状态表示——越蓝越快乐，越紫越悲伤，而且词语的大小由加权平均tf-idf值决定。

从这两张图中不难看出，希拉里的“标签”中，正面词汇与负面词汇相交织，其中比较重要的正面词汇有“经验”，“才能”，“女性”，“世界”；负面词汇有“犯罪”，“调查”和“谎言”，这也许是与电子邮件服务器丑闻相关。而特朗普的词图中，最大的词汇包括支持者形容的“前行”，以及现在的共和党初选中的“胜利”；负面词汇，或许来自于他的反对者，包括“羞辱”，“攻击”，“种族主义”，“骗子”和“危险”。

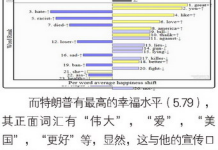
除此之外，还有可供我们参考的是词汇转移图，如图所示：

希拉里的词汇转移图与参照分布比较相似（5.76比5.77）。



负面词汇包括电子邮件调查及“监狱”，“犯罪”，“囚犯”，“丑闻”等。此外，“票据”是以负面形式呈现的词汇（被理解为支付票据），但是在希拉里这里则是指比尔·希拉里。

正面词汇主要有“她”，“女性”，“感谢”，“健康”，而负面词汇“憎恨”，“悲伤”，“失败者”，“诈骗”，“种族主义”较少被提及。



而特朗普有最高的幸福水平（5.79），其正面词汇有“伟大”，“爱”，“美国”，“更好”等，显然，这与他的宣传口号相关——使美国再次伟大。但是他的负面词汇包括了更多。比如“憎恨”，“种族主义”，“死亡”，“失败者”，“悲伤”，“禁止”，以及以轶读的方式，反映了他的反对者的观点。

通过分析大众趋势，民众可以通过数据了解到谁更符合国民的标准。而社交媒体运用大数据的统计和判断使得民众判断方向发生了一定的偏差，引导了整个舆论的导向，甚至改变了很多人的原始初衷，心中的那杆秤在不知不觉中发生了一定的偏差。

个人数据团队的关键性作用

当然，这个只是社交媒体和一些数据公司较为公正的数据统计，而对于特朗普和希拉里本人而言，个人背后的数据团队比较看来，简直就是一场大数据的盛宴。

传言希拉里有一支堪比硅谷公司的大数据团队——50名专业的程序员和开发者，大部分都是曾经供职于Facebook、Google、Twitter等大型的科技公司的高层人士。在他们的帮助下，如果想要用更多技术手段来帮助希拉里赢取更多选票和资金，简直是轻而易举的事情。比如，民主党对于竞选页面进行细微的调整，就可以让捐赠人储存信用卡信息。这种手段常用于电商公司将窗口用户变为付费用户的手段上，但现在在政治上同样适用，很多民众在不知不觉中便已经成为了希拉里“忠实”的支持者。

除此之外，这些大数据团队还能够处理一些突发的技术问题，比如，2015年，联邦竞选委员会报告竞选筹资截止日当天，外部邮箱系统突然崩溃。虽然当时场面一片混乱，但是希拉里的竞选团队竟在4个多小时内搭建了一个临时邮箱系统Balloon，使得危及顺利解决。

这样看来，特朗普似乎只有Twitter这样一个武器，实则不然，特朗普背后的数据团队对于他商人出身的身份进行了量身打造，为这样一位本该有着金钱光环的人打造了一副政客的钢筋铁甲，使得他的辩论能力突飞猛进。特朗普背后的大数据团队着重于希拉里过去的演讲，通过关键词和数据分析来洞悉希拉里演讲的漏洞和缺点，从而为特朗普提供有力的攻击武器。

还记得普经的辩论会吗？希拉里发言26分钟内便被特朗普打断了25次，这难道不是大数据团队支持的结果？

国内数据公司DataEye CEO汪祥斌认为，其实从上两届奥巴马的总统大选开始，大数据在整个总统大选过程中的应用已经越来越深入，从大选筹资阶段开始，精准的筹资邮件筛选到选情实时分析，选民人群精准定位，结果预测各个环节都已经开始数据化，整个总统大选已经变成一个典型的数据驱动的业务决策过程。不难看出美国的政治已经全面进入了大数据时代。

大数据公司是大选的幕后英雄

如果说到专门为民主党和共和党提供数据分析和服务的要数TargetSmart和DeepRootAnalytics这两大公司了。前者专门为民主党派和州民主党派以及他们的同盟提供大数据分析和服务；而后者则给共和党及其从属团队提供数据分析。

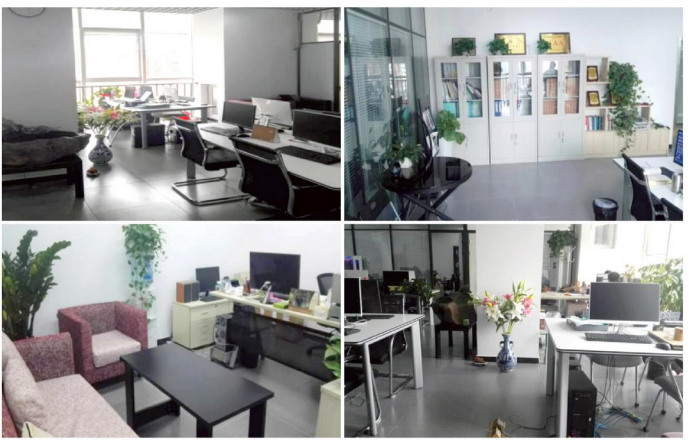
TargetSmart和DeepRoot都是利用Alteryx的软件来说明他们容纳、净化、混合以及分析来源不同的大规模资料。这种方法主要来分析选民的年龄结构，根据不同年龄段来分段并且打分，然后利用这些资讯来优化他们在媒体上的花销，特别是在非常重要的电视广告上，从而扩大宣传效应，使得事情的效果事半功倍。

资料将会指示客户该将他们的竞选广告放到哪，从而使得广告在目标人群的曝光率大大提升，同时还会提示他们花销的纪录，透过让客户在情景中能够意识到这个问题，不仅提供他们所做的与目标人群相关的理由，而且也会分析竞争对手或同盟所做的，对目标人群的影响，这就允许他们能够对正在进行的分配任务具有策略性，并对广告投放更聪明——把广告投放在最不显眼而又最高效的地方，同时根据其他人或组织的移动来及时做出反应。

这一届大选特朗普和希拉里总计为大选烧掉了将近1亿美金，这些钱烧在哪些地方？无外乎宣传公关，而这个其中的宣传打广告以及拉选票就占据了多数，包括电视广告，网络广告，直邮信件等等。在汪祥斌看来，如此高昂的广告投入也为大数据的应用提供了广阔的施展空间。不仅仅是筹资的多少，谁能更有效的进行精准的广告投放某种意义上来说也直接决定宣传战的结果。

/ 北京汇智方圆数据分析师事务所 /

编辑 / 协会会员处 李缘　日期 / 2016-10



北京汇智方圆数据分析师事务所是经北京市工商局批准注册，经中国商业联合会数据分析专业委员会（简称中数委）严格考察后批准入会的专业数据分析机构（中数委团证056号），公司于2011年正式成立。

事务所成立以来依托推进首都经济发展的良好时机以及京津冀一体化的大环境经济政策为全国百余家企业提供了专业、细致、客观、全面的数据分析服务，涵盖了数据采集、数据处理、经营数据分析、投资价值和收益分析等各个方面，得到了新老顾客的一致好评。

事务所还于2016年加入了北京企业评价协会成为会员单位，致力于诚信经营，信誉为本。在目前大数据发展背景下，我公司的数据存储技术、分析技术、处理技术等在不断的升级完善，与时俱进，不断创新，引进先进的经营理念，才能立于不败之地。大数据时代为企业带来了很好的发展契机，中小企业发展的空间最大，而帮助他们运用好大数据是我们首要责任，庞大的数据信息本身不能产生价值，只有对数据进行科学有效的分析、深入的整理才能彰显它的价值，从而为企业带来效益，这就是我们事务所未来的发展目标。

展望未来，我们要在中数委的监督指导下大力推进企业的数据利用效率，全力做到“五个发展支撑”，即决策与资源支撑、理论与技术支撑、团队与人才支撑、体制与机制支撑、配套性措施支撑。我们将竭诚与各界朋友倾力合作，为企业和大数据时代搭建专业高速的桥梁，为企业的数据提供科学的量化、引导和分析，挖掘潜在的经济价值。

我们拥有专业的数据分析师，我们拥有强大的技术团队，我们拥有蓬勃发展的公司，我们拥有优越的地理位置，我们期待着与您真挚的合作！

办公地址：北京市朝阳区常惠路北辰福第V中心E座1006
联 系 人：宋玉娅
联系电话：15300081507
010-53342661

大数据时代，你是企业需要的核心人才吗

CPDA® | CDA
数据分析学习培训

400-050-6600

欢迎访问数据分析师学习网 http://www.chinacpda.com/

数据与决策
DATA & DECISION

上海学员 蔡博生 哈森商贸数据中心 执行总监

“面对一堆数据，使用一个方法，找出一个说法，然后这方法与说法是可以被别人接受的”。不论你的专业与背景如何，如果你看懂这句话，你就具备成为CPDA数据分析师的能力。

河南学员 王永川 安钢自动化 高级项目总监

本着不断学习的心态报名了CPDA，没想到不仅学习了数据分析的技能，还认识了许多志同道合的学友。课堂上学习的许多模型，在工作中都进行了实战应用，真的是事半功倍！

四川学员 熊杰 中国电信四川分公司 项目总监

当下商业环境对数据方面的需求越来越多，数据分析也不再限于基础的描述性统计。如何能通过数据发现市场、规避风险、创造更高的利润成为企业的视线焦点，CPDA数据分析的课程能够帮助学员以数据思维了解商业环境!

北京学员 周哲 中国移动 高级大客户经理

通过CPDA的学习，充分了解数据分析同传统通信行业的融合，我感到传统通信行业对于大数据、数据分析知识的欠缺，因此，通过学习有助于提高传统通信行业对于大数据的应用，通过言传身教提高传统通信行业数据分析的能力。

我们的学员这样说

全国咨询热线：**400-050-6600**

办公地址：北京市朝阳区朝外SOHO-C座9层

办公电话：010-59000991 / 010-59000559

培训网址： www.chinacpda.con

协会网址： www.chinacpda.org