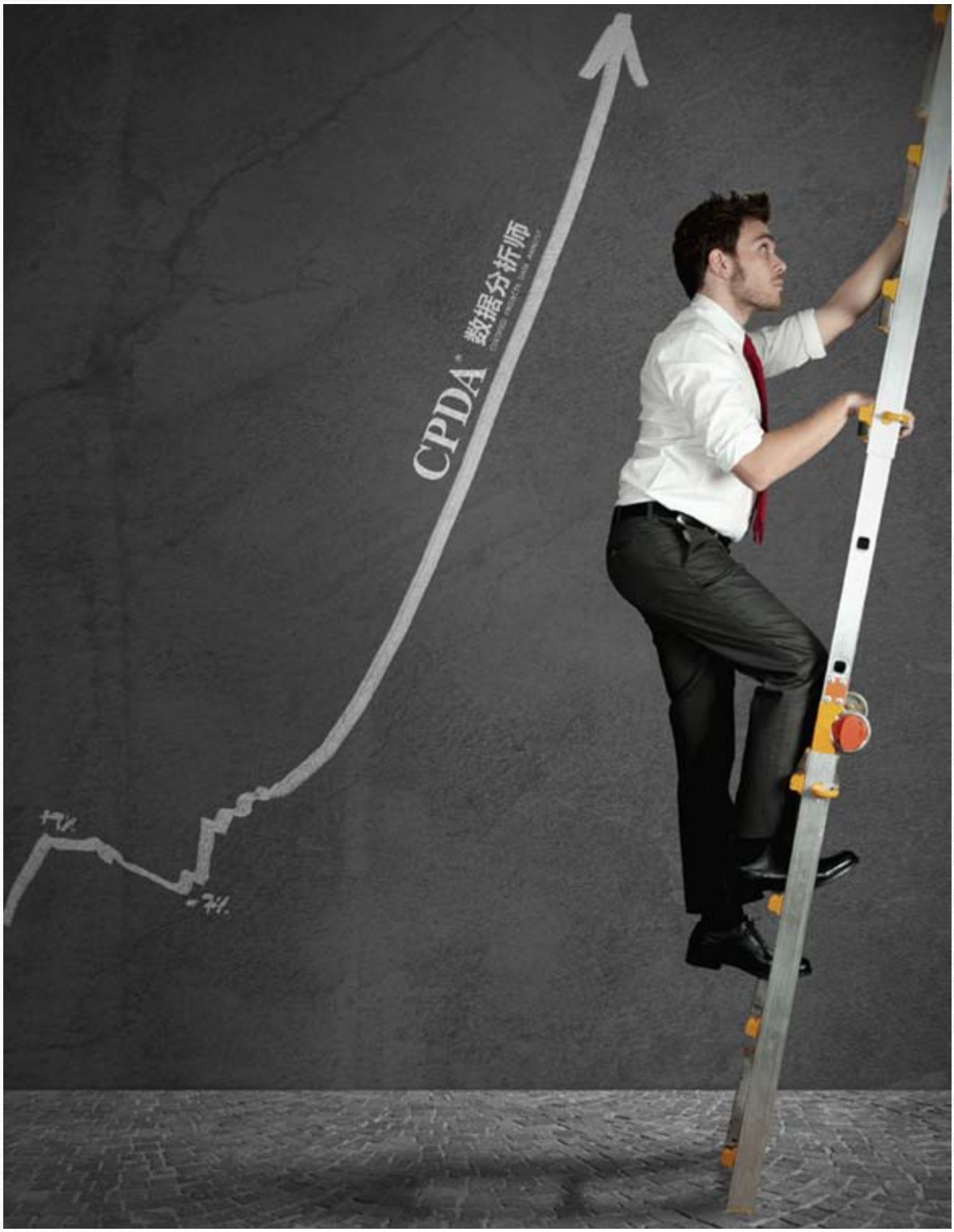
中 国 数 据 分 析 行 业 核 心 刊 物

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
2017年第01期 总第29期（季刊）
咨询热线：010-59000991 / 59000339
http://www.chinacpda.org/
投稿请至 xiehui@chinacpda.org



China Data Analysis

/ 大数据应用的持久战 /

/ 大数据应用的持久战 /

明代大儒王阳明主张“心外更无物”，他与友人有过一段精彩的对话：

先生游南镇，一友指岩中花树问曰：“天下无心外之物，如此花树，在深山中自开自落，于我心亦何相关？”先生曰：“你未看此花时，此花与汝心同归于寂；你来看此花时，此花颜色一时明白起来，便知此花不在你的心外。”

——《传习录·黄省曾录》

“存在就是被感知”， 这一命题在哲学界争议很大，但是今天不是要讨论哲学，而是想探讨一个类似的大数据问题：当前，很多企业在生产经营过程中产生和积累了大量的数据。从发展趋势看，不管企业数据处理能力提升到多么强，总是跟不上数据总量增加速度。必然导致的结果就是：很多数据将变成为无人到达的山中静静开放的花朵，美丽却无人欣赏。因此，对于企业来说，那些不被感知的大数据，是不是也如山中花，可以当它不存在呢?

原因只有一个：不被感知的数据，与可感知的数据一样在消耗企业大量资源，存在无法忽视的成本。这一成本主要来自三个方面：资金成本、闲置成本，以及注意力成本。

另一个常常被忽视的成本是企业注意力的无谓消耗。当大数据汹涌而来，企业人员的注意力并没有作好准备，大数据会大量消耗企业稀缺的注意力资源。与外部世界的剧烈变化相比，人的进化无疑是停滞不前的。

今天很多人都同意，数据是企业的一种战略性资产。这句话表面上是说数据可以支撑战略，对于企业战略是重要的。但未能明白表达的意思是：战略性资产很多未必是用于当下，而是为了应对未来的不确定性。更为重要的，战略性资产的配备是有成本的。如果资产的战略性价值长期不能覆盖其成本，那就可能转化为战略性负债。

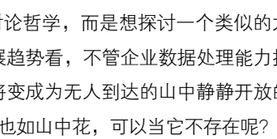
当我们接收的数据和信息越多，面临的选择就越多，如若不善于过滤、挖掘和处理，对各种决策就可能会造成负面影响，当然也会放大我们对未来不确定性的恐惧。小到个人命运大到国家前途，都是在这样一片混沌中煎熬着。

大数据本身不产生价值，大数据的根本用途是利用数据分析对我们的决策提供规律、知识和经验等科学依据，客观上减少面对未来决策的不确定性。以业务决策支持为分析目标，大数据不靠大，小数据也一样有大价值!

大数据的价值需要我们自己去定义！因为对于未来、对于未知领域，我们每个人或组织面临的不确定性问题是不一样的，有的偏个体（如疾病诊断，犯罪预测），有的偏大众（如广告营销、客户细分），有的偏微观（如基因序列，个性化教育），有的偏宏观（环境监测、天文数据处理），有的关注资源优化配置（如供需匹配，出行服务），有的关注宏观决策（如政府资产分析、综合管控）等。可以说大数据分析需求无处不在，而又大不相同。这就需要从自身实际需求和数据、技术现状出发，自行设定大数据分析的价值和应用目标，生搬硬套互联网公司那套做法，不可取!

所以，我们一定要看清我们属于什么级别的玩家？我们是否了解大数据风险与数据偏见？我们是否理解并能贯彻大数据思维？我们是搞技术驱动、业务驱动还是数据驱动?

来，我们一起准备好，打一场大数据应用的持久战!



中国数据分析
CHINA DATA ANALYSIS 用数据说话 · 做理性决策

中国数据分析 会员特刊 2017年第一期

本期目录 CONTENTS

主办
中国商业联合会数据分析专业委员会
编委
袁硕 、李缘
出版时间
2017年第一期 4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: 010-59000991 / 010-59000339
传真: 010-59000991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

/ 回望2016年，互联网公司的悲和喜 /

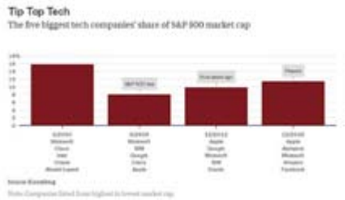
文 / 36大数据 编辑 / 协会会员处 袁硕 日期 / 2016-12

随着科技与我们的生活关系越来越密切，这个行业发展的一点一滴开始影响我们的生活。2016世界科技发展情况是怎样的呢?科技巨头垂食世界、广告也逐渐变成巨头之争、中国科技在世界大展拳脚这些都是正向的发展。然而这一年，还有苹果公司开启了10年来的第一次负增长，初创公司估值紧缩等等负面的消息。

一、科技巨头正在蚕食世界

今年世界上最具价值上市公司前五名第一次全部被科技公司占领。这五家公司分别是苹果、Alphabet(谷歌母公司)、微软、亚马逊和Facebook。截止至12月27日，这五家公司总值为2.4万亿美元(约合人民币16.6万亿)，S&P500市场价值占比超过11%。这是科技龙头企业在2000年3月互联网泡沫顶峰之后，又一次向占据S&P500 16%的份额靠近。坏消息是：大型科技公司日益增长的力量使得他们有了参与全球政治的野心。

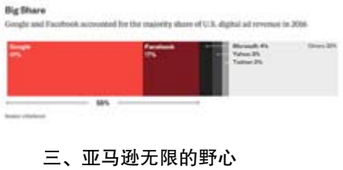
全球前5大科技公司S&P500市场价值占比：（S&P500—标准普尔500指数，是记录美国500家上市公司的一个股票指数，这个股票指数由标准普尔公司创建并维护。其所覆盖的所有公司都是在美国主要交易所的上市公司。与道琼斯指数相比，S&P500包含的公司更多，因此风险更为分散，能够反映更广泛的市值变化。）



二、广告变成了两强之争

Alphabet旗下的谷歌和Facebook是数十亿广告投资的热门目标。他们的技术便于汽车制造和清洁剂公司寻找到他们的的目标受众以进行产品宣传。因此，仅他们俩就占据了美国在线或移动广告投入58%的额度。相较于能够产生很大广告销售数字的谷歌和Facebook，从电视网络到新闻机构等依赖广告收入的其他公司都在反思其现有的业务。

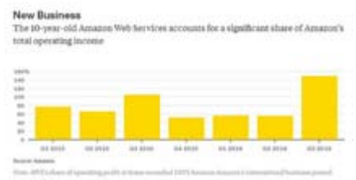
2016谷歌和Facebook广告收入额度远超其他公司：



三、亚马逊无限的野心

显然，在2016年，没有一个产业能够逃离出亚马逊的扩张。它是涉及到所有产业和服务的超大型零售商。在娱乐方面它的力量也在增长，同时，它还是准运输行业巨头，可能控制海陆空，甚至新的方式的运输业务。亚马逊网络服务云业务是它在十年前创造出来的一种计算业务，在第三季度构成超过总营业利润的100%(在计算国际损失后)。亚马逊的网络服务改变了亚马逊和高新技术产业的方向这一说法毫不夸张。

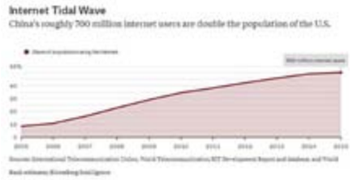
亚马逊网络服务营业利润占总利润比重：（亚马逊网络服务营业利润有时能够超过其总营业利润是因为亚马逊国际商业运营的亏损。）



四、中国科技大展拳脚

中国科技巨头百度、阿里巴巴和腾讯的体量之大都是难以想象的。在残酷竞争磨炼下的下一代，比如滴滴出行，以及中国的许多新技术想法都在被复制。中国的科技力量不仅在国内扩展他们的优势，也正在将其触角延伸到世界的其他地方。

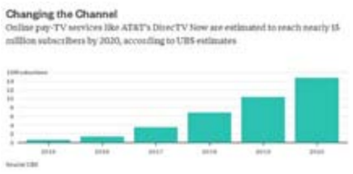
近年来中国互联网用户的增长：



五、电视终于与技术接轨

电视主导人们的休闲时间且可以称为广告商的“钱包”的时代已经过去了。电视的基本性质正在慢慢变化。智能手机上的非传统流行视频、发展起来的新的网络电视服务等正在重塑娱乐。数字电视只是简单的复制以前电视上的东西还是会成为完全不同的东西呢?

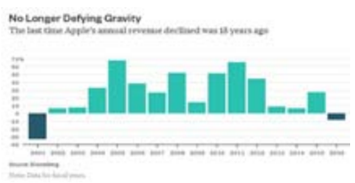
UBS估计的在线网络付费电视用户数量：



六、苹果进入了瓶颈期

长达10年的苹果时代就要结束了，他们的利润不再快速增长。今年苹果的收入是2001年以来的第一次负增长。这家公司不仅无法超越全球智能手机市场的变化，而且还要继续应对政府对其影响力的压制，比如执法、税务和生产方面。

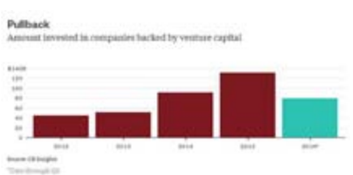
苹果公司2001年-2016年度收入情况如下：



七、初创公司估值紧缩

在2年看似对年轻科技公司的无限投资后，2016年这方面明显地紧缩了。科技初创公司投资的钱仍然处于历史高位，相较于2015年减少了很多。许多私人科技公司开始进行利润管理，而不是不惜一切代价地投入资本，否则投资回落的影响将可能更糟。

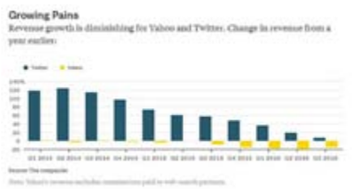
初创科技公司投资情况：



八、雅虎和推特的困境

互联网公司必须保持增长，否则就会陷入困境。2016年，雅虎和推特都经历了漫长的销售努力，他们不得不处理他们不再增长的收入和用户量。

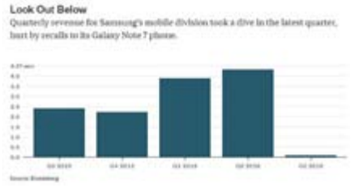
雅虎与推特近3年的收入增长情况：



九、电池的噩梦

三星在因电池上的错误造成火灾或爆炸后被迫停止生产galaxy note 7。美国也因电池过热召回了hoverboards，苹果处理新MacBook Pro生产线的电池寿命问题。2016年电池的“集体出事”显示了这一从智能手机到无人驾驶汽车的重要组件之一的脆弱性。

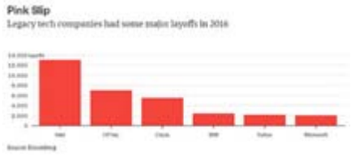
三星近5个季度收入情况：



十、旧技术紧缩

科技产业对落后者是残酷的，比如雅虎和推特。在某些情况下，这导致了2016年老牌公司为了抵消下降的收入或转移的资源而进行裁员，如英特尔公司，思科公司。惠普公司以及其他公司都在持续削减人员。

传统科技公司裁员情况：



十一、互联网公司中赚得最是盆满钵满的金融科技

金融科技的流行是因为它能够解决传统银行业务的低效率问题。任何人都想要快速、廉价以及更加个性化的金融服务。科技让这些需求得以实现，这也是当前世界拥抱金融科技革命的原因。近几年全球科技金融无论是投资额还是交易量都迅速爬升。

2011-2015全球金融科技投资额及交易情况：

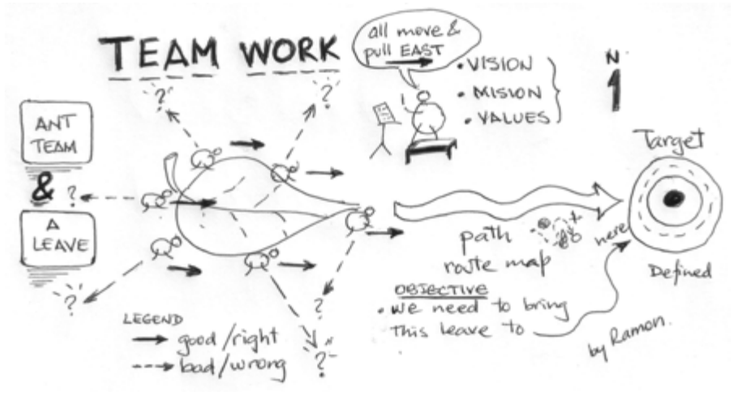
对于投资者来说，能够影响世界金融的全球27家金融科技公司中，美国占14家，资产总估值为310亿美元；中国占8家，资产总估值为964亿美元(约合人民币6693亿元)；世界其他国家还有5家比较有影响力的金融科技公司，其资产总估值为115亿美元。

27家金融科技公司的全球分布情况：美国东部5家分布在纽约、芝加哥和亚特兰大三个城市，其中纽约的OSCAR是美国硅谷以外最大的金融科技巨头，主要运营保险业务；西部9家都在硅谷，其中Stripe是美国估值最高的金融科技公司，主要提供在线支付服务；中国的八家金融科技公司都分布在东部沿海地区，其中阿里巴巴的蚂蚁金服无疑是体量最大的，估值达600亿美元(约合人民币4165亿元)。世界上最大的四个金融科技公司都是中国的。这是因为我们有超过5亿智能手机使用者，同时，移动支付和P2P信贷市场比较完善。除中美外，其他地区的金融科技公司仅有5家较有影响力。

以不同的经营类别来看，目前为止，支付和信贷是金融科技的两个最大市场。他们尽管比其他金融科技类别要成熟，但仍然有很大的发展空间。因为人们会逐渐适应移动支付以及其他新的商业模式。

十二、赚钱的除了金融科技公司，还有超级大IP

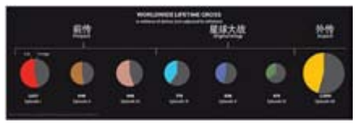
近年来，IP的赚钱能力大家有目共睹，而全球性的超级IP更甚。近期《侠



盗一号》上映，我们来看看《星球大战》这个IP的吸金能力究竟有强。

最新一部有完整数据的《原力觉醒》自上映首周即得到几部中上映第一周的最高票房，此后涨势也极好，最终拿到9.37亿美元(约合人民币64.7亿)的票房成绩。这也是目前《星球大战》的最好成绩。

下图为其全球总累计票房(单位：百万)。其中前传3部的票房成绩美国本身占大部分，而77年上映的第一部到83年的第三部《星球大战》及16年的外传在美国以外收益更高。在不计算通货膨胀的情况下，83年上映的《绝地归来》票房最低，但是也有4.75亿美元(约合人民币32.8亿)，可见其“吸金”能力之强。



一个《星战》IP赚了我们这么多钱，可我们的华侨都帮我们把钱汇了回来。

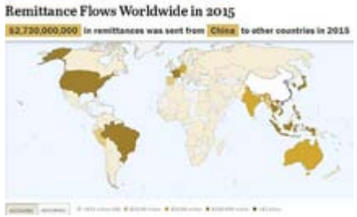
根据联合国数据，2016年有2.44亿人没有住在自己的国家。许多人离开了他们的家，去海外找工作。越来越多的家庭将他们的收入汇回国。在中国的移民把在中国赚的钱都汇到哪里了?海外华侨又汇了多少钱回来呢?

据世界银行数据，2015年总直接汇款将近5820亿美元(约合人民币4万亿元)。美国拥有世界上19%的移民，2015年他们汇了1335亿美元回家。这其中最大的受益者是墨西哥、中国和印度三个国家，分别占了243亿美元、162亿美元和100亿美元。

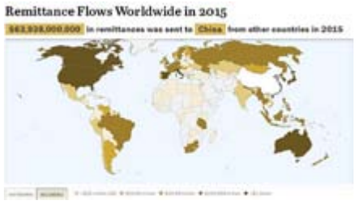
德国移民数量排在第二，2015年，他们也从德国汇了228亿美元回家，他们主要是欧洲国家。主要受益国是波兰、法国、意大利。

接下来两个图显示了中国2015年移民资金的流入流出状况。从中国流向其他国家的资金有27.3亿美元(约合人民币189.16亿元)，而流入中国的有639.38亿美元(约合人民币4千多亿元)。其中从中国获益最大的几个国家是韩国、菲律宾和日本，他们在这近200亿的汇款中，他们分别拿到了37.55亿元、31.38亿元以及19.88亿元。

外国移民将资金汇出中国的情况：



华侨汇款回国的情况：



这两个流入流出数据还不包括香港、澳门两个特别行政区在内，若将其计入那么这个数值将会更大。皮尤研究中心的数据中并没有找到台湾省这一选项，也没有说明有无计入中国数据之中。而香港、澳门两个特别行政区则能够查到有单独出来的数据统计。

这些汇款对于接收国家来说影响很大。世界银行数据显示，2015年，对于埃及来说流入的汇款相当于苏伊士运河收入的四倍；而对于尼泊尔，这类收入占了他们GDP总量的近1/3。

大部分移民寄回家的钱都用在食物和家庭应急上。有证据显示，许多接受者使用这些钱来提供给家人更安全、更健康的未来。

/ 2017年值得关注的5家深度学习初创企业 /

文 / 数据分析网　编辑 / 协会会员处 李缘　日期 / 2017-01

2016年是深度学习之年。但是那些喧嚣似乎更多属于大公司，这一方面是因为大公司占据了大部分的资源，也是由于AI的初创企业还没开始弄出点“大物件”就已经被大公司给收购了。那么今年又会有哪些公司被“盯上”呢？如果说人工智能（AI）此前还没有进入主流的话，2016年情况变了。

对此Google CEO Sundar Pichai发出了最有力的呼声，他说这个世界即将从“移动优先”进入“AI优先”。

苹果把AI挤进了iPhone里面，Google把它放到Pixel里面。Facebook把它带进了新闻流里面，微软把它植入到Word里面。三星收购了初创企业Viv以跟上苹果的Siri虚拟助手。而像Skype和Messenger这样的聊天应用现在都提供了聊天机器人。

目前大部分对AI的关注都落到了深度学习上面，也就是通过用大量数据对人工神经网络进行训练，然后让它们对新的数据做出推断。在过去5年的时间里，越来越多的深度学习初创企业在不断冒出来。

2016年，芯片巨头英特尔收购了深度学习软硬件制造商Nervana，企业软件公司Salesforce收购了开发深度学习软件用于迅速处理大量文本和图像的MetaMind。这两家都是去年列举的5家值得关注的深度学习初创企业之一。与此同时，2017的名单也新鲜出炉了。

Bay Labs

Bay Labs是应用深度学习到医疗成像的初创企业之一。其偏重工程的团队成员包括Johan Mathe在内，后者曾经在Google的Project Loon团队工作过。Facebook人工智能研究部门负责人Yann LeCun以及Khosla Ventures都对这家公司进行了投资。

Cerebras Systems

Cerebras是Andrew Feldman领导的一家秘密初创企业，后者曾经以3.34亿美元把微服务器公司SeaMicro卖给AMD。Feldman的这家新的初创企业正在开发AI硬件，据知情人士透露，知名风投机构Benchmark领投了对这家公司的一轮超过2000万美元的融资。

Deep Vision

总部位于Palo Alto的Deep Vision正在为深度学习开发低功耗的芯片。公司的两位联合创始人Rehan Hameed和Wajahat Qadeer还在斯坦福大学的时候曾经共同写过一篇有趣的论文，题目叫做《卷积引擎芯片多处理器》。

Graphcore

Graphcore开发了一种智能处理单元（IPU） PCIe加速器，神经网络可以利用这种硬件来训练和做出推断。该初创企业还开发软件来配合其基础设施利用现有的MXNet和TensorFlow深度学习框架开展工作。投资者包括Bosch Venture Capital、Foundation Capital以及Samsung Catalyst Fund等。

ViSenze

在2016年的ImageNet图像识别竞赛的个别竞赛单元上，成立于2012年的ViSenze表现要比好几个小组更出色。获得Rakuten Ventures投资的ViSenze是新加坡国立大学和清华大学联合成立的研究中心NExT拆分出来的初创企业。其软件可以对图像和视频进行对象识别和打标签，并且提供外观类似的内容。

/ 个人信息“黑市”日益猖獗，看国外如何保护公民隐私 /

文 / 中国网　编辑 / 协会会员处 袁硕　日期 / 2017-01

春节将至，电信网络诈骗犯罪活动再度呈现高发趋势。近年来，在互联网时代背景下，我国公民的个人信息遭到大量泄露，这无疑对诈骗犯罪活动起到了推波助澜的作用。

如今，个人信息贩卖“黑市”的日益猖獗，为诈骗分子提供了更多、更精确的公民个人信息，一定程度上增加了公民遭遇诈骗、损失钱财的可能性。

信息泄露问题严重

去年11月，中国青年政治学院互联网法治研究中心与封面智库联合发布国内首份《中国个人信息安全和隐私保护报告》(以下简称《报告》)。

《报告》通过对全国100多万份调查问卷进行分析研究，揭示了我国个人信息安全面临的严峻形势：超七成的受访者认为个人信息泄露问题严重;26%的人每天收到2至3条甚至更多的垃圾短信；多达81%的人经历过对方知道自己个人信息的陌生来电；租房、购房、购车等信息泄露后，被营销骚扰或诈骗的比例高达36%。

如今，个人信息的贩卖“黑市”已形成“地下大数据”。从信息来源看，个人隐私信息主要来自两方面：一是黑客通过入侵计算机系统非法获取信息；二是员工利用职务之便，将信息非法贩卖。

《中国青年报》上月援引公安部的数据报道称，2016年以来公安机关网络安全保卫部门共侦破侵犯公民个人信息犯罪案件1800余起，抓获犯罪嫌疑人4200余人，查获各类公民个人信息300余亿条，其中，抓获涉及40余个行业和部门的内部人员390余人、黑客近100人。

各国如何保护隐私

互联网时代和全球化背景下，个人信息保护早已成为世界各国关注的焦点。从西方国家最先开始立法保护个人信息至今，已有数十年历史。面对日益严峻的个人信息安全形势，世界各国在不断完善立法的同时，也在寻求新办法解决难题。

美国

1974年，美国参众两院就通过了《隐私权保护法》。这是美国行政法中保护公民隐私权的一项重要法律。以此为基础，美国采取分散法律的模式，依靠联邦和州政府的各类条例，来维护个人信息安全。例如，针对数据保护的《电子通讯隐私法》、面向金融机构的《金融服务现代化法案》、以及保护健康隐私的《健康保险携带和责任法》。

另外，由于立法总是呈现滞后性，美国政府还倾向于行业自律政策，鼓励企业参与到公民信息保护当中。目前的美国行业自律政策主要有三种：建议性的行业指示、网络隐私认证、技术保护模式。

“在线隐私联盟”是美国利用行业自律模式来保护公民隐私的典范。该组织由超过80家国际公司和协会组成，致力于为商业行为创造互信的良好环境、推动对个人网络隐私权的保护。1998年该组织发布了“在线隐私指引”，旨在指导网络和其他电子行业的隐私保护。

当地时间2013年6月5日，德国柏林，赤裸的活动家将身体涂成欧盟旗帜的颜色，在内政部外举行示威，请愿要求对私有数据进行更好的保护。(Timur Emek 视觉中国)

德国

早在1970年，德国黑森州就颁布了保护公民个人信息的《数据保护法》，这是世界上最早的隐私保护法。随后，有16个州相继通过与个人信息保护相关的法律。1977年，德国联邦政府正式颁布《联邦数据保护法》，全面保护德国公民个人信息。

德国对信息泄露行为的惩处格外严厉。《联邦数据保护法》规定，个人信息包括消费者提供给企业的姓名、年龄、性别、收入情况、身份证号码等，一些企业，如银行、电信公司，若泄露其掌握的客户资料信息都是违法行为。对于这些违法行为，根据具体情况给予经济和刑事处罚，重者可使其倾家荡产。

此外，德国还设有数据保护专员制度。德国联邦数据保护与信息自由专员是德国个人信息保护与信息自由法律实施的监督机构。数据保护专员在履职时保持独立，并且只服从法律、只接受法律监督，以此进一步保证对公民个人信息保护。

新加坡

早在2001年，新加坡政府和互联网服务商制定了行业自律规范《行业内容操作守则》。该守则规定必须尊重用户个人隐私。新加坡互联网服务提供商已经全部采用该守则，并将其纳入与用户的合同当中。

2007年，新加坡国会通过了《垃圾邮件控制法案》，对垃圾电子邮件展开重点整治。该法案规定，消费者对违反规定的垃圾电子邮件发送者可要求赔偿损失，每条垃圾电子邮件的赔偿费为25新元(约合128元人民币)。

2012年，经过两年的酝酿，综合考虑多方反馈，新加坡国会最终出台《个人信息保护法案》。该法案主要涵盖两方面：一是保护个人资料不被滥用;二是拒绝行销来电和信息。公司必须在获得消费者允许后，才能收集和使用消费者的个人信息，也需要向消费者解释他们收集和披露消费者个人信息的原因。

我国立法亟待加速

早在2008年第十一届全国人民代表大会开会期间就有代表提出制定个人信息保护法的提案。但我国个人信息保护统一立法却一直难产。

2009年修订的刑法将泄露个人信息的行为入罪，民法通则也有关于个人隐私保护的条例。目前，我国有近40部法律、30余部法规以及近200部规章涉及个人信息保护。但这些法律法规零星、分散，尚未形成体系。

如今，我国非法窃取、贩卖个人信息的“地下黑市”不断发展壮大，呈现出产业化、集团化、跨境化、智能化的趋势，个人信息安全形势日益严峻。因此，个人信息保护统一立法步伐亟待加快，为我国公民个人信息安全提供更有力的法律保护。

全国人大代表、南京邮电大学校长杨震曾在全国两会上多次提案，建议国家尽早启动个人信息保护法立法。杨震表示，信息社会需要建立在对个人信息保护的基础上，只有个人数据在法律保护下安全迅速地收集和流通，才能有利于促进我国社会的信息化进程，才能推动我国信息产业与世界接轨，这是信息时代发展的必然要求。

/ Python、R、Java、C++等：从业界反馈看机器学习语言趋势 /

文 / 网络大数据　编辑 / 协会会员处 李缘　日期 / 2017-01

对于开发者来说，掌握什么编程语言能更容易找到机器学习或者数据科学的工作？这是个许多人关心的问题，非常实际，也在许多论坛被翻来覆去地讨论过。非常显著的是 “Python 是大趋势”这一论调，似乎它即将在机器学习领域一统天下。

那么这种说法到底有几分事实？

首先要指出的是，大多数对编程语言的讨论都比较主观。比如说，有的开发者(尤其是初学者)会因为一门语言的某个特性很契合自己的使用习惯、用着最顺手，就狂赞这门语言，而对其他语言的优点选择性失明。而这篇编译自IBM开发者论坛的文章，则尽量避免了主观判断，用数据来展示各门开发语言在工业界的实际使用情况，可以说是十分难得。毕竟，统计学习的核心就是用数据说话。AI 开发者应该更明白可靠数据相比主观臆测的价值。

我们来看看 2016 年开发语言使用情况统计，到底哪门语言的使用人数上升最快?居前几位的都是哪些？下面是机器学习与数据科学领域各语言的雇主招聘指数对比图：

如图所示，这是利用美国职位搜索引擎 indeed.com 得出的机器学习、数据科学招聘趋势：对这些领域内开发职位所列出的编程语言要求进行了统计。它展示出公司、雇主们都在寻找哪些语言技能。

但注意：这并不能精确体现各公司开发人员正在使用哪些语言。这是美国的机器学习业界趋势，与中国学界关系不是那么紧密。没有包涵在搜索结果内的语言，不代表它们的招聘职位比上述语言要少。

我们可以清楚看出，美国雇主最需要的前四大语言排名是 Python，Java，R，C++。其中，Python 在 2015 年中超过 Java 跃升至第一。然后，把搜索结果限制在机器学习领域(去掉数据科学)，数据其实差不多：

这张折线图中包含了 Lua，但由于它的招聘职位实在太少，代表 Lua 的线与坐标轴重合。我们能从这两组数据中推断出什么？

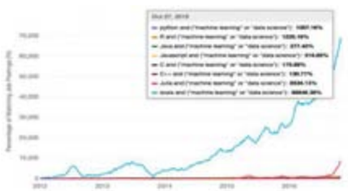
1、Python 是市场的领先者，作为最受欢迎的机器学习语言当之无愧。 另外，Python 与 Java 之间的差距正在被拉开。但是 Java 与 R 之间的差距正在被缩小。有业内人士对居第二位的语言是 Java 而不是 R 感到惊讶。通常，大家的主观感受是除了 Python，使用 R 语言开发机器学习应用最普遍。

2、Python 并没有成为霸主。各主流语言的招聘需求都很多。对小众语言如 Lua 和 Julia ，目前业界的需求确实小些，但其他语言都占有了相当的比例。

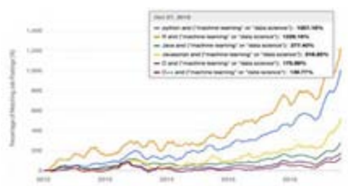
3、进入 2016 年后，市场对所有语言的需求都大幅上涨。Python 并没有与其他语言拉开明显差距。这表现出，最近一年里业界对机器学习和数据科学整体的兴趣和重视。

4、Scala 在 2014 年之后的增长十分惊人。2014 年之前，对它几乎没有招聘需求。但那年之后一直在稳定增长。2016 年，它赶上了 Javacript，达到主流语言阵营的水平。

作为一门口碑不错的新兴语言，Julia 的普及程度还很低。但在 2016 下半年有了巨幅增长。现在还看不出来它是否会成为主流语言。Scala，Julia，Lua 在机器学习、数据科学领域的雇主招聘指数增长率图中，我们可以很明显的看出来：



然后，当我们省略掉 Scala、Julia 和 Lua，统计主流语言的增长率，可以明白无误看出 Python 和 R 的增长速度远超其他主流语言：



R 的增长率始终高于 Python，位居第一，更是远超 Java。以此来看，在可预测的将来，R 不但不会消失，还会成为更受欢迎的主流机器学习语言。

因此，说“Python 是未来大趋势”肯定是不对的。但是，当我们聚焦于细分领域“深度学习”，数据就变得很不一样：

在深度学习市场，对 Python 的招聘需求仍然最高。但前五大语言的排序变成了Python，C++，Java，C，R。这里有很明显的对高性能计算语言的侧重。

而且，Java 的增长速度惊人，它可能很快成为深度学习市场的第二位。在可预期的将来，R 还不会成为最受欢迎的深度学习语言。令人惊讶的是 Lua 的存在感之低。要知道，开源框架巨头之一的 Torch 便是基于 Lua，许多开发者因此会认为它在深度学习市场占有特殊地位。

对于文章开头提出的问题——雇主需要掌握什么语言的开发者，答案已经很明确了：在机器学习和数据科学市场，Python, Java, 和 R 的招聘需求最大；在深度学习领域， Python, Java, C++ 以及 C 更被公司欢迎。

但请注意，这只是私营公司的招聘需求。学界研究人员的偏好会有很大出入。另外，自学 AI 的业余爱好者、暂时没打算在这行谋生的，也不需要对这些数据太过在意。

对于新入门、正犹豫选择哪门语言的技术宅，关键还是在投入大量时间之前多听、多看、多了解；比较不同人的观点，选择最适合自己的。

/ 达沃斯论坛，马云被问“与特朗普见面谁先找的谁?” /

文 / 凤凰国际　编辑 / 协会会员处 袁硕　日期 / 2017-02

据凤凰国际iMarkets报道，北京时间1月19日凌晨，马云在达沃斯论坛特别对话环节接受了纽约时报专栏作家安德鲁·罗斯·索尔金采访。这也是达沃斯论坛人气最为火爆的一场对话，会场外挤满了没有拿到票的听众，而媒体记者干脆在门口用手机开始直播。所有人都想从马云这里知道，即将上任的特朗普对于全球化和国际贸易将有何种举措，而马云的阿里巴巴将在美国和中国这世界前两大经济体的贸易往来中扮演什么角色。

对话即聚焦于此，索尔金从马云与特朗普的会面内容问起，谈论内容延展至贸易保护主义抬头背景下的全球化前景。马云表达了对全球化的坚定支持，同时他表示，全球化需要向更加普惠的方向发展，即升级现有为跨国巨头服务的国际贸易规则，更好的支持全球2000万家中小企业的进入国际贸易、分享全球化红利。而这也是马云所倡议的eWTP（全球电子商务平台）的根本目的。

马云表示，自己对未来十年中国和世界的重大变化感到兴奋。对中国和世界而言都将是个巨大的转变。

在对话的最后还发生了有趣的一幕，马云这位在去年创造了超过3万亿销售额，管理超过4万人、业务遍及全球的企业家在回答提问时表达了对加勒比海岸的阳光和沙滩的向往，他说：我来世界走一遭，并不仅仅为了工作。

马云、纽约时报记者索尔金

本次对话环节共持续30分钟，全程英语。

索尔金：我们会进行问答环节，并且谈谈阿里巴巴、中国、特朗普和美国、全球化与贸易等话题，再次欢迎马云先生！

索尔金：你刚刚到访了特朗普大厦，与侯任总统唐纳德·特朗普会面，可否介绍一下这次会面?

马云：这次会面富有成效，比我预期好得多。

索尔金：你有什么预期?

马云：我和大家一样也听说了很多、看到了很多关于他的新闻。进去之后，感到他心态开放且乐于倾听我的看法。我对于谈话的成果非常高兴，他主动提出要送我下楼，我想他对此次会面的成果应该也非常高兴。

索尔金：其实是你打给他、还是他打给你，这次会面是如何促成的?

马云：这也是我问自己的问题。有一天，有人问我你想不想与侯任总统见面，我说真的吗，我还没准备好，我不知道要谈些什么。过了几天，又收到了几次，有个朋友发邮件给我问同一件事，我想了想也许我应该去谈一谈，至少特朗普可能会对我说的东西感到高兴，所以我就去了。我们讨论了中小企、农产品、中美贸易，特别是聚焦让美国企业通过我们的网络面向亚洲销售，而这将为他们创造大量的就业。

阿里巴巴集团的董事兼总裁迈克·埃文斯（J. Michael Evans）

此前外界猜测，促成马云与特朗普会谈的人可能有两位，并且都是阿里巴巴的董事。第一个是阿里巴巴的投资人、董事会董事孙正义。在马云和特朗普会谈前，孙正义已经与特朗普进行过会面。第二个是阿里巴巴集团的董事兼总裁，主管国际业务的迈克·埃文斯（J. Michael Evans）。迈克·埃文斯是个中国通，也是华尔街上的传奇人物之一。在马云和特朗普的会面中，此人频繁出现在媒体镜头中。

阿里巴巴集团的董事兼总裁迈克·埃文斯（J. Michael Evans）

索尔金：你承诺未来五年为美国创造100万个职位，不是阿里巴巴直接雇佣？

马云：不是我们雇佣，阿里巴巴有4.5万名员工，我们没办法请100万人，我无法想象我们可以管理100万人。

索尔金：请谈谈你如何看中美关系，如何评论特朗普之前对于中国操纵货币的言论？你们在会面中谈到这些了吗？

马云：首先在美国言论自由，他可以说任何他想说的，我尊重而且也理解。但我也有自己的观点，我们不会辩论中美贸易、操纵这些，但我们达成了一些共识：中小企、开发中西部，帮助当地农民和中小企出口至中国。我们不会谈美国的就业职位流失到墨西哥、中国等等。我可以分享一下我的看法吗？首先，30年前当我刚刚大学毕业时，我们听说的是美国的美好战略，将制造就业外包给墨西哥、中国，把服务业外包给印度。有本书叫做《世界是平的》，作者是托马斯·弗里德曼。我觉得这是完美战略，美国说只想控制知识产权、科技、品牌，而将较低层次的工作交给世界其他地方，这是伟大的战略。

第二，美国的国际公司通过全球化赚了数以百万计美元，美国100强企业令人惊叹。我刚刚大学毕业时，当时想买摩托罗拉的BB机，售价是250美元，我的工资只有每月10美元，而制造BB机的成本仅仅8美元。过去30年，微软、思科、IBM这些公司赚的钱数以千万美元计，比中国四大行赚的钱加起来都多，比中国移动、中国联通等等加起来都多，他们的市值在过去30年增长了超过100%。那么赚来的钱都去哪了呢？作为商人我很关心资产负债表，关心钱由何而来、去往何处。

过去30年，美国在13场战争中花费了14.2万亿美元，如果这些资金有一部分用于投资基建、帮助白领和蓝领呢？无论你们的战略有多好，你们应该为民众而投资。不是每个人都能有机会上哈佛，像我就不行，我们应该为那些无力上学的人们投入资金。另外令我好奇的是，我年轻时听说的是美国有福特、波音等大型制造企业，而过去20年听到的都是硅谷和华尔街，资金流向了华尔街。

然后，2008年金融危机发生了，损失了19.2万亿美元，这是一笔巨资，洗劫了白领、毁灭了全球3400万就业。如果这些钱不是流向华尔街，而是投资了中西部、开发那里的产业，则会带来很大的改变。不是其他国家偷了你们的就业机会，这是你们的战略，是你们没有合理思考、分配资金。这是我的看法。

索尔金：现在我们看到了关于重构全球化的强烈抵制，实际上在达沃斯的多数讨论都是围绕这一话题。抵制发生于美国，但是习近平主席昨天来到了达沃斯，你也是与他同行的一员，对于他的讲话，你怎么看？

马云：全球化是很好的东西，美国是教育我们如何进行全球化的发达国家。记得2001年前后我们加入WTO时，我们都很担心——如果国际品牌和产品来到中国，毁灭我们的产业，让我们失去工作呢？当时说服了我们，20年之后你们却说这是很可怕的东西。我认为全球化是好的，但全球化需要优化，这是侯任总统特朗普希望解决的问题，我认为全球化应该是普惠的全球化。过去30年，全球化由6万家大企业控制；100年之前，是由几位国王控制。如果未来30年我们能够支持600万家企业跨境运营？如果未来30年我们能够支持2,000万家中小企跨境运营呢？我们相信全球化应该是普惠的。

索尔金：你认为习主席提到的事情会发生吗？中国代表了东方很多年，现在如果说美国将继续代表所有人而行动？

马云：习主席昨天提到，这是最好的时代，也是最坏的时代。世界需要新的领袖，但新的领袖意味着携手共进，这是我的理解。

新的领袖并不需要是特定的某一个，来教导大家什么可以做、什么不能做，但全世界需要团结在一起。作为一个中国的商人，我很喜欢也很自豪习主席昨天所提到的——作为一个商人，我希望全世界能够共同担负起责任、携手合作；作为一个中国人，我也对他所作出的承诺感到高兴，昨天他提及中国作为世界第二大经济体应该担负起责任。这是我第一次听到中国领导人作出量化承诺，他说未来10年我们的进口将达到8万亿美元，这让我感到兴奋，因为中国正从出口向进口转型，如果能够达成一个具体的数据目标，这对中国和世界而言都将是个巨大的转变。

索尔金：今天中国对全球化相对会更有兴趣，因为其带来的益处将继续支持中国朝着“发达国家”成长，对此你怎么看？

马云：首先，WTO规则不是中国制定的，也不是为中国而制定的。我想改变的是，过去WTO是为大企业而设计的，只有大企业能够参与。中国也从开放中受益良多，我认为中国应该学会一件事——过去中国能够增长，是因为我们面向世界开放，如果我们能够继续开放……（但不是完全开放，比如美国企业要进入中国就需要和当地企业合作？）这就是为什么我说中国也存在它的问题，这个世界存在着问题，中国当然也很多自己的问题，中国应该开放、应该自信等等。昨天习主席的话让我很有信心，他已经准备好让中国面向世界进一步开放。这是我的建议，我们应该通过商业团体、通过谈判来解决问题。中国已经加入WTO十几年，我想无论作为企业、作为国家、还是整个世界，都需要重新进行审视。不仅仅是因为不平衡的事物——我们可以喊停。

索尔金：你提出了eWTP，具体是什么？

马云：WTO很伟大，但它主要是为发达国家及其企业所设计的，对中小企而言没有机会。我们希望建立eWTP，Electronic World Trade Platform，来支持年轻人和中小企，他们可以通过手机和互联网在网上进行跨境买卖。此外，WTO也是一家非常有意思的机构，它能够20国政府走到一起、就一件事达成共识，这简直是不可能，我无法想象各方能够达成共识。商业应该由商人决定，我认为eWTP应该由商界人士们坐下来讨论、谈判、达成共识，而后获得政府支持的这么一个事物。

索尔金：关于阿里巴巴、关于你们自身的商业模式，我想大部分西方人可能不太理解。我能否尝试让你们与亚马逊做一下比较？这样比较可能你们会觉得不太公平。不过令我感觉很有意思的一点是，亚马逊所追求的，我感觉比较像是重资产的商业模式，他们购买飞机、想拥有整个供应链；而阿里巴巴就零售部分来看，相反地，你们并不想自营仓库、不想自营物流公司。对此你怎么看？杰弗里·贝索斯（亚马逊创始人及首席执行官）正确，还是你正确，还是你们会在中间地带会合？

马云：我希望双方都是正确的，因为世界不是只有一种商业模式，如果世界只有一种“正确”的商业模式，这个世界将非常乏味。我们需要各种各样的模式，为某种模式而努力的人们必须相信这种模式，我相信我所做的。

至于和亚马逊的不同，亚马逊更像是一个帝国，自己控制所有环节，从买到卖；我们的哲学则是希望打造生态系统，我们的哲学是赋能其他人，协助他们去销售、去服务，确保他们能够比我们更有力量，确保我们的伙伴、10万个品牌和中小企们能够因为我们的科技和创新，而拥有与微软、IBM竞争的力量。我们相信通过互联网技术，我们能够让每一家企业都成为亚马逊。

去年我们的GMV（商品交易额）超过5,500亿美元，如果要雇佣员工来负责这些商品的运送，我们需要500万人。我们不可能请500万人来运送我们平台上销售的商品，我们唯一能采取的方式就是赋能服务公司、物流公司，确保他们能够高效运作、能够盈利、能够雇佣更多人。

索尔金：如果自己不拥有供应链，是否能保证高效？人家亚马逊现在已经能够在几个小时内送货到家。

马云：去年我们在中国的125个城市实现了当日达。10年前从北京到杭州的邮寄就要8天，现在12个小时就能从北京送货到内蒙古城市，物流效率提升了。你不可能一天期望达到这样的进步，我们有足够的耐心。2016年双11我们平台上卖出了170亿美元的商品，3天内我们就派送了总共6亿个包裹。这就是正在发生，也是我们所骄傲的，不是我们挣了多少钱，而是我们具有多大的能量。我们可以使科技变得更有包容性，每一个小企业都可以使用，这是我的梦想。我1992年在中国创立我的第一家公司，一家小公司，为了向银行借5,000美元，花了3个月申请，仍然失败了，做小企业真的很困难。今天随着科技的发展，我们可以做到赋能，这是我想做的事情。

索尔金：一个对阿里巴巴目前仍在持续的批评是侵权问题。在中国知识产权是一个很大的问题，但阿里巴巴是个重要的批评对象。您认为阿里巴巴取得了哪些进步，如何看待其他国家的监管机构，包括美国，仍在质疑阿里巴巴？

马云：首先，当你拥有那么大规模的商业规模的时候，你必须学会接受批评。你必须倾听，再来判断哪些是对的，哪些是错的。

第二，作为一个赋能1,000万小商家规模的电商平台，我们不会像亚马逊Buy一样，特别是价值5,500亿美元的交易商品，你不可能全部检查，这是电子商务模式本身的问题。

第三，在过去17年，我们在打假和知识产权保护方面一直是领军者。但我们是互联网公司，没有执法权，我们发现了某人在卖假货，我们可以把他从平台上移除，但不能逮捕他。

去年一年，我们将400名涉假分子送进监狱，下架了3.7亿件假货。我们不但是打假的领军者，我们用大数据来监测谁在买、谁在制造、谁在售卖、地址在哪儿。我们现在对全世界尤其中国政府机构意识到这个问题感到高兴。好事是今天你去问这些“犯罪团伙”，这些制假者、售假者，他们说，他们可以去任何一个平台但现在他们不敢上淘宝天猫，因为我们的大数据科技可以查出他们是谁、地址在哪儿，并提交给警方，对他们进行捕获。

索尔金：您之前说过有关假货质量的话？

马云：我说的假货质量的话不是对假货进行赞扬，而是说经过这么多年，这些品牌商家必须非常小心，因为假货的质量正在大幅提高让人感到非常恐怖。这就是区别。你找到造假者，有人说，这是假货，你去找第三方鉴定到底是不是假货，发现有时假货质量更好。另一个更吓人的是，一家品牌说你们在卖假货，我们找了很久，想发现问题但找不到，后来从旗舰店卖了一个商品送过去检验，他们说这是假货。这很令人困惑。打假是同人性的贪婪作斗争，一点都不容易，也不可能结束，但必须继续战斗。我们每年投入2000名专职人员，每年投入10亿元人民币在打假中，不可能两年内结束战争。如果人们还在继续批评，重要的是我们自己对进展是高兴的。

如果人们赞扬我，说马云你很棒！我知道并不是很棒。或者阿里巴巴很棒！我们并非很棒，我们只是个17年历史的公司而已。但如果他们说，你们在打假上什么都没做，不，我们在做很多事情，但你不用去争辩，你只要去做自己相信的事就好。

索尔金：您谈到了在打假上运用了大数据，另一方面你们也在信用体系上运用了大数据，使贷不到款的人能够得到贷款。我们谈到芝麻信用，假如一个可能一些人没有信用纪录的交易市场上，怎么判断谁到底可以得到贷款，谁不该得到呢？

马云：首先，此前，我们有一个系统，教计算机学会怎么甄别假货，以及支付宝上的诈骗。我们已经做了十年。现在人们把这叫做AI人工智能。我们是一家数据公司，8年前，我们对自己说，我们不能做一家电商，我们要做一家数据公司，我们有消费者、制造商、物流、交易等等数据。

但如何运用好这些数据造福社会？很多中国小企业都非常好，有很好的信用，但是没有一个适合他们的信用系统。怎么用我们的数据，打造一个信用体系，使所有人都能获及这个信用体系？这在过去四年非常强大，所有用我们服务的人，我们都给他一个信用评级，在过去5年，我们放了500万个小企业商业贷款，即使他们只要求5000美元的贷款。3分钟可以决定是否放贷，给多少钱，1秒钟到账，不需要任何人去跟他们接触，我们叫这个是310。

芝麻信用也可以做谈恋爱的资本，丈母娘对未来女婿说，你要和我女儿谈恋爱，给我看看你的芝麻信用评分。如果人们要去租车，也会被要求看芝麻信用评分，如果他们不还钱，信用评分将会降低，可能无法租房。这是我们想要打造的系统。如果你买卖假货，芝麻信用也会体现。

索尔金：最后一个问题，对你进军好莱坞有很多猜测。您和阿里巴巴的名字出现在年初的几个大片中。阿里巴巴进军好莱坞的雄心是什么？

马云：每隔五年，我们都会做战略回顾，展望未来10年、30年。所有战略问题都问自己一个问题，是否解决社会问题？我们相信，解决的社会问题越多，你越成功。第二个问题，这个项目十年内会成功？那我们就做。如果一个月或一年就能成功？那就不用做了。怎么可能在一年和一个月成功？

五年前，我们有过一个大辩论，对未来10年、20年中国最需要什么。最后决定是happiness和health，双H战略。好莱坞电影能带给人快乐。现在没有人快乐，富有的人不快乐，穷人也不快乐。至少看电影能让人快乐，我觉得我们应该和好莱坞合作。

中国有很多英雄，中国英雄总有死，美国英雄永远不会死。如果所有英雄都死了，谁愿意做英雄，我想要我的英雄活下来，这个我们应该多多学习。目前我们只做了2年，还有8年。我想让我的公司不止是电商，而是给人启示。我从电影中得到很多启示。

我最喜欢的电影是《阿甘正传》。生活是艰难的，这是我从中学到的，得到了很多启示。在过去的17年别人说我是疯子笨蛋，你疯了，去做不可能做到的事情！你是个笨蛋，怎么能做这种模式，亚马逊是这个模式，eBay这个模式，阿里巴巴为什么这个模式？我对我自己说，阿甘说，继续干，别在意别人的想法。阿甘还说，没有人能挣钱，人们靠抓小虾挣钱。所以，我们服务小企业。

观众：你如何保证你不会搞砸人们的生活，你是有权力做决策的人，你如何保证你不会控制整个信用体系？

马云：首先我不确定，这是一个不确定的世界。每一天都不确定，唯一能确定是昨天。我不知道我会不会变独裁，或者会变愚蠢，这就是为什么我认为我应该趁年轻的时候退休。我有很多事情想做，我想做慈善、想做老师、想做环保。世界如此美好，我为什么总是要作为阿里巴巴集团CEO，我来到世界不是为了工作，而是来这个世界享受我的人生。我不想死在办公室里，而想死在阳光沙滩上。

观众：你觉得中国会进入贸易战吗？如果特朗普政府跟中国打贸易战，阿里巴巴会受到影响吗？

马云：我认为中美不应该打贸易战，永远也不应该有贸易战，我认为我们应该给特朗普政府一点时间， 他是一个思想开明的人，他在听大家的声音，我认为发动战争是非常容易，但结束战争是很困难的，甚至是没有可能。你看伊朗战争，阿富汗战争，它结束了吗？没有。我相信一件事，当贸易停止时，世界将陷入困境。

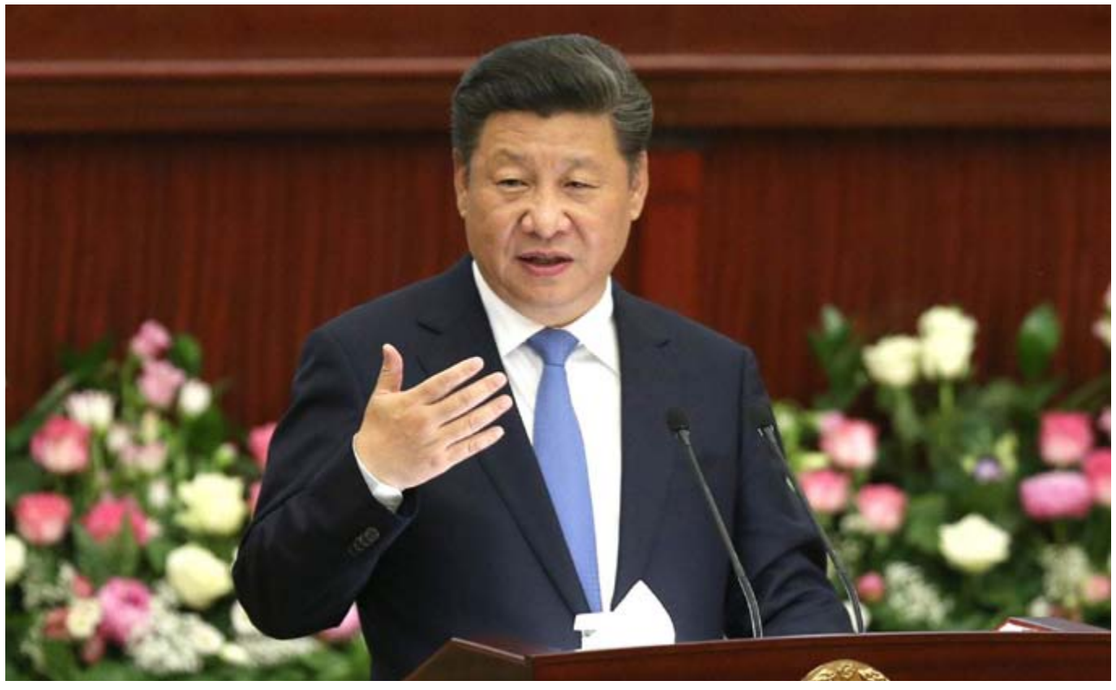
贸易让人们开始沟通，大家交流文化和价值。如果中美两国达成一致，阿里巴巴的商业模式将被摧毁，如果这样是可以停止贸易战的话，我也乐意去毁灭阿里巴巴的商业模式。你怎么能想象世界最大的经济体和世界第二大经济体有贸易战争，这将是一场对世界的灾难。如果我们能够把战争停下来，我们应该做任何事情来阻止它。

马云：我可以最后谈几句吗？我想向在座各位提出最后一个建议，所有政府都需要注意——未来30年对世界而言很重要。每次科技革命都需要50年，前20年科技公司出现、后30年科技得以应用。所以让我们关注未来30年，在此前的20年中出现了eBay、亚马逊、Facebook、阿里巴巴、谷歌……很好，但最重要的是让科技具有包容性、改变世界，这是未来30年。让我们留意那些30岁的人们，因为他们是互联网世代，他们会改变世界、会成为世界的建造者。第三，让我们留意那些雇员可能尚不足30人的小企业。30岁的人们、30人组成的企业，将会让世界更加美好。

索尔金：非常感谢马云先生，谢谢！

/ 习近平的2016步履：“只要路走对了，就不怕遥远” /

文 / 中国新闻中心　编辑 / 协会会员处 李缘　日期 / 2017-02



新年前夕，国家主席习近平通过中国国际广播电台、中央人民广播电台、中央电视台、中国国际电视台和互联网，发表二零一七年新年贺词。

岁月之海的奔涌从不停歇，砥砺奋进的征程永远向前。

“天上不会掉馅饼，努力奋斗才能梦想成真”、“大家撸起袖子加油干，我们就一定能够走好我们这一代人的长征路”……国家主席习近平在2017年新年贺词中的话语，温暖人心，擘画未来方向。

脚步丈量着大地和时光，一程又一程。

在2016年的300多个日日夜夜里，以习近平同志为核心的党中央，运筹帷幄、总揽全局，改革发展稳定、内政外交国防、治党治国治军全方位推进，治国理政新成就举世瞩目。

不管风吹浪打，胜似闲庭信步，从“进京赶考”到“党面临的‘赶考’远未结束”、从岁月峥嵘伟大远征到新长征之路，从“一带一路”壮美蓝图到行动画卷，这一年腾飞发展的新篇章徐徐展开。

新的起点，新的征程。习近平的话语言犹在耳、意蕴深长——“只要路走对了，就不怕遥远。”脚步已迈出，永远在路上。大国领袖迈着坚定而矫健的步履，带领全国人民一步一烙印，砥砺前行，驰而不息，久久为功。

“赶考”之路，砥砺前行迈向复兴“赶考”新征途。

1949年3月23日，毛泽东振臂一挥“今天是进京的日子，进京赶考去！”党中央离开西柏坡向北平进发。彼时，老一辈无产阶级革命家怀着忐忑之心，带着跳出“历史周期律”的命题应试。时隔67年，这场考试还远未结束，还在继续。

“60多年的实践证明，我们党在这场历史性考试中取得了优异成绩。同时，这场考试还没有结束，还在继续。今天，我们党团结带领人民所做的一切工作，就是这场考试的继续。”

2016年7月1日，在庆祝中国共产党成立95周年大会上，习近平对中国共产党的历史性“赶考”做了总结和展望。面对新一轮“赶考”，誓言铮铮掷地有声。中国共产党的发展奋斗史，就是一部领导中国革命、建设和改革，战胜一切艰难险阻、不断超越自我的“赶考”史。在这一年里，在以习近平为核心的党中央领导下，全国一盘棋，改革、创新、经济、外交、军事、扶贫等领域，齐头并进，全面发力。

新一轮“赶考”，全面从严治党战略蓝图渐次展开。

党的十八大以来，以中央八项规定为肇始，以作风建设为突破口，以党的群众路线教育实践活动为抓手，以反腐倡廉为动力，战略定力不断增强、规划蓝图日渐清晰、全面从严治党的战略思想在实践中不断成熟。

明者因时而变，知者随事而制。10月24日至27日，十八届六中全会聚焦全面从严治党，坚持思想建党和制度治党相结合，以新制定《关于新形势下党内政治生活的若干准则》和新修订的《中国共产党党内监督条例》为全面从严治党新作出重大部署,为深入推进党的建设新的伟大工程提供了行动指南。

“继续推进全面从严治党，共同营造风清气正的政治生态，确保党团结带领人民不断开创中国特色社会主义事业新局面。”十八届六中全会上的这一伟大号召，向全党再次吹响全面从严治党集结号，彰显了全面从严治党的实践新发展、时代新特征，铭刻了“四个全面”战略布局的历史新高度。

“道虽迩，不行不至；事虽小，不为不成。”全面从严治党迎来了新局面，未来之路如何走？“石可破也，而不可夺坚；丹可磨也，而不可夺赤。”建党95周年庆祝大会上，习近平郑重指出，严肃党内政治生活是全面从严治党的基础。我们要加强和规范党内政治生活，严肃党的政治纪律和政治规矩，增强党内政治生活的政治性、时代性、原则性、战斗性，全面净化党内政治生态。百代兴盛依清正，千秋基业仗民心。全面从严治党从无“休止符”永远在路上。

新一轮“赶考”，披荆斩棘铁腕反腐，行动脚步更加坚决。

2016年，我国反腐利剑再出鞘，力度不减节奏不变重拳频出，拍蝇惩贪、天网追逃追赃全面推进。

以“小切口”推动“大变局”，从推动中央八项规定精神落地生根破题。一组数据清晰表明了党中央坚决纠正四风的决心和恒心：2016年1月至11月，全国各级纪检监察机关共查处违反中央八项规定精神问题3.58万起，处理党员干部5.08万人，给予党政纪处分3.75万人。证明执纪越往后越严，绝非一句空话。

除了常规的中央巡视组进驻巡视，“回头看”已经渐渐成为常态。打破少数腐败分子的“避风”心态，以不定期“回头看”和时不时“回马枪”，能够防止“破窗效应”。中纪委机关报刊文指出，“巡视‘回头看’，不是‘回眸一笑’，而是‘回马一枪’。这一枪，要点中要害、枪枪见血”。

2016年中央巡视组三度杀出“回马枪”，对辽宁、安徽、山东、湖南、天津、江西、河南、湖北、北京、重庆、广西、甘肃等12个省市进行“回头看”。向全体党员干部明确传递信号：中央正风肃纪是常态工作，“打虎拍蝇”不仅剑指每一处角落，而且坚持久久为功，千万莫伸手，伸手必被捉。

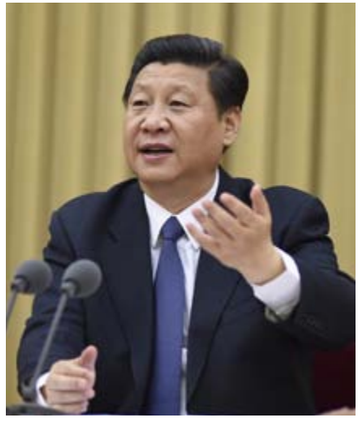
法立，有犯而必施；令出，唯行而不返。持之以恒对抗腐败，就必须依靠制度、坚持法治。事实上，随着《中国共产党巡视工作条例》、《中国共产党廉洁自律准则》、《中国共产党纪律处分条例》等法规条例的颁布或修订，在全面从严治党的时代要求下，党内法规体系逐步健全，反腐迈入制度化、法治化的轨道。

知易行难，任重道远。这是一场必须赢的生死仗。方此之时，唯有高举反腐的利剑，扎牢制度的笼子，从细处着手，向实处出发，在永不停歇的“赶考”路上坚定前行。

新一轮“赶考”，风正扬帆正当时，交出时代合格答卷。

“一切向前走，都不能忘记走过的路；走得再远、走到再光辉的未来，也不能忘记走过的过去，不能忘记为什么出发。面向未来，面对挑战，全党同志一定要不忘初心、继续前进。”习近平在建党95周年庆祝大会上说。

深化改革的“最后一公里”如何打通？创新驱动的引擎怎样点燃？“不让一个人掉队”的脱贫硬仗如何打赢？以发展方式的转变如何实现？跋涉在民族复兴之路上崛起中的大国，面对的沟坎与险滩前所未有。



事不避难者进。在自我书写、自我完善的路上，以吐故纳新的智慧全面从严治党，以壮士断腕的勇气深化改革，以爬坡过坎的意志奋力转型，以登高望远的胸怀走向世界，则必将在实现“两个一百年”的奋斗征程中交出优异答卷。

浩荡激流，百转千回。如今走到一个关键的历史大分水岭上，比任何时期都更接近中华民族伟大复兴的目标，却又面临着前所未有的巨大风险和挑战。以习近平同志为核心的党中央挺立时代潮头，知难勇毅向前，引领人民砥砺前行，迈向复兴的“赶考”新征途；引领“中国号”巨轮行稳致远，向着光辉的彼岸破浪前行。

新长征之路——不忘理想初心跃马新的征程。

80年前，中国共产党领导的工农红军完成了人类历史上旷世罕见的战略大转移，以向死而生的勇气，开创出一个革命新局面的起点。星移斗转，光阴荏苒，长征精神始终闪耀火热的光芒。

新的起点，新的长征。“长征这一人类历史上的伟大壮举，留给我们最可宝贵的精神财富，就是中国共产党人和红军将士用生命和热血铸就的伟大长征精神。”在纪念红军长征胜利80周年大会上，习近平深刻总结了长征的伟大意义和精神内涵，生动阐释了长征精神跨越时空的时代价值。

时空坐标下的改革创新，催生新的活力。

长征胜利80周年，历史翻开了新的一页，深化国防和军队改革也面临着一个难得的“机会窗口”，对人民军队而言，正如同一次新的长征，是一场回避不了的时代大考。

中流击水惟创新者强。八一前夕，习近平在中共中央政治局第三十四次集体学习时，深刻论述深化国防和军队改革的重要意义，高度肯定这轮改革“解决了一些多年来想解决但一直没有很好解决的问题，解决了许多过去认为不可能解决的问题”。

一年间，首艘国产航母主船体合拢成型，新一代隐身战斗机歼—20震撼亮相，执行长距离运输任务的运—20列装空军；一年间，“军委管总、战区主战、军种主建”的格局确立；新成立的陆军领导机构、火箭军、战略支援部队和军委机关15个部门相继亮相；七大军区调整为五大战区，中央军委联勤保障部队成立，能打胜仗成为主攻方向；一年间，《中央军委关于深化国防和军队改革的意见》、《关于深化国防和军队改革期间加强军事法规制度建设的意见》、《加强实战化军事训练暂行规定》等纲领性文件和制度政策相继出台……

军改一年来，开新图强的宏伟蓝图以恢弘磅礴之势铺开。人民军队迈步踏上了深化国防和军队改革的“新长征”，科学擘画，寄寓未来，在星夜兼程的改革强军征途中奋勇前行。

心系群众苗得土，推动民生跨越。

挽住云河洗天青。心中有人民，行动有方向。从长征中走来的红军队伍，极其珍视血火间铸就的军民情。

“小康路上一个都不能掉队。”习近平向世界庄严宣布：“全党全社会要继续关心和帮助贫困人口和有困难的群众，让改革发展成果惠及更多群众，让人民生活更加幸福美满。”2017年的新年贺词里，习近平“最牵挂”的还是困难群众，站稳群众立场、践行群众路线。回首这一年，习近平考察调研了7省市，足迹遍及重庆、江西、安徽、黑龙江、宁夏、河北、青海。扶贫、改革、创新、民生、生态等一直是他记挂心头的事。

1月，习近平在重庆调研。拳拳之心，爱民之情。在这次考察中他指出，扶贫开发成败系于精准，要找准“穷根”、明确靶向，量身定做、对症下药，真正扶到点上、扶到根上。“在扶贫的路上，不能落下一个贫困家庭，丢下一个贫困群众。”春节前，习近平来到江西吉安、井冈山等地调研考察。辞旧迎新时，总书记与贫困乡亲的心紧相连。

4月，安徽调研。“全面建成小康社会,一个不能少,特别是不能忘了老区。”五个小时的奔波辗转，总书记就是要了解农村脱贫特别是革命老区扶贫的真实情况。

5月，习近平到黑龙江考察，他十分关心林区全面停止商业性采伐后，产业转型发展和职工就业安置情况。“转型发展，民生为要。”总书记的讲话为东北振兴之路注入新内涵。

7月，习近平到宁夏考察并专门主持召开东西部扶贫协作座谈会。他强调，东西部扶贫协作和对口支援，是推动区域协调发展、协同发展、共同发展的大战略。同月，习近平还来到河北唐山市，就实施“十三五”规划、促进经济社会发展、加强防灾减灾救灾能力建设进行调研考察。

“要通过改变生存环境、提高生活水平、提高生产能力实现脱贫，还要有巩固脱贫的后续计划、措施、保障。”8月，习近平考察青海省海东市互助土族自治县班彦村指出，移民搬迁是脱贫攻坚的一种有效方式。

精神打底树立航标，引领伟大复兴新征程。

“今天是实现‘两个一百年’奋斗目标的新长征”“长征永远在路上”“不忘初心，走新的长征路”。习近平前不久到宁夏考察参观三军会师纪念馆时作出上述表述。在人类物质与精神文明高速发展的今天，回望长征初心，在一代又一代中国人心中刻下深深的烙印。如今，长征这一精神的源流，依然滋养着一代代中国人的心灵。

新中国成立前夕，毛泽东同志冷静告诫：夺取全国胜利，这只是万里长征走完了第一步。67年后的六盘山上，习近平同志郑重强调：我们每代人都要走好自己的长征路。

“当今世界，要说哪个政党、哪个国家、哪个民族能够自信的话，那中国共产党、中华人民共和国、中华民族是最有理由自信的。”在实现伟大梦想的伟大征途上，习近平这样说。

走好新长征路，在振奋精神中凝聚中国力量。67年艰苦创业，38年革故鼎新，无不印证一个道理：用共同理想凝聚民族意志，振奋亿万人民的爱国热情和昂扬斗志，没有什么能阻挡我们实现伟大梦想。

不忘理想初心继续前进，跃马新的征程谱写诗篇。今天的中国，不再只是“中国之中国”“亚洲之中国”,更与世界前途息息相关。淬炼坚如磐石的理想信念，百折不挠的英雄气概，中华儿女将以中国精神激发力量，在新的征程上续写新时代的壮丽史诗。

一带一路——奏响合作共赢曲引领新繁荣。

回首2016年，习近5次出国访问，推动双边关系，引领多边进程，足迹遍布十多个国家，在世界舞台上发出中国声音，为全球治理提出中国方案。追寻习近平的足迹发现，这一年来，在习近平密集的出访活动中，“一带一路”这一“热词”频频出现。世界由接纳到响应，再到积极行动，中国的“一带一路”朋友圈正在做实做大，承前启后的外交格局日渐清晰。

脚步不停，奏响中国特色大国外交新乐章。

2016年伊始，习近平访问沙特、埃及、伊朗。中东三国行，从推进“一带一路”谅解备忘录到五大领域交往合作，打开了中国梦与中东梦相融合的“筑梦空间”，揭开了新的一年“一带一路”新篇章。

3月出访捷克。中捷合作规划纲要启动，依托捷克传统工业强国的背景优势，“中国制造2025”与“捷克工业4.0”的战略对接全面展开，以点带面，为“一带一路”再提速。

6月，对传统友好国家的“走亲戚”之行。17日至24日，习近平访问塞尔维亚、波兰、乌兹别克斯坦并出席上合组织峰会，这是一次“一带一路”建设的提速之旅。习近平此行以“一带一路”建设为主线，依托传统友好国家，辐射中东欧、

中亚两大区域和上合组织重要平台，是完善中国总体外交布局的重要一步。

10月，脚步不停，习近平访问柬埔寨、孟加拉国并出席金砖国家领导人会晤。同饮一江水，共取一瓢饮。再次彰显了中国对周边外交的重视。中国周边国家在“一带一路”建设中正在迎来早期收获。

11月，访问拉美三国并出席APEC领导人非正式会议，站立历史潮头，习近平提出要推动建立覆盖整个亚太的全方位、复合型互联互通网络，调动起包括拉美成员在内各方参与互联互通合作的更大热情。

正如习近平所说，“一带一路”建设不是封闭的，而是开放包容的；“一带一路”不是中国一家的独奏，而是沿线国家的合唱；“一带一路”不是某一方的私家小路，而是大家携手共进的阳光大道。

深耕细作，硕果累累，务实合作带来新进展。

从中巴经济走廊、孟中印缅经济走廊到中俄蒙经济走廊，一个个区域合作新倡议应运而生；从俄罗斯欧亚经济联盟建设、欧盟“容克计划”到英国“英格兰北部经济中心”，都在积极探索与中国“一带一路”的对接；中老铁路、中泰铁路、印尼雅万高铁……“一带一路”合作共赢的理念，在东南亚地区生根发芽。

“一枝独秀不是春，百花齐放春满园”。作为人类“命运共同体”的重要组成部分，“一带一路”不仅造福中国人民，更造福沿线各国人民”，彰显出中国倡导的坚持共商、共建、共享的原则。

“一带一路”建设在探索中前进、在发展中完善、在合作中成长。沿线各国聚焦政策沟通、设施联通、贸易畅通、资金融通、民心相通，不断深化合作，已经在多个方面取得积极成果。

寻求世界经济突围路径，习近平强调“知者善谋，不如当时”。因时而生，共谋发展，“一带一路”正是促进国际经济合作的中国方案。

筚路蓝缕，春华秋实，携手共进擘画新蓝图。

在2016年8月17日召开的“一带一路”建设工作座谈会上，习近平指出，目前，已经有100多个国家和国际组织参与其中，中国同30多个沿线国家签署了共建“一带一路”合作协议、同20多个国家开展国际产能合作，联合国等国际组织也态度积极，以亚投行、丝路基金为代表的金融合作不断深入，一批有影响力的标志性项目逐步落地。“一带一路”建设从无到有、由点及面，进度和成果超出预期。

“世界那么大，问题那么多，国际社会期待听到中国声音、看到中国方案，中国不能缺席。”习近平的2016年新年贺词展现大国气度、大国担当。从满载荣光与历史的丝绸之路，到放飞希冀与梦想的“一带一路”，再到充满喜悦与自豪的G20杭州峰会，9月，在这样一座与丝绸缘分匪浅的城市，杭州呈现给世界一种历史和现实交汇的独特韵味，让世界感受到了互利共赢的“中国担当”与合作互联的“中国智慧”。

8年前，在国际金融危机最紧要关头，二十国集团临危受命，把正在滑向悬崖的世界经济拉回到稳定和复苏轨道让团结战胜了分歧，共赢取代了私利。

8年后，世界经济又走到一个关键当口，二十国集团承载着世界各国期待，使命重大。

杭州峰会期间，“一带一路”成为与会嘉宾讨论的关键词。习近平向世界郑重承诺：“中国的发展得益于国际社会，也愿为国际社会提供更多公共产品。我提出‘一带一路’倡议，旨在同沿线各国分享中国发展机遇，实现共同繁荣。”

钱塘江畔，弄潮儿向涛头立；钱塘江畔，杭州峰会与“一带一路”琴瑟和鸣。德国《世界报》驻华记者约宁·埃林说：“提出‘一带一路’倡议、倡议设立亚投行、举办G20杭州峰会，中国为世界经济提供了越来越多的公共产品，为推动全球经济治理改革作出了贡献。”

从愿景擘划到落地生根再到深耕细作，“一带一路”从历史深处走来，留下一行行坚实的足印。当今日自信的中国再次阔步走向世界舞台中央，古老丝路再次被唤醒，展望前方，“一带一路”倡议合奏时代交响，携手促共赢，必将向着未来创造新繁荣。

/ 回望2016——国务院常务会议大数据 /

文 / 人民网 编辑 / 协会会员处 李缘 日期 / 2017-01

“大众创业、万众创新”、“简政放权、放管结合、优化服务”、“督查”。2016年，国务院共召开39次常务会议。以下从“去产能”、“促进消费”、“新型城镇化”等10个方面，对这一年来的国务院常务会议进行梳理，为公众呈现“国务院常务会议大数据”。

一、12次提及“大众创业、万众创新”

1月13日	确定完善高新技术企业认定办法，使更多创新型企业得到政策支持
2月17日	确定支持科技成果转移转化的政策措施，促进科技与经济深度融合
3月30日	决定新设一批国家自主创新示范区，部署推进上海加快建设科技创新中心
4月20日	决定建设一批大众创业万众创新示范基地，推动双创迈向更高层次和水平
5月18日	确定持续推进商事制度改革的措施，营造有利创新的市场环境
6月1日	决定再取消一批职业资格许可和认定事项持续降低就业创业门槛
6月8日	决定建设福厦泉与合芜蚌两个国家自主创新示范区，引领带动体制创新和科技创新
7月20日	通过“十三五”国家科技创新专项规划，以创新型国家建设引领和支撑升级发展
9月1日	确定促进创业投资发展政策措施，部署建设北京全国科技创新中心，推进创新型国家建设
10月14日	部署持续深化商事制度改革，更大降低创业创新制度性成本
11月23日	决定再取消一批职业资格许可和认定事项，更加便利就业创业
12月21日	批准2016年度国家科学技术奖励评审结果

2016年4月25日下午，李克强总理来到四川成都菁蓉创客小镇，应邀与创客团队设计的羽毛球机器人“切磋”球技。

二、9次提及“互联网+”

1月27日	决定推动《中国制造2025》与“互联网+”融合发展
3月18日	部署推进“互联网+流通”行动，促进降成本扩内需增就业
5月4日	部署推动制造业与互联网深度融合，加快“中国制造”转型升级
6月8日	确定发展和规范健康医疗大数据应用的措施，通过“互联网+医疗”更好满足群众需求
7月20日	部署推进“互联网+物流”降低企业成本便利群众生活
9月14日	部署加快推进“互联网+政务服务”以深化政府自身改革更大程度利企便民
10月14日	运用“互联网+”在全国推进企业登记网上办理，鼓励对企业登记全程电子化优先
12月21日	会议通过“十三五”卫生与健康规划，促进“互联网+医疗”更大范围应用
12月28日	确定要利用大数据、云计算等推动“互联网+教育”发展，促进优质教育资源共建共享

2016年5月25日下，李克强总理在贵阳出席中国大数据产业峰会暨中国电子商务创新发展峰会开幕式并致辞。这是开幕式前，李克强总理参观贵州大数据成果展并与参展商交流。

三、7次提及“督查”

5月4日	决定对促进民间投资政策落实情况开展专项督查，着力扩大民间投资
6月22日	听取民间投资政策落实专项督查工作汇报，要求以不断深化改革调动民间投资积极性
7月7日	部署各地各部门整改审计查出的问题，通过改革完善制度提高公共资金绩效
7月27日	听取关于地方和部门推进重大项目落地审计情况汇报，完善奖惩机制
8月16日	部署对钢铁煤炭行业化解过剩产能开展专项督查，确保完成既定目标任务
10月31日	听取国务院第三次大督查情况汇报，推动改革发展和民生改善政策措施切实落地见效
11月29日	听取中央企业监督检查情况汇报，强化外部监督促进国资经营提质增效

2016年5月23日，李克强总理考察武汉钢铁公司。总理强调，化解过剩产能过程中，要保证企业多余人员转岗不下岗、转业不失业，确有困难的人员社保要兜底。

四、6次提及“简政放权管结合优化服务”

1月13日	决定再推出一批简政放权改革措施，让市场活力更大释放
5月11日	决定进一步精简投资项目报建审批，以改革营造更加便利的投资环境
6月15日	失效一批与现行法律法规不一致，不利于办事创业，不适应经济发展需要的政策性文件
10月08日	确定进一步精简政府核准的投资项目，简化外资企业等审批管理
10月31日	确定全面推进政务公开相关实施细则，促进政府施政更加透明高效
12月28日	以深化改革促开放，降低制度性交易成本

2016年11月21日，李克强总理考察上海自贸区市场监督管理局。该局将原工商、质监、食药监、价格监督检查职能合并为“四合一”网上综合执法平台，开展“双随机一公开”等市场监管。

五、6次提及“惠民生”

2月3日	决定实施新一轮农村电网改造升级工程，以补短板、调结构、促精准长惠民生
2月14日	部署推动医药产业创新升级，更好服务惠民生稳增长
3月18日	确定2016年经济体制改革重点工作，为促发展惠民生防风险提供保障
4月6日	确定2016年深化医药卫生体制改革重点，让医改红利更多惠及人民群众
8月16日	确定完善社会救助和保障标准与物价上涨挂钩联动机制，更好保障困难群众基本生活
12月21日	加快灾后水利薄弱环节和城市排水防涝能力“补短板”建设，促进民生改善和生态修复

2016年2月1日，恰逢农历小年，李克强总理来到宁夏固原市原州区中心敬老院，给老人们带来年货大礼包，亲手把“福”“寿”挂件挂在墙上。

六、5次提及“去产能”

1月22日	决定进一步化解钢铁煤炭行业过剩产能的措施，促进企业脱困和产业升级
5月18日	抓紧淘汰钢铁、煤炭等行业央企落后产能，加快重组整合和市场出清
7月27日	部署建立法治化市场化去产能机制，推动产业升级
8月16日	部署对钢铁煤炭行业化解过剩产能开展专项督查，确保完成既定目标任务
11月23日	听取2016年钢煤行业去产能工作基本完成情况汇报，派国务院调查组严查违法违规

2016年1月5日，李克强总理来到山西焦煤集团官地矿，下井前遇到刚出井的矿工。

七、5次提及“新型城镇化”

1月22日	部署深入推进以人为核心的新型城镇化，更大释放内需潜力
2月3日	结合推进新型城镇化、农业现代化和扶贫搬迁等，实施新一轮农村电网改造升级工程
3月30日	通过《成渝城市群发展规划》引领西部新型城镇化和农业现代化
5月4日	确定培育和发展住房租赁市场的措施，推进新型城镇化满足群众住房需求
9月14日	按照供给侧结构性改革和新型城镇化发展要求，大力发展钢结构、混凝土等装配式建筑

2016年5月6日，李克强总理考察人力资源社会保障部，与正在会商农民工就业形势的工作人员展开讨论。

八、3次提及“营改增”

1月22日	部署深入推进以人为核心的新型城镇化，更大释放内需潜力
2月03日	结合推进新型城镇化、农业现代化和扶贫搬迁等，实施新一轮农村电网改造升级工程
3月30日	通过《成渝城市群发展规划》引领西部新型城镇化和农业现代化

2016年4月1日，李克强总理就全面实施营改增到国家税务总局、财政部考察并主持召开座谈会。

九、3次提及“促进消费”

5月11日	部署促进消费品工业增品种提品质创品牌更好满足群众消费升级需求
8月24日	部署促进消费品标准和质量提升，增加“中国制造”有效供给满足消费升级需求
10月14日	确定进一步扩大国内消费的政策措施，促进服务业发展和经济转型升级

2016年3月25日，李克强总理考察海南博鳌乐城国际医疗旅游先行区，称赞这里把医疗养生度假结合，使传统医疗和旅游跃上新层次，形成新业态。

十、3次提及“关爱儿童”

1月27日	部署全面加强农村留守儿童关爱保护
4月13日	听取山东济南非法经营疫苗系列案件调查处理情况汇报，决定对一批责任人实施问责
6月1日	部署加强困境儿童保障工作，对他们的成长给予更多关爱帮助

2016年2月1日，李克强总理到宁夏固原西吉县半子沟村看望特困村民王进宝一家。

/ 大数据寒冬，如何冰解的破？ /

文 / 协会市场处 冯雪 编辑 / 协会会员处 袁硕 日期 / 2017-02

新年伊始，寒风料峭。有一股风扑面而来，大数据行业寒冬已至！一些公司乱了阵脚；一些公司沾沾自喜；行业的日渐成熟，自然去芜存菁！市场的专业性要求，自然大浪淘沙……大数据寒冬已至，如何冰解的破？行业大咖、事务所专家、资深学者，煮酒论剑。

嘉宾：邹东生 会长

中国商业联合会数据分析专业委员会会长；中国数据分析行业发起人、奠基人

丰富的企业经营管理咨询经验，自身数据分析专家，主持编写《投资数据分析》、《经营数据分析》等书。

嘉宾：孙雪 老师

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。

嘉宾：杨军 先生

重庆传晟数据分析师事务所主任

重庆大数据“十三五”产业规划课题组成员，擅长项目运营管理，具有丰富的实战经验，对房地产、医疗等行业有深入了解，担任当地人社劳动局、政府应用规划设计的特别顾问。拥有大数据行业多年从业经验，对行业未来发展有独到见解。

主持人：大家晚上好，从抛出今天直播的主题开始我就不断听到各种质疑的声音：有学员、有老师、有公司领导。当下这么热的行业怎么会和寒冬扯上关系？收到这些质疑其实我们是开心的，证明大家都认真思索并热爱这个行业！今天我们有请到中国商业联合会数据分析专业委员会会长，邹东生先生，北京犀数科技有限公司，首席数据官孙雪女士，重庆传晟数据分析师事务所主任——杨军先生和大家解读大数据寒冬的深意以及身处这个行业又将如何冰解的破？

主持人：就如刚才提到的，大家看到“寒冬”这个词都会心头一怔，为什么在大数据越来越热的时候，大数据圈子里却有“寒冬”的说法，各位嘉宾是否听说这个说法？又是如何看待这个“寒冬”的呢？

邹东生会长：“寒冬”这个说法，其实对于大数据领域的行业协会（2016年听好多人聊过）。这个说法是有一定理由的，2016年大数据很热，从国家政策到社会认知。但是也有一些相对冷的现象，比如很多人不懂什么是大数据，也就不敢深入接触大数据，这也就是市场上所谓的“冷”。

大数据业务不像互联网行业最初的那样，因为互联网很多是从开源入手，从宣传，从功能扩张开始，但大数据从一开始就给企业带来商业价值，优质的大数据企业就应该会有相应的营收，因为大数据他可以帮企业更好的决策。但是这也说明另外一个问题，为什么现在有的大数据企业相对偏冷，其实这是对于只有大数据概念，没有营收，不能解决实际问题的大数据企业，对他们来说，也许就是很“冷”。

这种大数据概念公司非常多，比如：各种名义上的大数据公司，如：以前做软件开发的、做市场调研、做硬件的、做BI的、做ERP,做数据交易的,只要和数据沾边都是做大数据的，但大家忽略了一点，客户不关心你是如何给自己企业贴上大数据标签的，大家只关注大数据能给自

身带来什么，如果你不能认清大数据的价值，怎么能让企业更好的满足客户需求？

杨军先生：大数据产业开始运营的时候，那个时候我觉得方向有一定的问题，很多企业很多公司，他们把重心放在了他们的产品研发上，把最后的分析、建模其他软件开发放在前面，但事实上，当时很多企业不具备做大数据产业的方法，我们发现从数据员开始做，反而阻力和困难小一些。

孙雪老师：其实，说到“大数据寒冬”，除了会长和杨总刚才提到的之外，我觉得，还有一部分原因就是随着大数据体系的不断完善，使得大数据产业创业成本变得很低，所以导致中小企业和创业者，他们进入大数据产业的门槛大幅度降低了，由于对这种新领域和新兴技术的过度的追捧和投入，形成了过度的竞争，所以很多公司开始收缩。在可以预见的未来，可能会出现更多的公司业绩下滑和倒闭的现象。

杨军先生：所以目前的状态下，很多公司倒闭，很多公司业绩下滑，很多公司出现了瓶颈，其实我们从目前跟政府的合作上看，它们目前想解决的方法或者路径都没搞清楚，根本谈不上大数据应用场景来做，所以，我们觉得有几个方法可以和大家交流一下：第一，以大数据咨询的方法来做，规划我们的路径和方法；第二，把我们的数据服务拿出来，从ETL建数据仓库开始；第三点才是通过技术手段解决目前的应用。

邹东生会长：现在的很多的大数据公司，更多偏重在一些概念上的炒作，比如大数据很多新兴的概念、新兴的大数据技术的炒作，还有一些大数据平台的概念，等等。但很多企业家，心里一直想问的一个问题就是无论是概念、技术还是平台也好，企业跟大数据有多大关系，其实客户想关注的问题就是，大数据到底能给企业带来什么变化。

孙雪老师：我觉得还有一个原因就是由于大数据人才的匮乏，随着大数据概念的热炒，好多从事互联网、IT的技术人员，也看到这个行业的前景，深深介入这个行业，他们都声称自己是大数据行业从业者，但我们知道，这种没有经过专业数据分析实验的技术工程师，他并不能胜任这种工作，那么长此以往，就会造成很多企业，他们期待的大数据价值实现不了，一些不懂大数据的人来做大数据，那么客户是不会买单的。那客户的需求长期得不到满足，就会使他们对大数据丧失信心，所以这个行业就会呈现逐渐“降温”的趋势。

邹东生会长：其实，在2016年随着国家政策的带动，包括很多企业各方面在推大数据的概念，很多企业也关注大数据了，我们作为行业组织，也跟这些企业进行过横向交互，大家很关注的一个问题就是：这么多大数据用语，企业并不明白，他的数据体量不够大，但只要你帮他做分析，让他看到数据带来的价值，他就会觉得大数据对我很有用，我可以更好地使用大数据，那么我就会积攒更多的数据，发挥更大的价值。

杨军先生：所以，我们这边也开始进行和高校一些合作，例如和重庆理工大学建大数据实验室，那么这个实验室就是满足现阶段，去培养一些人才，就重庆人才需求来说，目前没有相关专业，那么人才从哪来，那么这是制约重庆大数据发展的很大的问题。

孙雪老师：我们在和客户对接项目的时候，其实也发现了一个可喜的变化，

就像会长所说，从最初的企业里边没有人意识到数据的价值，需要我们业务人员不断培养大家意识观念，到现在很多企业会主动找到我们，让我们帮助他们解决实际的问题，他们已经将数据作为企业核心发展的竞争力，所以这点转变是让人非常欣慰的。

主持人：看来重庆的数据分析师有很大的缺口啊！

杨军先生：在目前的这种状态下，比如大数据寒冬的时候，我们反而觉得机会出来了，因为看公司的发展方向，我们从最开始的数据服务开始，转向大数据咨询，然后按照大数据建模的方式应用组合，可以给客户带场景的感受，在方式上加以组合，把核心竞争力凸现出来，我觉得这是突破“寒冬”的最好方式。

主持人：嘉宾们谈及的这些现象在行业内的确存在，去年底我了解到：业内部分大数据公司开始大幅裁员或调整，这是不是寒冬将至的现象呢？

杨军先生：行业内部不乏烧钱的公司，靠曝光炒作，虽然一时间名声大噪。据我了解到涉及到具体的业务需求时却交不出完美的答卷，捧场者寥寥无几。

孙雪老师：大数据发展已经上升到国策，很多公司都需要转型；目前市场上开源的技术使得开发成本低，为IT公司、互联网公司转型提供了一定便利，但是专业的数据服务需要的是根据客户需求，结合行业特征不断优化你的算法和模型帮你更精准去解决问题。

邹东生会长：大数据产业链大，我们不是一个点而是要从数据的采集开始，数据清洗，数据库引入，算法模型的改进，并且要带着对行业的认知和理解，对客户的真实需求的理解，为客户提供最直接的解决办法。

主持人：孙老师，我知道犀数科技致力于大数据深度分析，倡导为客户提供完整的数据分析服务，并提出了咨询+技术的解决方案，你们是如何应对现在的问题，或换一个说法，你们是如何把握现在的机遇的？

孙雪老师：犀数是专业做大数据咨询的专业公司，以数据分析为主导，协助客户提供完成的产业服务。我们有大数据平台，我们是免费给行业内的数据分析人才使用的，我们不认为技术是大数据的核心内容，我们是想要在最后一公里为大家提供和客户对接的服务。技术只是辅助，更主要的是数据分析师对客户需求的解读。我们如何运用数据为客户解决问题谈一两个实际案例。凸显出犀数科技的商业价值。

主持人：各位嘉宾聊了这么多，概括起来就是：我们说的寒冬是有指向性的，寒冬仅仅是对不切实际的大数据企业的一种良性淘汰，这种泡沫的挤出才能迎来真正大数据的繁荣！

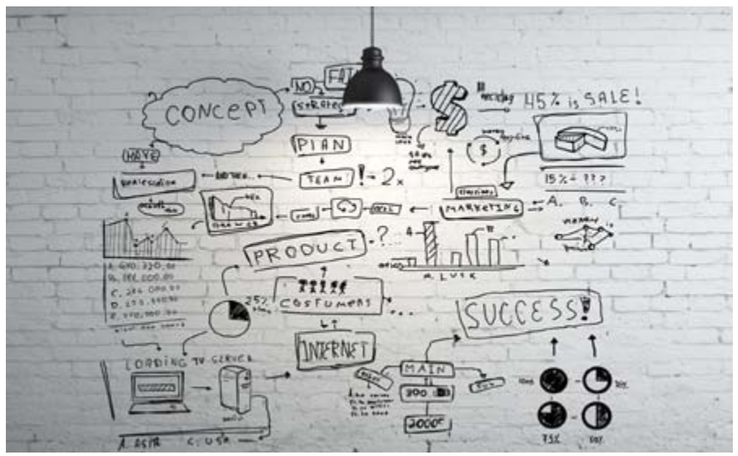
邹东生会长：建议大家要更加注重大数据分析的应用价值，通过自身的专业性满足客户深挖数据价值的需求。大数据要用起来才能真正实现数据价值的回归！当然，主角还是对于专业性人才的需求。

杨军先生：练好内功，不断提高专业性。当整个行业更加趋于理性，对每个数据分析项目的标准更高的时候，那些真正具有市场价值的事务所才会脱颖而出，换句话说，大浪淘沙，只有“实力派”才能熬过寒冬。媒体的宠儿或者穿梭于各个会议论坛的偶像将在这个寒冬亮出真实面孔。具备专业分析能力的事务所是推进整个行业发展的原动力。

事务所要想走的更久远，必须具备对数据的深度分析能力，这个能力能够帮企业提供更精准的决策依据，还要具备利用大数据为帮助企业盈利的能力，这个能力能够最直接的让企业感受到大数据的价值所在，当然事务所的盈利模式和造血能力也是不可或缺的因素。

孙雪老师：在未来的全世界，行业专家和技术专家的光芒都会因为数据分析专家的出现而变暗，因为后者不受旧观念的影响，能够聆听数据发出的声音。大数据革命正在瓦解已经建立起的产业和商业模式，并与各类传统产业加速融合。面对大数据带来的无限商机，我们专业性的数据分析公司更加关注专业能力强的人才，只有不断有新鲜血液的进入才能带来更强大的动力支持。

主持人：感谢各位嘉宾从行业的角度、事务所的发展、专业公司的运维方方面面和大家分享了安然过冬的好建议，大数据应用价值的回归，核心不是冰冷的技术和一堆计算机硬件，也不是开源的算法集合，更多的是大数据的优质研究人才！这让我们更加充满了信心，大数据寒冬来的更猛烈些吧，这样的去芜存菁正是我们各位专业性人才，专业性公司发的拐点！



/ 应用数学博士带你优选数据分析工具 /

编辑 / 协会会员处 李缘　日期 / 2017-01

前言：Excel、SPSS、R、Python、SAS？无论你是数据分析大咖还是不断苦练的菜鸟，面对如此多分析工具能精准找到最佳的吗？Excel的强大功能你了解多少？聚类分析首选SPSS吗？R和Python掌握其中一个就可以了吗？冯艳宾老师带大家迅速了解Excel、SPSS、R、Python、SAS的应用特点，并以移动用户细分实例，向大家展示如何快速掌握SPSS操作，实现客户的聚类分析。下面来听应用数学博士娓娓道来。

主讲老师：冯艳宾

个人简介：副教授，应用数学博士，长期从事数据挖掘、教育测量研究，以及经管运筹、统计、数学建模等课程的讲授，有很强的数据分析和模型构建能力，善于从数据中挖掘规律，并提炼总结成数学模型。曾获得教育部自然科学二等奖、拥有实用新型专利一项以及获得多项教学科研奖励。

大家好！今天我先介绍一下数据分析中的五个软件Excel、SPSS、SAS、R和Python主要特点。目标人群是数据分析的入门用户，因为很多入门用户困惑于选哪一种分析工具来作为数据分析的首选工具。针对这种情况，我把这五种工具的使用情况分别给大家简单介绍一下。

一、Excel

Excel是每个人基本上都会用到的，那对这款软件的功能都是基本了解的，我在这里帮大家罗列了一下和数据分析相关的功能，它可以进行许多的数据处理，以及统计的操作，它应用于管理、统计财经、金融等众多领域，所以它的应用是非常广泛的。

比方说：

（1）数据透视表，一个数据透视表演变出10几种报表，新手只要认真使用

1–2小时就可以进行一些初步分析；

（2）统计分析功能，常用的统计检验很容易实现；

（3）图表功能，常规分析中用到的大部分图都可以实现；

（4）高级筛选，这是Excel提供的高级查询功能，操作简单；

（5）自动汇总功能，这个功能其他程序都有，但是Excel简便灵活；

（6）高级数学计算，却只要一两个函数轻松搞定。

所以，Excel简单易学，且能完成大部分的日常数据处理和传统统计处理任务。

二、SPSS

SPSS是集数据录入、整理、分析、作图功能于一身的软件。SPSS的特点有如下几点：

（1）操作简便，界面非常友好,除了数据录入及部分命令程序等少数输入工作需要键盘键入外,大多数操作可通过鼠标拖曳、点击“菜单”、“按钮”和“对话框”来完成；

（2）功能强大，具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。自带11种类型136个函数。SPSS提供了从简单的统计描述到复杂的多因素统计分析方法,比如数据的探索性分析、统计描述、列联表分析、二维相关、秩相关、偏相关、方差分析、非参数检验、多元回归、生存分析、协方差分析、判别分析、因子分析、聚类分析、非线性回归、Logistic回归等；

（3）针对性强，对于初级人员其易用性很好，操作界面友好，对于高级使用人员可以采用编程方式更加灵活的满足自己的需求。

三、R

R是一套完整的数据处理、计算和制图软件系统。

主要优点如下：

（1）数据存储和处理系统；

（2）数组运算工具（其向量、矩阵运算方面功能尤其强大）；

（3）完整连贯的统计分析工具；

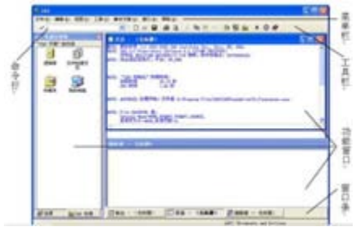
（4）优秀的统计制图功能；

（5）简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。

所以，R的思想就是它可以提供一些集成的统计工具，但更大量的是它提供各种数学计算、统计计算的函数，从而使使用者能灵活机动的进行数据分析，甚至创造出符合需要的新的统计计算方法。而且它是免费的自由软件，可以免费下载和使用的，可以下载各种外挂程序和文档。但对于初学者，入门会花费比较多的时间来了解基础的编程语句和命令格式。

四、SAS

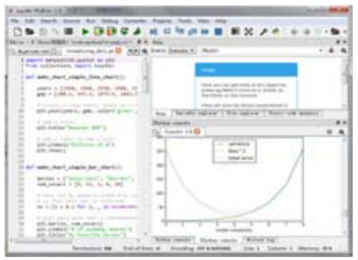
SAS 面对的主要还是大规模的数据，在药企中用的大多数也是SAS 。启动速度来说，SAS 更快，运行速度和运算速度上要比spss快。



SAS主要是针对专业用户进行设计，在编程操作时需要用户最好对所使用的统计方法有较清楚的了解，非统计专业人员掌握起来较为困难。

五、Python

Python是一套比较平衡的高级编程语言。



常用的数据统计包有：

（1）Numpy与Scipy，Numpy封装了基础的矩阵和向量的操作，Scipy提供了各种统计常用的分布和算法；

（2）Matplotlib，主要是用来提供数据可视化的；

（3）Scikit Learn，非常好用的MachineLearning库，封装几乎所有的经典算法，易用性极高；

（4）Python标准库，主要用于处理文本。Python数据统计是主要是通过第三方软件包来实现的，所以对于初学者Python的学习成本还是比较高的。

综合以上几种软件的基本使用特点，可以看到，如果你是一个入门级的数据分析师，而且想快速解决问题，建议用Excel或SPSS更快一些；如果你是一个开发人员，你就可以直接使用R或者Python都可以。

下面，我重点介绍一下SPSS的应用，因为我们还是从解决问题、快速上手为出发点，SPSS应用的界面和快捷的操作，这种角度我们去做市场分析、初步的市场调研等，SPSS更适合初学者做来解决实际业务问题。

我把SPSS常用的方法模块展示给大家：

数据处理分了数据录入、数据转换、缺失值处理等几种情况；描述统计比较强大的功能，分了频率统计、交叉统计、数据探索等几种情况；数据挖掘分了相关分析、对应分析、聚类分析、因子分析等几种数据处理常用情况，当然还有很多……

接下来，通过电信客户分析的案例来演示SPSS在软件处理当中的应用。这个案例的任务是想：通过对电信客户的通话记录数据，实现对客户不同类别的划分。（采用聚类分析和Spss软件操作）

我把原始数据的几种定义罗列了出来，大家先看一下都是什么意思：

变量名称	变量标签	数据类型
Customer_ID	用户编号	字符串
Peak_mins	工作日上班时间电话时长	数值（N）
OffPeak_mins	工作日下班时间电话时长	数值（N）
Weekend_mins	周末电话时长	数值（N）
International_mins	国际电话时长	数值（N）
Total_mins	总通话时长	数值（N）
average_mins	平均每次通话时长	数值（N）

我们的目的是把客户划分为类别，那么分几类？有两种，一种是你事先知道，另一种是你事先不知道，这是业务问题，根据研究调研及经验，我把移动用户分为5个主要消费群体，这是聚类目标。

在做最终的数据划分之前，我们还要进行数据本身的探索，下面是Spss操作过程，我们数据探索主要采用描述统计部分。点击菜单中的【分析】–>【描述统计】–>【描述】，我们可以看到每个客户电话的属性已经都罗列出来，按住【shift】健选中所有属性，添加到右侧的【变量】栏中，最右侧【选项】中，点开之后会看到输出的一些统计量的值，我们就默认上面已经选中的一些包括【均值】、【标准差】、【最大值】、【最小值】就可以–>点击【继续】–>点击【确定】，之后得到的就是6个变量的基础统计值：

描述统计量

	N	极小值	极大值	均值	标准差
工作日上班时间电话时长	498	23.43	2940.4	1066.902	606.3606
工作日下班时间电话时长	498	3.2	896.1	300.0306	189.7828
周末电话时长	498	1.23	339	93.9488	56.96957
国际电话时长	498	0.24	892.56	315.0327	141.37812
总通话时长	498	54.81	3423.3	1420.9414	669.53906
平均每次通话时长	498	0.7	31.19	5.6318	2.81932
有效的 N （列表状态）	498				

根据上表描述统计结果显示，就可以分析出各变量没有缺失值，除平均每次通话时长标准差较小外，其他变量的分布差异较大，所以对数据作之前要进行标准化处理。

那么怎么进行标准化处理？我们还要返回到Spss操作界面，点击菜单中的【分析】–>【描述统计】–>【描述】，然后我们勾选左下角的【将标准化得分另存为变量】–>【确定】，就自动生成标准化变量了，然后原始数据中重新出现了6个标准化之后的变量。

标准化的核心就是消除数据量纲的影响，避免不同量级数据造成的分析误差。

接下来要做的就是离群值和极值分析，更多用箱形图画，具体操作还是通过Spss界面，点击菜单中的【图形】–>【箱图】–>点击【简单】，选中【各个变量的摘要】–>点击【定义】，然后选中标准化之后的6个变量，放到右边的【框的表征】中–>点击【确定】，之后得到的就是箱图。

中间是箱体，两端的点是龙须，龙须之上，或者龙须之下的点还有标星号的点叫离群点，离群点怎么处理？由于我们选取的是局部数据作为示例，因此，离群点在这里我们可以先不处理它，但实际任务当中，我们要特别小心，因为离群点或者极值点都会对我们的分析方法产生影响的，所以我们在这里先不处理它。

我们聚类分析方法的软件操作：点击菜单中的【分析】–>【分类】–>比较常用的【K–均值聚类】，选中标准化之后的6个变量，放到右边的【变量】中，点击【确定】，最终得到不同类别不同的情况分布。

最后就是分析，具体操作：点击菜单中的【分析】–>【分类】–>比较常用的【K–均值聚类】，为了说明方便，我们选中原始变量，选好之后，我们要把所有对象分成5类–>【保存】–>选中【聚类成员】和【与聚类中心的距离】–>【继续】–>【确定】，我们可以看结果了，它把所有输出类别都进行了归纳标记。

类中心点如下，我们用不同颜色对类中心点进行了标注：

	Peak_mins	OffPeak_mins	Weekend_mins	International_mins	Total_mins	average_mins
2	2240.24	392.30	84.80	330.99	2602.20	3.33
3	1416.31	408.92	57.38	380.30	1879.61	3.49
1	1033.35	292.28	56.32	317.75	1381.95	3.54
5	667.62	230.97	51.37	246.10	949.96	4.17
4	52.70	39.38	18.23	15.56	110.31	2.79

我们对类进行特征分析（已排序）：

2类：总通话时间(Total_mins)长，上班通话通话时间(Peak_mins)长，周末通话时间(Weekend_mins)长，国际通话（International_mins）长，命名为高端商用客户；

3类：总通话(Total_mins)较长，下班通话时间（OffPeak_mins）最长，上班通话通话时间(Peak_mins)比较长，命名为中端商用客户；

1类：总通话时间(Total_mins)居中，上班通话通话时间(Peak_mins) 居中，周末通话时间(Weekend_mins)居中，国际通话（International_mins）居中，命名为中端日常客户；

5类：平均每次通话（average_mins）时长最长，命名为长聊客户；

4类：在各项中均较低，命名不常使用客户。

我们最终要把数据分析要落到业务分析上来，解决实际问题。

/ 经济景气指数实证研究 /

文 / 高晨、周余庆　编辑 / 协会会员处 李缘　日期 / 2017-02

摘要：本文以温州市社会经济主要月度统计指标为基础，初步构建了温州市经济景气指数（包含一致指数和先行指数），并利用两者建立了温州市季度GDP预测回归模型，为温州市季度GDP的预测提供了有效的定量工具。

关键词：温州；一致指数；先行指数；GDP；回归分析

经济景气指数是研究经济周期波动、进行国民经济监测预警常用的方法。它采用若干指标编制出的指数作为描述周期波动的主要形式，通过对一系列宏观指标的分析来测定、评价经济波动状况，预测经济循环波动的未来趋势。经济景气指数不仅可以客观描述国民经济运行状况，而且可以预测宏观经济未来的发展趋势[1]。因此，经济景气指数的研究对于国家或地区预判经济形势和及时调整宏观经济政策，具有十分重要的意义。

经济景气指数根据经济指标的非同步变动，把指标分成先行、一致、滞后三种状态，用先行指标反映并预测经济景气变化的未来态势，用一致指标反映并监测经济景气变化的当前形势，用滞后指标进行事后验证并作为修订前一轮政策的依据。鉴于先行指数的预见性以及一致指数的时效性，经济景气指数的研究通常以先行指数和一致指数为主。

目前，已有不少学者对先行指数与一致指数进行了大量的探索，得到了许多有意义的启示。然而，各地区的宏观经济状况与当地的政策、环境、市场行为等因素息息相关，导致了不同地区的宏观经济景气指数很可能相差甚远。本文以温州市为研究对象，研究温州市的宏观经济先行指数和一致指数，判断温州市宏观经济总体趋势，对政府制定经济政策具有较强的指导意义。

一、景气指标的选取

（一）指标预处理

经济景气指数涉及的指标非常多，根据经济上的重要性、统计上的充分性、统计的适时性等原则，结合国内外有关经济景气指数的研究成果确定了18个备选指标。指标数据为月度数据，时间跨度从2006年1月到2011年6月。

对于部分指标各年度1月份数据缺失的情况，采用线性插值法进行预处理。对于每个经济指标的月度时间序列来说，都包含着四种变动要素：长期趋势要素（T）、循环要素（C）、季节变动要素（S）和不规则要素（I），而季节变动要素和不规则要素往往遮盖或混淆了经济发展中的客观变化，给研究和分析经济发展趋势带来困难，因此需要剔除这两种变动要素。本文采用SPSS 17.0软件对指标进行季节性分解，所有数据都是剔除季节变动要素和不规则要素后的趋势循环数据（TC）。

（二）指标的分类

考虑到GDP仅有季度数据，数据量太少，而规模以上工业总产值月度同比增速为基准指标，将其他备选指标与基准指标波动的时间先后顺序来确定该指标的类别（先行或一致）。本文采用时差相关分析法[2]对18个经济指标进行分析与归类，分析结果如表1所示，其中先行指标有4个，一致指标有6个。

先行指标	一致指标
1、M1 同比增速	1、工业总产值同比增速
2、M2 同比增速	2、工业用电量同比增速
3、1/CPI	3、产品销售率同比增速
4、贷款余额累计增速	4、实际利用外资同比增速
	5、工业企业利润累计增速
	6、主要港口货物吞吐量同比增速

表1 温州市先行与一致指标

二、景气指数的编制

（一）景气扩散指数DI

扩散指数DI是（先行、一致或滞后）指标组内第t月扩张（上升）指标个数占组内所采用指标个数的比率（取值在[0，1]之间）。当DI大于50%时，代表经济扩张，反之则代表经济收缩。

下图1和图2分别描绘了2006年2月至2011年6月的温州市先行DI和一致DI。先行DI最近连续16个月均在50%之下，一致DI最近连续4个月均在50%之下，表明温州市经济处于收缩阶段。

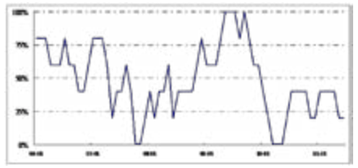


图1 先行扩散指数趋势图

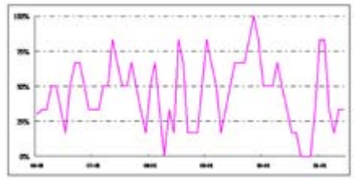


图2 一致扩散指数趋势图

（二）一致合成指数

本文采用目前最常用的是基于增长率循环的计算方法来构建合成指数，这种计算方法的具体步骤见文献[3]。

温州市一致合成指数（2006.01—2011.06）如下图3所示。2006年以来我市经历了2轮经济周期，第一次截止到2008年12月，第二次从2009年1月至今。本轮经济周期已有30个月，波峰出现在去年1月，之后转头向下，尽管指数自去年12月起连续3个月微幅上扬，但今年4月再次下行，表明本次周期的谷底并未来临，经济仍处在下滑阶段。

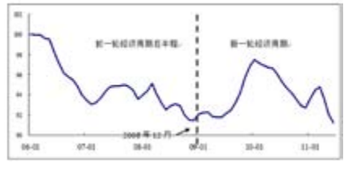


图3 一致合成指数趋势图

三、景气指数的检验

（一）一致合成指数检验

一致合成指数的有效性需要将其走势与经济实际运行情况进行对比，检验其一致性。图4为一致合成指数与GDP的季度累计数据趋势图，一致合成指数与GDP在整体趋势上基本保持一致。对两者进行相关分析，得出两者的相关系数为0.55，并通过了显著性检验。由此可以认为，本文得出的一致合成指数是有效的。

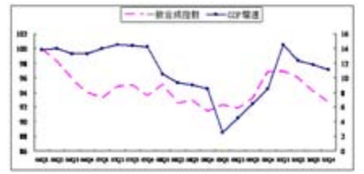


图4 一致合成指数与GDP趋势图

（二）先行合成指数检验

1、先行合成指数与一致合成指数

先行合成指数与一致合成指数的趋势如图5所示。由图5可以看到，先行合成指数的走势比一致合成指数大致领先3-6个月。同时，先行合成指数在2011年度前6个月内一直处于下滑趋势，按照先行指数领先一致指数三个月的判断，温州市在未来3个月内一致指数仍将继续下滑，也就意味着温州市未来三个月的GDP增速将减缓，这与温州市2011年三季度GDP增速减缓的实际基本吻合。

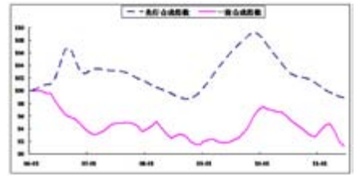


图5 修订后的先行合成指数与一致合成指数趋势图

2、先行合成指数与GDP

图6为先行合成指数与GDP的季度累计数据趋势图，可以看到，先行合成指数领先GDP大约1-2个季度。对超前1个季度和超前2个季度的先行合成指数分别与GDP进行相关分析发现，超前1个季度的相关系数为0.28但没通过显著性检验，而超前2个季度的相关系数为0.56且通过了显著性检验，说明先行合成指数领先GDP大致2个季度。但由图6可知，最近4个季度以来，先行指数的领先期有缩小的迹象。

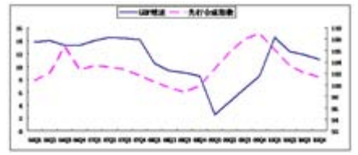


图6 先行合成指数与GDP趋势图

四、基于景气指数的GDP预测模型

（一）基于景气指数的GDP预测模型

1、基于一致合成指数的GDP预测模型

运用SPSS 17.0对两者进行回归分析，得出GDP关于一致合成指数的计算公式如下：

$$GDP(t) = -66.938 + 0.825 \times \text{一致合成指数}(t)$$

其中，(t)表示第t时刻（某季度）的数据。回归模型的F值通过了α为0.05的显著性检验，说明回归方程的可信度比较高，且标准化残差序列基本服从标准正态分布（见图5），且K-S检验表明残差与正态分布无显著性差异，可以认为残差满足线性回归模型的前提要求，但回归模型的R2值仅为0.3，说明回归方程拟合优度不太高。

2、基于先行合成指数的GDP预测模型

运用SPSS 17.0对超前2季度的先行合成指数与GDP进行回归分析，得出GDP关于先行合成指数的计算公式如下：

$$GDP(t) = -58.911 + 0.675 \times \text{先行合成指数}(t-2)$$

其中，(t-2)表示超前t时刻2个季度的数据。回归模型的F值通过了α为0.05的显著性检验，说明回归方程的可信度比较高，且残差序列通过了正态分布检验，可以认为残差满足线性回归模型的前提要求，但回归模型的R2值仅为0.31，说明回归方程拟合优度不太高。

3、基于先行与一致合成指数的GDP混合预测模型

对GDP做先行合成指数与一致合成指数的线性回归分析，得出数学方程如下：

$$GDP(t) = -88.846 + 0.521 \times \text{一致合成指数}(t) + 0.471 \times \text{先行合成指数}(t-2)$$

回归模型的F值通过了α为0.05的显

著性检验，说明回归方程的可信度比较高，且残差序列通过了正态分布检验，可以认为残差满足线性回归模型的前提要求，但回归模型的R2值仅为0.35，说明回归方程拟合优度不太高。

（二）GDP预测模型之比较

现对三个模型的优劣进行比较，图7显示了三个模型的残差图（即预测值与实际值的差距）。由此可见，三个模型的误差整体上基本一致，局部波动较大，难以分出优劣。因此，可综合利用三个模型预测GDP之走势。

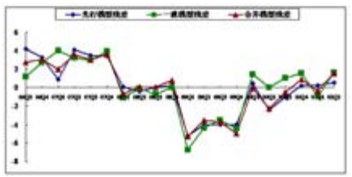


图7 三个模型的残差图

五、小结

通过对2006.01－2011.06温州市宏观经济的实证研究，可以看出本文编制的温州市国民经济综合指数能够较好地刻画我市的经济波动情况，构建的GDP预测模型也能够较好地预测GDP的走势，综合判断经济冷热状态，为经济决策提供参考。

本文研究的温州市经济景气指数系统具有如下几个特点：

一致合成指数与GDP变化趋势基本吻合；先行合成指数预测性较强，早于一致合成指数3个月左右；可综合利用三个回归模型预测GDP的走势。

当然，本文构建的景气指数还存在许多不足之处，一是基准指标的选取并非特别适合，由于GDP和工业增加值均只有季度数据，故只能退而求其次选择规上工业总产值；二是统计数据的时间跨度不够，由于数据缺失等原因，时间序列只能从2006年开始，制约了指标数据长期规律性的挖掘；三是统计方法的局限性，如季节调整并没有考虑五一、国庆等假日因素，经济形式的判定可能产生偏差；四是指标分类具有动态性，由于指标数据要受到政策、环境等的影响，先行指标、一致指标、滞后指标的分类可能会随时间的推移而变化，并非一成不变。因此，需要对指标数据进行定期跟踪，不断完善景气指数。

参考文献

[1] 冯韵, 何跃. 结合景气指数的GDP组合预测模型研究[J]. 统计与决策, 2010年, (20): 19–21.

[2] 徐国祥. 统计指数理论及应用[M]. 北京: 中国统计出版社, 2004.

[3] 张永军. 经济景气计量分析方法与应用研究[M]. 北京: 中国经济出版社, 2007.

/ HBase原理——数据读取流程解析 /

文 / 推酷网　编辑 / 协会会员处 袁硕　日期 / 2017–01

和写流程相比，HBase读数据是一个更加复杂的操作流程，这主要基于两个方面的原因：其一，是因为整个HBase存储引擎基于LSM–Like树实现，因此一次范围查询可能会涉及多个分片、多块缓存甚至多个数据存储文件；其二，是因为HBase中更新操作以及删除操作实现都很简单，更新操作并没有更新原有数据，而是使用时间戳属性实现了多版本。删除操作也并没有真正删除原有数据，只是插入了一条打上”deleted”标签的数据，而真正的数据删除发生在系统异步执行“Major_Compact”的时候。很显然，这种实现套路大大简化了数据更新、删除流程，但是对于数据读取来说却意味着套上了层层枷锁，读取过程需要根据版本进行过滤，同时对已经标记删除的数据也要进行过滤。

Client–Server交互逻辑

运维开发了很长一段时间HBase，那为什么客户端配置文件中没有配置RegionServer的地址信息？这里，客户端与HBase系统的交互阶段主要有如下几个步骤：

1、客户端首先会根据配置文件中zookeeper地址连接zookeeper，并读取/<hbase–rootdir>/meta–region–server节点信息，该节点信息存储HBase元数据“hbase:meta”表所在的RegionServer地址以及访问端口等信息。用户可以通过zookeeper命令“get /<hbase–rootdir>/meta–region–server”查看该节点信息；

2、根据“hbase:meta”所在RegionServer的访问信息，客户端会将该元数据表加载到本地并进行缓存。然后在表中确定待检索rowkey所在的RegionServer信息；

3、根据数据所在RegionServer的访问信息，客户端会向该RegionServer发送真正的数据读取请求。服务器端接收到该请求之后需要进行复杂的处理；

通过上述对客户端以及HBase系统的交互分析，可以基本明确两点：

1、客户端只需要配置zookeeper的访问地址以及根目录，就可以进行正常的读写请求。不需要配置集群的RegionServer地址列表；

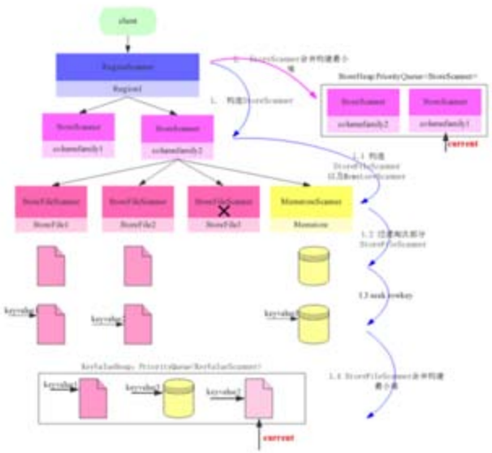
2、客户端会将“hbase:meta”元数据表缓存在本地，因此上述步骤中前两步只会在客户端第一次请求的时候发生，之后所有请求都直接从缓存中加载元数据。如果集群发生某些变化导致“hbase:meta”元数据更改，客户端再根据本地元数据表请求的时候就会发生异常，此时客户端需要重新加载一份最新的元数据表到本地。

RegionServer接收到客户端的get/scan请求之后，先后做了两件事情：构建scanner体系（实际上就是做一些scan前的准备工作），在此体系基础上一行一行检索。举个不太合适但易于理解的例子，scan数据就和开发商盖房一样，也是分成两步：组建施工队体系，明确每个工人的职责；一层一层盖楼。

构建scanner体系 – 组建施工队

scanner体系的核心在于三层scanner：RegionScanner、StoreScanner以及StoreFileScanner。三者是层级的关系，一个RegionScanner由多个StoreScanner构成，一张表由多个列族组成，就有多少个StoreScanner负责该列族的数据扫描。一个StoreScanner又是由多个StoreFileScanner组成。每个Store的数据由内存中的MemStore和磁盘上的StoreFile文件组成，相对应的，StoreScanner对象会雇佣一个MemStoreScanner和N个 StoreFileScanner 来进行实际的数据读取，每个StoreFile文件对应一个 StoreFileScanner， 注意：StoreFileScanner和MemstoreScanner是整个scan的最终执行者 。

对应于建楼项目，一栋楼通常由好几个单元楼构成（每个单元楼对应于一个Store），每个单元楼会请一个监工（StoreScanner）负责该单元楼的建造。而监工一般不做具体的事情，他负责招募很多工人（StoreFileScanner），这些工人才是建楼的主体。下图是整个构建流程图：



1、RegionScanner会根据列族构建StoreScanner，有多少列族就构建多少StoreScanner，用于负责该列族的数据检索。

1.1、构建StoreFileScanner：每个StoreScanner会为当前该Store中每个HFile构造一个StoreFileScanner，用于实际执行对应文件的检索。同时会为对应Memstore构造一个MemstoreScanner，用于执行该Store中Memstore的数据检索。该步骤对应于监工在人才市场招募建楼所需的各种类型工匠。

1.2、过滤淘汰StoreFileScanner：根据Time Range以及RowKey Range对StoreFileScanner以及MemstoreScanner进行过滤，淘汰肯定不存在待检索结果的Scanner。上图中StoreFile3因为检查RowKeyRange不存在待检索Rowkey所以被淘汰。该步骤针对具体的建楼方案，裁撤掉部分不需要的工匠，比如这栋楼不需要地暖安装，对应的工匠就可以撤掉。

1.3、Seek rowkey：所有StoreFileScanner开始做准备工作，在负责的HFile中定位到满足条件的起始Row。工匠也开始准备自己的建造工具，建造材料，找到自己的工作地点，等待一声命下。就像所有重要项目的准备工作都很核心一样，Seek过程（此处略过Lazy Seek优化）也是一个很核心的步骤，它主要包含三步：①定位Block Offset：在Blockcache中读取该HFile的索引树结构，根据索引树检索对应RowKey所在的Block Offset和Block Size；② Load Block：根据BlockOffset首先在BlockCache中查找Data Block，如果不在缓存，再在HFile中加载；③Seek Key：在Data Block内部通过二分查找的方式定位具体的RowKey。

1.4、StoreFileScanner合并构建最小堆：将该Store中所有StoreFileScanner和MemstoreScanner合并形成一个heap（最小堆），所谓heap是一个优先级队列，队列中元素是所有scanner，排序规则按照scanner seek到的keyvalue大小由小到大

进行排序。

这里需要重点关注三个问题：

首先，为什么这些Scanner需要由小到大排序?

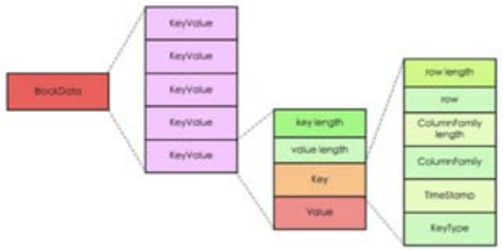
最直接的解释是scan的结果需要由小到大输出给用户，当然，这并不全面，最合理的解释是只有由小到大排序才能使得scan效率最高。举个简单的例子，HBase支持数据多版本，假设用户只想获取最新版本，那只需要将这些数据由最新到最旧进行排序，然后取队首元素返回就可以。那么，如果不排序，就只能遍历所有元素，查看符不符合用户查询条件。这就是排队的意义。工匠们也需要排序，先做地板的排前面，做墙体的次之，最后是做门窗户的。做墙体的内部还需要再排序，做内墙的排前面，做外墙的排后面，这样，假如设计师临时决定不做外墙的话，就可以直接跳过外墙部分工作。很显然，如果不排序的话，是没办法临时做决定的，因为这部分工作已经可能做了。

其次，keyvalue是什么样的结构?

HBase中KeyValue并不是简单的KV数据对，而是一个具有复杂元素的结构体，其中Key由RowKey，ColumnFamily，Qualifier ，TimeStamp，KeyType等多部分组成，Value是一个简单的二进制数据。

Key中元素KeyType表示该KeyValue的类型，取值分别为Put/Delete/Delete Column/Delete Family四种。

KeyValue可以表示为如下图所示：



最后，keyvalue谁大谁小是如何确定的?

上文提到KeyValue中Key由RowKey，ColumnFamily，Qualifier ，TimeStamp，KeyType等5部分组成，HBase设定Key大小首先比较RowKey，RowKey越小Key就越小；RowKey如果相同就看CF，CF越小Key越小；CF如果相同看Qualifier，Qualifier越小Key越小；Qualifier如果相同再看Timestamp，Timestamp越大表示时间越新，对应的Key越小。如果Timestamp还相同，就看KeyType，KeyType按照DeleteFamily –> DeleteColumn –> Delete –> Put 顺序依次对应的Key越来越大。

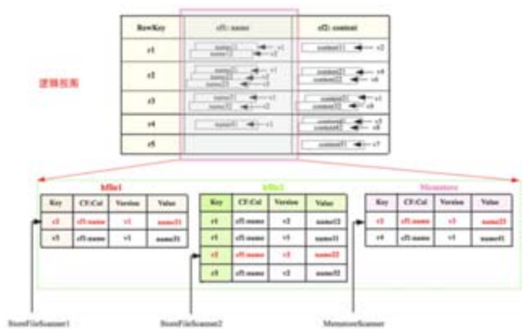
2、StoreScanner合并构建最小堆：上文讨论的是一个监工如何构建自己的工匠师团队以及工匠师如何做准备工作、排序工作。实际上，监工也需要进行排序，比如一单元的监工排前面，二单元的监工排之后……StoreScanner一样，列族小的StoreScanner排前面，列族大的StoreScanner排后面。

scan查询 – 层层建楼

构建Scanner体系是为了更好地执行scan查询，就像组建工匠师团队就是为了盖房子一样。scan查询总是一行一行查询的，先查第一行的所有数据，再查第二行的所有数据，但每一行的查询流程却没有什么本质区别。盖房子也一样，无论是盖8层还是盖18层，都需要一层一层往上盖，而且每一层的盖法并没有什么区别。所以实际上我们只需要关注其中一行数据是如何查询的就可以。

对于一行数据的查询，又可以分解为多个列族的查询，比如RowKey=row1的一行数据查询，首先查询列族1上该行的数据集合，再查询列族2里该行的数据集合。同样是盖第一层房子，先盖一单元的一层，再改二单元的一层，盖完之后才算一层盖完，接着开始盖第二层。所以我们也只需要关注某一行某个列族的数据是如何查询的就可以。

还记得Scanner体系构建的最终结果是一个由StoreFile Scanner 和 MemstoreScanner 组成的heap（最小堆）么，这里就派上用场了。下图是一张表的逻辑视图，该表有两个列族cf1和cf2（我们只关注cf1），cf1只有一个列name，表中有5行数据，其中每个cell基本都有多个版本。cf1的数据假如实际存储在三个区域，memstore中有r2和r4的最新数据，hfile1中是最早的数据。现在需要查询RowKey=r2的数据，按照上文的理论对应的Scanner指向就如图所示：



这三个Scanner组成的heap为<MemstoreScanner，StoreFileScanner2，StoreFileScanner1>，Scanner由小到大排列。查询的时候首先pop出heap的堆顶元素，即

MemstoreScanner，得到keyvalue = r2:cf1:name:v3:name23的数据，拿到这个keyvalue之后，需要进行如下判定：

1、检查该KeyValue的KeyType是否是Deleted/DeletedCol等，如果是就直接忽略该列所有其他版本，跳到下列（列族）；

2、检查该KeyValue的Timestamp是否在用户设定的Timestamp Range范围，如果不在该范围，忽略；

3、检查该KeyValue是否满足用户设置的各种filter过滤器，如果不满足，忽略；

4、检查该KeyValue是否满足用户查询中设定的版本数，比如用户只查询最新版本，则忽略该cell的其他版本；反正如果用户查询所有版本，则还需要查询该cell的其他版本。

现在假设用户查询所有版本而且该keyvalue检查通过，此时当前的堆顶元素需要执行next方法去检索下一个值，并重新组织最小堆。即图中MemstoreScanner将会指向r4，重新组织最小堆之后最小堆将会变为<StoreFileScanner2, StoreFileScanner1, MemstoreScanner>，堆顶元素变为StoreFileScanner2，得到keyvalue＝r2:cf1:name:v2:name22，进行一系列判定，再next，再重新组织最小堆。

/ 大数据集群部署与管理 /

文 / 36G大数据　编辑 / 协会会员处 李缘　日期 / 2017-01

一、大数据集群技术的概述

还记得“啤酒与尿布”的故事吗？在美国沃尔玛连锁超市，人们发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然摆在一起，但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这并非一个笑话，而是一个真实案例。

原来，美国的妇女通常在家照顾孩子，所以她们经常会嘱咐丈夫在下班回家的路上为孩子买尿布，而丈夫在买尿布的同时又会顺手购买自己爱喝的啤酒。这个发现为商家带来了大量的利润，但是如何从浩如烟海却又杂乱无章的数据中，发现啤酒和尿布这个看似不相干的物品销售之间的联系呢？这就是大数据的威力。

大数据在我们的生活中，发挥着越来越明显的作用。比如，大数据辅助购物平台推荐适合客户的产品，大数据辅助避免堵车，大数据辅助做健康检查，大数据娱乐等。对于很多公司来说，数据是有的，但是”死”数据，并不能发挥作用，或者产生的价值不到实际价值的冰山一角。如果想从大数据中获利，数据的采集、挖掘和分析等环节缺一不可，其中，大数据分析技术是重中之重，目前的大数据分析技术有 Hadoop、Spark、Strom 中。

要想从一大堆看似杂乱无章的数据中总结出规律，需要对这些数据进行一番非常复杂的计算分析。由于数据量之大，对计算的速度和精度要求都比较高，单纯的通过不断增加处理器的数量来增强单个计算机的计算能力已经达不到预想的效果，那么，大数据处理的方向逐渐的朝着分布式的计算集群来发展，将分布在不同空间的计算机通过网络相互连接组成一个有机的集群，然后将需要处理的大量数据分散到这个集群中，交由分散系统内的计算机组，同时计算，最后将这些计算结果合并得到最终的结果。

尽管分散系统内的单个计算机的计算能力不强，但是由于每个计算机只计算一部分数据，而且是多台计算机同时计算，所以就分散系统而言，处理数据的速度会远高于单个计算机。

那么如何部署和管理大数据集群，则是业界持续讨论的话题，这里以IBM Platform Converge为例，来阐述大数据集群部署、架构以及管理。

IBM Platform Converge是一种复杂的大数据处理平台(方案)，此方案可以从若干个物理机/虚拟机(可能在云端)开始，可以比较方便的部署一个大数据集群，并且管理和监控此集群。此平台包括了若干个大数据技术和集群技术，比如 xCAT、Spark、ELK、GPFS等。此集群的优点是节点的数量和存储的空间都具有弹性，也就是说，可以随时根据业务和应用的需求，来增加或者删除集群中的节点和存储空间，依次来节省成本。

二、大数据集群技术的架构与分析

一般来说，大数据集群的构架，主要分为几层：硬件层、OS 层、基础设施管理层、文件系统层、大数据集群技术层以及上层应用，如下图大数据集群的架构所示：

首先最下层为硬件，这些硬件可能为不同的厂商机器，比如 IBM、HP、DELL 或者联想等服务器，也有可能包括

不同的构架，比如 System X 或者 IBM POWER 等机器。这些机器有可能在机房，也有可能在云端(包括公有云和私有云)。硬件之上，需要安装运行操作系统(OS)，一般为 Linux OS，比如 Redhat、SUSE、Ubuntu 等。

在基础设施管理层，主要管理资源(更多的是软件资源)以及资源的虚拟化等，比如网络资源/设备、计算资源、内存、Slots 等的统一管理和优化分配，在此层，同时肩负着部署大型 Cluster 的任务，也就是将各个分散的节点通过 IBM SCF(Spectrum Cluster Foundation)软件，统一部署为一个整体。

在IBM SCF集群中，分为管理节点和计算节点。部署的顺序为，需要首先安装管理节点，然后按照不同的硬件、网络、OS 等配置集，来部署出计算节点。为了提高集群的鲁棒性，IBM SCF 本身支持高可用性(HA)，在安装完管理节点之后，使用类似的方法，来部署出备份的管理节点。

并行的文件系统，是大数据集群的重中之重，因为大数据有两个主要的特征，其一是数据量比较大，起步可能就以 PB 为单位，如此巨量数据的存储成为了集群需要解决的关键问题之一;另外一个特征是处理速度要快，随着集群技术的发展，并行化的思想尤为明显，并行化的计算产品和工具也层出不穷。

所以，并行的文件系统是大数据集群中不可或缺的一部分。比如，在 Hadoop 时代，HDFS 就在 Hadoop 阵营中，贡献了中流砥柱的作用。另外一个出色的并行文件系统为 IBM Spectrum Scale，其前身为 IBM GPFS，经过近来的版本迭代和发展，已经完美的支持目前流行的大数据计算模式，比如 Spark 等。

在资源管理和大数据集群层，主要部署两方面的组件，一是大数据分析处理组件，二是资源调度和管理组件。

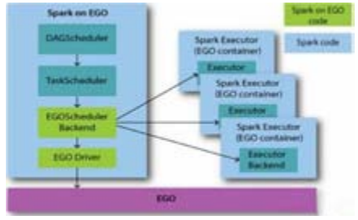
在一般情况下，这二者都是有机的结合在一起，组成一个产品。随着大数据的发展，大数据的分析和处理技术如井喷一般涌现出来。比如有 Hadoop, Spark, Storm, Dremel/Drill 等大数据解决方案争先恐后的展现出来，需要说明的是，这里所有的方案不是一种技术，而是数种，甚至数十种技术的组合。

就拿 Hadoop 来说，Hadoop 只是带头大哥，后面的关键的小弟还有：MapReduce, HDFS, Hive, Hbase, Pig, ZooKeeper 等等，大有”八仙过海，各显神通”的气势和场面。资源调度管理，主要是维护、分配、管理、监控软硬件资源，包括节点、网络资源、CPU、内存等，根据数据处理的需求来分配资源，并负责回收。

在此模型中，我们使用了 Spark 来处理大数据，使用 IBM Platform Enterprise Grid Orchestrator (EGO)来管理和监控资源。IBM Platform 是一种资源管理和调度、服务管理的工具。类似于大家熟知的 Mesos 或者 Yarn。

由于IBM Platform EGO目前并非开源产品，在此做一简单介绍。在 IBM Platform EGO 中，VEM kernel daemon (VEMKD)是 VEM 内核后台程序，一般运行在管理节点上，会启用其它后台程序并对分配请求做出响应。EGO Service Controller (EGOSC)属于 EGO 服务控制器，负责向 VEMKD 申请相应资源并控制服务实例。流程处理管理器(简称 PEM)负责 VEMKD 中的启用、控制以及监控活动，同时收集并发送运行时资源的使用情况。EGO 中 Consumer，代表的是能够从集群处申请资源的一个实体。

单一 Consumer 可以是业务服务、包含多种业务服务的复杂业务流程、单一用户或者一整条业务线。和开源的 Spark 一样，Spark 和 EGO 使用相同的 DAGScheduler 和 TaskScheduler，如下 Spark on EGO 构架图所示：



EGOSchedulerBackend 根据 Taskscheduler 提供的 task 和 task stage 等信息，负责从目前的 EGO 框架中获得资源。用户可以自定义资源分配方案，通过 Consumer 来分配资源。EGOSchedulerBackend 一旦获得资源，就可以通过 EGO Container 接口开始运行 Spark Executor。

EGOSchedulerBackend 监控 Spark Executor 运行的生命周期，以及资源使用情况和 task 状态等，比如当 task 完成时，EGOSchedulerBackend 触发调度逻辑来满足更多资源的获取或者资源的释放。

最上层为应用和业务，客户只需要提交 Spark Application 即可，集群负责统一的管理和调度，并返回执行结果。

三、大数据集群的部署

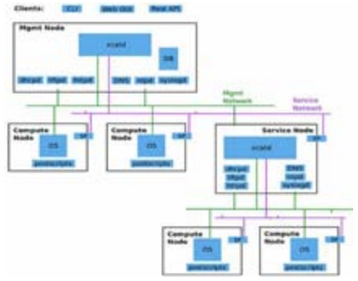
3.1 、硬件的部署

在此集群部署中，借助了比较成熟的硬件部署工具 Extreme Cloud Administration Toolkit (xCAT), xCAT 是一个开源的集群管理工具，能用于裸机部署，其架构如xCAT 构架图所示。

xCAT 可以自动发现硬件，开机之后，可以由 xCAT 从裸机自动引导安装，当然，也可以提前导入 client node 信息，xCAT 可以基于 IPMI 进行远程硬件控制，

如开关机，如收集 CPU 的温度等状态信息，支持 X86_64、POWER、System Z 等硬件类型。

支持的目前所有主流的操作系统，如 RHEL,CentOS, Fedora, Ubuntu, AIX, Windows, SLES, Debian 等。xCAT 各个组件的结构和流程如下图所示。在 xCAT 部署的集群中，主要有三种 Node: 管理节点(Management Node)、服务节点(Service Node)、计算节点(Compute Node)，如果并非特别大的集群，一般情况下，服务会被省略掉，只有管理节点和计算节点。管理节点上启动 DHCPD、tftpd、httpd、DNS、ntpd、syslogd、DB 等服务。

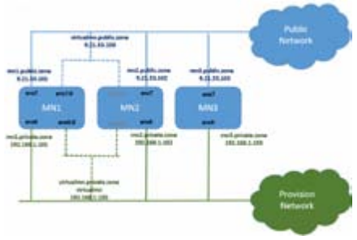


3.2 、软件的部署

软件部署主要在集群已经建立完成的基础上，并行在各个节点上安装大数据分析处理系统，在”资源管理和大数据集群”层，部署 Spark Cluster，并和 Platform EGO 深度集成，一些管理和监控等方面的程序也相继安装。还有，在提交应用之前，需要先创建 SIG(Spark Instance Group)，并启动 SIG，在创建 SIG 之后，也为 Platform EGO 来管理和控制其相关的服务。

3.3 、高可用性(HA)部署

在 IBM Platform Converge 中，高性能部署的构架如下图所示。通常有三个节点构成，分别为主管理节点 Management Node 1(MN1)、次管理节点 Management Node 2(MN2)和第三管理节点 Management Node 3(MN3)。但是需要说明的是，在 failover 切换的过程中，必须保证 MN1 和 MN2 其中一个健在，因为 MN3 只是负责 IBM Spectrum Scale 的 HA 过程，主要的服务和进程只运行在 MN1 和 MN2 上，在这二者之间进行切换。如下高可用性部署图所示：



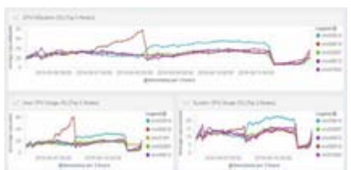
四、大数据集群的管理与监控

在大数据集群中，管理和维护是一件非常麻烦的事情，有可能会出现各种各样的问题，如果出现了，最好的办法是分析 LOG 和监控，在运维过程中，管理员需要不时的查看监控，并善于从监控中找到问题，及时的分析和解决 Cluster 中的报警(Alert)。以下展示了基本的 Cluster 的监控指标，比如 CPU、内存、存储资源、网络等。

在此集群中，监控主要采用的是 ELK 的日志监控分析系统，大致流程为，有 Beats 来收集日志和数据，然后发给 Logstash 来分析和处理日志再由 Elasticsearch 存储和检索，最后由 Kibana 来在 Web GUI 页面上展示出来。接下来，我们展示出几个方面的集群的监控。

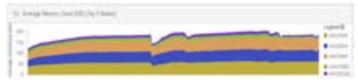
4.1 、CPU 的监控

下面的CPU 监控图展示了 Spark 集群中的 CPU 利用率的监控。如果 Spark 集群中的节点可能较多，可以使用 Kibana 的功能，来展示出 CPU 利用率最高的几个节点(图中展示的是 5 个节点的情况)，以便了解哪些节点的负载较重，当然也可以展示出整个系统平均的负载情况。



4.2 、内存的监控

众所周知，Spark 是一种内存利用率非常高的技术，换句话说，Spark 集群对内存的要求较高。Spark 集群的管理者需要实时的掌握内存的使用情况。如下图内存监控所示，展示出了集群中内存占用率比较高的节点的情况。



4.3 、磁盘和文件系统的监控

磁盘监控图展示了总体磁盘的个数，有问题磁盘的个数，和总体磁盘的使用率，对磁盘利用率的监控可以有效的防止因存储空间不够而影响应用的运行。

近几年来，数据的价值正得到越来越多的人的重视，如何让数据”活起来”，一直是 IT 界持续讨论的话题，在这种利益的驱动下，大数据的分析技术可谓是”遍地开花”，大数据集群的部署方案也层出不穷，针对不同的场景和不同的需求，各大 IT 公司都在争先恐后的提出各种各样的方案和技术。

如何选择合适的方案，主要可以从技术选题、稳定问题、高可用性、可扩展性、监控等方面入手。IBM Platform 致力于大数据的分析和部署的研究工作，从以上几个方面来看，IBM Platform Converge 是较为出色的大数据集群部署解决方案。

/ 在大型金融组织选择大数据和数据科学技术 /

文 / 推酷网　编辑 / 协会会员处 李缘　日期 / 2017-02

数据科学正快速成为各行各业开发人员和管理人员的关键技能，同时它似乎也非常有趣。但它也相当复杂——有太多的工程分析技术，你很难知道自己做得是否正确或者哪里存在陷阱。在该系列文章中，我们将探讨如何利用数据科学——从已经采用并成功实施数据科学的人们那里，了解数据科学的适用场景，以及如何让它成为你的资产。

企业组织现在越来越多地采用数据科学和高级分析技术，也越来越多地影响着决策、产品和服务。因此经常有人问到：数据科学最好的工具集是什么？从表面上看，这个问题似乎是关于技术之间的比较。结果你可能需要审阅一长串关于R、Spark ML及其相关技术（如：Jupyter或Zeppelin）的利弊列表。然而，对企业组织而言，首要问题是什么功能能够支持其未来的业务目标。关注这些可以让技术选型变得更容易，并且降低浪费时间和精力的风险。

我们如何才能达成共识，以务实和富有成效的方式进行有关技术选型的讨论？在这篇文章中，我们通过实际案例来探讨什么才是合适的框架。对企业组织来说，最典型的切入点是那些大量存在的数据孤岛（silos）和过度采用的技术。你不想仅仅因为利益相关者的要求而增加更多的技术和数据孤岛。新的技术和基础设施应取代现有技术并替换数据孤岛。但在现今的大环境下要做到这点并不容易，因为传统分析技术和商业智能供应商声称他们拥有针对新挑战的解决方案，同时还有大量的新技术出现，其中许多是开源的，这提供了更多的选择。新技术通常都宣称能取代传统工具，并提供传统工具无法企及的功能。而传统技术则反驳说它们能提供更好的企业品质，比如安全和支持。

我们在这里讨论的现实案例中的客户在一年多前与我的雇主联系，他们在短期和长期的战略需求方面面临着巨大挑战。这家FTSE 100公司正处于其生命周期中的转型时刻。它的整个组织结构发生了显著变化，需要重新改造其部分现有平台，因为它分裂的组织结构和依赖项不可维护，无法创造商业价值。在我们来看，客户的迫切需求是：在极短的期限内，用一种完全透明的方式，混合集成历史数据，解决高级报告和新数据平台分析技术所面临的问题。客户现有的数据仓库技术基于应用技术，十分昂贵且有局限性。如果不投入大量资金并且增添新兴的分析功能，新的报告和高级分析功能执行起来会极其缓慢甚至无法执行。

成本和局限性是重点关注对象。我们的客户意识到由于可预见的突破性技术变革，市场竞争正变得越来越激烈，从长远来看，源于核心业务活动的价值将不可避免地缩减。企业组织的领导者意识到他们迫切需要开发新的功能，以便在处理完当前的紧急需求后立即为企业的未来发展做好准备。

我们与主要利益相关者合作制定了一个计划，将主要数据集集中在一个中心区域，便于在企业未来的新一轮变革中灵活处理和分析。值得注意的是，我们并没有放弃核心数据仓库，只是把它还原到原先的角色。然而，我们仍然会逐步淘汰大量的旧系统，这些系统大多数存有数据并且难以访问。同时，要保证数据在不同平台上正常流动，以确保监管和安全。我们因此把高级分析技术和数据科学技术问题延后讨论。这是可行的，因为新平台可以在必要时根据需要采用那些相关技术。采用这种方法给客户带来的好处是显而易见的。未来的业务仍在不断变化，而眼前的业务需求需要马上得到解决。将决策和实施分阶段实施，且不阻碍平台的创新，这是一个双赢的解决方案。

第一个教训是避免在跟不上需求变化的技术上加倍投入。此外，尤为重要的是不要进行一对一的技术匹配。比如不要用一种相似的技术替换原有技术，这样做得到的效益十分有限。我们要考量这些技术给组织带来的成本支出和它们所能为组织提供的功能。大家总是希望借由更少更便宜的技术来降低成本，并指望它们能提供更多业务功能。理想情况是我们可以两者兼顾。在这个案例中，我们在淘汰旧系统的同时减少了数据仓库占用的空间，节省下来的资源可用于新的分析技术平台，这反过来取代了一些原有功能并增加了相关的新功能。

有了这个概念，我们就可以专注于我们正在努力实现的目标。现在的企业和以前的企业所面临的挑战是相同的。他们必须降低成本，提高盈利能力，不断改进以保持合规，并且在这个被服务自动化和商品化所驱动的环境里，可能还需要重新定义其核心业务。例如，过去几年中，数据和对数据的有效利用正在成为应对这些挑战的关键机会。

问题在于大多数企业组织不知道该如何寻求答案甚至不知道问题出在哪里。在各个业务领域内通常都有一些唾手可得的短期机会，它们将给现状带来完全可预期的改进。但大多数利益相关者已经习惯于自身的局限性，他们需要打破这种局限。当问及他们想要实现什么时，他们要么把思考局限在企业组织现有的功能范围内，要么为了解决未来的未知需求而要求那些不切实际的东西。

因此那些包括重新定位自身核心业务在内的长期基础性需求通常很难甚至无法得到满足。所以第二个教训是不要着眼于办不到的事情上，不要试图去预测未来，而是应该对眼下出现的需求灵活以对。在我们的案例中，你可以看到我们在不限制条件或不返工的情况下，为平台将来的迭代扩展留下空间。这是通过规划多个增建（buildout）步骤做到的。可以在合适的时机往这些步骤里添加一系列的功能。这里从诸多的功能中列出其中的两项，比如流处理功能或键值存储（key values stores）功能。

然而，如果我们完全以技术为驱动，指望使用各种技术来取代事后的内部反思（inward reflection）和需求收集，这是有风险的。我们可能最终采用了没有任何商业目的或价值的技术，导致高额的成本和高度复杂性，更糟糕的情况是导致项目完全失败。大数据和数据科学的流行促使利益相关者在这种情况下容易陷入炒作陷阱。他们认为采用技术可以解决业务目标、功能和需求方面的问题。对利益相关者来说，至关重要的是必须在大数据和数据科学方面提出正确的问题，以避免困惑和失望。这些问题是先决条件，包括具体的战略业务目标和需求。虽然战略目标必须从一开始就明确，但是如我们的案例所示，需求可以随着时间反复推导。

企业组织可以使用适当的大数据战略来评估当前形势，明确需求，并采用有关数据存储、处理和分析的新功能。事实上，这种敏捷性是以数据为驱动的现代组织的基础，它让企业能够在快速发展的技术环境中良好运作。数据科学可以利用组织在评估和采用这些技术方面所具备的能力。数据科学还为来自两方面的挑战提出了深入的见解，并给出了恰当的解决方案。这两方面的挑战一个是更多、更快、更多元化的数据，另一个是人们对这些数据在驱动产品、服务、洞察力和决策方面无限增长的期望。

在我们的案例中，传统的数据仓库解决方案正面临着挑战，因为它在单独完成第一个任务时，缺乏足够的灵活性来解决任何未知需求。不过这种解决方案也不是一无是处，因为这项特定业务在金融行业中运作，带有敏感数据并且受到高度监管。这项业务需要得到更深入的挖掘，而这又必须允许众多数据科学家和商业用户访问数据。大多数企业组织都存在这种矛盾，既要让所有潜在消费者都能接触到所有数据，但同时要确保数据的安全，不被滥用或泄漏。

对政府、医疗保健和金融客户来说，他们还得经受得住新闻媒体的考验，因为任何数据安全方面的问题，不管是真实发生了抑或是有发生的迹象，都可能成为灾难性的新闻头条。因此，安全问题不仅存在于现实中，也存在于意识中。有趣的是，这也是为什么许多客户对云技术犹豫不决的原因，因为在云技术里，随着安全的改进，感知和现实越来越互相偏离。有些公司可能要顾虑合规性，比如在哪里存储数据。另外，云服务供应商把越来越多的区域纳入监管范围来满足合规需求。

我们的客户选择了使用本地部署方案，我们为他们列出了解决当前问题需要的关键性功能，并为他们设计了一个将来可灵活扩展的平台。首要目标是构建一个平台，这个平台以Hadoop及其生态系统为核心，获取新旧数据，使用掩码和加密确保数据安全，然后基于这些数据生成报告。该方案所需的分析工具很简单，通常会利用SQL接口把那些遗留工具接入Hadoop生态系统，并使用Apache Hive。Hive是第一选择，因为它是整个分布式系统不可分割的一部分，它稳定而且对SQL支持良好，遗留系统可以通过标准连接访问它，它还跟分布式安全模型紧密集成。此外，第一阶段的性能要求与用于分析和报告的大小批次的数据更为相关。

核心平台的构建和集成，以及必要的PCI合规性，是现阶段的关键挑战。由于时间紧迫，我们必须立即开展工作，所有利益相关者都很乐意通过“快速失败”（fail fast）这个手段，对平台关键要素的落地实施进行验证，以迅速找到组织性阻碍和技术限制。自然而然地，只有当所有发现的问题都得到了解决，“快速失败”才是有效的。因此，无论是否能够达到某个里程碑，我们都需要在工作中举办一些研讨会，比如学习一些新的知识、引入新的业务，让技术利益相关者参与进来，一起解决问题或者为下一步的发展制订计划。

虽然有时也会遭遇困难，但是这种方法在高层领导支持下会比较有效。现有的流程和技术以及已建立合作的供应商可能需要

—— 《中国数据分析》 2017年 第一期 运数有道 ——

被作为解决方案的一部分进行评估。有时候这会导致与供应商和企业利益相关者在如实处理失败情况时对话困难，无论问题是来自于组织自身还是来自供应商和合作伙伴。高层利益相关者要强势进行战略审查和问题分析，因为身处数据驱动的发展最前沿，他们也是少数几个应该负责找出问题根源的人。这是唯一可行的建设性合作方法。因此必须让利益相关者加入研讨会并倾听他们的需求和进程，能在概念验证环境下进行反复验证，进而探讨各种可行或不可行的方法，这是极其重要的，这才能使我们迅速在工作上获得进展。

对敏感数据加密工具和屏蔽工具的选择是快速失败的一个很好的例子。一个有名的市场参与者推出了他们的解决方案，并坚称他们在金融方面的成功案例让他们成为客户的第一选择。然而事实证明，市场已经远离了他们。同时，Hadoop生态系统的新功能，比如透明数据加密与多租户模式的结合，对他们的产品和安全机制来说改变太大，无法适用。快速失败的良好运作以及在概念验证环境中引入新供应商的能力让延迟变得可控，并且这项选择工作在新一轮对另一提供商的评估之后取得了进展。

随着第一阶段的工作即将完成，整个组织的需求增加了，比如，访问平台和数据，增加工具以便更好地支持数据科学家和高级业务分析师。这些需求涵盖了探索性分析、几近实时的高级报告以及智能应用和产品。满足这些需求需要许多功能和工具。此外，许多数据科学家偏好不同的工具，包括R，Python（scikit-learn），Spark ML（使用Python，Scala或Java），以及各种商业解决方案和笔记工具（比如Jupyter或Zeppelin）。还有很多还不是很明确的初步需求和偏好，需要跟能够达成它们的工具进行匹配。我们还要注意监管、安全性、业务持续性、软件和数据集开发生命周期以及成本、复杂性和风险等这些常被忽略的问题。简而言之，组织要么在低风险的情况下以一种及时且可盈利的方式持续创新，要么被技术淹没。

创新灵活性太高和肆意采用技术会带来风险，使组织瘫痪。组织里的数据可能由于缺乏监管和安全性不足而泄漏或质量下降。当企业组织需要支持太多技术时，可能会导致资源缺乏和集成不可控。另一方面，紧密而简约且只考虑安全性的技术选型将会扼杀组织创新，造成人才流失、功能缺失，组织将最终发现自己无法应对新的机会和风险。另一种与上述完全不同的理念是通过漫长的瀑布迭代流程来制订完美解决方案。这种理念在无法收集需求、技术能力不断改变的创新环境下不占优势。

当我们将组织设想为一个拥有有限资源并旨在从中获得最大相关功能的实体时，敏捷式方法将成为最佳选择。其发展框架类似于我们用来评估技术选型和解决核心平台开发和构建时所出现问题的研讨会。我们可以将相关业务部门的各种数据科学和分析技术的利益相关者汇聚到一起进行讨论。什么是易于理解的用例？它们的优先等级和对组织的影响是什么？实施它们需要具备哪些条件？还有不太为人所了解的未来创新理念和潜在的功能需求？第二部分是技术问题。团队的技术偏好和现有技能是什么？对于各种必须得到满足的要求和组织标准，它们在开发生命周期方面有什么样的需求？理想情况下，技术问题能得到来自安全、基础设施、运营以及软件开发等部门的利益相关者的支持。

我们的客户比较先进，已经有显著的独立性，因为它的一些重要高层领导是大数据和分析技术专家。然而，他们也希望得到外部支持，得到同一领域专家的独立指导和评估。对于顾问而言，当客户接受你作为权威和值得信赖的独立顾问，这是梦寐以求的结果。我们一起举办了一个为数据科学工作做准备的研讨会。我们收集了各类信息，并且在研讨会期间，我们就能够做到对各个业务部门的工作按照优先级排序，并淘汰不合适的技术。

练习的效果是立竿见影的。所有利益相关者都互相认识，了解彼此的愿望和喜好，这本身就是有价值的。此外，基于几近实时流数据，我们还能够识别重要工作和决策服务。这可以为各方所用，也就是说，每个人在某些情况下都需要用到这类服务。我们能够避免类同开发，集中精力并将其作为试点项目优先安排。在缺乏监管的状态下，会出现不同业务部门使用不同的技术开发出同一个服务的不同版本。而采用上述的方法，我们就能够整合精力和工具选择。

我们的下一步计划是选择第一组要添加到数据科学工作平台的技术，特别是用于流数据的Spark ML、Java、Python以及Kafka。这些技术引入了现有用户案例所需的功能，并且还将涵盖一些未来和次要的用例。这个选择是在研讨会讨论最终候选技术并且考虑了运营和组织方面的问题之后做出的。例如，我们需要确定哪些技术受到最为广泛的支持和采用，并且最为成熟。是否得到广泛采用是在我们在这个阶段选择Java而不是Scala的一个影响因素。

重要的是不要排除任何可能性，并让利益相关者参与建设性讨论。如果备选方案看起来不可行，我们可以通过上述框架来降低它们的优先级。

我们即将参与服务的开发。可预见的好处是这为组织带来了一系列技术及其功能。我们会立即在关键业务项目中评估其非功能性能力，例如，围绕安全性、可靠性和性能来评估。此外，如果能证明这些技术有效和可用，它们可能被业务利益相关者采纳，减少对重叠替代方案的需求。有了正确的选择和成功的表现，持续采用更多技术的需求将逐渐淡去，而采用现有成熟可用的解决方案将变得越来越普遍。

我们将来的计划是继续使用该框架，并收集利益相关者和用户反馈，以便在现有功能不足的情况下进行评估和进一步采用技术。随后的研讨会将自然地从广泛的技术选型讨论转移到维护问题的讨论，最终我们将讨论在市场不断发展的情况下逐步淘汰技术的话题。

/ 厦门诚晟数据分析师事务所 /

文 / 厦门诚晟数据分析师事务所　编辑 / 协会会员处 李缘　日期 / 2017-02

厦门诚晟数据分析师事务所有限公司是经厦门市工商局批准注册，经中国商业联合会数据分析委员会（团证：第111号）严格考察后批准入会的专业数据分析机构，公司于2016年正式成立。

事务所成立以来依托推进厦门经济发展的良好时机以及厦门经济特区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，这就是我们公司未来的发展目标。

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：厦门火炬高新区火炬园火炬东路20-24号火炬新天地5号楼301B
联 系 人：吴加强、刘贵兰
联系方式：13850086686（吴加强）、13606082939（刘贵兰）、0592-5612966

新起点
大不同

CPDA®

www.chinacpda.com 400-050-6600