中国数据分析

CHINA DATA ANALYSIS 数据分析·因你而不凡

——中国数据分析行业核心刊物——

《中国数据分析》行业特刊
2021年第02期 总第46期(季刊)
咨询热线：400-050-6600
http://www.chinacpda.org

中国商业联合会数据分析专业委员会 主办

庆祝建党百年 初心如磐 奋勇前进

“度之往事，验之来事，参之平素，可则决之。”历史车轮滚滚向前。一百年来，中国共产党人勇担历史重任，始终坚持马克思主义基本原理，深刻把握历史发展规律，制定正确发展战略，作出科学决策。我们党的一百年，是矢志践行初心使命的一百年，是筚路蓝缕奠基立业的一百年，是创造辉煌开辟未来的一百年。在百年接续奋斗中，党团结带领人民开辟了伟大道路，建立了伟大功业，铸就了伟大精神，积累宝贵经验，创造了中华民族发展史、人类社会进步史上令人刮目相看的奇迹。

当今世界正经历百年未有之大变局，我国正处于实现中华民族伟大复兴关键时期，我们还会遇到各种各样的风险考验甚至是难以想象的惊涛骇浪，各行各业都必须站在党和国家工作全局看自身，立足新发展阶段，贯彻新发展理念，服务和融入新发展格局，落实好高质量发展的国家重大战略。因此，作为数据分析行业，努力推动数字产业化和产业数字化，加快数据中国建设是我们责无旁贷的使命。

同样，百年大计，教育为本。在我国开启全面建设社会主义现代化国家新征程之际，党和国家事业发展对教育的需要，对科学知识和优秀人才的需要，比以往任何时候都更为迫切。青年人才是祖国的未来、民族的希望，也是我们党的未来和希望。心怀“国之大者”，为服务国家富强、民族复兴、人民幸福贡献力量。协会自 2008 年成立至今，带着对大数据行业的热爱，本着“用数据说话、做理性决策”的信念，始终倡导大数据的应用与分析是执业的核心价值，而我们对数据分析师认证及执业管理的科学化以及促进数据分析师事务所组织化、规范化、标准化工作的推进，始终持之以恒。十多年来，更是以为党和国家培养卓越的大数据复合型人才为己任，坚持不懈，不忘初心，勇于创新。为了推动数据分析人才培养朝更高效、更优质的方向变革，今年我们成功完成了第九次课程改革，勠力同心，砥砺前行，力求因材施教，使应用价值最大化，培养人才更贴近国家发展用人标准，更符合数据分析行业人才发展标准。与此同时，我们正在组织的行业标准化制定工作，从国家发展战略、市场应用等多维度，凝聚广大行业专家和企业的真知灼见，在探索中不断深入，力求促进数据分析行业的健康发展。

“道固远，笃行可至；事虽巨，坚为必成。”今天，在庆祝我们党百年华诞的重大时刻，在“两个一百年”奋斗目标历史交汇的关键节点，初心如磐、锚定目标、奋勇前进。在党和国家领导下，永远保持谦虚谨慎、不骄不躁，永远保持艰苦奋斗，勇于创新，以新的理念、新的思考和新的探索，合力开启数据分析行业高质量发展的崭新篇章，在数据化变革中传播实用、真实、健康、发展的大数据理念和专业技能，挖掘大数据的实际应用价值，为中国的大数据行业奉献更多的力量，向建党百年献礼。

中国商业联合会数据分析专业委员会

本期目录 CONTENTS

卷首语
01 庆祝建党百年 初心如磐 奋勇前进

协会动态
03 探索融合发展——落实党建工作进程
04 中国数据分析行业《会员执业资质证书》改版升级啦!
05 如何打破发展瓶颈 高效推动业务突破
07 公布入选行业 " 标杆事务所会员单位 " 名单
08 数据分析行业创业指导活动

政策向导
09 以大数据助推新时代党建工作
12 安徽省大数据发展条例 5 月 1 日起施行
13 北京国际大数据交易所成立 探索数字经济新产业新模式
14 大数据驱动大未来
15 国家发改委等部门联合印发方案
—— 大数据中心布局有新调整

行业动态
17 Gartner 公布 2021 年十大数据和分析技术趋势
19 云计算、大数据与人工智能三者的关系
27 舆情大数据的社会科学应用
29 未来时代大数据技术与企业决策相伴而生
30 如何做好数据分析?

学数交流
32 栅格地图在高纬度区域误差过大及修正方法
34 我是如何在教培行业精进我的数据分析能力的
37 大数据分析工具 BI 的应用
39 AWS 宣布推出财务数据分析工具 Amazon FinSpace
40 商品零售购物篮分析实战案例：Apriori 关联规则算法
47 分享 | B 端运营需要关注数据指标
51 全球企业 TOP150 数字化转型成功要领!

事务所专栏
55 泓睿数据：企业数据化管理作用有多大?

主办单位
中国商业联合会数据分析专业委员会

编委成员
会员处 李苗苗

出版时间
2021 年 06 月出版 < 总第 46 期 >

美工设计
市场处 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 : 400-050-6600 / 010-59000991 转 652
传真 : 010-59000991 转 607

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

探索融合发展——落实党建工作进程

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

2021 年 3 月，中国商业联合会组织召开分支机构工作会议，会议中对 2021 年度服务工作提出新的要求。 为庆祝中国共产党百年华诞，昂扬姿态、拼搏奋进 ，具备条件的分支机构特别是行业性分支机构要在中国商联党委的领导下，发挥行业优势独立开展相关活动。

在党的十九大报告就提出了“网络强国”、“数字中国”和“智慧社会”的概念，大数据已上升为国家战略，数据成为新时代下政府最重要的资产。为推动政策落地，促进大数据产业发展，全国各地都制定了大数据发展政策扶持。教育部也公布了《2016 年普通高等学校本科专业备案和审批结果的通知》，继 2016 年北京大学、中南大学、对外经贸大学首批设立大数据相关学科后，中国人民大学、北京邮电大学、复旦大学等 32 所高校成为第二批成功申请“数据科学与大数据技术”本科新专业的高校。

技术发展催生下的新兴学科和专业，该怎样培养人才？培养什么样的人才？人才培养与学科研究又该如何处理定制与创新引领、交叉融合与专业建设的关系？是现今面临的问题，同时又亟待解决的问题。 中国商业联合会数据分析专业委员会，在中国商联党委的领导下，加强了党的建设、贯彻党的领导、践行“两个维护”，把握自身行业的属性，从国家对行业的要求中，制定规范的发展计划，在大数据分析人才培养方面，为国家输送更多社会急需的具备大数据处理及分析能力的高级复合型优秀人才，践行数据强国发展战略。

随着数字化转型进程的加快，政府需要搭建大数据平台，完成数据的整合；企业需要数据分析进行有利决策，提升竞争力；数据分析师需要一个合适的舞台，展现自己的价值。作为中国商业联合会数据分析专业委员会的会员单位——数据分析师事务所也凸显出自身的价值，在疫情常态化控制时代 , 践行以“大数据思维”助力政务数字化发展、企业数字化转型、教育数字化应用 , 围绕数据的深度分析、业务场景构建、深层次的咨询等 , 以大数据思维帮助企业、事业单位实现数字化转型并提供行之有效的战略决策。

在党建工作中，中国商业联合会数据分析专业委员会认真学习贯彻落实党中央工作部署，深刻理解行业利益、企业利益、会员利益与人民整体利益是局部和全局的关系，既要充分反映行业投诉，也要引导行业和企业做到围绕中心，服务大局。

在企业融合发展方面，中国商业联合会数据分析专业委员会明确了，全面建设社会主义现代化新征程，立足大数据人才培养发展的新阶段，贯彻新发展理念，构建新发展格局，落实落细国家纲要部署，推动大数据产业高质量发展，扎实推动行业和平发展、融合发展，坚定推进国家大数据行业的新进程。

中国商业联合会数据分析专业委员会，将进一步践行，细化行业服务、建设更专业的能力，强化协会整体意识，加强部门和机构的组织联系，人员交往、信息交流、业务合作、资源共享等优势，充分发挥行业协会大平台的作用。

中国数据分析行业《会员执业资质证书》改版升级啦！

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

中国商业联合会数据分析专业委员会（以下称“我会”）作为中国数据分析行业的管理协会，为提升数据分析行业影响力、进一步规范行业从业行为，提升我会会员服务，经我会研究决定，对我会颁发的《会员执业资质证书》自 2021 年起进行全面升级。新版资质证书增加了一定功能性，强调了资质证书与线上电子证书查询的联动，优化会员年检办理流程，方便广大会员进行保存和展示。

那么，相较旧版证书，新版证书有哪些更新呢？

01、新版资质证书更加方便社会各界在中国数据分析行业官方网站 （www.chinacpda.org）快速查询会员电子证书，以及其从业状况、资质有效期与年检情况。电子证书与纸质版证书信息一致，具有同等效力。

02、新版证书的出台，更加提升会员服务，优化年检办理流程。新版资质证书比旧版证书更完善且具设计感。每年年检，旧版证书须完成手动贴标，操作较为繁琐，而新版资质证书上印有二维码，只须扫描二维码即可在线进行查询年检情况。无须贴标。使查询更具便捷性。

03、新版证书的线上查询资质功能，有效杜绝社会上伪造我会颁发的资质情况，从而更加规范市场。无论是会员单位还是客户，都可以通过扫描新版资质证书上的二维码，进入到我会官网进行在线查询。我会会员每年须按时按要求完成年检，只有会员单位且会籍在有效期内，才能在网站查询到信息。

随着国家大数据战略的实施，这二年随着广大政府和企业对大数据的重视，数据会越来越多，技术门槛会越来越低，但是大量的数据进行深层次的分析就成为企业竞争的核心、成为企业大数据变现的核心，而我们的数据分析师事务所，正是在这样的背景下不断发展和壮大，助力社会的数字化转型。

我们希望这次资质的改版升级，能够让市场看到我们数据分析行业更加规范，从而使我们的数据分析师事务所在良性竞争的环境下发挥更多的专业性，为社会服务。

如果想进一步了解数据分析师事务所，可联系我会会员处咨询：

4 月 11 日 CPDA 数据分析师北京公益沙龙活动
——如何打破发展瓶颈 高效推动业务突破

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

面对日益调整的 KPI，你是否为用户增长而发愁？

各种思路、模型聊起来都会，接到项目却无从下手？

后疫情时代，数据分析的应用场景里智慧生态如何发力？

金融行业如何预测客户还款能力？跳槽旺季 HR 怎么分析到岗率？

……

4 月 11 日，后疫情时代北京地区首场公益沙龙活动精彩揭幕！来自互联网、金融、房地产、医疗、联合办公等行业的 CPDA 数据分析师和广大数据分析爱好者们共聚一堂，协会邀请到京东集团户增长中心 - 运营经理何正松，智慧城市专家、大数据科研专家赵迅多以及协会特聘讲师赵玉莲，和大家分享如何跳出舒适圈，用大数据思维快速找到破题思路，高效推动业务突破。

本次活动亮点

京东师兄何正松带来职场经验分享：

高标准 ROI 下，用户增长破题思路及如何运用 CPDA 所学，科学评估用户生命周期促进指标飙升。

数据分析如何助力智慧生态规划：

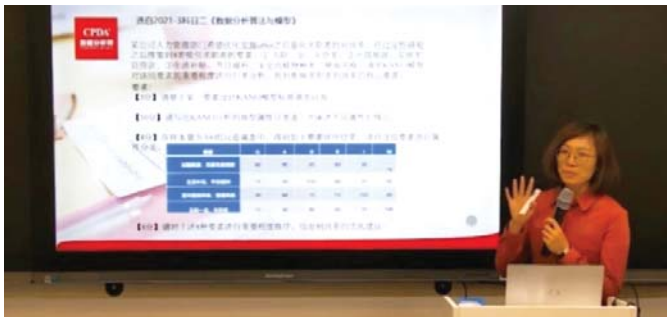
赵迅多老师解读市政联动的智慧生态园区如何以数据分析为驱动力，引入先进技术场景落地打造智慧生态系统。

协会特聘讲师赵玉莲解析考题难题：

赵玉莲老师深度剖析考试难题，分享思路和技巧，同时引入应用场景，学以致用：

1、金融机构如何构造客户质量的预测型，预测客户还款能力？

2、金三银四应聘旺季，HR 如何利用预测模型分析到岗率？



此次公益活动既有实操经验的分享，也有前沿规划的解读，大家纷纷表示此行收获颇丰，受益匪浅。期待下期活动能够早日举办！

……

欢迎 CPDA 数据分析师和广大对数据分析感兴趣的人群加入 Club，行业协会将集聚中国数据分析行业从业人员，整合大数据资源，助力来自全国各地的不同行业背景、多样分析能力的数据人才，共同找寻大数据实操宝典，呈现更优秀的数据思维、彰显数据应用的巨大价值。

何正松 (CPDA 学员)

京东集团，用户增长中心 - 线上用户运营经理

7 年互联网数据分析及大数据建模项目经验；4 年产品设计及运营经验。

擅长领域：互联网用户运营分群模型，用户路径链路分析，Leads(销售线索转化) 评级模型等；

曾就职于搜狐网、易车网、蓝色光标、京东等互联网企业，服务 B 端企业包括国家电网、奥迪、JEEP、欧莱雅、美赞臣等国内外知名企业。

赵迅多

慧创科技（北京）有限公司 智慧城市专家、大数据科研专家。

十余年丰富实战经验，主导智慧生态、智慧园区项目落地。推动市政联动，引入先进技术促进数据分析成果在园博会场景中的应用。

赵玉莲

中国商业联合会数据分析专业委员会高级顾问、专家组成员、CPDA 数据分析师导师，组织并参与 CPDA 数据分析师第八次课程体系改革工作。具有多年数据分析课程研发经验与教育培训经验。

公布入选行业＂标杆事务所会员单位＂名单

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

为了更加促进行业发展，协会今年在全国范围内，制定扶持重点会员单位的行业政策，这一举措将选拔具有代表性的事务所会员成为标杆，通过对重点地区重点企业会员的扶持，帮助会员在经营理念、发展方向、宣传推广等方面都有更好的提升，从而达到树立榜样，以点带面促进更多企业会员健康发展。

2021 年 5 月中旬，协会通过对申请标杆会员单位进行审核与筛选，从数据分析师团队的研究能力、事务所本身具备的发展水平，以及从业经验等多方面进行考量，最终选举出两家个会员单位，并在行业内进行公示。

数据分析行业创业指导活动——5 月 12 日首站北京专场、5 月 19 日山东专场、5 月 26 日上海专场

来源 / 中国商业联合会数据分析专业委员会　编辑 / 协会会员处　李苗苗　日期 / 2021-01

十四五规划中 7 次提及“数字经济”，24 次提及“数字化”，并明确将数字经济独立成篇，描绘出未来 5 年数字中国建设的崭新蓝图。作为推动经济社会转型升级、培育经济增长新动能和构筑国际竞争新优势的重要途径，数字经济将是“十四五”时期经济社会发展的重要推动力。

在此背景下，中国商业联合会数据分析专业委员会作为数据分析行业协会，将发挥出更强劲的发展动能，为大数据人才提供更加广阔的发展空间，在大数据应用价值不断凸显的今天，数据分析师事务所作为一个全新的第三方服务行业，围绕数据的深度分析、业务场景构建，为国内企业数字化转型产生了巨大推力，在企业成长的各个阶段起到重要作用。协会通过在全国各地开展创业指导活动，帮助数据分析师更加具象的了解数据分析师事务所，认识到在面对巨大的市场需求和数据分析行业未来发展不断精细化的背景下，创业成立数据分析师事务所巨大的商业价值。

5 月，连续三周协会分别在北京、山东、上海数据分析师社群举办创业指导线上活动。每次活动均吸引百余名数据分析师参与其中。他们都是来自各个行业的数据精英翘楚，借此次活动“云”聚一堂，共同探讨，从大数据行业热点，到行业创业方向，大家踊跃交流想法，积极请教事务所创业环境、发展前景等相关问题……

数据人永远对未来，对创新持有一种热情。

三期活动圆满完成，参与活动的数据分析师分别表示受益匪浅。之后，协会还将在全国各地社群举办创业指导活动。

5 月 12 日，北京创业指导活动“现场”：

5 月 19 日，山东创业指导活动“现场”：

5 月 26 日上海群专场：

以大数据助推新时代党建工作

来源 / 来源：《红旗文稿》　编辑 / 协会会员处　李苗苗　日期 / 2021-01

2019 年 7 月 9 日，习近平总书记在中央和国家机关党的建设工作会议上的讲话中指出：“要紧扣机关党建时代特点和党员思想行为特征开展工作，积极探索有利于破解难题的新途径新办法，积极探索信息化条件下开展工作的新载体新路数。”当前，随着信息技术的迅猛发展，以大数据为代表的信息技术深刻改变着人们的思想认识、行为方式，充分利用好大数据的优势，努力提高党建工作的广度、深度、精度，对探索新形势下党建工作新路径具有重要意义。

大数据对推进新时代党建工作的重要意义

《中共中央关于加强党的政治建设的意见》中提出，积极运用互联网、大数据等新兴技术，创新党组织活动内容方式，推进“智慧党建”。处于信息化大潮中，我们要探索大数据在创新党建工作中的特殊作用。

大数据是全面落实新时代党的建设总要求的重要推力。

中国共产党的领导优势和组织优势是做好各项工作，战胜各种困难和风险的重要保证。当今世界正经历百年未有之大变局，世界进入动荡变革期，我们党面临的困难和挑战不同以往。面对新形势新挑战，如何全面落实新时代党的建设总要求，做到管党有方、治党有力、建党有效，把我们党建设得更好、建设得更强，大数据运用在信息资源、传播范围、工作效率等方面均为基层党建工作提供了有利契机。大数据已经成为世界各国取得信息化发展的有力武器，大数据的便捷性、时效性、延展性、包容性，决定了其在今天全面落实新时代党的建设总要求中具有不可忽视的作用。

大数据是提升党建科学化、精准化、标准化的应有之义。

大数据技术很重要的一个特点就是将对象数据化。党建工作可以利用大数据将党组织建设、党员教育管理服务等工作网络化、智能化，提供规范便捷的党务管理、创新多样的组织活动、一站式党员服务、多角度学习宣传等，实现网络发展到哪里，党建工作就覆盖到哪里。大数据还能够使集中学习和组织生活突破时间和空间的限制，为党员学习教育提供实时化、高质量、大容量素材，为党员学习成果检测提供数字化支撑。可以通过对数据的研究和分析，掌握所有党员的思想水平、工作表现、行为习惯、党性修养等信息，升级和优化队伍建设模式，使队伍建设从粗放式管理转变为精细化管理模式。同时，在流动党员管理、领导干部考核、党代表履职情况评估、党的基层组织建设等方面，大数据的价值也可以进一步挖掘。

大数据是加强对党员领导干部日常监督的有效手段。

坚持党的领导，必须从严管党治党，“从严”很重要的一个方面就是强化对干部的监督管理。加强对党员领导干部的日常管理监督，尽早发现苗头性问题，把纪律挺在前面，把全面从严治党向末端延伸、向深处推进。从当前的情况来看，在精准地加强对干部、对权力的日常监督，仍然存在着许多不足，而大数据广泛研究运用为我们完善干部日常监督管理提供了解决方法和路径。在互联网迅猛发展的当下，推进传统监督手段与现代信息手段深度融合，探索运用大数据思维和信息化手段，强化综合分析研判，推动干部监督由单一化向立体化转变，使日常监督更有操作性、实效性，是织密权力运行的“笼子”和加强对领导干部日常监督的有效手段。坚持把党的建设与信息技术的运用结合起来，可以全面、深入、具体了解党员干部信

息和党建中存在的“隐藏”问题，干部日常管理工作不易被发现的小问题更有可能在数据分析中暴露出来，进而可以第一时间作出预警反应，给干部“咬耳朵”“扯袖子”，将干部监督关口前移，把问题解决在初始之时、萌芽状态。

运用大数据推进党建工作的路径

运用大数据推进党建工作是一项系统工程，只有坚持政治引领、用好技术手段、强化队伍建设，才能更好地发挥这一新兴基础战略性资源的作用，不断提高党建工作质量。

政治引领是前提

大数据是一把双刃剑，在实现数据安全融合与数据交易流通带来的巨大价值同时，数据大规模使用也不可避免地带来了潜在风险，存在公开内容引发舆论风波的风险，还存在数据泄露、丢失、滥用等风险。在充分运用大数据推动党的建设过程中，强化政治引领是前提。我们要深化对于党建中大数据应用的认识，进一步明确大数据助推大党建的根本要求，始终坚持正确的政治立场、政治方向、政治原则，从创新党建工作形式入手，为党组织和党员群体提供优质高效的信息和技术服务，为提升党组织的组织力、凝聚力、战斗力服务。要运用马克思主义哲学原理辩证看待大数据与党建工作结合中的利与弊，实现科技与党建的相互促进，将互联网这个最大变量，变为推动党建工作的最大增量。

技术手段是关键

数据本身并没有意义，数据只有经过处理解释后才有意义。从海量的数据出发，经过收集、整理、加工、分析等一系列技术处理，深度提取有效信息并运用到具体目标，才体现数据本身的价值。从这个意义上看，大数据的内涵更在于处理这些大规模数据的思维、工具及手段等能力。对于新时代党的建设工作而言，要重视先进技术、思想理念的融入，打破常规工作模式，推进党务工作模式向信息化、智能化方向发展。特别是需要打破传统“就党建抓党建”思维，以大党建思维为导向，统筹各个职能部门，依托大数据、物联网、VR 和 AR、云计算等先进的技术手段，以数据和硬件为基础，以传统党建工作方式和智慧党建新模式的有效结合为抓手，以强化政治功能、突出服务功能、体现创新功能为目标，构建一个高效、畅通、安全的党建信息化管理平台。

队伍建设是基础

习近平总书记强调，善于获取数据、分析数据、运用数据，是领导干部做好工作的基本功；强调要打造多层次、多类型的大数据人才队伍。在移动智能时代，人才在推动党建工作转型与创新中发挥着重要作用，不仅要具有较强的党建工作业务能力，还需要具有较高的信息技术运用能力，能够有效地运用大数据技术、物联网技术、云计算技术开展党务工作，加强党的思想引领。目前，从事党建工作的人才队伍，总体上年龄偏大，对新技术的运用不熟练，有的甚至存在排斥心理。这就需要遴选政治素质好、网络技术强同时又善于做党建工作和思想政治工作的中青年人才，充实到党建工作队伍中来。还可以在智能知识、信息安全维护和应用能力等方面开展党务工作人员培训工作，打造党建大数据工作队伍，统一部署指导各地区的党建大数据融合工作，并结合工作落实效果作出反馈和评价，以推动党建大数据融合工作的稳步前进。

运用大数据推进党建工作的保障措施

运用大数据推进党建工作是极具创新的工作，需要各方面的支持和保障。这就要求我们必须树立用互联网思维谋划党建、用信息化技术推进党建的战略思维，明确统领机构、出台建设标准、健全各项机制，确保运用大数据推进党建工作取得实效。

加强统筹谋划

习近平总书记明确指出，各级党委要高度重视信息化发展对党的建设的影响，做到网络发展到哪里党的工作就覆盖到哪里，充分运用信息技术改进党员教育管理、提高群众工作水平，加强网络舆论的正面引导。运用大数据推进党建工作的建设是一项系统工程，没有各方面的配合，很难推行下去，需要构建起党委统一领导，职能部门分工负责、齐抓共管的智慧党建领导体系。各级党组织必须从战略和全局的高度认识推进“大数据 + 党建”的重要性和紧迫性，不断增强网络党建的自觉性和坚定性。相关职能部门统领制定“党建 + 大数据”有关政策文件、平台设计，统筹协调和督促检查工作，同时强化成员单位及有关机构的职责任务，构建党委政府统一领导、有关部门各司其职的“大数据 + 党建”发展格局。党员领导干部具有带头示范作用，需要自觉树立新理念、学习新技术、运用新媒体，带头参与到“大数据 + 党建”的各项实践中来，积极学网、懂网、用网，真正成为智慧党建工作的带头人、引路人。

出台建设标准

目前，运用大数据推进党建工作处在探索阶段，没有统一的建设标准，各地都有自己的特色和侧重点，这不利于规范运行，不利于高质量发展。我们需要紧紧围绕新时代党建工作要求和工作特点，制定包含网站、软件、APP、微信公众平台、二维码应用等一套“大数据 + 党建”工作的技术标准和协议，明确各项业务数据的交互标准，定义各级党组织进行交换的数据规范，并积极推动上升为统一标准。同时，从党员教育、管理和服务三方面，制定统一的“大数据 + 党建”业务流程，如发展党员工作流程、“三会一课”流程、谈心谈话流程。

健全各项机制

我们需要加快建立数据收集、分析、归类整理和使用的管理机制，解决运用大数据推进党建项目多头领导、平台重复设计、数据标准不一等问题，实现指挥统一、责权明确、运用到人。需要建立健全个人信息安全法规、党员数据信息使用监管机制，加强对个人隐私和数据安全的法律保护。需要建立健全传播工作机制，实现党组织间跨地域、跨行业间的知识资源共享，通过新技术手段实现关键技术要素的学习和传递。需要建立健全经费保障制度，推进“大数据 + 党建”基础设施建设。需要有针对性地制定相应的绩效考核标准，周密部署各级党组织平台建设工作，对建设成效进行评估，让“大数据 + 党建”新模式尽快落地，发挥实效。

安徽省大数据发展条例 5 月 1 日起施行

来源 / 来源：安徽日报　编辑 / 协会会员处　李苗苗　日期 / 2021-01

4 月 30 日，记者从安徽省政府新闻办举行的新闻发布会上获悉，《安徽省大数据发展条例》于 5 月 1 日起施行。《条例》就数据资源管理、大数据开发应用、促进大数据领域技术和产业发展、数据安全管理等作出系统规定。

数据资源实行统筹管理。《条例》明确，县级以上政府数据资源主管部门应统筹协调、督促指导本行政区域内数据资源管理工作，推动政务数据等公共数据归集，推进数据资源汇聚融合、共享开放、有效流动和开发应用。公共数据按规定向江淮大数据中心平台归集，实现公共数据资源的汇聚和集中存储。

省政府数据资源主管部门建设和运行管理江淮大数据中心总平台，省有关部门、单位建设和运行管理江淮大数据中心分平台，设区的市政府数据资源主管部门统筹所辖县、市、区建设和运行管理江淮大数据中心子平台。省、设区的市政府数据资源主管部门依托江淮大数据中心总平台或子平台，统筹建设本行政区域公共数据共享交换平台、开放平台。

《条例》规定，安徽省、设区的市政府应推动建设大数据产业发展集聚区、数字经济创新发展试验区、线上经济创新发展试验区，建设大数据重点实验室等技术创新平台。各级政府和有关部门应按有关规定依托“皖事通办”平台，推进政务服务事项一网通办、全程网办，开发大数据应用场景，促进政务服务跨地区、跨部门、跨层级数据共享和业务协同，提升政务服务规范化、便利化、智慧化水平。县级以上政府及其有关部门应运用“互联网 + 监管”系统，汇聚整合、关联分析监管执法和市场领域数据资源，支持事中事后监管和服务，提高监管和服务的效能。省、设区的市政府应用大数据赋能城市治理，统筹建设城市大脑，推动城市管理和服务数字化、智能化、智慧化。

促进大数据发展和保障数据安全方面，《条例》规定，安徽省政府统筹财政资金，整合信息化、电子政务等专项资金，设立省大数据中心专项资金 。省政府数据资源主管部门会同有关部门统筹大数据交易服务机构的设立，搭建数据要素交易平台，建立数据产权交易机制，推动建立行业自律机制。网信部门、公安部门、通信管理部门和其他有关部门在各自职责范围内落实国家数据安全审查制度，建立健全数据安全保护体系。

北京国际大数据交易所成立 探索数字经济新产业新模式

来源 / 来源：中国新闻网　编辑 / 协会会员处　李苗苗　日期 / 2021-01

为推动数据要素市场化配置和数字经济高质量发展，助力推进首都“两区”建设，3 月 31 日，北京国际大数据交易所成立，北京数据交易系统上线。北京金控集团联合 48 家单位，共同发起成立北京国际数据交易联盟，并发布北京数据交易服务指南。

在北京市人民政府的大力推动下，北京市经济和信息化局会同北京市金融局、北京市商务局、北京市委网信办等部门，组织北京金控集团牵头发起成立北京国际大数据交易有限公司 (北京国际大数据交易所或北数所)。这是国内首家基于“数据可用不可见，用途可控可计量”新型交易范式的数据交易所，定位于国内打造领先的数据交易基础设施和国际重要的数据跨境流通枢纽。

加快数据开放流通，建设全球数字经济标杆城市

当前，数据已经成为社会的核心经济资源和基本生产要素。作为数据要素安全保护、市场化配置、价值生成、跨境传输的基础设施，数据交易所的重要性不言而喻。

北京国际大数据交易所是北京市落实建设“国家服务业扩大开放综合示范区”和“中国 (北京) 自由贸易试验区”数字经济领域的重点项目，是北京市创建“全球数字经济标杆城市”的重要内容。

据介绍，北数所将在保护个人隐私和信息安全的前提下，发挥技术密集和数据密集的双轮驱动，以技术促流通，以流通促创新，探索数字经济的新产业、新业态、新模式。

“当前，北京市正在构建多层级、安全、负责任的数据交易体系。北数所将以培育数据交易市场、释放数据要素价值为核心，打造立足京津冀、辐射带动全国、面向全球提供服务的金融科技基础设施。”殷勇在致辞中强调。

突出“一新三特色”，探索破解数据交易难题

数据交易课题由来已久，然而不同于土地、劳动力、资本等生产要素，数据有成本极低、再生性强、难以建立排他性等特点，权属界定不清、要素流转无序、定价机制缺失、安全保护不足等问题，一直是掣肘数据要素高效配置的痛点。如何解决这些问题，是决定北数所未来发展的关键所在。

据介绍，北数所将突出“一新三特色”，在推进底层技术创新的基础上，以数据使用价值为基本交易对象，探索特色模式、特色规则、特色生态，打造全国数据交易探索的新样板。

创新技术为支撑。依托北京在隐私计算、区块链等领域的技术先发优势，将数据要素解构为可见的“具体信息”和可用的“计算价值”，对其中“计算价值”进行确权、存证、交易，实现数据流通的“可用不可见、可控可计量”，为数据供

需双方提供可信的数据融合计算环境。

特色模式为引领。准入方面，北数所将实行实名注册的会员制，对数据来源进行合规审核，对数据交易行为进行规范管理。管理方面，实行数据分级分类管理，创新免费开放、授权调用、共同建模、加密计算等多种融合使用模式。流转方面，探索从数据、算法定价到收益分配的涵盖数据交易全生命周期的价格体系，形成覆盖数据全产业链的数据确权框架。产业链延伸方面，培育数据来源合规审查、数据资产定价、争议仲裁等中介机构，推动产业链创新发展。

特色规则为保障。今天，北数所正式发布了《北京数据交易服务指南》，囊括架构、方式、机制、安全等数据交易服务细则。同时，还将探索建立大数据资产评估定价、交易规则、标准合约等政策体系，积极推动数据创新融通应用纳入到“监管沙盒”；构建数据交易市场风险防控体系，建立数据安全备案机制和数据市场安全风险预警机制，强化关键领域数字基础设施安全保障。

特色生态为延展。与北数所同时成立的，还有北京国际数据交易联盟，包括大型商业银行、电信运营商、头部互联网企业以及数据中介服务等 50 多家机构或企业参与其中。这是打造北数所交易生态的重要部分。

此外，北京还将积极探索跨境数据安全流通，积极吸引跨国企业和国际机构加入，构建立足中国、面向国际的国家级数据资源流通生态体系。

国际货币基金组织原副总裁、中国人民银行原副行长，清华大学国家金融研究院院长、北京国际数据交易联盟理事长朱民指出，在数据使用过程中，存在着技术、法律、经济、监管等一系列挑战，需要我们共同来扩大“朋友圈”、拓展交易场景，探讨交易模式、探索交易规则、推动跨界流转，从而把中国海量的数据转换成真正的资源和生产力，成为价值创造的巨大源泉。这正是成立数据交易平台和数据交易联盟的重大意义所在。

描绘数据流通新图景，打造数字经济时代新基建

仪式当天，北京数据交易系统成功上线。这是基于区块链和隐私计算技术支持的全链条交易服务体系，将为市场参与者提供数据清洗、供需撮合、法律咨询、价值评估、权属认证等一系列专业化服务。

北京金控集团党委书记、董事长范文仲称，北数所将依托北京自贸区国际商务服务片区的优势资源，建立集数据登记、评估、共享、交易、应用、服务于一体的数据流通机制，打造国内领先的数据交易基础设施和国际重要的数据跨境流通枢纽。

大数据驱动大未来

来源 / 人民日报　编辑 / 协会会员处　李苗苗　日期 / 2021-01

用好大数据这个时代赋予我们的强大引擎，就能抓住新机遇、培育新动能、塑造新优势，推动中国经济在数字化大潮中乘风破浪，驶向高质量发展的美好未来

催生新业态、畅通产业链，让万千企业点“数”成金，大数据是经济高质量发展的推动力；数据多跑路、百姓少跑腿，让“一网通办”“一次办好”成为常态，大数据是优化营商环境、提升服务效能的“加速器”；动态反映经济社会各指标发展趋势，多维度多层面反映政策落地效果，让社会管理更加精细、更具智慧，大数据是提升社会治理体系和治理能力现代化的“金钥匙”……

伴随数字经济快速发展，数据这座巨大“宝藏”正显示出前所未有的使用价值和发展潜力。疫情防控期间，交通、通信、消费、税收、金融等众多领域的数据有机整合，为疫情监测、防控救治、资源调配等提供了有效指引，助力疫情防控和复工复产，成为大数据促进经济社会发展的一次成功实践。“健康宝”等大数据产品的广泛应用，更是对全社会进行了一次生动的数据科普，彰显了大数据作为基础性战略资源的重要意义。

人类社会发展的历史经验表明，每一次经济形态的重大变革，往往催生并依赖新的生产要素。正如劳动力和土地是农业经济时代主要的生产要素，资本和技术是工业经济时代重要的生产要素，进入数字经济时代，数据正逐渐成为驱动经济社会发展的新的生产要素。去年发布的《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》将数据列为五大生产要素之一，强调要加快培育数据要素市场。将数据作为生产要素，表明随着经济活动数字化转型加快，数据对提高生产效率的乘数作用凸显。这也是最具时代特征的生产要素的重

要变化。作为全球数字经济发展较为领先的国家之一，我国要高度重视数据这一新型生产要素的重要价值，加快将大数据转化为现实生产力，为经济社会发展注入新动能、增添新活力。

用好大数据，开放共享是基础。大数据的价值，首先体现在“大”。数据的积累是一个从量变到质变的过程，把数据的开发、利用局限于某个单位、某个行业、某个领域时，就会形成信息孤岛，难以发挥有效价值。实现全社会数据信息的有序有效开放，要建立“部门间”数据共享、“政企间”数据开放、“企企间”数据融通的数据要素流通公共服务体系。

用好大数据，高效配置是关键。数据作为生产要素，就要按照市场化原则进行配置。在社会主义市场经济条件下，要充分发挥市场配置资源的决定性作用，畅通数据要素流动渠道，保障不同市场主体平等获取大数据，按照市场规则、通过市场竞争实现效益最大化和效率最优化。与此同时，要更好发挥政府作用，完善政府调节与监管，创造公平的市场竞争环境，确保数据要素自由流动、安全可控、公平竞争。

用好大数据，依法规范是保障。大数据是把“双刃剑”，在信息技术不断升级的背景下，隐私泄露、大数据“杀熟”等问题日渐突出，让很多人望“数”生畏，甚至开始抵制数据收集和使用。应对大数据带来的新问题，必须加快制定相关法律法规，为大数据的收集、存储、使用、发布立好规矩。信息安全没有局外人，大数据关系每一个社会成员的切身利益，唯有依法规范，才能扬长避短，不断营造健康高效和有序可控的大数据发展环境。

大数据蕴含着大机遇、驱动着大未来。用好大数据这个时代赋予我们的强大引擎，就能抓住新机遇、培育新动能、塑造新优势，推动中国经济在数字化大潮中乘风破浪，驶向高质量发展的美好未来。

国家发改委等部门联合印发方案
—— 大数据中心布局有新调整

来源 / 人民网 - 人民日报海外版　编辑 / 协会会员处　李苗苗　日期 / 2021-01

京津冀、长三角、粤港澳大湾区、成渝以及贵州、内蒙古、甘肃、宁夏都“榜上有名”。近期，国家发改委等部门联合印发的《全国一体化大数据中心协同创新体系算力枢纽实施方案》（以下简称《方案》）向社会公布，明确将在这些地区建设全国算力网络国家枢纽节点。未来中国数据中心建设将迎来调整，推动实现合理布局、供需平衡、绿色集约和互联互通。

数据中心成为重要新型基础设施

随着各行业数字化转型升级进度加快，特别是 5G、人工智能、物联网等新技术的快速普及应用，数据存储、计算、传输和应用的需求大幅提升，数据中心已成为支撑各行业“上云用数赋智”的重要新型基础设施。

《方案》着眼大数据中心背后，有着突出的现实需求。

数据显示，目前我国数据增量年均增速超过 30%，数据中心规模从 2015 年的 124 万家增长到 2020 年的 500 万家。数据应用正从消费互联网向工业互联网加速渗透，我国已经成为全球大数据应用最为活跃、最具潜力、环境最优的国家之一。

数据中心建设发展仍存在进一步优化的空间。“我国数据中心存在一定程度的供需失衡、失序发展等问题。”国家发改委高技术司有关负责人说，一些东部地区应用需求大，但能耗指标紧张、电力成本高，大规模发展数据中心难度和局限性大；一些西部地区可再生能源丰富，气候适宜，但存在网络带宽小、跨省数据传输费用高等瓶颈，无法有效承接东部需求。

我国数据中心年用电量已占全社会用电的 2% 左右，且仍在快速增长，需要进一步挖掘节能减排潜力，处理好发展和节能的关系；各行业纷纷建设数据中心，但互不联通，出现了“数据中心孤岛”“云孤岛”等苗头，需加快推动数据中心、云、网络之间的协同联动，提高资源利用率。《方案》提出的全国算力网络国家枢纽节点建设，正着力重点推动在数据中心布局、网络、电力、能耗、算力、数据等方面进行统筹规划。

建设全国算力网络国家枢纽节点

为推动数据中心合理布局、供需平衡、绿色集约和互联


中华人民共和国国家发展和改革委员会
National Development and Reform Commission
热门搜索：油价
请输入关键字
首页 | 机构设置 | 新闻动态 | 政务公开 | 政务服务
首页 > 政务公开 > 政策 > 通知
关于印发《全国一体化大数据中心协同创新体系算力枢纽实施方案》的通知
发改高技〔2021〕709号
各省、自治区、直辖市及计划单列市、新疆生产建设兵团发展改革委、网信办、工业和信息化主管部门、通信管理局、能源局：
根据《关于加快构建全国一体化大数据中心协同创新体系的指导意见》（发改高技〔2020〕1922号）部署要求，为加快推动数据中心绿色高质量发展，建设全国算力枢纽体系，国家发展改革委会同有关部门研究制定了《全国一体化大数据中心协同创新体系算力枢纽实施方案》。现印发给你们，请结合实际，认真抓好贯彻落实。

互通，《方案》明确在京津冀、长三角、粤港澳大湾区、成渝以及贵州、内蒙古、甘肃、宁夏建设全国算力网络国家枢纽节点。

据介绍，全国一体化算力网络国家枢纽节点，是我国算力网络的骨干连接点。传统上，我国通信网络主要围绕人口聚集程度进行建设，网络节点普遍集中于北上广等一线城市。数据中心对网络依赖性强，随之集中于城市部署。

近年来，随着数据中心规模快速扩张，对土地供应、能源保障、气候条件等提出了更高要求，现有城市资源，特别是东部一线城市资源，已难以满足持续发展要求。“通过国家枢纽节点，统筹规划数据中心建设布局，引导大规模数据中心适度集聚，形成数据中心集群。”上述负责人表示。

对于京津冀、长三角、粤港澳大湾区、成渝等用户规模较大、应用需求强烈的节点，将重点统筹好城市内部和周边区域的数据中心布局，实现大规模算力部署与土地、用能、水、电等资源的协调可持续，满足重大区域发展战略实施需要；对于贵州、内蒙古、甘肃、宁夏等可再生能源丰富、气候适宜、数据中心绿色发展潜力较大的节点，将重点提升算力服务品质和利用效率，打造面向全国的非实时性算力保障基地。

统筹布局，完善标准，一体化实施

在具体实施上，《方案》明确国家枢纽节点建设的总体思路是统筹布局、完善标准、一体化实施推进。

要做好统筹布局——在可再生能源丰富和气候、地质等条件适宜的区域，建设数据中心集群，实现数据中心绿色、集约、高效发展。将规模适中、对网络实时性要求极高的边缘数据中心，在城市城区内部合理规划布局。城市内部原则上不再大规模发展数据中心。加强对已有数据中心的改造升级，提升效能。

完善标准制度——对于服务金融交易、车联网等领域，网络时延要求极高的数据中心，允许在城市内部发展；对于服务工业互联网、人工智能推理等领域，网络时延要求相对较高的数据中心，鼓励在数据中心集群发展；对于服务后台加工、存储灾备等，网络时延要求不高的数据中心，要优先向贵州、内蒙古、甘肃、宁夏节点转移。

国家发改委方面介绍，未来将推动相关政策试点、工程试点优先在国家枢纽节点实施。将加强网络、能源等方面的政策支持力度，重点围绕国家枢纽节点布局新型互联网交换中心、互联网骨干直连点等网络设施，积极协调安排能耗指标予以适当支持。同时，有关部门也将加强工作统筹力度，推动各枢纽节点尽快细化时间表、路线图。

Gartner 公布 2021 年十大数据和分析技术趋势

来源 / 人民网 - 人民日报海外版　编辑 / 协会会员处　李苗苗　日期 / 2021-01

Gartner 近日公布了 2021 年十大数据和分析技术趋势，这些技术趋势将帮助企业组织应对这一年中的各种变化、不确定性和机遇。

Gartner 杰出研究副总裁 Rita Sallam 表示：“疫情给企业组织带来颠覆的速度，迫使数据和分析领导者必须采用恰当的工具和流程应对这些关键技术趋势，对那些可能会给他们竞争优势带来最大潜在影响的技术趋势设置更高优先级。”

数据和分析领导者应该把以下 10 个技术趋势作为他们的关键投资方向，加强他们预测、转移和响应的能力。

趋势 1：更智能、负责任的、可扩展的 AI

人工智能（AI）和机器学习（ML）正在带来更大的影响，这就要求企业采用新技术构建更智能的、消耗数据更少的、符合道德原则的、更具弹性的 AI 解决方案。企业组织通过部署更智能、更负责任的、更可扩展的 AI，将利用学习算法和可解释的系统，加速价值实现，给业务带来更大影响力。

趋势 2：可组合式的数据和分析

开放的、容器化的分析架构让数据分析功能可组合性更强。可组合式的数据分析利用来自多个数据、分析和 AI 解决方案的组件，快速构建灵活且用户友好型的智能应用，从而帮助数据分析领导者将洞察和行动连接在一起。

随着数据重心转移到云端，可组合式的数据分析将成为一种更加敏捷的方式，开发支持云市场、低代码和无代码解决方案的分析应用。

趋势 3：数据架构是基础

更高程度的数字化和不再受约束的消费者，推动着数据分析领导者越来越多地使用数据架构来一个对企业组织数据资产日益加剧的多样化、分布式、规模和复杂性。

数据架构利用分析功能来持续监控数据管道，通过对数据资产的持续分析，支持各种数据的设计、部署和使用，缩短集成时间 30%，缩短部署时间 30%，缩短维护时间 70%。

趋势 4：从大数据到小数据、宽数据

疫情给企业带来的极端变革，导致那些基于大量历史数据的机器学习和人工智能模型变得不那么重要了。同时，由人类和 AI 做出的决策变得更加复杂和苛刻，要求数据分析领导者拥有更多种类的数据才能更好地了解态势。

因此，数据分析领导者应该选择那些可以更有效地利用可用数据的分析技术。数据分析领导者依赖于所谓的“宽数据”和“小数据”，宽数据可以对各种小型的、大型的、非结构化的、结构化的数据源进行分析和协同，小数据指的是那些需要较少数据但仍提供有用见解的分析技术应用。

Sallam 表示：“小数据和宽数据提供强大的分析和人工智能功能，降低了企业组织对大数据集的依赖性，而且通过使用宽数据，企业组织还可以获得更丰富、更完整的、全方位的态势感知，使他们能够运用分析来做出更好的决策。”

趋势 5：XOps

XOps（包括 DataOps、MLOps、ModelOps 和 PlatformOps）的目标是利用 DevOps 最佳实践来实现效率和规模经济，确保可靠性、可重用性和可重复性，同时减少技术和流程的重复，实现自动化。

大多数分析和人工智能项目都因为仅仅在事后才能解决可操作性问题而失败了。如果数据分析领导者利用 XOps 进行大规模运营，将实现分析和人工智能资产的再生性、可追溯性、

完整性和可集成性。

趋势 6：工程决策智能

工程决策智能不仅适用于单个决策，还适用于决策序列，可将其分为多个业务流程，甚至是突发决策和结果构成的多个网络。随着决策得到增强并且越来越自动化，工程决策让数据分析领导者有机会做出更准确、可重复、透明和可追溯的决策。

趋势 7：数据和分析是一项核心业务功能

数据分析不再是一个次要项目，而是变成了核心的业务功能。在这种情况下，数据分析变成与业务成果一致的共享业务资产，而且因为中央和联合数据分析团队之间能够更好地展开协作，数据分析孤岛问题也得到了解决。

趋势 8：关联一切的图形技术

图形技术构成了很多现代数据分析功能的基础，可以在各种数据资产之间找到人、地方、事物、事件和位置之间的关系。数据分析领导者依靠图形技术快速回答复杂的业务问题，而这些问题往往需要上下文感知，以及理解多个实体之间的关联本质。

Gartner 预测，到 2025 年图形技术将被用于 80%的数据分析创新项目中，高于 2021 年的 10%，从而促进整个企业组织的快速决策。

趋势 9：增强型消费者的崛起

如今大多数企业用户使用的是预定义的仪表板和手动数据浏览功能，这可能导致结论、决策和操作失误，而自动的、对话式的、移动且动态生成的洞察将取代预定义的仪表板，可根据用户需求进行定制，交付给消费方。

Sallam 表示：“这将推动分析能力转移到信息消费者——增强型消费者，让他们具备那些以前只有分析师和数据科学家才能拥有的能力。”

趋势 10：边缘位置的数据和分析

数据、分析和其他支持技术正在被越来越多地运用于边缘计算环境中，并且这些技术更靠近物理资产所在的位置，位于 IT 权限范围之外。Gartner 预测，到 2023 年超过 50%的数据分析领导者的主要职责将涉及到在边缘环境中创建、管理和分析的数据。

数据分析领导者可以利用这一趋势来提高数据管理的灵活性、速度、治理和弹性。从支持实时事件分析到实现“物”的自主行为，各种各样的使用场景正在吸引着人们对数据分析边缘能力的兴趣。

云计算、大数据与人工智能三者的关系

来源 / CIOT　编辑 / 协会会员处　李苗苗　日期 / 2021-01

当今智能行业最为热门的话题无非就是云计算、大数据与人工智能。它们之间好像互相有关系，一般谈云计算的时候会提到大数据，谈人工智能的时候会提大数据，谈人工智能的时候会提云计算……它们相辅相成、不可分割。若非技术人员，可能会较难理解这三者间的相互关系，所以有必要解释一下。

一、云计算最初是实现资源管理的灵活性

首先来说云计算，云计算最初的目标是对资源的管理，管理的主要是计算资源，网络资源，存储资源三个方面。

1.1 管数据中心就像配电脑

什么叫计算，网络，存储资源呢？就说你要买台笔记本电脑吧，你是不是要关心这台电脑什么样的 CPU 啊？多大的内存啊？这两个我们称为计算资源。

这台电脑要能上网吧，需要有个网口可以插网线，或者有无线网卡可以连接我们家的路由器，您家也需要到运营商比如联通，移动，电信开通一个网络，比如 100M 的带宽，然后会有师傅弄一根网线到您家来，师傅可能会帮您将您的路由器和他们公司的网络连接配置好，这样您家的所有的电脑，手机，平板就都可以通过您的路由器上网了。这就是网络。

您可能还会问硬盘多大啊？原来硬盘都很小，10G 之类的，后来 500G，1T，2T 的硬盘也不新鲜了。(1T 是 1000G)，这就是存储。

对于一台电脑是这个样子的，对于一个数据中心也是同样的。想象你有一个非常非常大的机房，里面堆了很多的服务器，这些服务器也是有 CPU，内存，硬盘的，也是通过类似路由器的设备上网的。这个时候的一个问题就是，运营数据中心的人是怎么把这些设备统一的管理起来的呢？

1.2 灵活就是想啥时要都有，想要多少都行

管理的目标就是要达到两个方面的灵活性。哪两个方面呢？比如有个人需要一台很小很小的电脑，只有一个 CPU，1G 内存，10G 的硬盘，一兆的带宽，你能给他吗？像这种这么小规格的电脑，现在随便一个笔记本电脑都比这个配置强了，家里随便拉一个宽带都要 100M。然而如果去一个云计算的平台上，他要想要这个资源的时候，只要一点就有了。

所以说它就能达到两个方面灵活性。

第一个方面就是想什么时候要就什么时候要，比如需要的时候一点就出来了，这个叫做时间灵活性。第二个方面就是想要多少呢就有多少，比如需要一个很小很小的电脑，可以满足，比如需要一个特别大的空间，以云盘为例，似乎云盘给每个人分配的空间动不动就就很大很大，随时上传随时有空间，

永远用不完，这个叫做空间灵活性。

空间灵活性和时间灵活性，也即我们常说的云计算的弹性。

为了解决这个弹性的问题，经历了漫长时间的发展。

1.3 物理设备不灵活

首先第一个阶段就是物理机，或者说物理设备时期。这个时期相当于客户需要一台电脑，我们就买一台放在数据中心里。物理设备当然是越来越牛，例如服务器，内存动不动就是百 G 内存，例如网络设备，一个端口的带宽就能有几十 G 甚至上百 G，例如存储，在数据中心至少是 PB 级别的 (一个 P 是 1000 个 T，一个 T 是 1000 个 G)。

然而物理设备不能做到很好的灵活性。首先它不能够达到想什么时候要就什么时候要、比如买台服务器，哪怕买个电脑，都有采购的时间。突然用户告诉某个云厂商，说想要开台电脑，如果使用物理服务器，当时去采购啊就很难，如果说供应商啊关系一般，可能采购一个月，供应商关系好的话也需要一个星期。用户等了一个星期后，这时候电脑才到位，用户还要登录上去开始慢慢部署自己的应用，时间灵活性非常差。第二是空间灵活性也不行，例如上述的用户，要一个很小很小的电脑，现在哪还有这么小型号的电脑啊。不能为了满足用户只要一个 G 的内存是 80G 硬盘的，就去买一个这么小的机器。但是如果买一个大的呢，因为电脑大，就向用户多收钱，用户说他只用这么小的一点，如果让用户多付钱就很冤。

1.4 虚拟化灵活多了

有人就想办法了。第一个办法就是虚拟化。用户不是只要一个很小的电脑么？数据中心的物理设备都很强大，我可以从物理的 CPU，内存，硬盘中虚拟出一小块来给客户，同时也可以虚拟出一小块来给其他客户，每个客户都只能看到自己虚的那一小块，其实每个客户用的是整个大的设备上其中的一小块。虚拟化的技术能使得不同的客户的电脑看起来是隔离的，我看着好像这块盘就是我的，你看这呢这块盘就是你的，实际情况可能我这个 10G 和您这个 10G 是落在同样一个很大很大的这个存储上的。

而且如果事先物理设备都准备好，虚拟化软件虚拟出一个电脑是非常快的，基本上几分钟就能解决。所以在任何一个云上要创建一台电脑，一点几分钟就出来了，就是这个道理。

这个空间灵活性和时间灵活性就基本解决了。

1.5 虚拟世界的赚钱与情怀

在虚拟化阶段，最牛的公司是 Vmware，是实现虚拟化技术比较早的一家公司，可以实现计算，网络，存储的虚拟化，这家公司很牛，性能也做得非常好，然后虚拟化软件卖的也非常好，赚了好多的钱，后来让 EMC(世界五百强，存储厂商第一品牌) 给收购了。

但是这个世界上还是有很多有情怀的人的，尤其是程序员里面，有情怀的人喜欢做一件什么事情呢？开源。这个世界上很多软件都是有闭源就有开源，源就是源代码。就是说某个软件做的好，所有人都爱用，这个软件的代码呢，我封闭起来只有我公司知道，其他人不知道，如果其他人想用这个软件，就要付我钱，这就叫闭源。但是世界上总有一些大牛看不惯钱都让一家赚了去。大牛们觉得，这个技术你会我也会，你能开发出来，我也能，我开发出来就是不收钱，把代码拿出来分享给大家，全世界谁用都可以，所有的人都可以享受到好处，这个叫做开源。

比如最近蒂姆·伯纳斯·李就是个非常有情怀的人，2017 年，他因“发明万维网、第一个浏览器和使万维网得以扩展的基本协议和算法”而获得 2016 年度的图灵奖。图灵奖就是计算机界的诺贝尔奖。然而他最令人敬佩的是，他将万维网，也就是我们常见的 www 的技术无偿贡献给全世界免费使用。我们现在在网上的所有行为都应该感谢他的功劳，如果他将这个技术拿来收钱，应该和比尔盖茨差不多有钱。

例如在闭源的世界里有 windows，大家用 windows 都得给微软付钱，开源的世界里面就出现了 Linux。比尔盖茨靠 windows，Office 这些闭源的软件赚了很多钱，称为世界首富，就有大牛开发了另外一种操作系统 Linux。很多人可能没有听说过 Linux，很多后台的服务器上跑的程序都是 Linux 上的，比如大家享受双十一，支撑双十一抢购的系统，无论是淘宝，京东，考拉，都是跑在 Linux 上的。

再如有 apple 就有安卓。apple 市值很高，但是苹果系统的代码我们是看不到的。于是就有大牛写了安卓手机操作系统。所以大家可以看到几乎所有的其他手机厂商，里面都装安卓系统，因为苹果系统不开源，而安卓系统大家都可以用。

在虚拟化软件也一样，有了 Vmware，这个软件非常非常的贵。那就有大牛写了两个开源的虚拟化软件，一个叫做 Xen，一个叫做 KVM，如果不做技术的，可以不用管这两个名字，但是后面还是会提到。

1.6 虚拟化的半自动和云计算的全自动

虚拟化软件似乎解决了灵活性问题，其实不全对。因为虚拟化软件一般创建一台虚拟的电脑，是需要人工指定这台虚拟电脑放在哪台物理机上的，可能还需要比较复杂的人工配置，所以使用 Vmware 的虚拟化软件，需要考一个很牛的证书，能拿到这个证书的人，薪资是相当的高，也可见复杂程度。所以仅仅凭虚拟化软件所能管理的物理机的集群规模都不是特别的大，一般在十几台，几十台，最多百台这么一个规模。这一方面会影响时间灵活性，虽然虚拟出一台电脑的时间很短，但是随着集群规模的扩大，人工配置的过程越来越复杂，越来越耗时。另一方面也影响空间灵活性，当用户数量多的时候，这点集群规模，还远达不到想要多少要多少的程度，很可能这点资源很快就用完了，还得去采购。所以随着集群的规模越来越大，基本都是千台起步，动辄上万台，甚至几十上百万台，如果去查一下 BAT，包括网易，包括谷歌，亚马逊，服务器数目

都大的吓人。这么多机器要靠人去选一个位置放这台虚拟化的电脑并做相应的配置，几乎是不可能的事情，还是需要机器去做这个事情。

人们发明了各种各样的算法来做这个事情，算法的名字叫做调度 (Scheduler)。通俗一点的说，就是有一个调度中心，几千台机器都在一个池子里面，无论用户需要多少 CPU，内存，硬盘的虚拟电脑，调度中心会自动在大池子里面找一个能够满足用户需求的地方，把虚拟电脑启动起来做好配置，用户就直接能用了。这个阶段，我们称为池化，或者云化，到了这个阶段，才可以称为云计算，在这之前都只能叫虚拟化。

1.7 云计算的私有与公有

云计算大致分两种，一个是私有云，一个是公有云，还有人把私有云和公有云连接起来称为混合云，我们暂且不说这个。私有云就是把虚拟化和云化的这套软件部署在别人的数据中心里面，使用私有云的用户往往很有钱，自己买地建机房，自己买服务器，然后让云厂商部署在自己这里，Vmware 后来除了虚拟化，也推出了云计算的产品，并且在私有云市场赚的盆满钵满。所谓公有云就是虚拟化和云化软件部署在云厂商自己数据中心里面的，用户不需要很大的投入，只要注册一个账号，就能在一个网页上点一下创建一台虚拟电脑，例如 AWS 也即亚马逊的公有云，例如国内的阿里云，腾讯云，网易云等。

亚马逊呢为什么要做公有云呢？我们知道亚马逊原来是国外比较大的一个电商，它做电商的时候也肯定会遇到类似双 11 的场景，在某一个时刻大家都冲上来买东西。当大家都冲上买东西的时候，就特别需要云的时间灵活性和空间灵活性。因为它不能时刻准备好所有的资源，那样太浪费了。但也不能什么都不准备，看着双十一这么多用户想买东西登不上去。所以需要双十一的时候，创建一大批虚拟电脑来支撑电商应用，过了双十一再把这些资源都释放掉去干别的。所以亚马逊是需要一个云平台的。

然而商用的虚拟化软件实在是太贵了，亚马逊总不能把自己在电商赚的钱全部给了虚拟化厂商吧。于是亚马逊基于开源的虚拟化技术，如上所述的 Xen 或者 KVM，开发了一套自己的云化软件。没想到亚马逊后来电商越做越牛，云平台也越做越牛。而且由于他的云平台需要支撑自己的电商应用，而传统的云计算厂商多为 IT 厂商出身，几乎没有自己的应用，因而亚马逊的云平台对应用更加的友好，迅速发展成为云计算的第一品牌，赚了很多钱。在亚马逊公布其云计算平台财报之前，人们都猜测，亚马逊电商赚钱，云也赚钱吗？后来一公布财报，发现不是一般的赚钱，仅仅去年，亚马逊 AWS 年营收达 122 亿美元，运营利润 31 亿美元。

1.8 云计算的赚钱与情怀

公有云的第一名亚马逊过得很爽，第二名 Rackspace 过的就一般了。没办法，这就是互联网行业的残酷性，多是赢者通吃的模式。所以第二名如果不是云计算行业的，很多人可能都没听过了。第二名就想，我干不过老大怎么办呢？开源吧。如上所述，亚马逊虽然使用了开源的虚拟化技术，但是云化的代码是闭源的，很多想做又做不了云化平台的公司，只能眼巴巴的看着亚马逊挣大钱。Rackspace 把源代码一公开，整个行业就可以一起把这个平台越做越好。

于是 Rackspace 和美国航空航天局合作创办了开源软件 OpenStack，如图所示 OpenStack 的架构图，不是云计算行业的不用弄懂这个图，但是能够看到三个关键字，Compute 计算，Networking 网络，Storage 存储。还是一个计算，网络，存储的云化管理平台。

当然第二名的技术也是非常棒的，有了 OpenStack 之后，果真像 Rackspace 想象的一样，所有想做云的大企业都疯了，你能想象到的所有如雷贯耳的大型 IT 企业，IBM，惠普，戴尔，华为，联想等等，都疯了。原来云平台大家都想做，看着亚马逊和 Vmware 赚了这么多钱，眼巴巴看着没办法，想自己做一个好像难度还挺大。现在好了，有了这样一个开源的云平台 OpenStack，所有的 IT 厂商都加入到这个社区中来，对这个云平台进行贡献，包装成自己的产品，连同自己的硬件设备一起卖。有的做了私有云，有的做了公有云，OpenStack 已经成为开源云平台的事实标准。

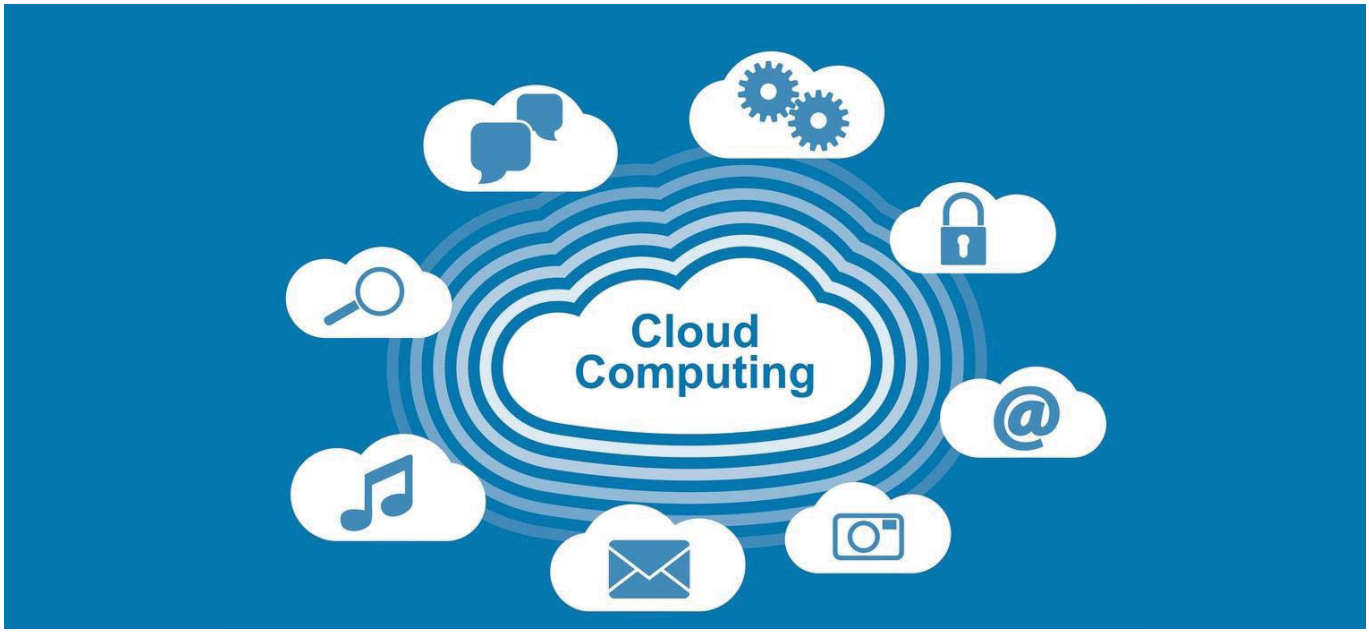
1.9IaaS, 资源层面的灵活性

随着 OpenStack 的技术越来越成熟，可以管理的规模也越来越大，并且可以有多个 OpenStack 集群部署多套，比如北京部署一套，杭州部署两套，广州部署一套，然后进行统一的管理。这样整个规模就更大了。在这个规模下，对于普通用户的感知来讲，基本能够做到想什么时候要就什么什么药，想要多少就要多少。还是拿云盘举例子，每个用户云盘都分配了 5T 甚至更大的空间，如果有 1 亿人，那加起来空间多大啊。其实背后的机制是这样的，分配你的空间，你可能只用了其中很少一点，比如说它分配给你了 5 个 T，这么大的空间仅仅是你看到的，而不是真的就给你了，你其实只用了 50 个 G，则真实给你的就是 50 个 G，随着你文件的不断上传，分给你的空间会越来越多。当大家都上传，云平台发现快满了的时候 (例如用了 70%)，会采购更多的服务器，扩充背后的资源，这个对用户是透明的，看不到的，从感觉上来讲，就实现了云计算的弹性。其实有点像银行，给储户的感觉是什么时候取钱都有，只要不同时挤兑，银行就不会垮。

这里做一个简单的总结，到了这个阶段，云计算基本上实现了时间灵活性和空间灵活性，实现了计算，网络，存储资源的弹性。计算，网络，存储我们常称为基础设施 Infranstracture, 因而这个阶段的弹性称为资源层面的弹性，管理资源的云平台，我们称为基础设施服务，就是我们常听到的 IaaS，InfranstractureAsAService。

二、云计算不光管资源，也要管应用

有了 IaaS，实现了资源层面的弹性就够了吗？显然不是。还有应用层面的弹性。这里举个例子，比如说实现一个电商的应用，平时十台机器就够了，双十一需要一百台。你可能觉得很好办啊，有了 IaaS，新创建九十台机器就可以了啊。但是 90 台机器创建出来是空的啊，电商应用并没有放上去啊，只



能你公司的运维人员一台一台的弄，还是需要很长时间才能安装好的。虽然资源层面实现了弹性，但是没有应用层的弹性，依然灵活性是不够的。

有没有方法解决这个问题呢？于是人们在 IaaS 平台之上又加了一层，用于管理资源以上的应用弹性的问题，这一层通常称为 PaaS（PlatformAsAService）。这一层往往比较难理解，其实大致分两部分，一部分我称为你自己的应用自动安装，一部分我称为通用的应用不用安装。

我们先来说第一部分，自己的应用自动安装。比如电商应用是你自己开发的，除了你自己，其他人是不知道怎么安装的，比如电商应用，安装的时候需要配置支付宝或者微信的账号，才能别人在你的电商上买东西的时候，付的钱是打到你的账户里面的，除了你，谁也不知道，所以安装的过程平台帮不了忙，但是能够帮你做的自动化，你需要做一些工作，将自己的配置信息融入到自动化的安装过程中方可。比如上面的例子，双十一新创建出来的 90 台机器是空的，如果能够提供一个工具，能够自动在这新的 90 台机器上将电商应用安装好，就能够实现应用层面的真正弹性。例如 Puppet,Chef,Ansible,CloudFoundary 都可以干这件事情，最新的容器技术 Docker 能更好的干这件事情，不做技术的可以不用管这些词。

第二部分，通用的应用不用安装。所谓通用的应用，一般指一些复杂性比较高，但是大家都在用的，例如数据库。几乎所有的应用都会用数据库，但是数据库软件是标准的，虽然安装和维护比较复杂，但是无论谁安装都是一样。这样的应用可以变成标准的 PaaS 层的应用放在云平台的界面上。当用户需要一个数据库的时候，一点就出来了，用户就可以直接用了。有人问，既然谁安装都一个样，那我自己来好了，不需要花钱在云平台上买。当然不是，数据库是一个非常难的东西，光 Oracle 这家公司，靠数据库就能赚这么多钱。买 Oracle 也是要花很多很多钱的。然而大多数云平台会提供 Mysql 这样的开源数据库，又是开源，钱不需要花这么多了，但是维护这个数据库，却需要专门招一个很大的团队，如果这个数据库能够优化到能够支撑双十一，也不是一年两年能够搞定的。比如您是一个做单车的，当然没必要招一个非常大的数据库团队来干这件事情，成本太高了，应该交给云平台来做这件事情，专业的事情专业的人来自，云平台专门养了几百人维护这套系统，您只要专注于您的单车应用就可以了。

要么是自动部署，要么是不用部署，总的来说就是应用层你也要少操心，这就是 PaaS 层的重要作用。

虽说脚本的方式能够解决自己的应用的部署问题，然而不同的环境千差万别，一个脚本往往在一个环境上运行正确，到另一个环境就不正确了。

而容器是能更好的解决这个问题的。

容器是 Container，Container 另一个意思是集装箱，其实容器的思想就是要变成软件交付的集装箱。集装箱的特点，一是封装，二是标准。

在没有集装箱的时代，假设将货物从 A 运到 B，中间要经过三个码头、换三次船。每次都要将货物卸下船来，摆的七零八落，然后搬上船重新整齐摆好。因此在没有集装箱的时候，每次换船，船员们都要在岸上待几天才能走。

有了集装箱以后，所有的货物都打包在一起了，并且集装箱的尺寸全部一致，所以每次换船的时候，一个箱子整体搬过去就行了，小时级别就能完成，船员再也不用上岸长时间耽搁了。

这是集装箱“封装”、“标准”两大特点在生活中的应用。

那么容器如何对应用打包呢？还是要学习集装箱，首先要有个封闭的环境，将货物封装起来，让货物之间互不干扰，互相隔离，这样装货卸货才方便。好在 Ubuntu 中的 LXC 技术早就能做到这一点。

封闭的环境主要使用了两种技术，一种是看起来是隔离的技术，称为 Namespace，也即每个 Namespace 中的应用看到的是不同的 IP 地址、用户空间、程号等。另一种是用起来是隔离的技术，称为 Cgroups，也即明明整台机器有很多的 CPU、内存，而一个应用只能用其中的一部分。

所谓的镜像，就是将你焊好集装箱的那一刻，将集装箱的状态保存下来，就像孙悟空说：“定”，集装箱里面就定在了那一刻，然后将这一刻的状态保存成一系列文件。这些文件的格式是标准的，谁看到这些文件都能还原当时定住的那个时刻。将镜像还原成运行时的过程（就是读取镜像文件，还原那个时刻的过程）就是容器运行的过程。

有了容器，使得 PaaS 层对于用户自身应用的自动部署变得快速而优雅。

三、大数据拥抱云计算

在 PaaS 层中一个复杂的通用应用就是大数据平台。大数据是如何一步一步融入云计算的呢?

3.1 数据不大也包含智慧

一开始这个大数据并不大，你想象原来才有多少数据？现在大家都去看电子书，上网看新闻了，在我们 80 后小时候，信息量没有那么大，也就看看书，看看报，一个星期的报纸加起来才有多少字啊，如果你不在一个大城市，一个普通的学校的图书馆加起来也没几个书架，是后来随着信息化的到来，信息才会越来越多。

首先我们来看一下大数据里面的数据，就分三种类型，一种叫结构化的数据，一种叫非结构化的数据，还有一种叫半结构化的数据。什么叫结构化的数据呢？叫有固定格式和有限长度的数据。例如填的表格就是结构化的数据，国籍：中华人民共和国，民族：汉，性别：男，这都叫结构化数据。现在越来越多的就是非结构化的数据，就是不定长，无固定格式的数据，例如网页，有时候非常长，有时候几句话就没了，例如语音，视频都是非结构化的数据。半结构化数据是一些 xml 或者 html 的格式的，不从事技术的可能不了解，但也没有关系。

数据怎么样才能对人有用呢？其实数据本身不是有用的，必须要经过一定的处理。例如你每天跑步带个手环收集的也是数据，网上这么多网页也是数据，我们称为 Data，数据本身没有什么用处，但是数据里面包含一个很重要的东西，叫做信息 Information，数据十分杂乱，经过梳理和清洗，才能够称为信息。信息会包含很多规律，我们需要从信息中将规律总结出来，称为知识 knowledge，知识改变命运。信息是很多的，但是有人看到了信息相当于白看，但是有人就从信息中看到了电商的未来，有人看到了直播的未来，所以人家就牛了，你如果没有从信息中提取出知识，天天看朋友圈，也只能在互联网滚滚大潮中做个看客。有了知识，然后利用这些知识去应用于实战，有的人会做得非常好，这个东西叫做智慧 intelligence。有知识并不一定有智慧，例如好多学者很有知识，已经发生的事情可以从各个角度分析的头头是道，但一到实干就歇菜，并不能转化成为智慧。而很多的创业家之所以伟大，就是通过获得的知识应用于实践，最后做了很大的生意。

所以数据的应用分这四个步骤：数据，信息，知识，智慧。这是很多商家都想要的，你看我收集了这么多的数据，能不能基于这些数据来帮我做下一步的决策，改善我的产品，例如让用户看视频的时候旁边弹出广告，正好是他想买的东西，再如让用户听音乐的时候，另外推荐一些他非常想听的其他音乐。用户在我的应用或者网站上随便点点鼠标，输入文字对我来说都是数据，我就是要将其中某些东西提取出来，指导实践，形成智慧，让用户陷入到我的应用里面不可自拔，上了我的网就不想离开，手不停的点，不停的买，很多人说双十一我都想断网了，我老婆在上面不断的买买买，买了 A 又推荐 B，老婆大人说，“哎呀，B 也是我喜欢的啊，老公我要买”。你说这个程序怎么这么牛，这么有智慧，比我还了解我老婆，这件事情是怎么做到的呢?

3.2 数据如何升华为智慧

数据的处理分几个步骤，完成了才最后会有智慧。

第一个步骤叫数据的收集。首先得有数据，数据的收集有两个方式，第一个方式是拿，专业点的说法叫抓取或者爬取，例如搜索引擎就是这么做的，它把网上的所有的信息都下载到它的数据中心，然后你一搜才能搜出来。比如你去搜索的时候，结果会是一个列表，这个列表为什么会在搜索引擎的公司里面呢，就是因为他把这个数据啊都拿下来了，但是你一点链接，点出来这个网站就不在搜索引擎它们公司了。比如说新浪有个新闻，你拿百度搜出来，你不点的时候，那一页在百度数据中心，一点出来的网页就是在新浪的数据中心了。另外一个方式就是推送，有很多终端可以帮我收集数据，比如说小米手环，可以将你每天跑步的数据，心跳的数据，睡眠的数据都上传到数据中心里面。

第二个步骤是数据的传输。一般会通过队列方式进行，因为数据量实在是太大了，数据必须经过处理才会有用，可是系统处理不过来，只好排好队，慢慢的处理。

第三个步骤是数据的存储。现在数据就是金钱，掌握了数据就相当于掌握了钱。要不然网站怎么知道你想买什么呢？就是因为它有你历史的交易的数据，这个信息可不能给别人，十分宝贵，所以需要存储下来。

第四个步骤是数据的处理和分析。上面存储的数据是原始数据，原始数据多是杂乱无章的，有很多垃圾数据在里面，因而需要清洗和过滤，得到一些高质量的数据。对于高质量的数据，就可以进行分析，从而对数据进行分类，或者发现数据之间的相互关系，得到知识。比如盛传的沃尔玛超市的啤酒和尿布的故事，就是通过对人们的购买数据进行分析，发现了男人一般买尿布的时候，会同时购买啤酒，这样就发现了啤酒和尿布之间的相互关系，获得知识，然后应用到实践中，将啤酒和尿布的柜台弄的很近，就获得了智慧。

第五个步骤就是对于数据的检索和挖掘。检索就是搜索，所谓外事不决问 google，内事不决问百度。内外两大搜索引擎都是讲分析后的数据放入搜索引擎，从而人们想寻找信息的时候，一搜就有了。另外就是挖掘，仅仅搜索出来已经不能满足人们的要求了，还需要从信息中挖掘出相互的关系。比如财经搜索，当搜索某个公司股票的时候，该公司的高管是不是也应该被挖掘出来呢？如果仅仅搜索出这个公司的股票发现涨的特别好，于是你就去买了，其实其高管发了一个声明，对股票十分不利，第二天就跌了，这不坑害广大股民么？所以通过各种算法挖掘数据中的关系，形成知识库，十分重要。

3.3 大数据时代，众人拾柴火焰高

当数据量很小的时候，很少的几台机器就能解决。慢慢的当数据量越来越大，最牛的服务器都解决不了问题的时候，就想怎么办呢？要聚合多台机器的力量，大家齐心协力一起把这个事搞定，众人拾柴火焰高。

对于数据的收集，对于 IoT 来讲，外面部署这成千上万的检测设备，将大量的温度，适度，监控，电力等等数据统统收集上来，对于互联网网页的搜索引擎来讲，需要将整个互联网所有的网页都下载下来，这显然一台机器做不到，需要多台机器组成网络爬虫系统，每台机器下载一部分，同时工作，才能在有限的时间内，将海量的网页下载完毕。

对于数据的传输，一个内存里面的队列肯定会被大量的数据挤爆掉，于是就产生了基于硬盘的分布式队列，这样队列可以多台机器同时传输，随你数据量多大，只要我的队列足够多，管道足够粗，就能够撑得住。

对于数据的存储，一台机器的文件系统肯定是放不下了，所以需要一个很大的分布式文件系统来做这件事情，把多台机器的硬盘打成一块大的文件系统。

再如数据的分析，可能需要对大量的数据做分解，统计，汇总，一台机器肯定搞不定，处理到猴年马月也分析不完，于是就有分布式计算的方法，将大量的数据分成小份，每台机器处理一小份，多台机器并行处理，很快就能算完。例如著名的 Terasort 对 1 个 TB 的数据排序，相当于 1000G，如果单机处理，怎么也要几个小时，但是并行处理 209 秒就完成了。

所以说大数据平台，什么叫做大数据，说白了就是一台机器干不完，大家一起干。随着数据量越来越大，很多不大的公司都需要处理相当多的数据，这些小公司没有这么多机器可怎么办呢？

3.4 大数据需要云计算，云计算需要大数据

说到这里，大家想起云计算了吧。当想要干这些活的时候，需要好多好多的机器一块做，真的是想什么时候要，想要多少就要多少。例如大数据分析公司的财务情况，可能一周分析一次，如果要把这一百台机器或者一千台机器都在那放着，一周用一次对吧，非常浪费。那能不能需要计算的时候，把这一千台机器拿出来，然后不算的时候，这一千台机器可以去干别的事情。谁能做这个事儿呢？只有云计算，可以为大数据的运算提供资源层的灵活性。而云计算也会部署大数据放到它的 PaaS 平台上，作为一个非常非常重要的通用应用。因为大数据平台能够使得多台机器一起干一个事儿，这个东西不是一般人能开发出来的，也不是一般人玩得转的，怎么也得雇个几十上百号人才能把这个玩起来，所以说就像数据库一样，其实还是需要有一帮专业的人来玩这个东西。现在公有云上基本上都会有大数据的解决方案了，一个小公司我需要大数据平台的时候，不需要采购一千台机器，只要到公有云上一点，这一千台机器都出来了，并且上面已经部署好了的大数据平台，只要把数据放进去算就可以了。

云计算需要大数据，大数据需要云计算，两个人就这样结合了。

四、人工智能拥抱大数据

4.1 机器什么时候才能懂人心

虽说有了大数据，人的欲望总是这个不能够满足。虽说在大数据平台里面有搜索引擎这个东西，想要什么东西我一搜就出来了。但是也存在这样的情况，我想要的东西不会搜，表达不出来，搜索出来的又不是我想要的。例如音乐软件里面推

荐一首歌，这首歌我没听过，当然不知道名字，也没法搜，但是软件推荐给我，我的确喜欢，这就是搜索做不到的事情。当人们使用这种应用的时候，会发现机器知道我想要什么，而不是说当我想要的时候，去机器里面搜索。这个机器真像我的朋友一样懂我，这就有点人工智能的意思了。

人们很早就在想这个事情了。最早的时候，人们想象，如果要是有一堵墙，墙后面是个机器，我给它说话，它就给我回应，我如果感觉不出它那边是人还是机器，那它就真的是一个人工智能的东西了。

4.2 让机器学会推理

怎么才能做到这一点呢？人们就想：我首先要告诉计算机人类的推理的能力。你看人重要的是什么呀，人和动物的区别在什么呀，就是能推理。我要是把我这个推理的能力啊告诉机器，机器就能根据你的提问，推理出相应的回答，真能这样多好。推理其实人们慢慢的让机器能够做到一些了，例如证明数学公式。这是一个非常让人惊喜的一个过程，机器竟然能够证明数学公式。但是慢慢发现其实这个结果，也没有那么令人惊喜，因为大家发现了一个问题，数学公式非常严谨，推理过程也非常严谨，而且数学公式很容易拿机器来进行表达，程序也相对容易表达。然而人类的语言就没这么简单了，比如今天晚上，你和你女朋友约会，你女朋友说：如果你早来，我没来，你等着，如果我早来，你没来，你等着。这个机器就比比较难理解了，但是人都懂，所以你和女朋友约会，你是不敢迟到的。

4.3 教给机器知识

所以仅仅告诉机器严格的推理是不够的，还要告诉机器一些知识。但是知识这个事儿，一般人可能就做不来了，可能专家可以，比如语言领域的专家，或者财经领域的专家。语言领域和财经领域知识能不能表示成像数学公式一样稍微严格点呢？例如语言专家可能会总结出主谓宾定状补这些语法规则，主语后面一定是谓语，谓语后面一定是宾语，将这些总结出来，并严格表达出来不久行了吗？后来发现这个不行，太难总结了，语言表达千变万化。就拿主谓宾的例子，很多时候在口语里面就省略了谓语，别人问：你谁啊？我回答：我刘超。但是你不能规定在语音语义识别的时候，要求对着机器说标准的书面语，这样还是不够智能，就像罗永浩在一次演讲中说的那样，每次对着手机，用书面语说：请帮我呼叫某某某，这是一件很尴尬的事情。

人工智能这个阶段叫做专家系统。专家系统不易成功，一方面是知识比较难总结，另一方面总结出来的知识难以教给计算机。因为你自己还迷迷糊糊，似乎觉得有规律，就是说不出来，就怎么能够通过编程教给计算机呢？

4.4 模拟大脑的工作方式

于是人类开始从机器的世界，反思人类的世界是怎么工作的。

人类的脑子里面不是存储着大量的规则，也不是记录着大量的统计数据，而是通过神经元的触发实现的，每个神经元有从其他神经元的输入，当接收到输入的时候，会产生一个输出来刺激其他的神经元，于是大量的神经元相互反应，最终形成各种输出的结果。例如当人们看到美女瞳孔放大，绝不是大脑根据身材比例进行规则判断，也不是将人生中看过的所有的美女都统计一遍，而是神经元从视网膜触发到大脑再回到瞳孔。在这个过程中，其实很难总结出每个神经元对最终的结果起到了哪些作用，反正就是起作用了。

于是人们开始用一个数学单元模拟神经元

这个神经元有输入，有输出，输入和输出之间通过一个公式来表示，输入根据重要程度不同 (权重)，影响着输出。

于是将 n 个神经元通过像一张神经网络一样连接在一起，n 这个数字可以很大很大，所有的神经元可以分成很多列，每一列很多个排列起来，每个神经元的对于输入的权重可以都不相同，从而每个神经元的公式也不相同。当人们从这张网络中输入一个东西的时候，希望输出一个对人类来讲正确的结果。例如上面的例子，输入一个写着 2 的图片，输出的列表里面第二个数字最大，其实从机器来讲，它既不知道输入的这个图片写的是 2，也不知道输出的这一系列数字的意义，没关系，人知道意义就可以了。正如对于神经元来说，他们既不知道视网膜看到的是美女，也不知道瞳孔放大是为了看的清楚，反正看到美女，瞳孔放大了，就可以了。

对于任何一张神经网络，谁也不敢保证输入是 2，输出一定是第二个数字最大，要保证这个结果，需要训练和学习。毕竟看到美女而瞳孔放大也是人类很多年进化的结果。学习的过程就是，输入大量的图片，如果结果不是想要的结果，则进行调整。如何调整呢，就是每个神经元的每个权重都向目标进行微调，由于神经元和权重实在是太多了，所以整张网络产生的结果很难表现出非此即彼的结果，而是向着结果微微的进步，最终能够达到目标结果。当然这些调整的策略还是非常有技巧的，需要算法的高手来仔细的调整。正如人类见到美女，瞳孔一开始没有放大到能看清楚，于是美女跟别人跑了，下次学习的结果是瞳孔放大一点点，而不是放大鼻孔。

4.5 没道理但做得到

听起来也没有那么有道理，但是的确能做到，就是这么任性。

神经网络的普遍性定理是这样说的，假设某个人给你某种复杂奇特的函数，f(x)：

不管这个函数是什么样的，总会确保有个神经网络能够对任何可能的输入 x，其值 f(x)（或者某个能够准确的近似）是神经网络的输出。

如果在函数代表着规律，也意味着这个规律无论多么奇妙，多么不能理解，都是能通过大量的神经元，通过大量权重的调整，表示出来的。

4.6 人工智能的经济学解释

这让我想到了经济学，于是比较容易理解了。

我们把每个神经元当成社会中从事经济活动的个体。于是神经网络相当于整个经济社会，每个神经元对于社会的输入，都有权重的调整，做出相应的输出，比如工资涨了，菜价也涨了，股票跌了，我应该怎么办，怎么花自己的钱。这里面没有规律么？肯定有，但是具体什么规律呢？却很难说清楚。

基于专家系统的经济属于计划经济，整个经济规律的表示不希望通过每个经济个体的独立决策表现出来，而是希望通过专家的高屋建瓴和远见卓识总结出来。专家永远不可能知道哪个城市的哪个街道缺少一个卖甜豆腐脑的。于是专家说应该产多少钢铁，产多少馒头，往往距离人民生活的真正需求有较大的差距，就算整个计划书写个几百页，也无法表达隐藏在人民生活中的小规律。

基于统计的宏观调控就靠谱的多了，每年统计局都会统计整个社会的就业率，通胀率，GDP 等等指标，这些指标往往代表着很多的内在规律，虽然不能够精确表达，但是相对靠谱。然而基于统计的规律总结表达相对比较粗糙，比如经济学家看到这些统计数据可以总结出长期来看房价是涨还是跌，股票长期来看是涨还是跌，如果经济总体上扬，房价和股票应该都是涨的。但是基于统计数据，无法总结出股票，物价的微小波动规律。

基于神经网络的微观经济学才是对整个经济规律最最准确的表达，每个人对于从社会中的输入，进行各自的调整，并且调整同样会作为输入反馈到社会中。想象一下股市行情细微的波动曲线，正是每个独立的个体各自不断交易的结果，没有统一的规律可循。而每个人根据整个社会的输入进行独立决策，当某些因素经过多次训练，也会形成宏观上的统计性的规律，这也就是宏观经济学所能看到的。例如每次货币大量发行，最后房价都会上涨，多次训练后，人们也就都学会了。

4.7 人工智能需要大数据

然而神经网络包含这么多的节点，每个节点包含非常多的参数，整个参数量实在是太大了，需要的计算量实在太大，但是没有关系啊，我们有大数据平台，可以汇聚多台机器的力量一起来计算，才能在有限的时间内得到想要的结果。

人工智能可以做的事情非常多，例如可以鉴别垃圾邮件，鉴别黄色暴力文字和图片等。这也是经历了三个阶段的。第一个阶段依赖于关键词黑白名单和过滤技术，包含哪些词就是黄色或者暴力的文字。随着这个网络语言越来越多，词也不断的变化，不断的更新这个词库就有点顾不过来。第二个阶段时，基于一些新的算法，比如说贝叶斯过滤等，你不用管贝叶斯算法是什么，但是这个名字你应该听过，这个一个基于概率的算法。第三个阶段就是基于大数据和人工智能，进行更加精准的用户画像和文本理解和图像理解。

由于人工智能算法多是依赖于大量的数据的，这些数据往往需要面向某个特定的领域 (例如电商，邮箱) 进行长期的积累，如果没有数据，就算有人工智能算法也白搭，所以人工智能程序很少像前面的 IaaS 和 PaaS 一样，将人工智能程序给某个客户安装一套让客户去用，因为给某个客户单独安装一套，客户没有相关的数据做训练，结果往往是很差的。但是云计算厂商往往是积累了大量数据的，于是就在云计算厂商里面安装一套，暴露一个服务接口，比如您想鉴别一个文本是不是涉及黄色和暴力，直接用这个在线服务就可以了。这种形势的服务，在云计算里面称为软件即服务，SaaS(SoftwareASAService)

于是人工智能程序作为 SaaS 平台进入了云计算

终于云计算的三兄弟凑齐了，分别是 IaaS，PaaS 和 SaaS，所以一般在一个云计算平台上，云，大数据，人工智能都能找得到。对一个大数据公司，积累了大量的数据，也会使用一些人工智能的算法提供一些服务。对于一个人工智能公司，也不可能没有大数据平台支撑。所以云计算，大数据，人工智能就这样整合起来，完成了相遇，相识，相知。

舆情大数据的社会科学应用

来源 / 中国社会科学网 - 中国社会科学报　编辑 / 协会会员处　李苗苗　日期 / 2021-01

在计算社会科学的发展进程中，多种形态的大数据类型不断涌现，比如书籍文本大数据 Google Books、网络百科大数据 Wikipedia 等。在众多的大数据类型当中，以 Twitter、Facebook 以及在线新闻舆情信息汇聚成的舆情大数据，构成计算社会科学的重要观测对象。本文试图就舆情大数据的主要特征及其在社会科学领域的应用场景做简要分析。

舆情大数据第一个重要特征是其话语属性。文本的内容表达了各种各样的观点、态度和立场，这些归结到一点，实际上就是话语，即各种各样的社会主体，基于其立场，表达各种各样的意见和看法。因此，话语分析应该是舆情大数据分析的第一层重要含义，借助于舆情大数据的高维属性，对文本进行话语分析，呈现话语背后的立场与观点、不同话语主体之间的交锋与博弈、不同话语的声量大小与社会影响等等。

在中国崛起的时代，可以分析西方政治话语与中国话语之间的博弈，还可以分析西方话语的建构逻辑，以及如何寻找西方话语的破解之道。对于中国话语，我们则可以分析中国话语的国际影响力，以及如何进一步讲好中国故事、建构中国话语。

话语分析的方法多种多样。从简单的主题分析、语义分析到词丛与搭配分析等，借助于这些技术，我们可以对文本表达做一些初步的分析；而借助于向量空间模型，比如说借助于词向量模型，可以对话语中的关键特征所嵌入的语境深入挖掘；借助于句向量空间模型，则可以对话语的类型进行分类，呈现话语的结构。

舆情大数据的第二个重要特征是其情感属性。舆情者，情绪也。舆情信息中，总是会充斥着丰富的情感表达，这是由舆情信息的属性所决定的。一方面，就新闻舆情而言，舆情需要与受众 “同呼吸、共命运”。舆情数据的一个重要特征就是共情，只有这样，舆情信息所表达的喜怒哀乐，才能与大众的喜怒哀乐保持共振，舆情才能够影响社会，才能够吸引观众。因此，在线新闻舆情信息的一个重要特征就是其情感属性。另一方面，就社交媒体信息而言，社交媒体的主体部分是大众直接在社交媒体上表达所思所想、生存状态与生存方式，在这些自我表达中，也往往是有感而发，分享的是或喜悦、或忧伤、

或震惊、或愤怒的情感。

正是因为无论是新闻媒体信息，还是社交媒体信息，都富含情绪表达，因此对舆情大数据进行情感计算，就成为一项非常重要的任务。这些年来，自然语言处理领域的情感计算技术飞速发展。从最初借助于 LIWC、WordNet 等情感词库开展情感词频统计，到现在基于机器学习和 BERT 模型等开展情感的精细描述，多种多样的情感分析技术在飞速发展。

就情感计算的内容而言，从最初计算正向和负向情感这样的初级分类，到现在可以计算喜、怒、哀、乐、爱、惧、憎等基本情绪。随着情感计算技术的进一步发展，未来进一步计算更加具体的情绪，比如羡慕、嫉妒、恨等都是大有可为的。正如李飞飞所言，人工智能的发展，在经历了“视觉计算”之后，下一个发展的重点就是情感计算。对海量的非结构化文本信息和图像进行情感计算，正是自然语言处理领域飞速发展的重要方向，而这为与情感计算相关的科学研究提供了坚实的技术支撑。

舆情大数据第三个重要特征是其传播属性。舆情大数据的受众和生产主体都是大众，信息、话语或者情绪的传播，构成舆情的一个重要景观，而某种话语或者观点在网络空间或者社交媒体空间能否传播开来，很大程度上取决于其传播属性。社交媒体平台上涉及非常丰富的传播现象，传播的要素不仅仅包括话语，还包括情绪的传播与扩散，比如疫情期间的恐慌情绪传播。

纵观这些形形色色的传播现象，我们可以发现，绝大多数传播信息最终是在浩瀚的信息海洋中归于寂灭，但也有一些有传播生命力的传播要素最终扩散开来，形成滔天巨浪。这里的关键问题在于，决定一些传播要素的传播力、传播景观的因素究竟是什么？比如说民粹主义思潮，为什么这些话语一时席卷全球的网络空间，构成了改写历史的重大社会思潮？再比如，有哪些力量在操纵着网络空间的信息传播？资本、政府、社会组织等利益主体在其中扮演着怎样的角色？

分析舆情传播特征的方法也多种多样。既可以从经典传播学的 5W 模型出发，描述信息传播过程与传播效果，也可以从网络分析和复杂网络分析的方法出发，分析社会网络和社会结构如何塑造信息传播的景观。

舆情大数据第四个重要特征是其社会属性。舆情大数据包含社会生活中各种各样的利益主体，涵盖社会系统中各个阶层，新闻舆情大数据描述和记录了社会生活中各个阶层和群体的生活方式与生存状态，而社交媒体大数据的参与主体也是各种各样。正是因为舆情大数据涵盖社会各阶层，包含多种事件，空间范围涵盖五湖四海。基于此，我们可以分析不同社会阶层与群体的生活方式与生存状态，可以运用阶层分析、利益主体分析、群体比较分析等多种传统社会科学的研究方法，分析不同群体的政治社会态度，分析社会各阶层之间的互动与博弈，分析国家与社会的关系模式，总结归纳社会运行逻辑和社会结构逻辑。

舆情大数据的第五个重要特征是其全球属性。网络无国界，舆情信息在全球层面越来越形成相互连通的局面，也有越来越多的社交媒体平台横跨全球多个国家，地球上某个地方发生的舆情事件极有可能波及遥远国度；同时，越来越多的舆情大数据，比如全球在线新闻舆情大数据 GDELT，汇聚了世界所有国家的舆情信息。

在这样的背景下，越来越多的舆情大数据具有全球性特征，为社会科学开展“环球航行”提供了观测数据的平台。正是因为这样，社会科学第一次可以借助于这些具有全球特征的数据库，对世界上多个国家开展比较研究，或者开展全球尺度的分析与研究，比如说 Golder 等人试图借助于 Twitter 的平台，分析欧洲、北美、非洲和大洋洲等多个地区人类情绪演变昼夜节律的全球普遍性。笔者认为，值得进一步深入分析的方向是，也可以从全球层面，分析不同文化背景下不同民族的生存方式与生活状态，分析地球不同角落大众的价值观和文化观念，开展大规模的跨文化比较研究。

在舆情大数据监测全球的大背景下，社会科学研究可以利用这些全球性的实证数据，对世界多个国家与社会开展实证分析，克服过去社会科学研究的“地方性知识”的局限，开展真正具有全球比较意义的实证分析。因此，将全球视野纳入社会科学的实证分析中来，通过对比多个社会系统的演变特征，或者将世界不同社会纳入同样的分析体系，或许能够为未来的社会科学研究拓展研究视野和开辟新的研究领域提供重要支撑。

未来时代大数据技术与企业决策相伴而生

来源 / 腾讯科技　编辑 / 协会会员处　李苗苗　日期 / 2021-01

大数据的核心价值是预测。大数据的关键并不在于数据的“大”，而在于挖掘其中隐藏的大价值。把数学算法运用到海量的数据上，通过分析与某事物相关的所有数据，大数据可以自动搜索出其中有价值的信号和模式，预测事情发生的可能性，进而作出准确研判。这种能力被视为人工智能，是大数据的核心。利用大数据对企业运营进行监控，可以帮助、指导企业对企业相关业务流程进行有效优化，帮助企业作出科学决策。企业“智商”的基础就是各式各样各类的数据。

未来，大数据与企业决策相伴而生

将来，数据将不再是简简单单的为决策服务的工具，大数据将和决策相伴相生，有大数据的地方就有决策的产生，甚至是不需要人类参与，大数据将根据之前人类设定好的参数自动化的制定策略，从而达到为人类服务的目的，当然，不管何时，都应以人为本大数据应以辅为主，可以看出，大数据对决策的影响是深远的，不仅仅是在当下，在了解了大数据对决策的影响，才能为企业决策提供一个方向，慢慢的改变，慢慢的适应，这样才能走的更远。

树立大数据思维意识

大数据时代的到来，对人类社会各个层面，特别是人们的决策思维带来巨大的影响。

一是看待大数据要用历史的眼光。相对于农业社会和工业社会，信息时代的生产要素与发展驱动力明显不同，大数据、云计算的应用将对社会发展产生前所未有革命性的颠覆性的影响，在推动社会进步的同时，对人的思维也将产生颠覆性影响，应积极跟上。

二是积极适应大数据时代，树立大数据决策思维。大数据不仅仅是先进信息技术的外在表现，更是对企业决策者世界观、方法论的改造。在新形势下，如若不及时转变决策思维则

会被时代说淘汰，要学会紧跟时代步伐，学会收集、分析、处理数据，挖掘数据潜在大价值，全方位高效地实施决策。

正确认知大数据价值与效益

处于经济社会中的任何人都期望中大数据中尽快挖掘出意想不到的“大价值”。对于企业决策而言，就是通过各类数据分析解决企业面临的困难或实现价值增值。

例如，非常经典的“啤酒 + 尿布湿”案例即告诉企业决策者，大数据中可以发现貌似极不相干的事件之间可能存在某种必然的有价值的商机值得去探索。在这之中，不是简单的看到二者之间的联系，而是通过现象发现问题存在的根本缘由，并通过缘由去发现新知识、新规律，事件相关性本身是无价值的。

大数据在重新定义企业“智商”同时，对企业核心资产同时也进行了重新定义，数据资产可以说必然成为现代商业社会的企业核心竞争力。当前，大数据已经在医疗、教育、零售、交通运输等行业得到应用，它的开发和利用将深刻改变传统行业的运作方式并大幅提高行业运作效率。未来，企业决策行为会日益基于数据分析而作出，大数据决策必将成为信息技术下一次重大突破的重要方向之一。

如何做好数据分析？

来源 / 一个数据玩家的自我修养　作者 / GClover　日期 / 2021-01

人人挂在嘴边的数据分析到底包含哪些方面？学好 Python 真的就能做好数据分析吗？

数据分析，拆开来看其实是几个方面：工具、理论、业务

工具，指的是我们从事数据分析所使用的具体工具，如 SQL、Excel、Python、R、SAS 等；

理论，指的是我们从事数据分析时所依赖的理论基础，如概率论、统计学、机器学习及相关的建模和分析框架；

业务，指的是数据分析落地的具体场景，输入和输出以及要解决的具体问题。

工具和理论都是比较容易速成的，这也是为什么各类网课主要集中在这些领域。

业务是依赖于在行业的经验，因此，转行最好先在同行业里面转，可以借用之前对于行业的业务理解快速上手。

以上三个方面固然重要，但并不是数据分析的全部，还需要再加一个维度，那就是思维模式。也就是说我们除了数据分析的工具、理论以及业务知识，还需要具备数据分析的思维。

那么什么叫做数据分析思维呢？

我认为可以分为三个方面：

01 定量思维

迪斯尼通过草坪规划道路的故事大家也许都听过：在迪斯尼乐园提前开放的半年里，草地被踩出许多小道，这些踩出的小道有宽有窄，优雅自然。第二年，格罗培斯让人按这些踩出的痕迹铺设了人行道。1971 年在伦敦国际园林建筑艺术研讨会上，迪斯尼乐园的路径设计被评为世界最佳设计。

迪斯尼后续还推出了 MagicBand，这个手环可以在园内支付，可作为酒店房卡，可以用来当 FastPass，可以用来停车等等，通过这些环节收集的数据，就可以知道哪几个项目最热门，哪几个项目不太热门，什么位置餐厅人满为患，说明还需要增加配置，什么地方餐厅无人问津，可能要做优化……时间一长，上述行为积累的数据就有了各种价值，看起来无法量化的东西，通过巧妙的收集数据，都可以量化。这就是数据思维第一条，万物皆可测。

02 相关思维

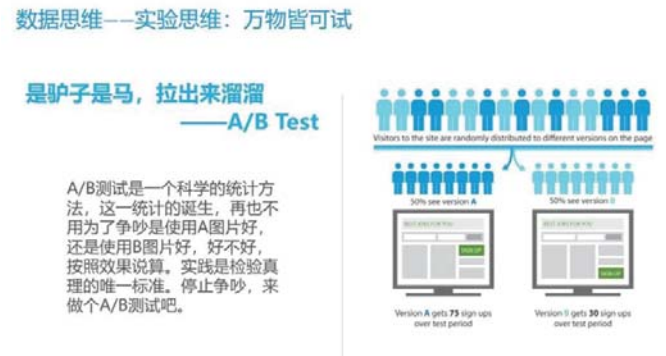
在大数据时代，随着算力的不断加强，原来小样本的计算已经可以升级为全样本计算，并且可以发现变量间的相关关系，用来代替原来小样本中推导出的因果关系。最经典的例子就是 08 年的 Google Flu：Google 流感趋势（Google Flu

Trends，GFT）是 Google 于 2008 年推出的一款预测流感的产品。Google 认为，某些搜索字词有助于了解流感疫情。Google 流感趋势会根据汇总的 Google 搜索数据，近乎实时地对全球当前的流感疫情进行估测一个搜索行为和一个疾病的发生，看似不相关的两件事情却存在强相关，这在原来是不可想象的。

不过，尽信数据不如无数据，一定要找到业务含义。就拿 Google Flu 来说，在研究成果公布以后，研究人员发现结果不再准确了。经过反复确认和调研，他们发现很多人得知了这项研究成果后，会抱着好奇的心态尝试搜索关键字——尽管他们周围并未出现相关病例，却导致了预测结果不再准确。

03 实验思维

告别拍脑袋决策，告别依赖个人审美决策，告别依赖个人经验决策。通过实际的数据表现来决策。同时根据实验结果不断的迭代和优化模型。当然，实验的前提是测量，必须先将所有实验的数据采集下来才能根据实验数据进行决策。



根据数据分析的结果，可能某些人群针对某个方案更加有效，这又会用到相关思维，即某些要素的相关性决定了最后的数据表现。通过以上三个思维模式，我们可以将实际中的业务问题进行拆解，从而转化为数据分析问题。这么说可能还是比较抽象，具体来看看如何应用。

在广告营销领域，有一个著名的说法：“营销行业的哥德巴赫猜想”：

这是相当长的一段时间广告营销行业最大的痛点：蒙着眼睛放广告。来了客户也不知道是广告带来的，还是自己找上门来的，或者其他渠道推荐来的。那么，用上数据分析思维的广告营销会变成什么样子呢？运用定量思维，那就是营销效果要可以度量。

一个广告投出去，我需要知道到底带来了多少转化，每个渠道的转化率怎样，以及这些客户的后续活跃程度如何，是不是假量？是不是羊毛党？是不是僵尸户？等等。那么如何度量呢？我们自然可以想到，要检测转化率，那就要对每个渠道进来的客户打标签，定期出报表，监控每个标签下客户的活跃情况等等，自然的就形成了客户分群经营，分群营销，分群活动投放等等策略。

运用相关思维，那就是通过相关性分析，使得广告的投放更加精准。减少无效的广告投放，在更相关的人群上投放他们感兴趣的广告，提升转化率，节省营销费用。

那么如何进行相关性分析呢？通过前期采集的数据，使用 Apriori 、Collaborative Filtering 等算法，找出用户特征、用户行为及其最终购买之前的相关关系，从而优化投放及推荐模型。运用实验思维，那就是通过实验，判断哪个投放模型更优，哪个投放渠道更优，同时根据反馈不断迭代和优化模型。那么如何进行实验呢？自然是通过 A/B Test 方法，随机均分流量到不同的投放模型上，同时采集客户的反馈，不断的根据反馈迭代和优化模型。

总的来说，做好数据分析，除了掌握工具、理论和业务，还需要具备数据分析的思维，有了数据分析的思维框架，更容易将业务、理论和工具贯通，形成自己的数据分析框架，更好、更有效的进行数据分析工作。

栅格地图在高纬度区域误差过大及修正方法

来源 / CPDA 数据分析师 刘程浩　编辑 / 协会会员处 李苗苗　日期 / 2021-01

前不久我和同事一起瞎聊，聊到制作数据可视化时，有个话题吸引了我的注意：如果不注意纬度的变化，高纬度地区地图上展示的栅格图，会存在因地图变形导致栅格内数据有很大误差的场景。

说者无心，听着有意，这大概就是一杯咖啡汇聚宇宙能量的某个具体表现吧。

说到栅格图，简单科普下，就是为了研究某个地理区域的各项指标分布状况，将该区域用若干的单位面积单元格覆盖，并通过色块的阶梯来呈现指标的分布情况。

例如下图，我采用了某个技术论坛上在百度地图上绘制的上海市区的一个栅格地图

在上面的栅格地图中，每个单元格可以表示 1×1km 的区域，或 N×Nkm 的区域。并且每个栅格的颜色由浅到深，可以表示很多指标的高低。例如人口密度、空气 PM2.5、用电量……等等。

由于栅格地图具有非常直观的多维地理呈现功能，因此单个栅格地图或多个栅格地图的对比使用，可以从数据中得到非常多的发现。

那我们聊天的话题中，为什么这个话题会这么吸引我呢？原因可能是受我中学时代当过地理课代表的原因吧。

当然了，由于我国国土大部分面积都处于中低纬度，我本次文章题目中出现的问题并不严重（准确的说地图绘制方法不同）。而且我们用栅格地图研究问题时，往往研究的是一个小范围区域，此时的地图变形误差影响并不大，也容易纠正。

但是当我们放眼全球时，特别的，在靠近两级的地方做研究时（例如俄罗斯的远东地区、智利狭长的跨中高纬度国土、阿根廷靠近南极圈区域等等），这种情况就比较明显了。而且由于地图的变形严重，用变形的地图来制作栅格地图，那么就相当站在巨大误差的肩膀上，得到的结果就匪夷所思了。

那为什么会出现地图变形呢？

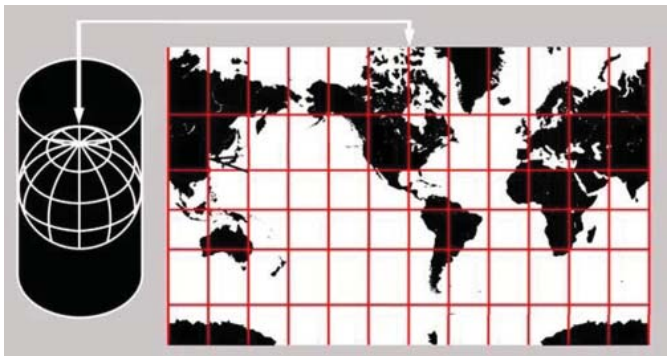
我们就回到中学地理来看看。

我们生活在地球，假设地球是个正球体。如果我们想把地球表面做成一个平面二维的地图，会发现因为正球体表面无法连续分割，导致做不出一个完整的地图。

另外，仔细观察地球仪，我们会发现高纬度地区，极点、经线和纬线闭合的区域其实是个曲面三角形。

如果按照经线把地球仪表面剥开并铺平，你是无法得到一个连续的地图的。那我们平时看到的世界地图是怎么画出来的？大部分用到的是圆柱投影法（又称墨卡托投影法 Mercatorprojection，或麦卡托投影法、正轴等角圆柱投影法）

假设用一张白纸卷成一个圆管，套在一个半透明地球仪上。地球仪的赤道和纸管相切。然后在这个半透明的地球仪中心放一个灯泡，点亮这个灯泡，那么地球仪表面的地图就会投影到这个纸管上面。



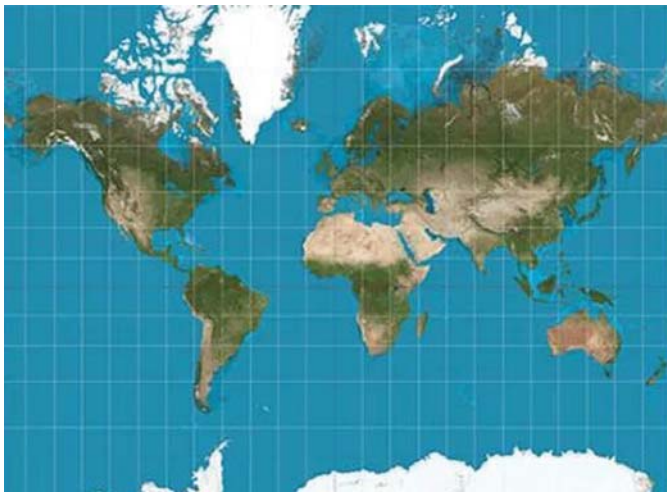
我们把纸管上留下的投影地图展开，就会得到我们常见的世界地图了。

这个世界地图有 2 个最大的特点：

1、每条纬线都一样长，和实际地球仪上存在较大的出入。

因为我们知道纬度越高，纬线圈就越小。

2、纬度越高，投影长度就被拉长的越厉害



这个出入就造成了地图变形。

也就是说，在投影法得到的地图上，纬度越高，看上去一定大小的地区，实际上在地球上是没有那么大的。就好像上图中的南极洲，它横向覆盖了一整条纬线，而我国在世界地图上的大小，才占据了 1/6 根纬线。而且整体看上去南极洲比我国大的多的多。但实际上南极洲只有 1400 万平方公里，只比我国大 40% 左右。同理，靠近北极圈的格陵兰岛，变形后看上去就很大，和南美洲差不多大。但实际上格陵兰岛只有 216 万平方公里，还不够我国的 1/4！因此，在这个投影法绘制的地图上，特别是高纬度地区实际面积放大了的前提下，用栅格图进行分析，就容易出问题。

纬度	纬度至赤道距离 KM	投影地图的放大倍数
0	0	1
15	1668	1.049
30	3336	1.122
45	5004	1.258
60	6672	1.549
75	8340	24.207

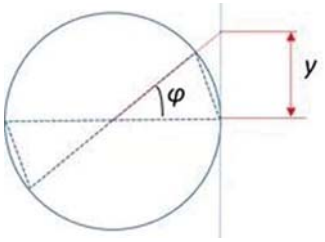
以上分析了问题的表现，以及问题的根本成因，那么如何进行修正呢？

我们采用了一个非测绘专业，但符合专业的数据分析师的思维的方法。

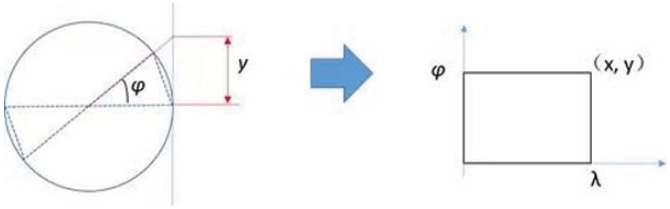
利用古德曼函数的逆推导得到面积缩放修正法。

1. 1. 古德曼函数简介

纬度实际上是一个球心角，假设不存在地球自转偏转角，赤道平面就是水平面。纬度实际上就是以赤道平面向上或向下的旋转的一个球心角 ϕ。这个球心角所对应的圆弧，在相切的平面上的笛卡尔投影为 y。



那么地理坐标经纬度（λ，ϕ）和平面笛卡尔坐标（x,y）之间的关系就如下图



通过古德曼函数的逆运算（也就是反函数），是可以得到纬度 ϕ 和地图上的纬度 y 之间的关系。

这个是古德曼函数

$$\varphi=2\tan^{-1}\left(e^{y}\right)-\frac{\pi}{2}$$
$$=\tan^{-1}\left(\sinh\left(y\right)\right)$$

我们做个对比，看看 2 种数值之间的差异，按地球平均半径 6371KM 计算

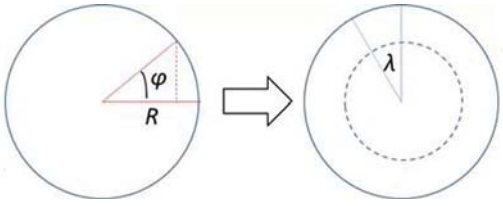
有了地球仪上的纬度 ϕ 和投影地图上的纬度 y 之间关系后，还不够，还需要考虑经度上的变形。

在这里我们就不搞那么复杂了，用一个简单的几何方法进行近似替代。

假如确定某个纬度 ϕ，那么在这个纬度上的纬线长度肯定比赤道短。

假设地球仪半径为 R，纬度 ϕ 上的纬线和赤道相比，这个比例很容易算出来

$$\frac{2\pi R\cos\varphi}{2\pi R}=\cos\varphi$$



之后就容易估算出地图上一个经纬网格的面积，和地球仪上一个经纬网格的变形比例了。

我们假设地球仪上一个正方形，纬度为 15°，经度为 15°（考虑每个时区 15°的实际业务场景），那么到了投影地图上，面积就会放大。

我们把具体的放大倍数做了一个渐进表，如下：

经度	纬度	经线等分长度 KM		经度0-15°上纬线长度 KM		面积KM²		
		地球仪	投影地图	地球仪	投影地图	投影地图	地球仪	比例
0-15	0-15	1668	1750	1611	1668	2918558	2734573	1.07
	15-30	1668	1872	1444	1668	3121926	2548217	1.23
	30-45	1668	2098	1179	1668	3498612	2188203	1.60
	45-60	1668	2584	834	1668	4309172	1679068	2.57
	60-75	1668	40376	432	1668	67343736	1934[illegible]2780	
	75-90	因无人或人烟稀少，不纳入分析范围						

这个表怎么看呢，其实就是看最右边的 3 列。

例如，在地球仪上 0-15°纬度和 0-15°经度包围的一个曲面四边形，它的表面积接近 2734573km^2，而在投影地图上，这个四边形的面积为 2918558km^2。两者差异不大。也就是说在地球仪上用 2734573 个 1×1km 的单元格做栅格，几乎能填满这个区域；而在投影地图上，则需要 2918558 个单元格。后者是前者的 1.07 倍。而在高纬度地区，例如 60-75°纬度地区，这个比例上升到了接近 64 倍！

所以，当我们用栅格图计算时，会发现高纬度地区的城市用的单元格多了，但是每个单元格下的指标，例如家庭密度一下子就小了很多。这样一来，同样是高纬度城市的市中心，就会造成这里也人烟稀少的情况，和业务实际差异很大。

因此有了这个比例倍数，我们在使用栅格图时，特别是计算总量或者单元格上的指标时，就可以根据城市所在的纬度区域，去缩小或放大相应的倍数。

当然了，地图使用的投影方法不同，我们计算得到的放大或缩小的倍数变化情况也不同，这个需要根据具体情况做调整。

我是如何在教培行业精进我的数据分析能力的

来源 / 中国商业联合会数据分析专业委员会特聘讲师 赵玉莲　编辑 / 协会会员处　李苗苗　日期 / 2021-01

先说说自身情况吧：09 年本科毕业，专业工商管理。在北京一家出版公司做金融财经方向的编辑和培训做了 5 年多（包括实习期）。

自从学校出来这 5 年半，变化在不经意间悄悄发生，不知道从哪一天开始我半夜不去人人开心网偷菜了，也不知从哪一天开始经常光顾百度糯米团购寻找各种吃喝玩乐了，互联网 + 的大风在我耳边不知不觉中越吹越大，越吹越响，越发觉得传统行业不借点风就要活不下去了……

有跳槽的念头是 2015 年 8 月，当时数据分析的风在我这个普通的打工人耳朵边只是轻轻的偶尔撩拨几下，当时我还是在前程无忧、智联上刷刷简历，求职目标也仅仅是更加聚焦在教培行业（没有出版这个传统业务模块的那种，为什么会这样？当时可能特别想拥抱互联网吧，去传统化？也许是）

如果你不觉得几年前的自己很傻，
那么说明这几年你没什么长进

与数据分析结缘

人有时候还真得信命，我手里拿着银行培训主管的 offer，在去入职的路上接到了 CPDA 课程设计的入职邀请，我在广渠门与朝阳门之间也不记得是哪个马路牙子上坐了下来，我需要静下来做这个决定，一个是自己经验可以延续的也是希望从事的职位，一个是没有专业背景加持，但吸引我去追的一个憧憬，选择的天平后来由于《发现数据之美》这本书而有了倾斜。

你遇到过面试还带送书的么？我遇到了！！！

在 CPDA 面完试，面试领导送了我这本书，让我回去看一看。

说实话，当年我回去翻了一遍这本发现数据美的书目录，然而我并没有发现什么美，因为我发现目录显示的内容很专业，虽然商科背景的我对字面内容不陌生，但要说心底的感觉，就是没啥感觉。

与众不同的面试体验，更容易激起人的好奇心和探索欲，最终的选择就是顺从人性，顺其自然，我来到了 CPDA 数据中心（当时还没有成立科研中心）。这么多年过去了，我曾留意观察过好多次，看领导会不会跟娱乐圈某晨批发石头一样批发一批图书，结果证明我想多了。为了避免领导看到这篇文章尴尬，必须要说一下，感谢领导发现我的潜力！！！

庆幸的很，入职之后一段时间我没有被逼去硬啃统计学相关知识，Excel 的报表关联，数据透视，SQL 语法，Python 或者 r，建模能力等等这些，也许领导知道，这不是我的强项，庆幸这样，不对，应该是感谢领导。

2015 年貌似招聘数据分析师的企业并不多，即使有在招聘数据分析师的，岗位职责也是千姿百态，出现频次比较高的词汇应该是“大数据”，关于这一点，当年我为了胜任自己的新工作还真有去招聘网站搜数据分析，新入职尽快熟悉公司业务，这个我还是自认为做的很到位的，而且这个办法今天也不过时，身边越来越多不知道数据分析做什么的朋友也都先去招聘网站看，这条经验大家随意拿走，各大行业通吃的妙招。

我的摸爬滚打

摸

2015 年下半年刚刚入职的我，最先面对的是一个数据分析培训项目，冀东水泥的数据分析内训，得嘞，我本是唐山人，对这些重工业不陌生，来吧，冀东水泥的需求，正好我也了解下数据分析到底怎么在企业中应用，我自己总是感觉命运很眷顾我，我需要什么，就来什么。当然也不乏有一些小插曲，因为我问冀东水泥的老乡，但老乡说他也说不清楚冀东水泥做数据分析内训的具体需求。怎么办？没有具体需求，就从有需求的人下手，谁在冀东水泥发起的这个需求，领导，领导是怎么有这个需求的？

请看这条新闻：

冀东水泥2015年亏损超17亿 系上市以来首亏

公告天天看2016年4月12日

2016年04月12日 07:03　来源：中国经济网

不记得当年是怎么打探领导要求做数据分析内训的原因的了，总而言之这条新闻说明了一切，这就是赤裸裸的有关数据分析的事实，谁会钱多到没地方花，要数据分析干什么？最根本的还不是要数据产生价值，解决实际面临的问题。

这是我摸到的关于数据分析影响最深，最重要的事情！

爬

CPDA 有一支与众不同的教研团队，由内部课程设计师与外部专家共同构成，当年我作为新入职的课程设计师，最先接触的工作就是与外部专家进行课程的沟通，印象比较深的有时任上市公司深圳雷柏科技副总裁王兴海博士；中国人民大学商学院 EMBA 市场与校友部部长胡旭老师；《数据分析 , 企业的贤内助》《活用数据：驱动业务的数据分析实战》的作者陈哲老师；以及从 CPDA 数据分析师，到高级数据经理再到国内知名互联网公司手游事业部海外发行数据总监的张炳出老师等。



我入职后面临的第一个挑战就是与这些老师的沟通无法深入，原因在于我对于数据分析的理解和知识架构不成体系，于是我先报名学习了 CPDA 课程，通过考试，继而又参加了对外经贸大学与慧科教育集团联合启动的”大数据分析与应用”硕士研修班。

这个过程用“爬”来形容还是很贴切的，在别人看不到的地方，一句话就能概括的学习过程背后是一点一点向自己想

要的目标挪动的过程，过程很枯燥，但不得不说，体系化的学习相当于我在数据分析领域爬过一遍，脑子里有了数据分析应用知识的大框架，并且形成了知识点和应用点之间的关联，这种学习方式让我在课程打磨中快速与老师建立起了良性的沟通，并且加速我从专家老师身上学习授课技巧，吸收数据分析工作经验和避坑指南，于是，我从课程设计师慢慢向讲师进化，这时距离我入职大概一年半的时间，第一次涨薪 20%。

滚

做讲师是我遇到的第二个挑战。说是挑战，主要是担心自己从讲台被动滚下去，因为这时候我的数据分析实战经验不多，课堂中的项目案例都是来自专家老师的经验，而数据分析课程最大的构成就是项目案例。

值得鼓励的是我在没有机会实践的时候，已经补充和学习了基础的理论，了解数据分析这个领域的知识体系和结构，也了解了这个领域所涉及到的关键技术。

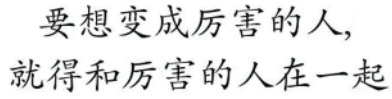
纸上谈兵有很多种，这个时候的我的纸上谈兵至少在理论上自圆其说，能够达到完整和严谨的标准，哈哈，主要是有别人的实践可参考，我后续实践的思路就是先模仿再超越，当然模仿本身没有错，但是不要知其然不知其所以然得模仿，不然模仿或实践完了还是一头雾水或无法贯通。

还是要炫耀一下我近水楼台先得月的优势，CPDA 是中国商业联合会数据分析专业委员会的一个数据人才认证项目，协会管理全国数据分析师事务所，但是受项目保密性限制，非项目组的成员没办法加入，怎么才能得到一些机会呢，主动请缨是上策，加入项目组，哪怕打杂呢，做的多，机会多，学的多，我也因此慢慢接触到东航、北京东城网格等数据项目。

数据分析一旦入了门，就会发现好玩的东西太多了，需要强调一下，我的入门可不是从会用 SPSS、SQL、r 和 Python 开始的，我是从数据最有吸引力的地方开始的，至于分析工具到现在我也是哪个方便用哪个，遇到操作卡壳了，百度一下基本都能解决，如果加点难度，就去找数据中心的工程师，数据体量不大的取数也就一行代码的工作量，一杯奶茶的心意就可以加深同事间的感情啦！

同时我还在网络上搜集特别好的数据分析案例和数据分析竞赛数据，看到优秀的文章后我会想尽办法联系上作者，好的案例分享者基本是数据分析领域的佼佼者，善于总结，乐于分享，因此我常常能要到脱敏数据，把案例重新跑一遍，这绝对是标准的模仿。

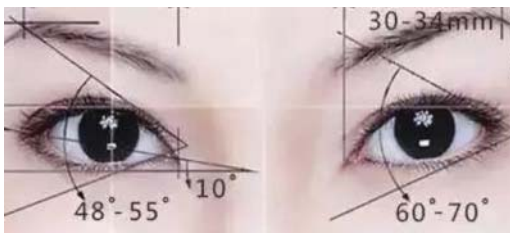
在这我也要特别感谢达观数据的高长宽老师，他是第一位乐于分享并给我脱敏数据支持我做进一步分析的人，在他的数据运营交流群得到很多数据运营有价值的东西，这波操作也让我意外发现数据分析的圈子其实很小，很多行业内的大咖都是有交集的，也特别感谢马世权老师、齐文光老师、吕雪芬老师、王鑫老师、冯艳宾老师、高松老师、夏龙老师、李妹老师等专业领域的高级选手，能与高手打交道，是一件十分难得的事情！



为什么要用“滚”形容这个阶段呢？我需要解释一下，不然容易产生误会。

可能用“滚动”更合适一些吧，不知道大家看到“滚”这个字，脑子里首先出现的是女朋友还是下坡的“滚动”感，我是积极向上的正经“滚”，是有牵引的“滚动“，是上坡，是动起来就能越来越省力气的那种，因为这个阶段的我真的有种得心应手的感觉了，感觉有根线在牵引着，动起来了，这个无形的线就是数据分析行业的力量，这时距离我入职大概三年的时间，第二次涨薪，不包括授课薪酬，那是单独额外结算哦，不过比授课课酬更有价值的是给未来带来了无限机会。无论是”前景“还是”钱景“，搭上数据分析这个技能的都算得上是”白富美，高富帅“的职位！

看过我这篇文章底稿的朋友说我凡尔赛，怎么那么会炫耀，但是我真的从心底觉得自己数据分析的入门起点很高（不拉仇恨），看我发现数据美的眼睛：



如果对一件事的了解不深、不透，总是浅尝辄止，那自然体会不到这件事的妙处。如果能长期坚持去做一件事，一定是这件事带来的丰盈感和满足感超过了我的所有付出，一定是这件事日日夜夜萦绕在我的心头让我欲罢不能，一定是这件事

唤起了我内心深处最强烈的兴趣。

我倾心在教育培训领域的职业理想一直未变，ABB 集团中国区总裁顾纯元先生曾得到过高人的指点：“如果你想到一家大的企业发展，首先自己要有一个平台，在这个平台上把工作做好，之后就会很顺。在一家大的企业，有两个职位是平台：一个是销售，一个是研发。在这两个平台职位上，如果要做好就必然要跟公司的各个部门打交道。销售方面要关心工厂和生产，要给研发反馈，要与财务等部门跟进沟通，研发也一样。”

从教育培训行业来看，两个平台型的职位是销售和教研。我从作为课程设计师入职到目前的科研中心管理岗，部门职位的平台特性越来越明显，向外接触学员，向内需要与专家老师、教研、教务、管理人员进行沟通，负责学员沟通答疑、教学教案设计、授课排课等多项工作，平台特性反超一线销售咨询。有了数据分析技能的加持，科研中心与数据中心的工作交集越来越大。

在实际工作中经常做的一件工作就是找可以驱动业务的数据指标。对于不同的业务需求，会有不同的衡量方法，同时也可能会有好几个，当 CPDA 说自己发展的好会说有多少净值用户数，但是如果想要把产品进一步提高，净值用户数作用就不大了，因为这是结果，不是手段，需要找到这么一个或几个指标，是跟长期目标发展相关的，同时又能通过驱动这些指标达到长期发展的目的。比如，CPDA 招生在最初期，市场广告投入带来多少意向学员（流量）是一个很重要的指标，流量进来就更会关注转化率，而有了用户积累之后又细化出了老带新转化率指标，作为科研中心更关注用户需求评价指标，如此各种。目的就是发现和识别业务问题，为业务问题排查原因并提供解决方案。

我是深知专注于单一技能的工匠型、专业化发展之路不是理想路线，毕竟岁月催人老，“小赵”做成“老赵”也是一种遗憾，毕竟我不是可以在专业化道路上走向极致的那少量的天才。社会越发展，分工越专业，一个人能做的事会越来越有限。走上管理岗是我人格转型和能力提升不可缺少的关键一步，数据化管理让我随时随地了解我想知道的数据，做到对业务状况了如指掌，还能够根据经营状况有依据的给出下一步发展方向的决策建议。

即使有一天我放弃教育培训，与几位有共同理想的 CPDA 的学员借着数据分析的东风成立一家数据分析师事务所，凭着自己做过这些含金量超高的岗位，招聘核心员工的高昂成本可以省下不少，还能提升我对项目管理的掌握能力，最重要的是数据化管理也会大大降低决策失误，是不是。

数据时代，我们企业、我们个人都在一场无硝烟的数据战场上打拼！我“摸爬滚打”到今天，相信插上数据分析的翅膀，我们都能越飞越高……本故事纯属真实，如有雷同，可来认亲！

大数据分析工具 BI 的应用

来源 / 51CTO　编辑 / 协会会员处　李苗苗　日期 / 2021-01

大数据分析工具 BI，是企业数据化管理的一整套方案，用于将企业中现有的数据进行有效整合，快速准确的提供决策依据，帮助企业做出明智的业务经营决策，解决企业管理问题。

大数据分析工具 BI 的应用

通信大数据行程卡

新冠疫情的发生，给人们的生活带来了不小的影响。直到如今，疫情的阴霾仍未消散，所以防疫工作不容松懈。而利用大数据监控每个人的行动轨迹对疫情防控是一个非常有必要的措施。当我们进出居民小区、企业园区、商场超市、机场车站等公共场所时，会用手机扫描一种“通信大数据行程卡”的二维码，在信息码服务中申报行程即可查询和证明本人近 14 天的到访地。

大数据能够准确识别用户是否到过疫情风险区，并能够溯查感染者相关的接触人员，有效支撑疫情精准防控，这是大数据在疫情防控中的一个典型应用。

那，什么是大数据呢？

随着全球数字化、网络宽带化、互联网应用于各行各业，累积的数据量越来越大，这些短时间内我们很难收集并利用的海量数据就是大数据。通过大数据分析能够获取很多智能、深入、有价值的信息，并提供决策支撑，能够为我们的生产生活、经营管理、社会治理、民生服务等各方面带来高效、便捷、精准的服务。

大数据在抗击疫情场景中的应用，除了刚刚提到的“通信大数据行程卡”之外，还体现在远程会诊、无接触式快检、智能语音助手、疫情动态和预警、提升医疗物资供应效率、物资供求信息精准对接、发展“非接触式”服务模式等方面。目前，大数据分析已是潮流，在各行各业中应用渐广。

01. 大数据支持智慧城市建设

随着全球城市化进程的加速，环境污染、交通拥堵等城市病随之到来，城市发展面临着巨大的挑战。为了应对这些难题，各国不约而同的以智慧城市建设作为解决问题的抓手。大数据在智慧城市中的落脚点是为智慧城市的智慧交通、智慧医疗、智慧生活等各个领域提供强大的决策支持，科学治理城市。

在智慧交通系统中，大数据通过对道路、车辆、天气、行人等大量交通信息的实时挖掘，能有效缓解交通拥堵，并快速响应突发状况，为城市交通的良性运转提供科学的决策依据。在智慧安防系统中，大数据通过平安城市、智能交通管理、环境保护、危化品运输监控、食品安全监控等信息挖掘，可以及时发现人为或自然灾害、恐怖事件，提高应急处理能力和安全防范能力等。在智慧城管系统中，大数据通过对不同时间段、不同区域、不同部门获得的大量监测数据进行实时采集、实时处理及深度挖掘，实现对城市管理实时监控与长期管理优化。

02. 大数据服务电子商务

大数据在电子商务行业的应用主要是在精准营销和个性化服务方面。

大数据支持下的营销核心在于，让企业的业务在合适的时间，通过合适的载体，以合适的方式，推送给需要此业务的用户。互联网企业使用大数据技术采集有关客户的各类数据，并通过大数据分析建立“用户画像”来抽象地描述一个用户的信息全貌，从而可以对用户进行精准营销和广告投放等。

随着电子商务规模的不断扩大，商品数量和种类快速增长，顾客需要花费大量的时间才能找到自己想买的商品。个性化推荐系统通过分析用户的行为，包括反馈意见、购买记录和社交数据等，以分析和挖掘顾客与商品之间的相关性，从而发现用户的个性化需求、兴趣等，然后将用户感兴趣的信息、产品推荐给用户。个性化推荐系统针对用户特点及兴趣爱好进行商品推荐，能有效地提高电子商务系统的服务能力，从而保留客户。

03. 大数据助力企业数字化转型

近年来，随着数字化转型的兴起，大数据已成为企业管理水平提升的主要推动力。数字化企业有三个特征：互联、精细、智能，从数据的角度来讲，则是产生大数据、利用大数据、挖掘大数据的过程，通过实时分析大量结构化和非结构化数据以获取见解能力，推动企业工作流程的数字化和产线的智能化。此外，大数据还可以帮助企业更好地了解客户的偏好和行为，从而创造更多个性化相关体验，并引入基于洞察力的产品和服务扩大企业营收，提升和延长企业价值链。

大数据成功应用落地的关键在于与企业日常运营的深度融合，由于很多企业业务涵盖范围广、信息系统繁杂、车间现场设备种类多，还存在各种形式的信息孤岛，要求企业构建一个模块化但具有凝聚力的数字平台，并以从各种来源收集的大数据作为分析动力。

目前，大数据分析工具 BI(BusinessIntelligence，商业智能) 在企业中的应用渐广。BI 是企业数据化管理的一整套方案，用于将企业中现有的数据进行有效整合，快速准确的提供决策依据，帮助企业做出明智的业务经营决策，解决企业管理问题。此外，BI 也可以生成各种分析预测报表、KPI 数据，方便高层管理者及时了解企业的业绩、市场、研发、制造等各方面所需的信息。

提供 BI、数据分析、数据可视化、大数据等方面的咨询与干货，开拓 BI 新观，让数据分析真正成为有意义的洞察。

AWS 宣布推出财务数据分析工具 Amazon FinSpace

来源 / siliconANGLE　编辑 / 协会会员处　李苗苗　日期 / 2021-01

亚马逊网络服务公司（Amazon Web Services Inc.）瞄准了金融分析师，为他们推出了一款专门设计的、新的数据分析工具。亚马逊表示这款新的工具将使他们能够比以往更有效地完成工作。

Amazon FinSpace 是今天发布的一项新服务，其工作原理是汇总、分类然后标记来自各种不同来源的数据，让客户组织内部的任何人都可以更轻松地对其进行搜索。

该工具是专门针对对冲基金、资产管理公司、保险公司等公司内的分析师设计的。它为他们提供了一种更简便的方法，让他们能够按照需要对手头上的各种数据进行分析，这些数据的来源既包括内部的来源——例如他们投资组合管理系统、订单管理系统和执行管理系统内的数据，也包括第三方馈送的数据——例如海量的股票历史价格数据、就业数据和收益报告等。

这家云计算巨头表示，Amazon FinSpace 很有必要，因为发现和准备分析需要的数据是一项非常耗时费力的工作，而且还非常复杂。数据财务分析师依赖的数据通常分散在多个部门之中，这些数据通常非常具体。

而且，对这些数据的访问通常都有非常严格的控制。一旦被授予了访问权限，分析师们就必须手动准备数据，并完成从数据中获取见解所必须的转换工作。AWS 表示，Amazon FinSpace 可以负责所有这些收集的工作，让数据可以立即被用于分析。

亚马逊的开发者“传教士”Sebastien Stormacq 在一篇博客文章中表示，它所提供的见解也同样出色。他解释说，金融分析师通常依赖布林线（Bollinger Bands）和指数平滑曲线（Exponential Moving Averages）之类的技术指标来识别趋势和模式，但是，他们用来执行此项操作的传统工具并非专门针对云规模的海量分析。

结果，分析师们通常只好依赖比较小的代表性数据集，而这种做法会限制他们的预测能力，或者他们可以手动将数据分解为较小的子集，把它们零碎化，然后再手动进行重组。Stormacq 表示，借助 Amazon FinSpace，他们可以使用比以往大得多的数据集，这意味着从总体上来说，能够获得更准确的见解。

使用 Amazon FinSpace，数据既可以通过 FinSpace 应用程序编程界面也可以通过基于 web 应用程序拖放界面被摄入到系统之中。然后这些数据会被以易于查找和共享的方式进行转换和组织。分析师们要做的一切只是浏览可视目录或者使用熟悉的业务术语进行搜索——例如“最近三年的期权交易”，然后他们就可以访问自己所需的信息了。

该服务依赖于 Apache Spark 分析引擎来完成数据转换。为了满足合规性要求，用户可以直接在服务中定义数据访问策略，这些策略将在数据搜索、可视化和分析等环节被执行。该服务还会跟踪数据的使用情况，并可以生成合规性和活动报告，以显示谁访问了特定的数据集以及他 / 她是在什么时候访问的。

Deloitte Touche Tohmatsu Ltd. 首席和金融服务数据与分析负责人 Jojy Mathew 表示，他的公司坚信基于云的分析的潜力，因为它使得几年前无法回答的问题变得可能得到答案了。

Mathew 表示：“当企业在构建算法、随即和预测模型的时候，大数据集非常关键。”他表示：“Amazon FinSpace 将允许用户以需要的规模处理 PB 级的数据。此外，FinSpace 还可以快速创建‘分析沙箱’，将高级分析功能带给平民数据科学家们。”

Constellation Research Inc. 的分析师 Holger Mueller 对 SiliconANGLE 表示，公共云正朝着针对特定行业提供垂直服务的方向发展，而对信息技术有着较高支出的金融服务则是其中的重点领域。

Mueller 表示：“AWS 将整个五月份专门用于金融服务领域，从逻辑上讲，它是从一种解决方案开始的，该解决方案解决了金融机构难以应对的大量数据泛滥的问题。”他表示：“他们需要克服数据延迟、数据重力和数据出口成本等挑战。借助 Amazon FinSpace，AWS 正在应对这方面的多项挑战，不仅仅是数据管理方面的挑战，还有赋能方面的挑战，该公司希望降低平民数据科学家的使用门槛。”

Amazon FinSpace 服务现在已经在美国东部的弗吉尼亚州和俄亥俄州、美国西部的俄勒冈州、加拿大中部和欧洲地区（爱尔兰）全面上市，不久之后，这项服务还将在更多地区上市。该服务的价格依据存储的数据量、用户数量和完成数据处理和分析所使用的计算能力来计算。

商品零售购物篮分析实战案例：Apriori 关联规则算法

来源 / 犀数苑　编辑 / 协会会员处　李苗苗　日期 / 2021-01

购物篮分析是通过发现顾客在一次购买行为中放入购物篮中不同商品之间的关联，研究顾客的购买行为，从而辅助零售企业制定营销策略的一种数据分析方法。

本案例使用 Apriori 关联规则算法实现购物篮分析，发现超市不同商品之间的关联关系，并根据商品之间的关联规则制定销售策略。

一、目标

通过对商场销售数据进行分析，得到顾客的购买行为特征，并根据发现的规律而采取有效的行动，制定商品摆放、商品定价、新商品采购计划，对增加销量并获取最大利润有重要意义。请根据提供的数据实现以下目标：

1. 构建零售商品的 Apriori 关联规则模型，分析商品之间的关联性。
2. 根据模型结果给出销售策略。

二、分析方法

购物篮关联规则挖掘的主要步骤如下：

1. 对原始数据进行数据探索性分析，分析商品的热销情况与商品结构。
2. 对原始数据进行数据预处理，转换数据形式，使之符合 Apriori 关联规则算法要求。
3. 在步骤 2 得到的建模数据基础上，采用 Apriori 关联规则算法调整模型输入参数，完成商品关联性分析。
4. 结合实际业务，对模型结果进行分析，根据分析结果给出销售建议，最后输出关联规则结果。

三、数据探索分析

查看数据特征以及对商品热销情况和商品结构进行分析

1、数据特征

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

%matplotlib inline
```

In [2]:

```
inputfile = '/home/kesci/input/data_act_combat5529/GoodsOrder.csv'  # 输入的数据文件
data = pd.read_csv(inputfile,encoding = 'gbk')  # 读取数据
data .info()  # 查看数据属性
print("-"*40)

print(' 描述性统计结果：\n',data.describe().T)  # 输出结果
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43367 entries, 0 to 43366
Data columns (total 2 columns):
id      43367 non-null int64
Goods   43367 non-null object
dtypes: int64(1), object(1)
memory usage: 677.7+ KB
----------------------------------------
描述性统计结果：
        count       mean        std  min   25%   50%   75%   max
id  43367.0  4908.589504  2843.118248  1.0  2455.5  4828.0  7380.5  9835.0
```

In [3]:

```
data.head()
```

Out[3]:

	ID	Goods
0	1	柑橘类水果
1	1	人造奶油
2	1	即食汤
3	1	半成品面包
4	2	咖啡

2、分析热销商品

销量排行前 10 商品的销量及其占比

In [4]:

```
group = data.groupby(['Goods']).count().reset_index()  # 对商品进行分类汇总
```

group_sorted = group.sort_values('id',ascending=False)

print('销量排行前10商品的销量:\n', group_sorted[:10]) # 排序并查看前10位热销商品

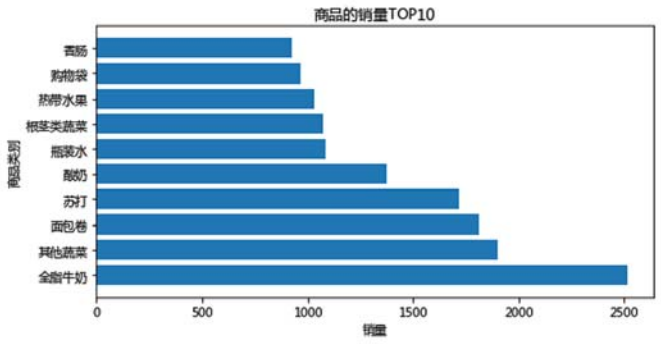
销量排行前10商品的销量:

	Goods	id
7	全脂牛奶	2513
8	其他蔬菜	1903
155	面包卷	1809
134	苏打	1715
150	酸奶	1372
99	瓶装水	1087
70	根茎类蔬菜	1072
85	热带水果	1032
143	购物袋	969
160	香肠	924

画条形图展示出销量排行前10商品的销量

In [5]:

```
x=group_sorted[:10]['Goods']
y=group_sorted[:10]['id']
plt.figure(figsize = (8, 4))
plt.barh(x,y)
plt.xlabel('销量')
plt.ylabel('商品类别')
plt.title('商品的销量TOP10')
plt.savefig('./top10.png')
```



In [6]:

```
# 销量排行前10商品的销量占比
data_nums = data.shape[0]
for idnex, row in group_sorted[:10].iterrows():
    print(row['Goods'],row['id'],row['id']/data_nums)
```

全脂牛奶 2513 0.05794728710770863
其他蔬菜 1903 0.0438812922268084
面包卷 1809 0.04171374547466968
苏打 1715 0.039546198722530956
酸奶 1372 0.031636958978024765
瓶装水 1087 0.025065141697604168
根茎类蔬菜 1072 0.024719256577582033
热带水果 1032 0.023796896257523
购物袋 969 0.022344178753430026
香肠 924 0.021306523393363617

In [7]:

```
inputfile1 = '/home/kesci/input/data_act_combat5529/GoodsOrder.csv'
inputfile2 = '/home/kesci/input/data_act_combat5529/GoodsTypes.csv'

# 读入数据
data = pd.read_csv(inputfile1,encoding = 'gbk')
types = pd.read_csv(inputfile2,encoding = 'gbk')

group = data.groupby(['Goods']).count().reset_index()
sort = group.sort_values('id',ascending = False).reset_index()

data_nums = data.shape[0] # 总量
del sort['index']

# 合并两个 datafreame,on='Goods'
sort_links = pd.merge(sort,types)

# 根据类别求和，每个商品类别的总量，并排序
sort_link = sort_links.groupby(['Types']).sum().reset_index()
sort_link = sort_link.sort_values('id',ascending = False).reset_index()
del sort_link['index'] # 删除“index”列

# 求百分比，然后更换列名，最后输出到文件
sort_link['count'] = sort_link.apply(lambda line: line['id']/data_nums,axis=1)
sort_link.rename(columns = {'count':'percent'},inplace = True)
print('各类别商品的销量及其占比:\n',sort_link)

# 保存结果
outfile1 = './percent.csv'
sort_link.to_csv(outfile1,index = False,header = True,encoding='gbk')
```

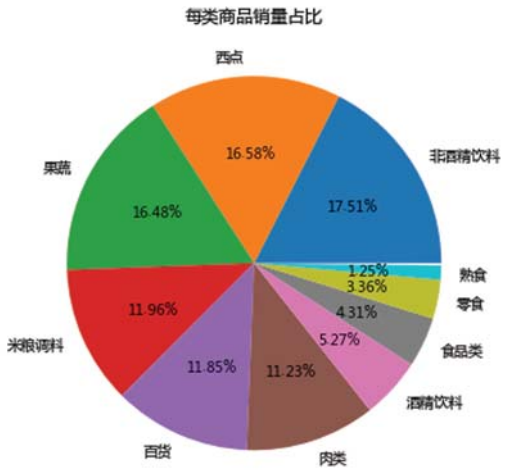
各类别商品的销量及其占比:

```
   Types   id   percent
0 非酒精饮料  7594  0.175110
1    西点  7192  0.165840
2    果蔬  7146  0.164780
3  米粮调料  5185  0.119561
4    百货  5141  0.118546
5    肉类  4870  0.112297
6  酒精饮料  2287  0.052736
7   食品类  1870  0.043120
8    零食  1459  0.033643
9    熟食   541  0.012475
```

画饼图展示每类商品销量占比

In [8]:

```
data = sort_link['percent']
labels = sort_link['Types']
plt.figure(figsize=(8, 6))
plt.pie(data,labels=labels,autopct='%1.2f%%')
plt.title(' 每类商品销量占比 ')
plt.savefig('./persent.png') # 把图片以 .png 格式保存
```



通过分析各类别商品的销量及其占比情况可知，非酒精饮料、西点、果蔬 3 类商品的销量差距不大，占总销量的 50% 左右。

进一步查看销量第一的非酒精饮料类商品的内部商品结构，并绘制饼图显示其销量占比情况

In [10]:

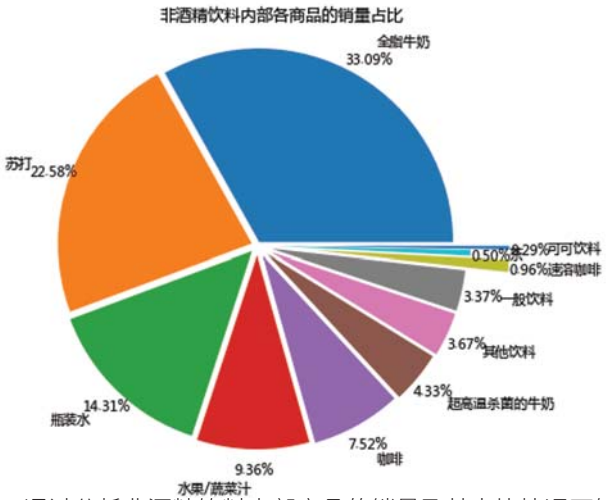
```
# 先筛选“非酒精饮料”类型的商品，然后求百分比，然后输出结果到文件。
selected = sort_links.loc[sort_links['Types'] == ' 非酒精饮料 ']
# 对所有的“非酒精饮料”求和
child_nums = selected['id'].sum()
# 求百分比
selected.loc[:,'child_percent'] = selected.apply(lambda line: line['id']/child_nums,axis = 1)
selected.rename(columns = {'id':'count'},inplace = True)
print(' 非酒精饮料内部商品的销量及其占比 :\n',selected)
outfile2 = './child_percent.csv'
sort_link.to_csv(outfile2,index = False,header = True,encoding='gbk') # 输出结果
```

非酒精饮料内部商品的销量及其占比 :

```
    Goods  count  Types  child_percent
0     全脂牛奶   2513  非酒精饮料       0.330919
3      苏打   1715  非酒精饮料       0.225836
5     瓶装水   1087  非酒精饮料       0.143139
16   水果 / 蔬菜汁    711  非酒精饮料       0.093627
22     咖啡    571  非酒精饮料       0.075191
38  超高温杀菌的牛奶    329  非酒精饮料       0.043324
45    其他饮料    279  非酒精饮料       0.036740
51    一般饮料    256  非酒精饮料       0.033711
```

```
101    速溶咖啡    73 非酒精饮料      0.009613
125      茶    38 非酒精饮料      0.005004
144    可可饮料    22 非酒精饮料      0.002897
In [11]:
# 画饼图展示非酒精饮品内部各商品的销量占比
data = selected['child_percent']
labels = selected['Goods']

plt.figure(figsize = (8,6))
# 设置每一块分割出的间隙大小
explode = (0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.08,0.3,0.1,0.3)
plt.pie(data,explode = explode,labels = labels,autopct = '%1.2f%%',
        pctdistance = 1.1,labeldistance = 1.2)
# 设置标题
plt.title(" 非酒精饮料内部各商品的销量占比 ")
# 把单位长度都变的一样
plt.axis('equal')
 # 保存图形
plt.savefig('./child_persent.png')
```



通过分析非酒精饮料内部商品的销量及其占比情况可知，全脂牛奶的销量在非酒精饮料的总销量中占比超过 33%，前 3 种非酒精饮料的销量在非酒精饮料的总销量中的占比接近 70%，这就说明大部分顾客到店购买的饮料为这 3 种，而商场就需要时常注意货物的库存，定期补货。

四、数据预处理

前面对数据探索分析发现数据完整，并不存在缺失值。建模之前需要转变数据的格式，才能使用 Apriori 函数进行关联分析。这里对数据进行转换。

In [12]:

```
inputfile='/home/kesci/input/data_act_combat5529//GoodsOrder.csv'
data = pd.read_csv(inputfile,encoding = 'gbk')

# 根据 id 对“Goods”列合并，并使用“，”将各商品隔开
data['Goods'] = data['Goods'].apply(lambda x:','+x)
data = data.groupby('id').sum().reset_index()

# 对合并的商品列转换数据格式
data['Goods'] = data['Goods'].apply(lambda x :[x[1:]])
data_list = list(data['Goods'])

# 分割商品名为每个元素
data_translation = []
for i in data_list:
  p = i[0].split(',')
  data_translation.append(p)
print(' 数据转换结果的前 5 个元素：\n', data_translation[0:5])
```

数据转换结果的前 5 个元素：

```
[[' 柑橘类水果 ', ' 人造黄油 ', ' 即食汤 ', ' 半成品面包 '], [' 咖啡 ', ' 热带水果 ', ' 酸奶 '], [' 全脂牛奶 '], [' 奶油乳酪 ', ' 肉泥 ', ' 仁果类水果 ', ' 酸奶 '], [' 炼乳 ', ' 长面包 ', ' 其他蔬菜 ', ' 全脂牛奶 ']]
```

五、模型构建

本案例的目标是探索商品之间的关联关系，因此采用关联规则算法，以挖掘它们之间的关联关系。关联规则算法主要用于寻找数据中项集之间的关联关系，它揭示了数据项间的未知关系。基于样本的统计规律，进行关联规则分析。根据所分析的关联关系，可通过一个属性的信息来推断另一个属性的信息。当置信度达到某一阈值时，就可以认为规则成立。

Apriori 算法是常用的关联规则算法之一，也是最为经典的分析频繁项集的算法，它是第一次实现在大数据集上可行的关联规则提取的算法。除此之外，还有 FP-Tree 算法，Eclat 算法和灰色关联算法等。本案例主要使用 Apriori 算法进行分析。

模型具体实现步骤 :

· 设置建模参数最小支持度、最小置信度，输入建模样本数据

· 采用 Apriori 关联规则算法对建模的样本数据进行分析，以模型参数设置的最小支持度、最小置信度以及分析目标作为条件，如果所有的规则都不满足条件，则需要重新调整模

型参数，否则输出关联规则结果。

目前，如何设置最小支持度与最小置信度并没有统一的标准。大部分都是根据业务经验设置初始值，然后经过多次调整，获取与业务相符的关联规则结果。本案例经过多次调整并结合实际业务分析，选取模型的输入参数为：最小支持度 0.02、最小置信度 0.35。其关联规则代码如代码所示。

In [13]:

```
from numpy import *

def loadDataSet():
    return [['a', 'c', 'e'], ['b', 'd'], ['b', 'c'], ['a', 'b', 'c', 'd'], ['a', 'b'], ['b', 'c'], ['a', 'b'],
            ['a', 'b', 'c', 'e'], ['a', 'b', 'c'], ['a', 'c', 'e']]

def createC1(dataSet):
    C1 = []
    for transaction in dataSet:
        for item in transaction:
            if not [item] in C1:
                C1.append([item])
    C1.sort()
    # 映射为 frozenset 唯一性的，可使用其构造字典
    return list(map(frozenset, C1))

# 从候选 K 项集到频繁 K 项集（支持度计算）
def scanD(D, Ck, minSupport):
    ssCnt = {}
    for tid in D:  # 遍历数据集
        for can in Ck:  # 遍历候选项
            if can.issubset(tid):  # 判断候选项中是否含数据集的各项
                if not can in ssCnt:
                    ssCnt[can] = 1  # 不含设为 1
                else:
                    ssCnt[can] += 1  # 有则计数加 1
    numItems = float(len(D))  # 数据集大小
    retList = []  # L1 初始化
    supportData = {}  # 记录候选项中各个数据的支持度
    for key in ssCnt:
        support = ssCnt[key] / numItems  # 计算支持度
        if support >= minSupport:
            retList.insert(0, key)  # 满足条件加入 L1 中
            supportData[key] = support
    return retList, supportData

def calSupport(D, Ck, min_support):
    dict_sup = {}
    for i in D:
        for j in Ck:
            if j.issubset(i):
                if not j in dict_sup:
                    dict_sup[j] = 1
                else:
                    dict_sup[j] += 1
    sumCount = float(len(D))
    supportData = {}
    relist = []
    for i in dict_sup:
        temp_sup = dict_sup[i] / sumCount
        if temp_sup >= min_support:
            relist.append(i)
            # 此处可设置返回全部的支持度数据（或者频繁项集的支持度数据）
            supportData[i] = temp_sup
    return relist, supportData

# 改进剪枝算法
def aprioriGen(Lk, k):
    retList = []
    lenLk = len(Lk)
    for i in range(lenLk):
        for j in range(i + 1, lenLk):  # 两两组合遍历
            L1 = list(Lk[i])[:k - 2]
            L2 = list(Lk[j])[:k - 2]
            L1.sort()
            L2.sort()
            if L1 == L2:  # 前 k-1 项相等，则可相乘，这样可防止重复项出现
                # 进行剪枝（a1 为 k 项集中的一个元素，b 为它的所有 k-1 项子集）
                a = Lk[i] | Lk[j]  # a 为 frozenset() 集合
                a1 = list(a)
                b = []
                # 遍历取出每一个元素，转换为 set，依次从 a1 中剔除该元素，并加入到 b 中
                for q in range(len(a1)):
                    t = [a1[q]]
                    tt = frozenset(set(a1) - set(t))
                    b.append(tt)
                t = 0
                for w in b:
                    # 当 b（即所有 k-1 项子集）都是 Lk（频繁的）
```

的子集，则保留，否则删除。

```
                if w in Lk:
                    t += 1
            if t == len(b):
                retList.append(b[0] | b[1])
    return retList

def apriori(dataSet, minSupport=0.2):
# 前 3 条语句是对计算查找单个元素中的频繁项集
    C1 = createC1(dataSet)
    D = list(map(set, dataSet))  # 使用 list() 转换为列表
    L1, supportData = calSupport(D, C1, minSupport)
    L = [L1]  # 加列表框，使得 1 项集为一个单独元素
    k = 2
    while (len(L[k - 2]) > 0):  # 是否还有候选集
        Ck = aprioriGen(L[k - 2], k)
        Lk, supK = scanD(D, Ck, minSupport)  # scan DB to get Lk
        supportData.update(supK)  # 把 supk 的键值对添加到 supportData 里
        L.append(Lk)  # L 最后一个值为空集
        k += 1
    del L[-1]  # 删除最后一个空集
    return L, supportData  # L 为频繁项集，为一个列表，1，2，3 项集分别为一个元素

# 生成集合的所有子集
def getSubset(fromList, toList):
    for i in range(len(fromList)):
        t = [fromList[i]]
        tt = frozenset(set(fromList) - set(t))
        if not tt in toList:
            toList.append(tt)
            tt = list(tt)
            if len(tt) > 1:
                getSubset(tt, toList)

def calcConf(freqSet, H, supportData, ruleList, minConf=0.7):
    for conseq in H:  # 遍历 H 中的所有项集并计算它们的可信度值
        conf = supportData[freqSet] / supportData[freqSet - conseq]  # 可信度计算，结合支持度数据
        # 提升度 lift 计算 lift = p(a & b) / p(a)*p(b)
        lift = supportData[freqSet] / (supportData[conseq] * supportData[freqSet - conseq])
        if conf >= minConf and lift > 1:
            print(freqSet - conseq, '-->', conseq, ' 支持度 ', round(supportData[freqSet], 6), ' 置信度：', round(conf, 6), 'lift 值为：', round(lift, 6))
            ruleList.append((freqSet - conseq, conseq, conf))

# 生成规则
def gen_rule(L, supportData, minConf = 0.7):
    bigRuleList = []
    for i in range(1, len(L)):  # 从二项集开始计算
        for freqSet in L[i]:  # freqSet 为所有的 k 项集
            # 求该三项集的所有非空子集，1 项集，2 项集，直到 k-1 项集，用 H1 表示，为 list 类型，里面为 frozenset 类型，
            H1 = list(freqSet)
            all_subset = []
            getSubset(H1, all_subset)  # 生成所有的子集
            calcConf(freqSet, all_subset, supportData, bigRuleList, minConf)
    return bigRuleList

if __name__ == '__main__':
    dataSet = data_translation
    L, supportData = apriori(dataSet, minSupport = 0.02)
    rule = gen_rule(L, supportData, minConf = 0.35)
frozenset({' 水果 / 蔬菜汁 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.02664 置信度：0.368495 lift 值为：1.44216
frozenset({' 人造黄油 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.024199 置信度：0.413194 lift 值为：1.617098
frozenset({' 仁果类水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.030097 置信度：0.397849 lift 值为：1.557043
frozenset({' 牛肉 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.021251 置信度：0.405039 lift 值为：1.58518
frozenset({' 冷冻蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.020437 置信度：0.424947 lift 值为：1.663094
frozenset({' 本地蛋类 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.022267 置信度：0.350962 lift 值为：1.813824
frozenset({' 黄油 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.020031 置信度：0.361468 lift 值为：1.868122
frozenset({' 本地蛋类 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.029995 置信度：0.472756 lift 值为：1.850203
frozenset({' 黑面包 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.025216 置信度：0.388715 lift 值为：1.521293
frozenset({' 糕点 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.033249 置信度：0.373714 lift 值为：1.462587
```

frozenset({' 酸奶油 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.028876 置信度：0.402837 lift 值为：2.081924

frozenset({' 猪肉 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.021657 置信度：0.375661 lift 值为：1.941476

frozenset({' 酸奶油 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.032232 置信度：0.449645 lift 值为：1.759754

frozenset({' 猪肉 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.022166 置信度：0.38448 lift 值为：1.504719

frozenset({' 根茎类蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.048907 置信度：0.448694 lift 值为：1.756031

frozenset({' 根茎类蔬菜 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.047382 置信度：0.434701 lift 值为：2.246605

frozenset({' 凝乳 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.026131 置信度：0.490458 lift 值为：1.919481

frozenset({' 热带水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.042298 置信度：0.403101 lift 值为：1.577595

frozenset({' 柑橘类水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.030503 置信度：0.36855 lift 值为：1.442377

frozenset({' 黄油 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.027555 置信度：0.497248 lift 值为：1.946053

frozenset({' 酸奶 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.056024 置信度：0.401603 lift 值为：1.571735

frozenset({' 其他蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.074835 置信度：0.386758 lift 值为：1.513634

frozenset({' 酸奶 ', ' 全脂牛奶 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.022267 置信度：0.397459 lift 值为：2.054131

frozenset({' 酸奶 ', ' 其他蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.022267 置信度：0.512881 lift 值为：2.007235

frozenset({' 全脂牛奶 ', ' 根茎类蔬菜 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.023183 置信度：0.474012 lift 值为：2.44977

frozenset({' 其他蔬菜 ', ' 根茎类蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.023183 置信度：0.48927 lift 值为：1.914833

根据输出结果，对其中 4 条进行解释分析如下：

1. {' 其他蔬菜 '，' 酸奶 '}=>{' 全脂牛奶 '} 支持度约为 2.23%，置信度约为 51.29%。说明同时购买酸奶、其他蔬菜和全脂牛奶这 3 种商品的概率达 51.29%，而这种情况发生的可能性约为 2.23%。

2. {' 其他蔬菜 '}=>{' 全脂牛奶 '} 支持度最大约为 7.48%，置信度约为 38.68%。说明同时购买其他蔬菜和全脂牛奶这两种商品的概率达 38.68%，而这种情况发生的可能性约为 7.48%。

3. {' 根茎类蔬菜 '}=>{' 全脂牛奶 '} 支持度约为 4.89%，置信度约为 44.87%。说明同时购买根茎类蔬菜和全脂牛奶这 3 种商品的概率达 44.87%，而这种情况发生的可能性约为 4.89%。

4. {' 根茎类蔬菜 '}=>{' 其他蔬菜 '} 支持度约为 4.74%，置信度约为 43.47%。说明同时购买根茎类蔬菜和其他蔬菜这两种商品的概率达 43.47%，而这种情况发生的可能性约为 4.74%。

由上分析可知，顾客购买酸奶和其他蔬菜的时候会同时购买全脂牛奶，其置信度最大达到 51.29%。因此，顾客同时购买其他蔬菜、根茎类蔬菜和全脂牛奶的概率较高。

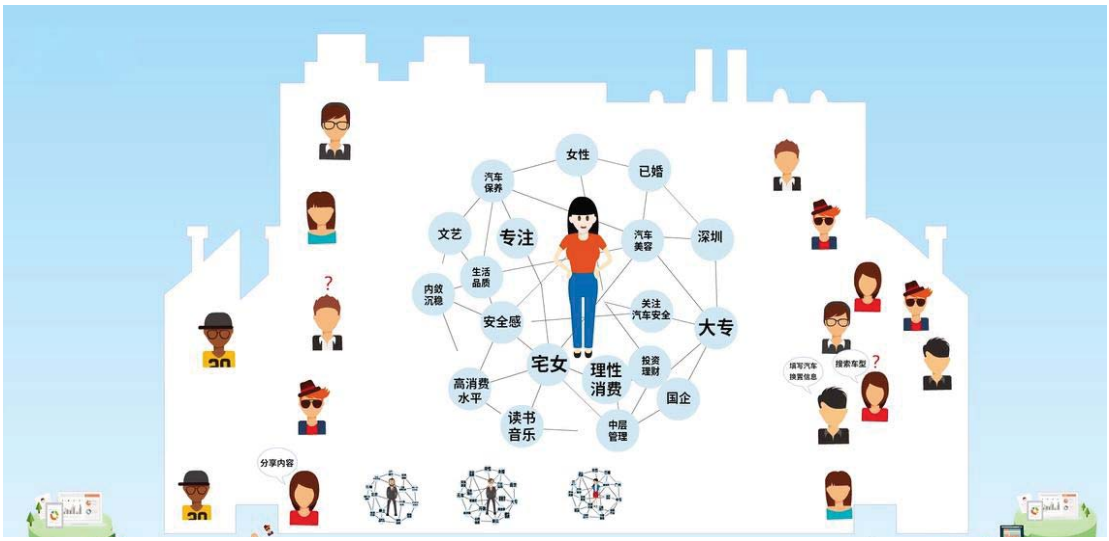
对于模型结果，从购物者角度进行分析：现代生活中，大多数购物者为“家庭煮妇”，购买的商品大部分是食品，随着生活质量的提高和健康意识的增加，其他蔬菜、根茎类蔬菜和全脂牛奶均为现代家庭每日饮食的所需品。因此，其他蔬菜、根茎类蔬菜和全脂牛奶同时购买的概率较高，符合人们的现代生活健康意识。

六、模型应用

模型结果表明：顾客购买其他商品的时候会同时购买全脂牛奶。因此，商场应该根据实际情况将全脂牛奶放在顾客购买商品的必经之路上，或是放在商场显眼的位置，以方便顾客拿取。顾客同时购买其他蔬菜、根茎类蔬菜、酸奶油、猪肉、黄油、本地蛋类和多种水果的概率较高，因此商场可以考虑捆绑销售，或者适当调整商场布置，将这些商品的距离尽量拉近，从而提升顾客的购物体验。

分享 | B 端运营需要关注数据指标

来源 / 木木自由公众号　作者 / 木兮月宝　日期 / 2021-01



导读：To B 运营的本质，其实就是让产品用得更好，体验更佳，且把软件做的比竞品更好。落到具体业务上，就是解决获客、转化、续签等问题。然而，随着近年来 B 端产品的持续升温，产品种类不断增加，针对客户和业务的精细化的数据分析能力成为了一名优秀 B 端运营最重要的核心竞争力之一，也让 B 端运营团队都有着深厚的数据驱动业务决策文化。

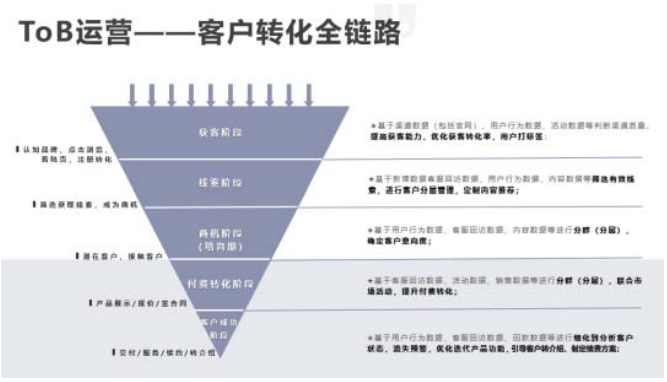
前言

在 ToB 领域，客户从认知品牌到点击浏览，着陆页，注册转化，获取线索，成为商机，潜在客户，接触客户，产品展示 / 报价，签合同，交付，客户服务 / 续约，转介绍，整个客户全链路都属于 B 端运营的阶段。

那么，To C 运营的用户转化全链路是什么？我想你应该会想到：AARRR 模型，获取用户 - 提高留存率 - 提高活跃度 - 获取收入 - 自传播。

同样，To B 整个客户全链路也可以归纳为五个阶段：获客阶段 - 线索阶段 - 商机阶段（培育期）- 付费转化阶段 - 客户成功阶段。

但是，B 端运营在每个阶段关注的数据指标，侧重点不同。

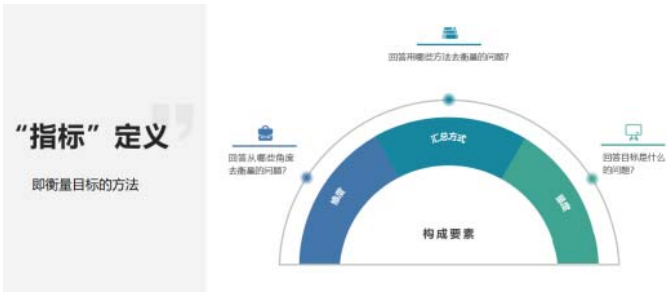


第一部分

【从 0 到 1 建立数据分析指标体系底层逻辑】中以及提到过数据指标的定义，我们现在来回顾一下。“指标”即衡量目标的方法。而构成要素有维度＋汇总方式＋量度。

维度：回答从哪些角度去衡量的问题?

汇总方式：回答用哪些方法去衡量的问题？量度：回答目标是什么的问题?



所谓的“数据指标”，简单来说就是可将某个事件量化，且可形成数字，来衡量目标，在日常工作中大家都会应用的到。在一定程度上，“数据指标”能揭示出产品用户的行为和业务水平状况。我们在工作中会关注一些数据指标，如转化率，留存率，日活，月活等。

然而，在 B 端运营中不同阶段，又有哪些数据指标，什么样的数据指标是值得我们去关注的，或者是有效的，并且能帮助产品业务线找到自己的提升方向呢。

第二部分

01、基于渠道数据（包括官网）、用户行为数据、活动数据等判断渠道质量，提高获客能力

▌渠道数据：

投放消耗：统计时间内花费的金额获客成本（roi）：统计时间内花费的金额 / 新增人数。曝光量：通过应用市场投放广告曝光的次数。点击量：广告被点击的次数，是 APP 被下载并激活的前提。下载量：通过应用市场等渠道，下载 APP 应用的用户数量。激活量：安装应用后，首次打开 APP 应用的用户数量。激活转化率：从下载到激活的用户转化。新增注册量：即注册 APP 的用户数。注册转化率：从激活到注册的用户转化。企业创建数量：创建企业的数量。日均自然量占比：自然量新增 / 新增人数。新增用户渠道来源占比各个渠道留存率：每个推广渠道来源，x 日留存率为 x 日前的新用户在今天还启动应用的比例。

如：渠道转化漏斗为例：

各个应用市场渠道统计为例：

渠道	本周数据					上周数据				
	本周下载激活	本周注册	本周企业用户数	本周注册转化率	本周费用	上周下载激活	上周注册	上周企业用户数	上周注册转化率	上周费用
OPPO应用商店	881	295	58	33.48%	0	1745	849	117	48.65%	0
VIVO应用商店	1704	646	105	37.91%	3751	2040	1017	117	49.85%	3757
华为应用商店	1194	657	85	55.03%	0	1993	1363	238	68.39%	0
小米应用商店	367	192	39	52.32%	0	618	362	63	58.58%	0
苹果市场	3614	1336	396	36.97%	3400	3056	2599	595	85.05%	0
对比	874	420	83	48.05%	0	1388	917	170	66.07%	0
总计	8634	3546	764	41.07%	7151	10840	7107	1300	65.56%	3757

▌官网数据：

pv，uv，ip，新访客，跳出率，SEO 关键词排名，商桥数、400 电话数、自然流量走势、展现量、点击量、点击均价、总消费；· · · · · ·

以某软件官网数据为例：

苹果

	当日	次日	第三天	第四天	第五天	第六天	第七天
2020-12-03	100.0%	35.3%	11.8%	5.9%	17.7%	23.6%	17.7%
2020-12-04	100.0%	33.4%	11.2%	22.3%	16.7%	27.8%	
2020-12-05	100.0%	16.7%	25.0%	16.7%	25.0%		
2020-12-06	100.0%	11.8%	11.8%	0.0%			
2020-12-07	100.0%	27.3%	18.2%				
2020-12-08	100.0%	21.5%					
2020-12-09	100.0%						

OPPO

	当日	次日	第三天	第四天	第五天	第六天	第七天
2020-12-03	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2020-12-04	100.0%	100.0%	0.0%	50.0%	0.0%	0.0%	
2020-12-05	100.0%	0.0%	0.0%	0.0%	0.0%		
2020-12-06	100.0%	100.0%	100.0%	100.0%			
2020-12-07	100.0%	0.0%	0.0%				
2020-12-08	100.0%	0.0%					
2020-12-09	100.0%						

苹果

	当日	次日	第三天	第四天	第五天	第六天	第七天
2020-12-03	100.0%	35.3%	11.8%	5.9%	17.7%	23.6%	17.7%
2020-12-04	100.0%	33.4%	11.2%	22.3%	16.7%	27.8%	
2020-12-05	100.0%	16.7%	25.0%	16.7%	25.0%		
2020-12-06	100.0%	11.8%	11.8%	0.0%			
2020-12-07	100.0%	27.3%	18.2%				
2020-12-08	100.0%	21.5%					
2020-12-09	100.0%						

OPPO

	当日	次日	第三天	第四天	第五天	第六天	第七天
2020-12-03	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2020-12-04	100.0%	100.0%	0.0%	50.0%	0.0%	0.0%	
2020-12-05	100.0%	0.0%	0.0%	0.0%	0.0%		
2020-12-06	100.0%	100.0%	100.0%	100.0%			
2020-12-07	100.0%	0.0%	0.0%				
2020-12-08	100.0%	0.0%					
2020-12-09	100.0%						

▌活动数据：

活动场次，活动城市，活动到场分布，参与人数，费用，活动成本，ROI；·····

02、基于用户行为数据优化获客转化率

▌用户行为数据：

页面访问路径：统计用户从打开应用到离开应用整个过程中每一步的页面访问和跳转情况。转化率：指进入下一页面的人数（或页面浏览量）与当前页面的人数（或页面浏览量）的比值。····

03、基于触点行为数据做线索质量权重归因，给用户打标签，用来后期针对性内容触达或者营销

▌触点行为数据：

新进 20 以上企业数来源渠道占比所属行业占比搜索关键词·····

01、基于新增数据进行客服回访，了解客户 7 要素，建立初步认知以及筛选有效线索

▌用户数据

新增用户：新用户每日进入的人数有效线索数有效接通电话数量拒接率用户增长率客户 7 要素（公司名称、所在城市、人员规模、主要需求、之前有无使用类似软件、有无付费经历、付费金额）·····

02、基于活跃度判断线索状态，进行客户分层管理

▌活跃数据

DAU（日活跃用户）：产品一个自然日的活跃用户数，定义活跃用户，各个产品都没有明确的界定。连续活跃 n 周用户：连续 n 周，每周至少启动过一次 APP 的活跃用户；重要用户：连续活跃 4 周及以上的用户；连续活跃用户：连续活跃 1 周及以上的用户；·····

03、基于搜索词给线索打标签，锁定用户搜索意图，定制内容推荐

▌内容数据

阅读量，阅读量走势，线索转化量阅读次数文章数量阅读人数完成阅读次数（用户滑到图文消息底部的次数）阅读完成率送达阅读率·····

01、基于用户行为数据和标签交叉查询进行分群（分层）

▌用户数据

客户健康度指标的监控是实时的，一般来说会是几个关键事件（比如平均登录次数，帮助页面的 PV，联系客服的次数，

使用核心功能的次数）整合后得出的数字。·····

02、基于内容数据和分群，包装最佳客户案例

▌内容数据

客户案例阅读量文章数量客户案例数量·····

03、基于分群分层，进行二次回访，确定客户意向度

▌客服回访数据

有效接通电话商机数拒接率·····

01、配合客户分层， 针对商机客户进行有效回访

▌客服营销数据

有效接通电话拒接率意向度·····

02、 联合市场活动，提升付费转化

▌活动数据

活动参与度活动成本新客户付费数老客户付费数总营收订单数客单价购买会员类型·····

03、关注销售数据，提升销售人员的能力

▌销售数据

销售额订单量完成率增长率重点产品的销售占比各平台销售占比利润成交率（转化率）人均产出·····

01、基于产品功能埋点，细化到分析客户状态，流失预警

▌用户数据

功能活跃指标：主要关注某功能的活跃人数，关注某个功能。客户流失率 =（在指定时间段内取消的客户数量）/（在同一时间段开始时的付费客户的数量），根据不同产品业务的特性，流失率的定义也都不同。·····

02、进行客户使用产品状态观察以及用户反馈，优化迭代产品功能

▌用户行为数据

留存率活跃数用户活跃率同比，环比·····

03、基于客户分群做定制化关怀，引导客户转介绍、制定续费方案

▌客户数据

续费率，也就是付费客户留存率，在前期数据并不充足的情况下，比流失率能够更准确的反映出产品被客户的接受程度的高低。客户续费率 = 完成续费的客户数量 / 当期到期的客户合同数量客户身价 LTV，LTV = ARPA / 客户流失率（*ARPA 平均客月价）平均每个客户的月度营业额计算公式：平均客月价 = 当月 MRR/ 当月活跃客户数转介绍（传播因子）投诉数量投诉率·····

第三部分 结尾

以上就是对 B 端运营在每个阶段关注的数据指标进行的梳理。不同行业不同产品的数据指标很多，北极星指标、定性指标、量化指标、虚荣指标等，但 B 端运营在每个阶段关注的数据指标，侧重点不同。可不管怎么样，To b 运营通过各种途径获取销售线索，留住用户进行付费 / 续费，使品牌最大化，从而树立品牌正面形象，创造更多营收。

然而，做 to b 运营，有一个现象，大部分用户都需要培养用户习惯，我们要先开拓市场、培养市场、之后才能占有市场，其实我们可以尝试去建立一个社群，帮助用户去体验产品，并认同产品的驱动，用伙伴的方式共同成长，特点就是由专家去运营社群，扶持非商业用户的成功。

这个时代，赢得客户需要付出更多的产品之外的东西，那就是用户成功，我们需要的是站在用户业务的角度去表达产品、站在解决用户业务困难的角度去制定解决方案，而不是在跟用户表达，我有什么，难道不是你的客户，你就不能以用户成功的心态去支持？只有一颗帮助用户成功的心态才能赢得客户，才有机会去做客户成功。

商业在于共赢，而不是一家的略显优势，此时，整个行业都将面临销声匿迹！

全球企业 TOP150 数字化转型成功要领!

来源 / Mike 聊经管　编辑 / 协会会员处　李苗苗　日期 / 2021-01

现在人人都谈数字化转型。向数字化转型已经成了企业和组织特别重视的一项变革行动。数字技术使得市场竞争更加激烈，它在颠覆众多行业的同时，也带来了许多机会。

数字技术带来的机遇也不言而喻。麦肯锡的研究表明，各大企业纷纷摩拳擦掌、满怀雄心壮志，希望在未来三到五年内通过数字举措实现 5%-10%，甚至更高的年增长率和成本效率。

为了更准确地了解当今企业所面临的数字挑战，麦肯锡对全球 150 家企业进行了深入诊断调查。经过对各大企业在数字领域表现的考察，可以将数字化转型成功与否归结为 4 点:

第一，传统型企业必须仔细考虑可供它们选择的战略。能够在行业中起到颠覆作用的全球性玩家始终是少数。而能够借助数字平台，制定实际标准，并能最终建设整个生态系统的公司更是寥寥无几。95% 至 99% 的传统型企业必须选择一条不同的道路，这条道路不能只在既定业务的边缘旁侧敲击，而是要全心全意地采取一套清晰的数字战略。

第二，数字化能否成功，取决于能否依据战略，培养出相应的数字能力，并达到可观的规模。一旦培养出合适的能力，即便消费者的消费心理和行为不断变化，公司也能适时、适当地做出调整。

第三，大数据分析、数字内容管理和搜索引擎优化等技术能力固然重要，但总有不完善的时候，而强大且灵活的企业文化可弥补该缺陷。

第四，公司需要调整其组织结构、人才发展、融资机制和关键绩效指标 (KPI)，让这些内容与自身的数字战略保持一致。

总体而言，要想在数字时代紧跟潮流，传统型企业的高管团队可以遵循上述四点建议构建企业未来发展规划。当然，除此之外，有很多其他因素也需要考量。如果没有正确的企业发展规划和与之匹配的管理思维，极有可能走错方向。即便走对了方向，也可能速度太慢，抑或停滞不前。

01 制定正确的战略

“数字化”对自己的组织意味着什么？这是高管们必须首先回答并达成共识的问题。之后，成功的第一步就是制定一个明确而连贯的数字化战略，并将其完全整合到整体的企业战略中去。如果整合不完善，任何后续措施都必然会出问题。然而，制定正确的数字战略对于很多公司都是个挑战。数字领先企业与一般企业之间的差距集中体现在战略方面。制定正确数字战略的一个难点，就是只有小部分有着高曝光率的龙头企业才能博得社会宣传与关注 (以及普遍走高的市场估值)，这其中就包括市场颠覆型的企业，如优步 (Uber)。

要想制定正确的数字战略，企业要回答三个关键问题：1) 最值得关注的数字机遇和威胁在哪里？ 2) 数字颠覆可能发生的速度有多快，规模有多大？ 3) 怎样才能更好地拥抱这些机会，更好地配置资源以规避主要威胁？绝大多数公司需要对症下药，采取战略措施来解决这三个问题。

这些措施包括：

小规模转型自身业务模式，以进入新市场或重新定义现有市场。例如 , 深圳的平安银行创立了橙子银行 , 聚焦数字化，针对年轻消费者。其所提供的简单、高回报的金融产品和一分钟开户的金融服务对年轻消费者的吸引力是传统网点或任何复杂的金融产品组合都难以望其项背的。

紧跟潮流，抓住行业发展所带来的价值。英国百货公司 John Lewis 推行“线上点击加线下实体”的方案，使客户能够在网站下单后，在门店或社区杂货网点提货。

积极重新配置资产，从受到数字威胁的领域转向受益于数字化的领域。举例来说，德国的 Bauer 媒体集团曾系统性地重新配置资源，抛弃过去的传统媒体模式，开发出更具数字特质的产品组合。虽然表面来看其整体收入有所缩减，但实际上其营收增长及市盈率均有所增长。

通过数字途径和工具，提高现有业务模式效率。例如，为了更好地服务迪士尼度假村和主题公园的游客，迪士尼公司开发了一系列数字工具，比如可帮游客预订主题公园项目的 Fast Pass+ 服务，或者方便游客在园区内预订和规划游览路线的 MagicBand 手环 (有一半迪士尼游客都选择佩戴这种手环)。更高效的游览路线，使得迪士尼的神奇王国在 2013-2014 年的假期高峰每天可多接待 3000 名游客。

找到最合适自己的数字战略至关重要。成功的数字战略与差异化的管理方法相辅相成。如果战略正确，管理起到的干预作用就会更清晰。因此，公司可以考虑以下做法 :

大胆将眼光放远，不过分追求短期的财务业绩，能够承担适当的风险，对数字化举措和 IT 架构进行大规模投资。

将数字化整合进公司战略，使数字化成为业务核心，自然形成内部协作，公司治理也会并重数字化需求。战略优先与投资决策属同一个流程。

坚持不懈关注客户需求，有助于公司在关键领域不断创新。虽然最初期客户的数据有时候会误导业务的方向，但是他们的行为往往能够在短时间内渗透大众市场。公司可以通过多种场景 (如视频会议、短信和线上聊天) 与消费者建立直接的连接。

一旦公司经过深思熟虑达成一项战略，它们就必须全心全意地投入其中。只对表面皮毛修修补补的日子已经一去不复返了。

02 大规模能力建设

要想数字化一举成功，一些特定的能力，尤其是那些夯实关键流程和工作内容的能力至关重要。其中，最重要的是模块化、IT 平台与敏捷技术交付能力，它们让公司在快速发展的世界中随时与客户保持同步。麦肯锡调研的大多数公司的 IT 平台均存在重大缺口，这也从侧面反映出目前企业普遍为在 IT 和投资中对数字举措进行优先考虑。

除此之外，调研中表现优异的企业往往能够通过数字方式与客户互动，并在以下四方面优化它们的成本效益。

基于大数据的决策。优秀的数字化企业善于紧跟顾客的数字化消费决策之旅。例如，它们会快速收集并形成结构化数据 (如人口结构、购买历史) 与非结构化数据 (如社交媒体、语音分析)，预测客户行为新模式，并由此调整与客户的互动方式。这些公司巧妙地利用自身业务内外的可用资源来应对市场中最至关重要的问题。

2012 年，感冒流感药剂制造商 Reckitt Benckiser 利用医疗网站 WebMD 的搜索数据 (当时该网站每月近 3200 万访客)，追踪美国的感冒流感症状，并预测可能爆发病情的地区。之后，该公司在这些地区发布了针对地域和症状的广告和促销活动 (包括免费送货上门)。在感冒和流感多发的季节，这项计划使 Reckitt Benckiser 一个月内的咳嗽感冒药品在全美的销量同比增加了 22%。

数据分析与复盘能力。任何一家企业都会追求对自身业务发展的总结，业绩越成功的企业越是如此。庞大的产品体系与繁杂的营销策略会产生大量存在规律价值的数据，依靠传统管理方式显然无法研究这些数据，数据分析手段与相关技术平台成为解决这个问题的最好方式。

能让业务持续良好发展的秘密是藏不住的。现在大数据搜集、数据分析逐渐在全世界所有行业成为风潮，中国已经是世界上每天产生数据最多，产生数据分析需求最多的国家之一。

与消费者建立联系。这也是不可或缺的一环。企业应该

积极拥抱能够加深品牌与客户联系的新技术 (如应用程序、个性化和社交媒体)，这些技术一方面能为客户提供更优体验，另一方面也能服务于产品开发。

2009 年，博柏利发起了“风衣艺术 (Art of the Trench)”活动，鼓励顾客访问其线上平台，并上传自己身着风衣的照片。其他买家和时尚专家会对照片评论并点赞 , 并通过电子邮件和社交媒体分享到自己的平台上。用户还可以直接点击进入博柏利官网购买照片上的款式。随着时间的推移，这些创新模式也与公司结合得愈发紧密。虽然博柏利的做法可能并非无可挑剔，但是总的来说，此番做法再结合其他创新举措，使得公司在六年内年收入翻了一番。

流程自动化。优秀的数字企业会将自动化的重点放在流程设计上，并在过程中不断进行试错和优化。要想实现成功的流程自动化，首先要将眼光放得长远，抛开眼下的限制，提前设计未来的每个流程。比如，将周转时间从几天缩短到几分钟。明确了未来想要达成的目标后，就可以对相关的约束条件 (如法律协议) 从长计议。

一家欧洲银行就使用该方法将其开户流程从两到三天缩短至不到十分钟。与此同时，该银行通过在其信用评分模型中添加一款线上计算器，成功实现了抵押贷款申请流程中部分环节的自动化，在短短一分钟内为客户提供初步报价。该系统在显著提高客户满意度的同时，大大地降低了成本。

双速 IT。当下，消费者的期望给 IT 带来了新的压力。传统的 IT 架构难以应对数字产品创新中快速变化的测试、失败、学习、调整和迭代机制。领先企业通常既有专业高速的新 IT 能力，以实现快速产出，亦对其传统的 IT 能力进行了优化，以支持传统的业务运营。

这种 IT 架构可以支持两种不同的运营速度。一方面，面向客户的技术灵活多变，能够快速反应。例如，这些技术可以在几天内就开发部署完成新的微服务，或几秒钟内为客户提供动态的个性化网页。另一方面 , 核心 IT 基础架构牢固，可以支撑高质量的数据管理和内置的安全保障，旨在确保交易与支持系统所需的稳定性和灵活度，保证核心业务服务的可靠性。

一家英国金融机构就是通过采用双速 IT 模式，改善了其线上零售银行服务。该银行开设了一个新的开发办公室，借鉴初创企业公司文化，执行敏捷的工作流程，快速完成新产品测试和优化。为了长期培养这种能力，该公司同时发展服务架构，以加速发布面向客户的新功能。

03 快速、敏捷的企业文化

强大的技能固然重要，但如果不能做到尽善尽美，公司也大可以将传统文化与速度、灵活性、开放度和学习能力相结合，弥补其技能的缺失。塑造这样的文化，方法不止一种，但 DQ 诊断得分较高的公司，都会采取 DevOps、持续交付和敏捷等边测试边学习的软件开发方法 , 并取得了不错的成效。曾经，这些方法仅处在工作环境的边缘地带，而如今它们的应用促进了核心人才的互动与沟通。以前各自为政的职能和业务部门也因此有了新的凝聚力。

这种边测试边学习的方法结合了自动化、监控、社区共享和跨部门协作 , 将各自为政的职能与流程融入到快速变化、以产品工作为核心的文化中去。在技术和产品产权共享的环境中，数据使用能快速得到普及，将复杂性降到最低，并且能快速进行资源再配置，建立一个可循环、模块化和可交互的 IT 系统。想要塑造这种协作、敏捷的文化，高管可以重点关注以下四个关键领域。

外部协作。通过发展鼓励协作的企业文化，企业能够参与到更广泛的生态系统中，与非本行业内的企业进行合作、深度学习和协同创新。然而，对于大多数企业来说，凭借一已之力构建这些网络或生态系统难度较大。但是，在一套复杂的生态系统中，企业可以另辟蹊径，树业有专攻 (如在生产或物流方面)，以此创造价值。

企业与外界的协作不一定非得在大生态系统里才能实现，与客户、技术商和供应商的小规模合作也能让企业受益匪浅。此外，巧妙利用自己员工队伍以外的资源，如在兴趣小组或网络中招揽人才，也不失为一种好的方法。比如 SAP 在推出 NetWeaver 软件时，就充分调动了用户社区资源。

以上所述的外部协作都要求数字领导者认识到自身所长和别人的过人之处，提高与个人和机构合作的能力。在各种宣传炒作中，他们还必须懂得区分真正的机会与威胁，辨别对方是敌是友。

风险偏好。数字行业的领军企业普遍敢于采用大胆的举措。相比之下，落后企业的高管则偏向于规避各种风险。虽然成熟企业不太可能打造或主导大型生态系统，但是它们仍会受到市场或行业中颠覆性力量的影响，需要面对随之而来的风险。当今世界大数据涌现，不确定性也日益增加，企业必须做出决策，尽早地对颠覆力量做出回应。

大规模地推广边测试边学习的策略。敏捷文化的核心是边测试边学习的思维方式和产品开发方法，这种模式可以在任何成熟企业的项目或流程中得到有效应用和转换。相较于坐以待毙、不听取市场反馈、被动等待热门产品的诞生，数字领先企业选择不断学习、不断追踪，并迅速地在市场中投放新产品。之后，它们会分析消费者兴趣，收集消费者反应，并不断改进产品。严格的数据监控可以帮助团队快速决定是完善还是放弃新举措。如此这般，失败固然常见，但成功的几率也大为上升。

例如，Nordstrom 的创新实验室就面向顾客，推出了一系列周期为一周的试点活动。为了开发太阳镜购买的 App，公司创新团队在西雅图的零售旗舰店设立临时工作点，搭建各种样板场景，模拟现实的线上购物情境，让购物者进行点击和选择。顾客可以指出他们认为最有用的功能，或样板中存在的问

题。利用这些信息，编程人员实时调整，当下发布新版 APP 供客户现场操作。经过一周的不断调整和再发布后，这款 APP 已经成为了门店销售的绝佳帮手。

内部合作。无论数字化与否，团队协作都格外重要。沃顿商学院的 AdamGrant 表示，最影响团队效率的因素，是同事在工作中互相帮助的程度。在企业提升 DQ 的过程中协作文化更显得尤为重要。许多公司缺乏必要的数字化业务作为主干来协同传统上各自为政的职能部门，无论是从客户服务到订单履行，还是供应链管理和财务报表，均缺乏必要的协同。

在麦肯锡调查的 150 家公司中，只有不到 30% 的公司表示它们拥有高度协作的企业文化，不过，这也说明其他公司的改进空间巨大，先进的科技能够在这里发挥较大的促进团队合作的作用。例如，将虚拟云作为跨职能、地域间的协作的平台，让各职能团队在云上协作开展实验、试错与创新。

04 组织与人才

除了战略、能力和文化之外，领先的数字公司在管理人才、流程和组织架构方面也采取了统一的举措。

吸引和培养数字分析人才。DQ 公司认为，高管团队里需要有一位主管数字化的领导，将业务、营销与技术专长方面的人才结合起来。但同时，中层的才干也很关键。他们才是脚踏实地，深入一线的骨干，数字化举措的成败，往往取决于他们，因为他们才是最终将产品和服务推向市场的中坚力量。

在当下的环境中挖掘到合适的数据分析人才并不容易。企业应该认识到数字分析能力往往跟行业知识同等重要，在数字化转型的早期阶段尤为如此，据统计，只有不到 35% 的人拥有除现阶段本职工作之外的数字分析能力。

DQ 公司在人才培训和发展方式上也独具匠心。比如，几年前宝洁与谷歌开展了员工互换项目，一方面旨在加强宝洁的搜索引擎优化技能，另一方面也让谷歌这家互联网巨头学习到更多的营销知识。这种互换项目不但能培养员工能力，也为公司带来了更多发展空间与可能性。

企业还必须采取适当的激励措施和明确的职业发展方向来培养数字人才。实际上，一些成熟的企业可能比想象的更有优势，因为年轻人更愿意帮助时尚服饰、豪华轿车、新闻杂志等行业的知名品牌建立数字化渠道。渠道一旦建立成功，就是良性循环的开端。对优秀人才的培养足够好，就能吸引更多的人才，组织也能快速地拓展，确保在数字行业的领军地位。当公司人才济济时，更多的人才也就蜂拥而至了。

对数字化进程进行实时监控。领先的数字企业都会经常性地，甚至实时地追踪数字化 KPI，并就此展开沟通。它们根据数字化的优先级衡量这些 KPI，并确保高层能监督和管理这些指标。

举例来说，当星巴克推出一套新的 POS 系统时，经理会对交易过程录像并询问员工意见，根据他们的反馈，再调整结账流程，最后成功将手机或银行卡交易缩短了十秒钟左右。这样一来，员工能更快地完成销售，每年为顾客节省 90 万小时的宝贵时间。

采用非传统结构。尽管没有适用于所有公司的万全之策，但是高 DQ 企业的组织结构都经过精心设计，完美贴合企业现在的数字转型阶段。有些公司也坦然接受现实，认识到无法通过快速转型自己的核心业务实现新的数字增长。例如，许多成功的传统媒体已经把数字化业务从现有的成熟内容中剥离出来。

Axel Springer 最近的重组中，就以数字业务模式为主导，发展数字业务所需的独特文化、构建相关的绩效管理体系和管控治理方法。与此同时，Axel Springer 传统业务能够适应新的数字环境，可以单独调整并发展。

举例而言，欧莱雅和 TD 银行集团等成熟企业创立了卓越中心并任命了首席数字官。博柏利等其他公司则下设委员会，负责制定宏观目标，并确保高管与数字项目保持步调一致。这些组织结构会随着公司发展而变化。社交媒体等新孵化的力量最终会趋于成熟，并融入到整体业务当中去。

写在最后：

数字化成熟之旅需要企业领导层的全心付出，并持续对人员、能力、技术和文化变革进行投入。企业成功的第一步，是清楚认识自身，明确长期战略机遇，并乐于试错，不断改进解决方案。

泓睿数据：企业数据化管理作用有多大？

来源 / Mike 聊经管　编辑 / 协会会员处　李苗苗　日期 / 2021-01

在市场竞争越来越激烈的背景下，越来越多的企业开始部署大数据采集，为业务创造新的价值。

应用不代表应用成熟，相当一部分企业对大数据、数据分析的理念、技术利用和应用模式还处于初级阶段，谈概念的多、真正用好的少。

传统企业在制定数据战略以业务数据分析方面还存在诸多问题，在尝试解决这些问题时，企业管理层很容易把问题源头指向外在因素，过于关注行业数据与同行数据，却忽略了企业内因的影响效应。

从企业内部思考并找到影响业务创新的关键因子，对于企业突破发展瓶颈至关重要。成都泓睿数据分析师事务所利用两年时间，通过对近 40 个企业单位量身定制数据管理体系，建立数据分析模型，研判每一个环节中的数据界点，帮助客户完成业务创新，每个案例都获得了显著成效，并深受企业客户好评。

四川某知名地暖集团，2019 年初在成都泓睿数据分析师事务所帮助下建立数据管理体系，对过去几年的业务数据进行了多维度分析，在听取事务所专家建议后对销售产品结构搭配


证书编号：20210420013075

中国数据分析行业

会员执业资质证书

企业名称：成都泓睿数据分析师事务所有限公司
注册地址：成都青羊区蜀金路1号3栋21楼2101号
法定代表人：宋利
注册资金：1000万
经济类型：有限责任公司(自然人投资或控股)
经营范围：数据处理服务；市场调查；会计、审计及税务服务；企业管理咨询；社会经济咨询；信息处理和存储支持服务。（依法须经批准的项目，经相关部门批准后方可开展经营活动）。

会员编号：数据委团证第130号
资质类别：事务所会员

发证单位：中国商业联合会数据分析专业委员会
发证日期：2021 年 4 月 20 日
（签章有效）

进行调整，最终实现 2019 年产品销量同比增长 60%。四川某供应链管理公司，2019 年底应收账款占总业绩比 70%，几次接近流动资金红线。

2020 年，成都泓睿数据分析师事务所开始为其提供数据分析与管理服务，通过以往经营情况梳理，数据化管理系统搭建，该公司实现了以月为单位的业务数据复盘。通过多次业务策略调整，2020 年 5 月起，该公司流动资金稳定维持在 150 万以上，当年企业应收账款占总业绩比下降至 25%。每一个服务案例都帮助客户快速搭建了产品数据模型，帮助客户更好的完成了大数据技术与自身业务模式的融合改造，让客户可以充分挖掘数据潜力实现业绩突破。

成都泓睿数据分析师事务所于 2016 年 9 月 6 日完成工商和税局手续注册成立。2020 年 4 月通过中国数据分析行业权威机构——中国商业联合会数据分析专业委员会考核，成为中国商业联合会数据分析专业委员会的事务所会员单位。

事务所核心顾问团队成员均有深厚的财务管理从业背景，致力于开创并推行企业数据管理体系。成立至今，事务所累计签约服务近 40 家单位，不仅为众多企业竖起了风险预防的有力屏障，更为企业精细化管理奠定了坚实的基础。帮助企业持续性发展中实现创收节能、提升管理效率、突破瓶颈、降低经营风险取得显著成效。

联系我们
成都泓睿数据分析师事务所
地　址：成都市青羊区蜀金路 1 号
联系人：宋利
电　话：13678049584
邮　箱：751294110@qq.com

CPDA® 数据分析师
Certified Projects Data Analyst

不甘平凡

UNWILLING ORDINARY

关注CPDA数据说