国联证券 GUOLIAN SECURITIES

2018 年 11 月 07 日

大数据时代，演绎第三次浪潮的华彩乐章

计算机行业

投资建议： 中性

上次建议： 中性

一年内行业相对大盘走势

投资要点：

➢ **大数据时代，演绎第三次浪潮的华彩乐章**

对于大数据，Gartner给出的定义是需要运用新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。1980年，著名未来学家阿尔文·托夫勒在其著作《第三次浪潮》中，将“大数据”描绘为“第三次浪潮的华彩乐章”。

➢ **大数据发展全球加码，广阔空间蕴含商机无限**

基于大数据对各行业的深入影响，美国、欧盟等主要发达经济体都积极推进各自的大数据战略，中国亦将其视为新经济的重要支撑。据信通院数据，2017年中国大数据相关产业规模为4700亿元，预计2020年有望赶超1万亿，年均复合增速近30%，其中，核心产业规模2017年为234亿元，同比增长39%，预计2018年可达329亿，空间广阔。同时，大数据投融资市场也持续升温，2012-2016年期间，国内共发生大数据投融资事件超1600件，统计公布金额的1300余起投资，其融资总额达1200多亿，2016年同比增长189.7%。

➢ **大数据产业链：数据为源、分析为核、应用为王**

分析大数据产业链，主要涵盖数据来源、数据管理与分析、数据应用。1）数据是行业发展的源泉，政府、BAT、运营商等是当前中国大数据的主要拥有者，另在细分领域拥有入口资源的公司也是稀缺标的。2）数据管理与分析是产业中游。数据管理负责数据的集成、存储、安全等环节，其中，数据存储是产业链的支撑，参与者以传统数据库企业为主；数据安全是产业发展的重要保障，渗透数据存储、传输、交互的各个环节。而产业链最核心的当属数据分析与挖掘，其能力直接决定着大数据应用的推广程度和范围，当前Hadoop、Spark是使用较为广泛的两种处理框架，算法方面受益人工智能，神经网络算法关注度再次高涨。3）应用为王，对大数据分析结果进行应用是完成产业商业化目标，实现价值的终点。对比市场空间、政策倾向及惠及民生等方面，我们更为看好政务大数据及医疗大数据市场，另从产品形态看，整体解决方案商更容易树立标杆案例，灯塔效应明显。

➢ **投资建议**

我们认为，产业链上数据是源泉、存储是支撑、安全是保证、分析是核心、应用是价值实现。建议关注拥有位置领域入口资源的**四维图新**，布局芯片及AI服务器的**中科曙光**，以及掌握视频数据分析能力的**海康威视**，外加应用领域的智慧公安解决方案商**美亚柏科**、智慧医疗解决方案商**创业软件**等。

➢ **风险提示**

技术遭遇瓶颈；政策有所延缓；订单低于预期；市场系统性风险

吴金雅 分析师
执业证书编号：S0590517020001
电话：0510-82833337
邮箱：wujy@glsc.com.cn

朱松 研究助理
电话：0510-82833217
邮箱：zhus@glsc.com.cn

相关报告

1、《政策暖风频吹，新技术重视度再次彰显》2018.11.05

2、《亚马逊、微软公布最新财报，再次彰显云计算高景气度》2018.10.29

3、《市场存反弹需求，短期关注超跌中长期关注高景气行业》2018.10.25

正文目录

图表目录

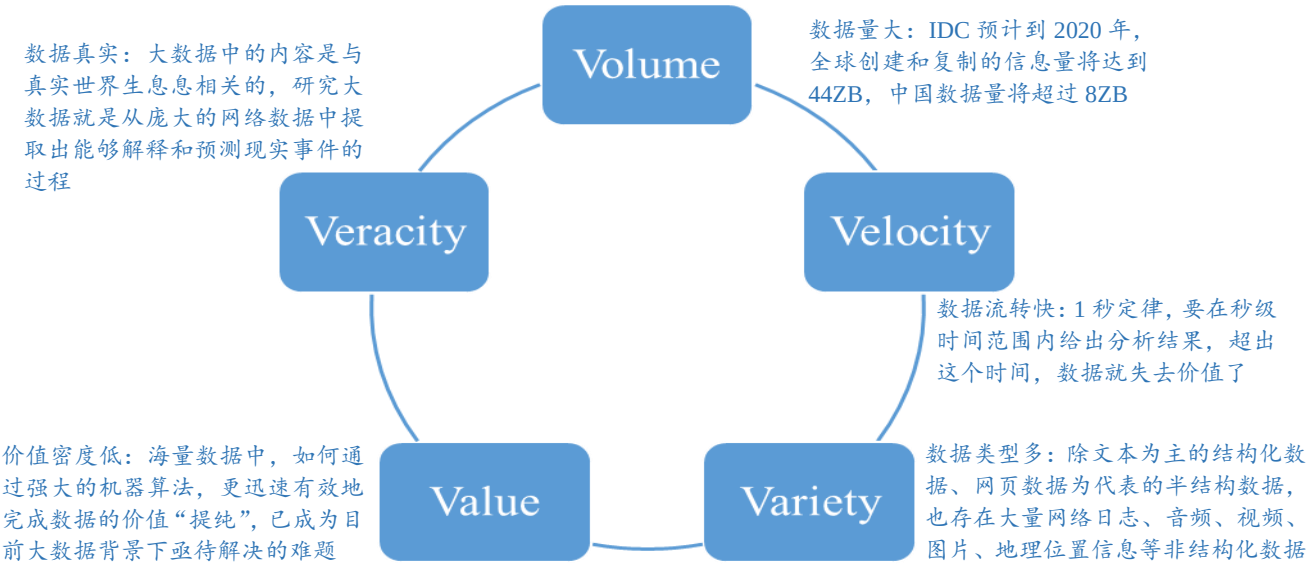
1. 大数据时代，演绎第三次浪潮的华彩乐章

1.1. 大数据的定义

对于大数据，Gartner 给出的定义是需要运用新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》提出，大数据不是随机样本，而是全体数据；不是精确性，而是混杂性；不是因果关系，而是相关关系。

大数据具备 Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）的特点（IBM）。随着信息技术不断发展，互联网快速普及，与人们的生产、生活日益紧密，全球数据亦呈现倍数级增长的特点，对经济发展、社会治理、国家管理、人民生活都产生了重大影响。

图表 1：大数据 5V 特性



来源：艾瑞咨询、百度百科、国联证券研究所

1980 年，著名未来学家阿尔文·托夫勒在其著作《第三次浪潮》中，将“大数据”描绘为“第三次浪潮的华彩乐章”。

2003 年《The Google File System》、2004 年《MapReduce: Simplified Data Processing on Large Clusters 》、2006 年《Bigtable: A Distributed Storage System for Structured Data》谷歌大数据三大论文发布，以及 2005 年 Hadoop 项目的诞生，使得大规模处理结构化、半结构化、非结构化数据[1]的廉价方案成为可能，为大数据产业的快速普及创造了基础条件。

2008 年，大数据得到部分美国知名计算机研究人员认可。业界组织计算社区联

[1] 结构化数据：能用数据或统一结构加以表示，如数字、符号。半结构化数据：介于结构化数据与非结构化数据之间，和普通纯文本相比，半结构化数据具有一定的结构性，但和具有严格理论模型的关系数据库数据相比，半结构化的数据结构变化又很大，如 HTML、XML 文档。非结构化数据：无法用数字或统一结构表示的信息，如图像、声音、视频等。（参考易观智库）

盟（Computing Community Consortium）发表白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》，详尽阐述了大数据对社会治理的推动作用，及其潜在的商业价值。大数据正式进入世界最具有价值和影响的技术行列。

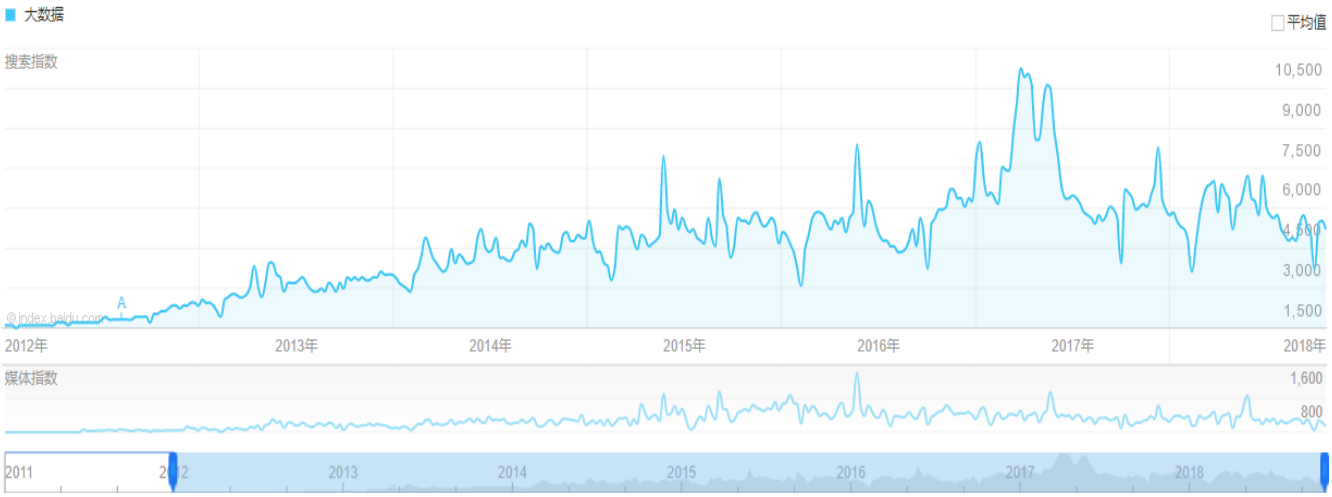
2009 年，美国政府为构建开放、透明机制，启动 Data.gov 网站向公众开放多种政府数据，包括交通、经济、医疗、教育和人口服务等。2012 年，Data.gov 已累积来自 172 个政府机构的数据集，数量从 2009 年的 47 个暴增至 40 万个以上，催化美国政府推出相关政策，加速大数据技术发展。

至此，大数据产业迎来其发展的大时代。

图表 2：大数据发展历程

来源：亿欧智库、国联证券研究所

图表 3：大数据搜索指数趋势（百度）



来源：百度、国联证券研究所

1.2. 为什么要研究大数据？

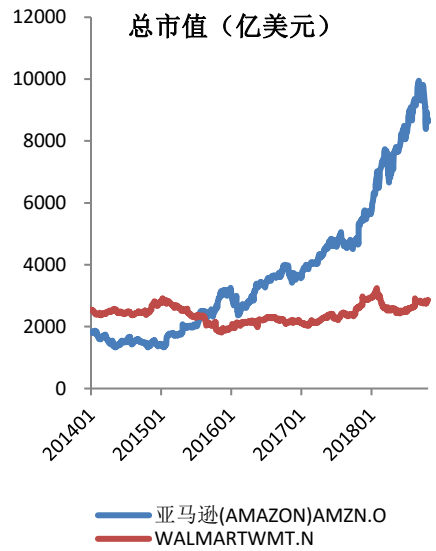
2015 年，亚马逊市值第一次超越沃尔玛，当前前者市值更是后者的三倍多，而亚马逊销售额中有 1/3 是依托大数据精准营销产生。通过记录顾客浏览网站时的行为数据，如所搜关键词、到访页面、关注商品、购买订单，以及不定期举行活动引导客户明确喜好，如主题投票，亚马逊搜集并分析客户属性、兴趣、需求，利用聚类等大数据模型为客户群体推荐合适商品。

以色列的环境比中国大西北更恶劣，但将大数据引入农业后，以色列成为了“欧洲的厨房”。凭借较高的信息化和数字化基础，以色列农业技术公司利用大数据帮助农民根据农场的具体情况采用更加个性化的耕种方案。如 Taranis 公司利用大数据分析法推出包括预测天气、灌溉和病虫害状植物模型技术，指导农民合理灌溉、杀虫；AKOL 公司更是将不同区域农民工作习惯等人为因素纳入农作物生长及环境状况的大数据分析范畴，进一步优化方案。

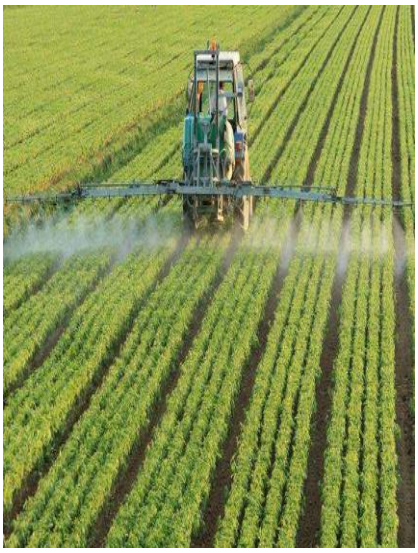
更甚者，在体育界，植入科技和大数据之后，美国金州勇士队在短短几年内就实现了从一个“烂”球队到 NBA 总冠军的飞跃。勇士队老板拉科布作为数据分析的坚实拥趸，把数据分析思想充分融入到球队的训练之中，最先引入球馆录像和分析系统，同时其团队统计历年 NBA 比赛，发现最有效的进攻是眼花缭乱的传球和准确的投篮，并创造了三分球新打法，助力勇士队快速成长。

图表 4：大数据贡献列举

亚马逊、沃尔玛市值对比



以色列农业



勇士队夺冠

来源：百度图片、钱塘数据、国联证券研究所

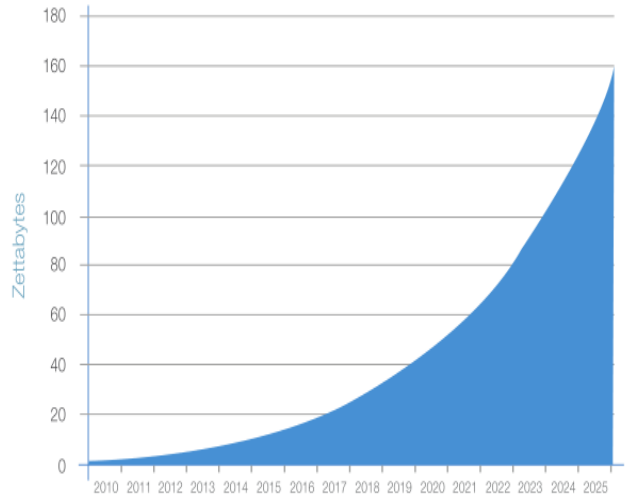
正如《大数据时代》所言，大数据开启了一场重大的时代转型，就像望远镜让我们感受到宇宙，显微镜让我们能够观测微生物，大数据收集、分析海量数据帮助我们更好地理解世界，是众多新发明和新服务的源泉。如今，数据已经成为重要的商业资本，可以作为前期投入创造实际经济价值，此外，大数据也撼动着医疗、教育、人文、

社交等世界的方方面面……其社会价值亦不可估量。

1.3. 大数据发展的基础：数据积累、算力提升、技术创新

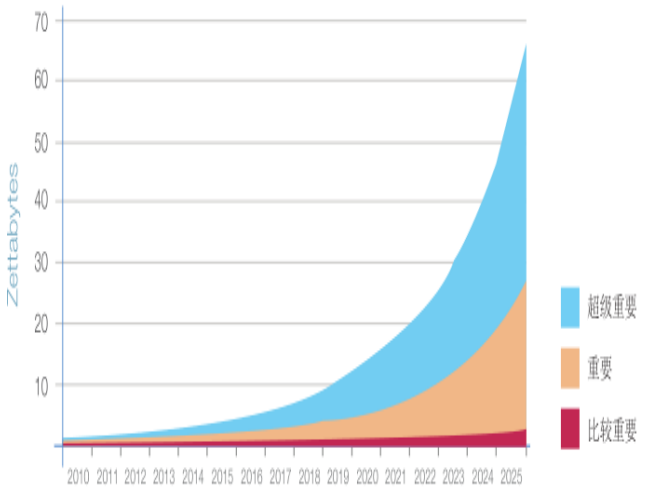
在互联网快速普及、物联网加速渗透的背景下，PC、手机、传感设备等全面兴起，推动全球数据呈现倍数增长、海量集聚的特点，为大数据产业发展奠定了庞大的数据基础。根据 IDC 统计，2011 年全球创建和复制的数据总量为 1.8ZB，2016 年这一规模为 16.1ZB，预计 2020 年将达到 44ZB，在其《数据时代 2025》白皮书（希捷赞助）中，更是预测到 2025 年，全球创建和复制的数据总量将扩展至 163ZB（1ZB 等于 1 万亿 GB）。

图表 5：全球数据规模



来源：IDC、希捷、国联证券研究所

图表 6：数据的重要性归类



来源：IDC、希捷、国联证券研究所

同时，处理如此规模的数据量也对算力提出了巨大的挑战。所幸，摩尔定律推动处理器性能不断提升，GPU、FPGA、TPU 等高算力芯片不断涌现，为大数据产业发展保障了迅速的处理能力。在 Google I/0 2018 开发者大会上，谷歌发布了第三代 TPU 处理器，基于 TPU 3.0 的新运算阵列 TPUv3 Pod 性能相比 TPUv2 Pod 有 8 倍提升，运算速度可超 100PFlops（PFlops：每秒千万亿次浮点计算）。

再者，云计算、人工智能等新技术的出现也为大数据产业发展提供了技术支撑。云计算为企业实现了更为便捷的大数据解决方案，其按用量付费、可扩展的存储计算能力、便捷易部署等特点，大大降低了企业应用大数据的难度与成本，促进大数据产业加快推广。人工智能通过深度置信神经网络等领先算法，自动处理、分析大规模数据，从而获得预测性的洞察，指导或直接替代人工决策，提高大数据核心——预测的效率性。

2. 大数据发展全球加码，广阔空间蕴含商机无限

基于大数据对各个行业的深入影响，近几年，美国、欧盟、日本等主要发达经济体都积极推进各自的大数据战略。2009 年，美国科学家委员会（NSTC）就发布了《开

发数字数据的威力》报告，初步提出发展大数据的框架，奥巴马政府亦对大数据行业大力支持，帮助美国取得世界领先地位。

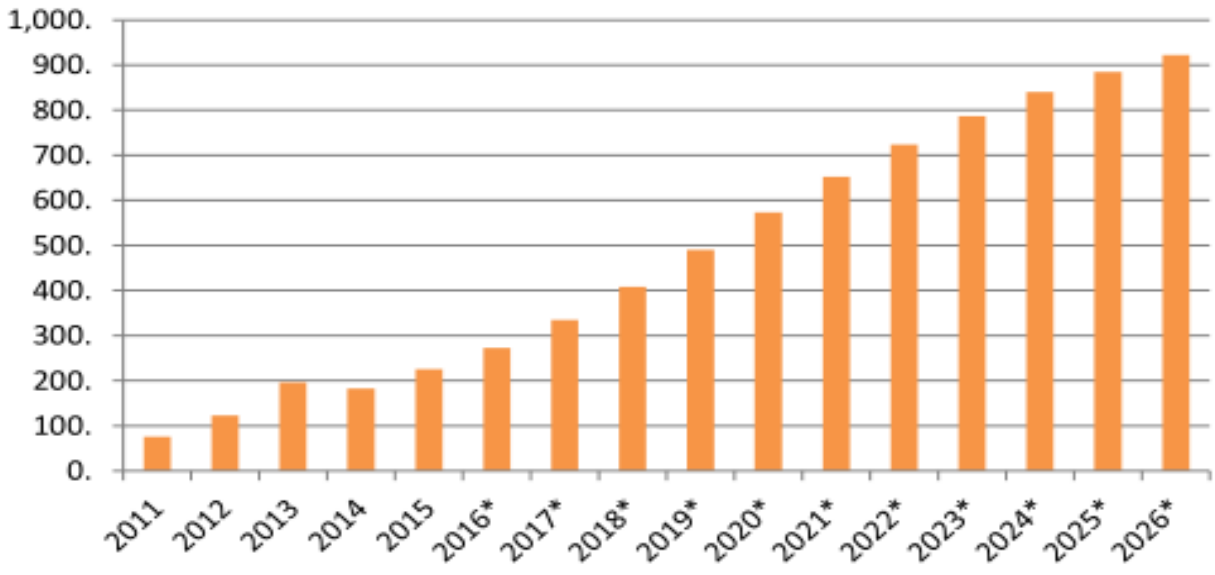
参考《大数据白皮书（2016）》，IDC、Wikibon 等咨询机构分析，2016 年全球大数据核心产业规模约为 300 亿美元，预计 2020 年有望达到近 600 亿美元。

图表 7：美欧日韩关于大数据的主要政策（非完全统计）

国家	政策
美国	2012 年 3 月，奥巴马政府宣布启动“大数据研究与开发计划”，投入 2 亿美元进行大数据相关技术研发 2013 年 5 月，奥巴马政府发布行政令，加大政府数据开放力度，以更有效地利用宝贵的公共信息资源 2014 年 5 月，白宫行政办公室与总统科技顾问委员会联合发布《大数据：抓住机遇，保护价值》与《大数据和隐私：技术视角》，分别从政策和技术的角度分析了大数据的发展对社会带来的影响，特别是对隐私的影响 2016 年 5 月，白宫又发布了《联邦大数据研发战略计划》报告，在已有基础上总结未来研发重点战略，指导大数据发展进程
欧洲	2012 年 9 月，欧盟委员会公布“释放欧洲云计算服务潜力”战略，旨在把欧盟打造成推广云计算服务的领先经济体，预计到 2020 年，大数据技术领域新增投资将为欧盟创造 9570 亿欧元产值，增加 380 万个就业岗位 2013 年英国政府发布《英国数据能力发展战略规划》，并建立世界首个“开放数据研究所”
日本	2013 年 6 月，安倍内阁正式公布《创建最尖端信息技术国家宣言》，这一以开放大数据为核心的 IT 国家战略，旨在把日本建成具有“世界最高水准的广泛运用信息产业技术的社会”
韩国	2012 年，韩国国家科学技术委员会就大数据未来发展环境发布重要战略规划 2013 年，韩国未来创造科学部提出“培育 1000 家大数据、云计算系统相关企业”的国家级大数据发展计划，以及出台《第五次国家信息化基本计划(2013-2017)》等多项大数据发展战略

来源：全球科技经济瞭望、中国经济报告、国联证券研究所

图表 8：全球大数据核心产业规模（亿美元）



来源：Wikibon（2016-03）、《大数据白皮书（2016）》、信通院、国联证券研究所

中国亦将大数据视为新经济的重要支撑。2012 年，《“十二五”国家战略性新兴产

业发展规划》明确提出支持海量数据存储、处理技术的研发和产业化，2014 年“大数据”首次出现在《政府工作报告中》，奠定了行业快速发展的政策基础。而 2017 年以来，党的十九大报告、中共中央政治局就实施国家大数据战略进行第二次集体学习、首届数字中国峰会召开等，均再次显示出领导层对加快建设数字中国的高度重视。

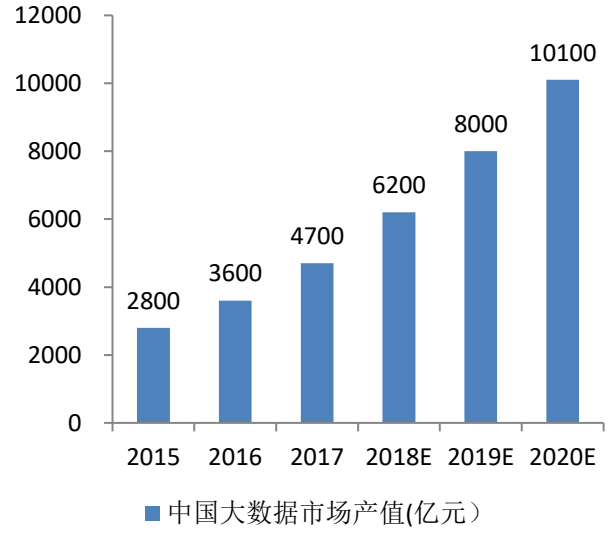
图表 9：国内大数据相关政策（非完全统计）

时间	政策
2012-07	国务院发布《“十二五”国家战略性新兴产业发展规划》，明确提出支持海量数据存诸、处理技术的研发和产业化
2013-07	《上海推进大数据研究与发展三年行动计划（2013-2015 年》发布，攻克关键技术，研制大数据核心装备，形成大数据领域的核心竞争力
2013-08	国务院《关于促进信息消费扩大内需的若干意见》，推动商业企业加快信息基础设施演进升级，增强信息产品供给能力，形成行业联盟，制定行业标准，构建大数据产业链，促进创新链与产业链有效嫁接
2014-02	贵州《关于加快大数据产业发展应用若干政策的意见》，打造大数据产业发展应用高地，建成全国领先的大数据资源中心和大数据应用服务示范基地
2014-03	大数据首次写入政府工作报告，设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展
2015-03	国务院制定“互联网+”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展，引导互联网企业拓展国际市场
2015-04	发改委《创新投资管理方式建立协同监管机制的若干意见》，提出运用互联网和大数据技术来创新监管的方式。
2015-08	国务院正式印发《促进大数据发展的行动纲要》，成为我国发展大数据产业的战略性指导文件
2016-03	《十三五规划纲要》提出“实施国家大数据战略”，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新
2017-01	工信部印发《大数据产业发展规划（2016—2020 年）》，到 2020 年技术先进、应用繁荣、保障有力的大数据产业体系基本形成。大数据相关产品和服务业务收入突破 1 万亿元，年均复合增长率保持 30%左右，加快建设数据强国，为实现制造强国和网络强国提供强大的产业支撑
2017-10	中国共产党第十九次全国代表大会报告，提出加快建设创新型国家，加强应用基础研究，拓展实施国家重大科技项目，突出关键共性技术、前沿引领技术、现代工程技术、颠覆性技术创新，为建设科技强国、质量强国、航天强国、网络强国、交通强国、数字中国、智慧社会提供有力支撑
2017-12	中共中央政治局就实施国家大数据战略进行第二次集体学习，习近平总书记在主持学习时强调深入了解大数据发展现状和趋势及其对经济社会发展的影响，分析我国大数据发展取得的成绩和存在的问题，提出要推动大数据技术产业创新发展；构建以数据为关键要素的数字经济；运用大数据提升国家治理现代化水平；运用大数据促进保障和改善民生；切实保障国家数据安全，加快建设数字中国
2018-04	首届数字中国建设峰会召开，会上国家互联网信息办公室发布了《数字中国建设发展报告（2017）》，《报告》总结了党的十八大以来数字中国建设取得的重大成就和基本经验，评估了“十三五”信息化发展主要目标、重大任务、重点工程和优先行动的进展情况，分析了数字中国建设面临的形势，提出了下一步努力方向

来源：全球科技经济瞭望、中国经济报告、政府网站、公开资料、国联证券研究所

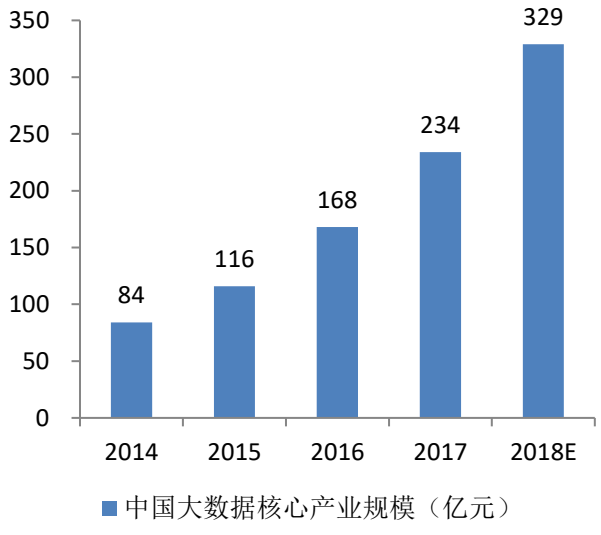
对于中国大数据产业的规模，目前各个研究机构均采取间接方法估算。根据信通院数据，2017 年中国大数据产业规模（包括数据资源建设、大数据软硬件产品的开发、销售和租赁活动，以及相关信息技术服务）为 4700 亿元人民币，同比增长 30%，且预计 2020 年这一规模有望赶超 1 万亿，年均复合增速近 30%。其中，大数据核心产业规模 2017 年为 234 亿元，同比增长 39%，预计 2018 年为 329 亿。

图表 10：中国大数据市场产值



来源：《大数据白皮书（2018）》、信通院、国联证券研究所

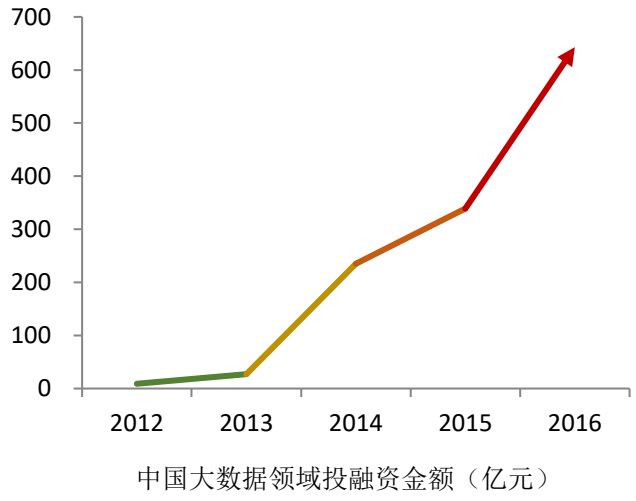
图表 11：大数据核心产业规模



来源：《数字中国建设发展报告（2017）》、信通院、国联证券研究所

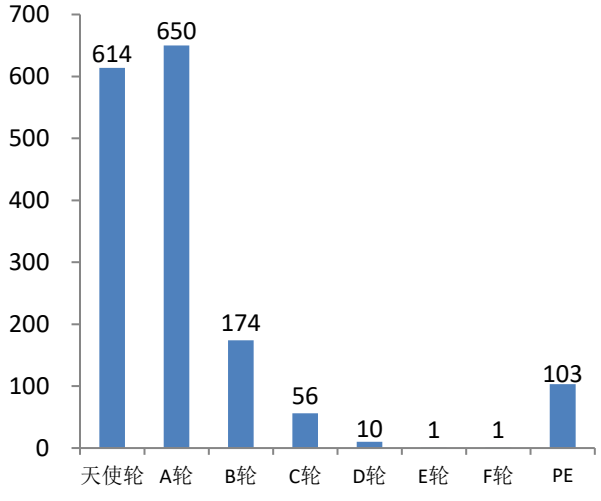
同时，大数据投融资市场也持续升温。根据信通院数据，2012-2016 年期间，国内共发生大数据投融资事件超 1600 件，统计公布金额的 1300 余起投资，其融资总额达 1200 多亿，2016 年同比增长 189.7%。轮次上，A 轮占比最高为 40%，天使轮次之为 38%；方向上，数据分析、应用项目等创新企业最受资本追捧。

图表 12：中国大数据领域投融资金额



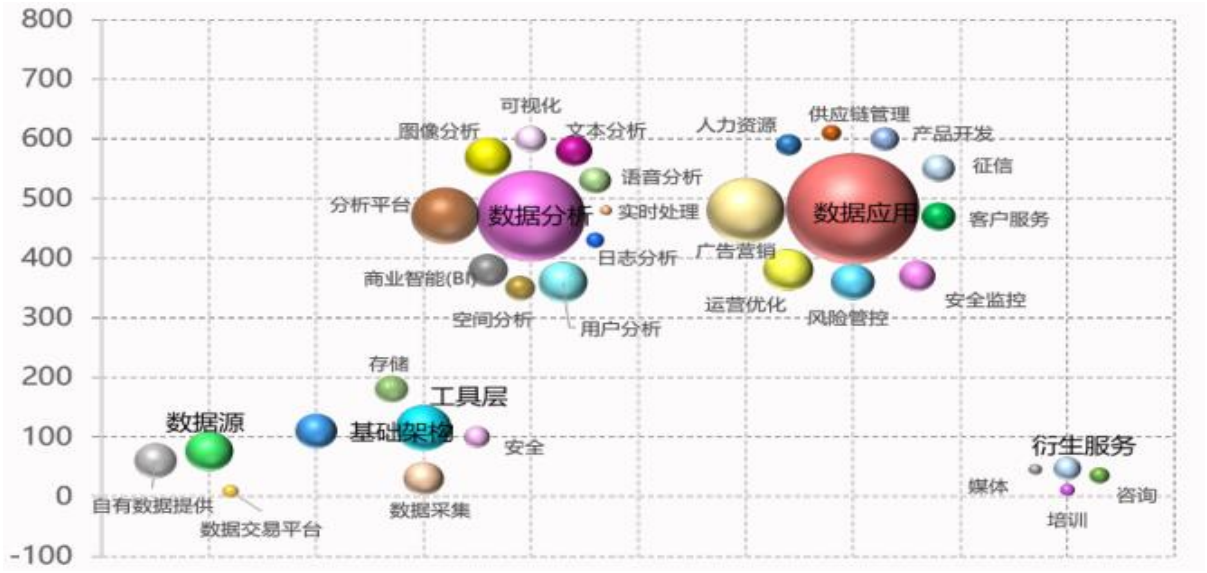
来源：信通院、数据猿、国联证券研究所

图表 13：中国大数据领域投融资轮次分布（次）



来源：信通院、数据猿、国联证券研究所

图表 14：2012-2016 各产业项目融资情况（单位：亿元）



来源：中国大数据和人工智能产业分析平台、信通院、数据猿、国联证券研究所

3. 大数据产业链：数据为源、分析为核、应用为王

分析大数据产业链，主要涵盖数据来源、数据管理与分析（包括集成、存储、安全、挖掘、分析等）、数据应用。

图表 15：大数据产业链图谱

来源：艾瑞咨询、国联证券研究所

注：此图仅为示意图，并未将所有企业列出，且排名不分先后

3.1.数据来源：政府、BAT、运营商等是当前大数据的主要拥有者

政府部门、BAT 为代表的互联网企业、运营商是当前中国大数据的主要拥有者。除此之外，利用网络爬虫或公开应用程序接口 API 等途径对网络数据进行采集也是一大重要来源。在大数据时代，拥有数据就拥了核心资源：工业时代，石油是最大的巨头，数据时代，BAT 等因为拥有最多、最全的搜索、电商和社交数据，也成为绝对的王者。此外，一些在细分领域拥有入口资源的公司也是稀缺标的，如已发布位置大数据平台的四维图新等。

3.2.数据管理与分析：存储是支撑、安全是保证、分析是核心

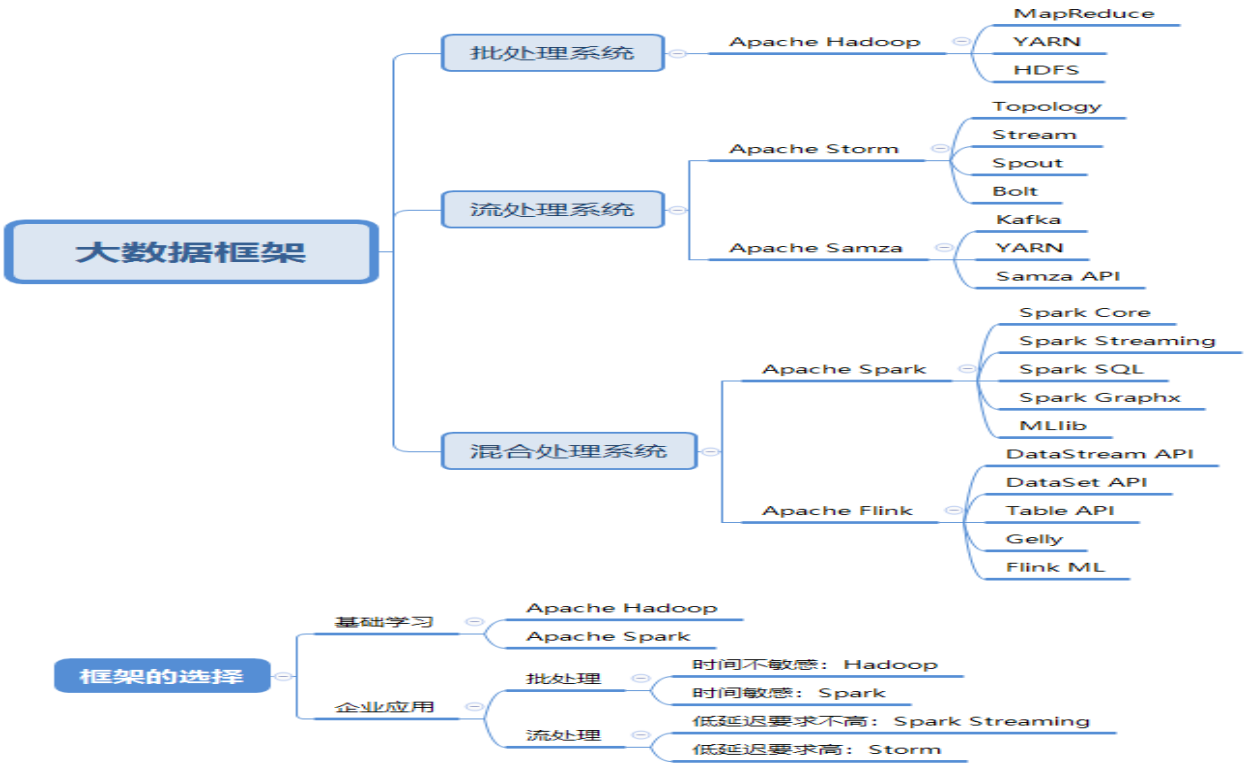
数据管理与分析位于产业中游，基于多种处理框架及算法，数据管理负责数据的集成、存储、安全等环节；数据分析按应用类型包括 AI、BI、可视化分析等，按数据类型包括图像、文本、视频、语音分析等。

其中，数据存储是产业链的支撑，参与者以传统数据库企业为主，国际上有 IBM、Oracle、Intel、Green-plum 等；国内主要有华为、中兴、同有、浪潮、中科曙光等，各家企业针对大数据应用的具体领域开展数据库架构和数据组织管理研究，形成各自的优势产品。数据安全是产业发展的重要保障，渗透数据存储、传输、交互的各个环节，主要参与方包括赛门铁克、360、启明星辰、绿盟科技、美亚柏科等。而**产业链最核心的当属数据分析与挖掘，其能力直接决定着大数据应用的推广程度和范围**。数据分析一是从大量的结构化、半结构化、非结构化数据中分析出计算机可以理解的语义信息或知识，二是对隐性的知识，如关联情况、意图等进行挖掘。常用的方法包括分类、聚类、关联规则挖掘、序列模式挖掘等，国际上主要参与者包括谷歌、亚马逊、Facebook、IBM、甲骨文、微软等，国内主要包括海康威视、科大讯飞、BAT、网易、智慧星光、思必驰等。

3.2.1 数据处理框架：Hadoop、Spark 是应用较为广泛的两种框架

数据处理框架按所处理的数据形式及得出结果的时效性分类，可分为批处理系统和流处理系统。批处理主要操作大规模数据集，包括将大任务分解为小任务，分别在集群中的节点上并行计算，可根据中间结果重新组合数据，再计算和组合最终结果。而流处理则是对由连续不断的单条数据组成的数据流进行计算，强调的是处理结果的时效性。典型的批处理框架是 Apache Hadoop，典型的流处理系统是 Apache Storm，还有一种同时具备批处理及流处理能力的混合系统，如 Apache Spark。其中，Hadoop、Spark 是应用较为广泛的两种框架。

图表 16：大数据处理框架（非完全统计）



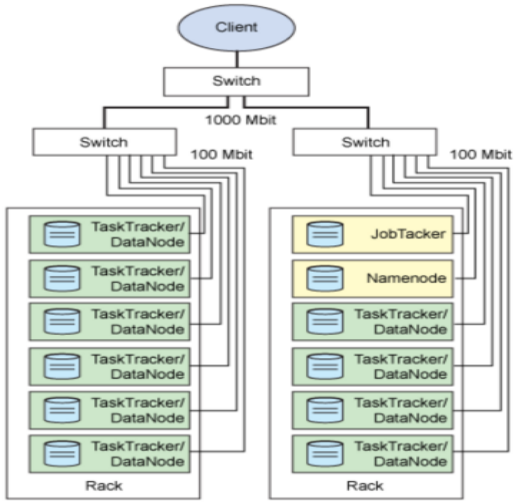
来源： 数据派 THU、国联证券研究所

➢ Hadoop

Apache Hadoop 是首个在开源社区获得极大关注的大数据处理框架，由 Apache 基金会于 2005 年秋作为 Lucene 的子项目 Nutch 的一部分正式引入。该项目最早用于探索网页搜索，Yahoo 在最初阶段做出了较大贡献，后 Hadoop 发展成能够为分布式数据提供各种服务的运算架构。

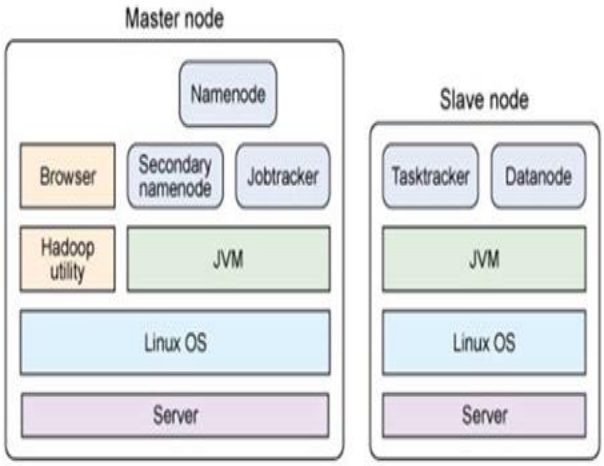
HDFS（Hadoop Distributed File System）和 MapReduce 是 Hadoop 的核心设计。两者分别是 Google File System（GFS）、Google MapReduce 的开源实现（谷歌三宝 MapReduce、GFS 和 BigTable，具体可见谷歌著名的三篇大数据论文，Hadoop 亦参考于此）。HDFS 是一种分布式文件系统层，可对集群节点间的存储和复制进行协调；MapReduce 是适合海量数据处理的编程模型，基本思想是“分而治之、然后归约”，可将大任务分解为多个小任务并行执行，其工作分 Map、Reduce 两个阶段：Map（映射）函数可理解为初略归类、分解任务，包括加载、解析、转换、过滤数据；reduce（归约）函数可理解为精简结果得到最终结果，负责把分解后多任务处理的结果汇总起来，处理的是 Map 输出的一个子集。

图表 17：Hadoop 物理结构



来源：与非网、千锋、国联证券研究所

图表 18：单点物理结构

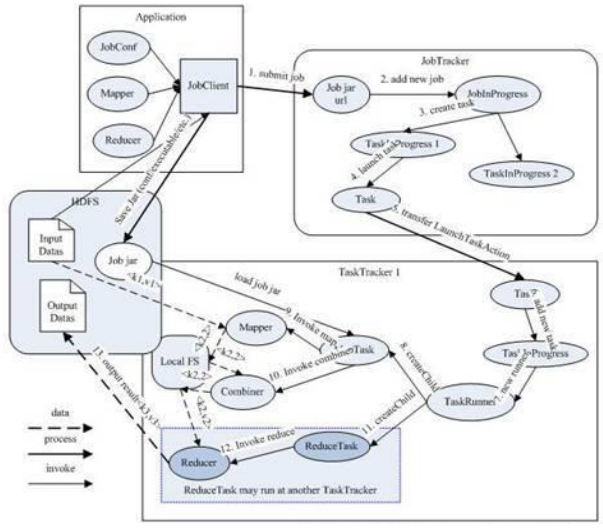


来源：与非网、千锋、国联证券研究所

备注：黄色为主节点；绿色为从节点；Switch：交换机；Rack：机柜；JVM：Java 虚拟机；NameNode[2]：名称节点； Secondary NameNode[3]：辅助名称节点； DataNode[4]：数据节点； JobTracker[5]：作业跟踪器； TaskTracker[6]：任务跟踪器。

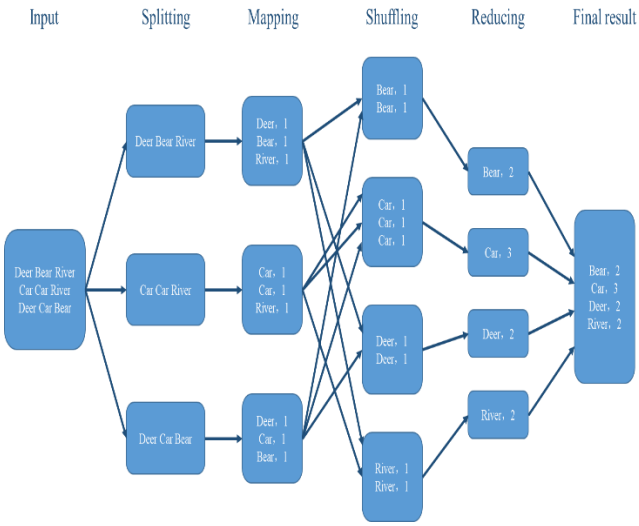
NameNode、Secondary NameNode、DataNode 组成 HDFS 体系； JobTracker、TaskTracker 是 MapReduce 的两个后台进程。

图表 19：Hadoop MapReduce 运行流程



来源：百度图片、CSDN、国联证券研究所

图表 20：MapReduce 示例（统计单词）



来源：国联证券研究所

Hadoop 一出现就受到众多大公司的青睐，Yahoo 、LinkedIn、Fox 互动媒体、默多克传媒、MySpace 等均有运用，同时也引起了研究界的普遍关注。随后，一系列围

[2] NameNode，对整体分布式文件系统进行总控制，记录全部元数据分布存储的状态信息，如文件如何分割成数据块、数据块会存储到哪些节点，以及对内存和 I/O 进行集中管理。实际应用中用户首先访问 Namenode，通过该总控节点获取文件分布的状态信息，再与具体的数据节点通信。每个集群只有唯一一个 NN。

[3] Secondary NameNode，监控 HDFS 状态的辅助后台程序，可以作为备用 NameNode 使用，但目前还不能自动切换。

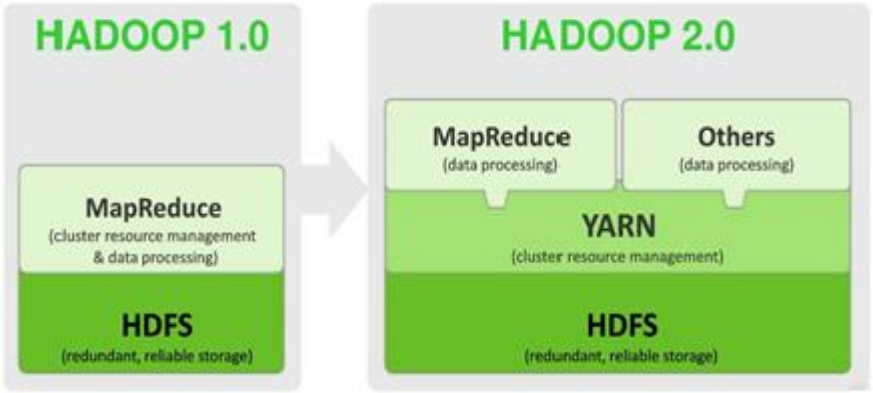
[4] DataNode，负责把 HDFS 数据块读、写到本地文件系统。

[5] JobTracker，用于处理作业（用户提交的代码）的后台程序，决定有哪些文件参与作业的处理，然后把作业切割成多个小 task，并把它们分配到所需数据所在的子节点。每个集群只有唯一一个 JT。

[6] TaskTracker，位于每个从节点，与 DN 结合（程序、数据物理节点就近原则），管理各自节点上的 task（由 JT 分配），每个节点只有一个 TT，但一个 TT 可启动多个 JVM，用于并行执行 map 或 reduce 任务，它与 JT 交互通信，可告知 JT 子任务完成情况。

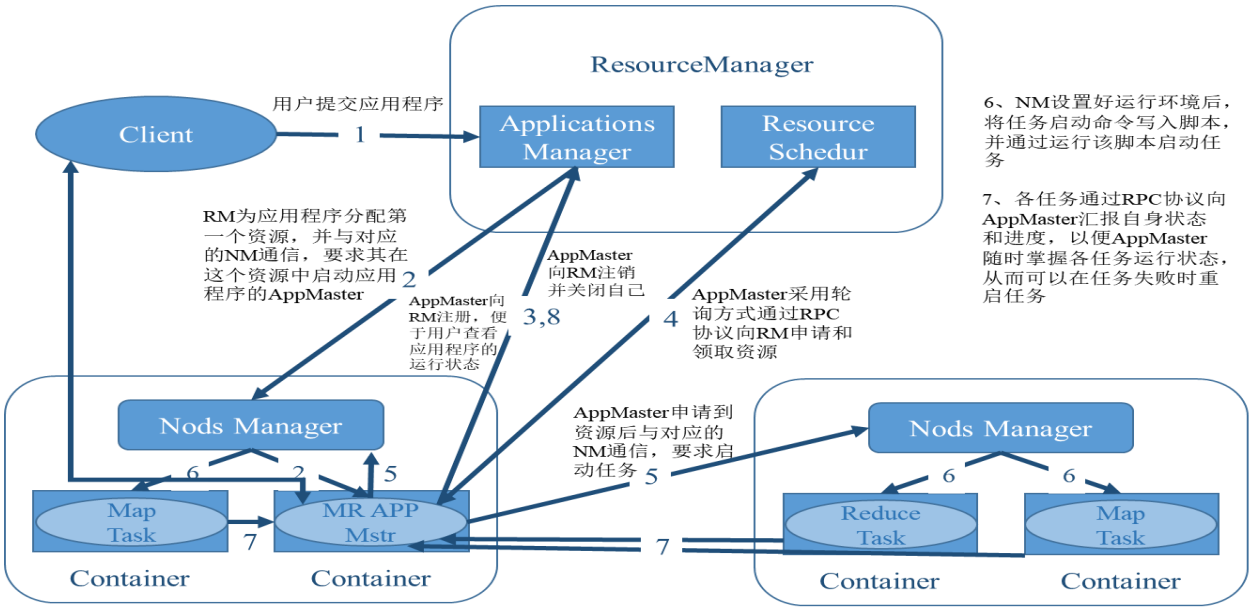
绕 Hadoop 的开源技术得到开发，生态不断丰富。如：Hive 提供数据仓库功能，包括数据抽取、转换、装载（由 Facebook 贡献）；HBase 实现海量结构化表的实时读写访问功能，类似谷歌的 BigTable；Cassandra 通过复制数据来提供容错数据存储功能。而 YARN[7]（Yet Another Resource Negotiator）的引入，更是让 Hadoop 超越 MapReduce 程序，支持其他更多的分布式应用。

图表 21：HaDoop2.0 引入 YARN



来源：比特网、国联证券研究所

图表 22：YARN 运行流程



来源：CSDN、itboth、国联证券研究所

备注：MRAppMaster：类似于 1.0 的 JT，但不包含资源管理。负责任务划分、资源申请并将之二次分配给 Map 和 Reduce Task、任务状态监控和容错

[7] YARN 是 Hadoop 集群的资源管理系统，其出现源于 MRv1 的 JT 单点故障、4000 节点上限、资源利用率低等缺陷。YARN 重构的思想是将 JT 的资源管理及作业调度/监控功能分离，主要方法是创建一个全局的 ResourceManager 和若干个针对应用程序的 ApplicationMaster。

Hadoop 具备拓展性、容错性和高效性等优点，更为重要的是其低成本。在这之前，大数据功能通常只能从商业软件供应商处依靠专门的硬件获取，而开源的 Hadoop 使数据存储和处理能力——这些本只有像谷歌或其他商用运营商类公司才具备的能力，在普通商用硬件上也得到应用，大大降低了使用大数据的先期投入，并且具备了使大数据接触到更多潜在用户的潜力。(《大数据云图》，大卫·芬雷布著)

图表 23：Hadoop 特性

特性	内容
拓展性	Hadoop 的分布式存储和分布式计算基于集群节点完成，决定了其可以扩展至更多的集群节点
容错性	通过分布式存储，Hadoop 能够自动保存多份副本，当数据处理请求失败后，会自动重新部署计算任务
高效性	接收到数据请求后，Hadoop 可以在数据所在的集群节点上并发处理
低成本	可运行在普通商用硬件，降低了使用大数据的先期投入

来源：CSDN、国联证券研究所

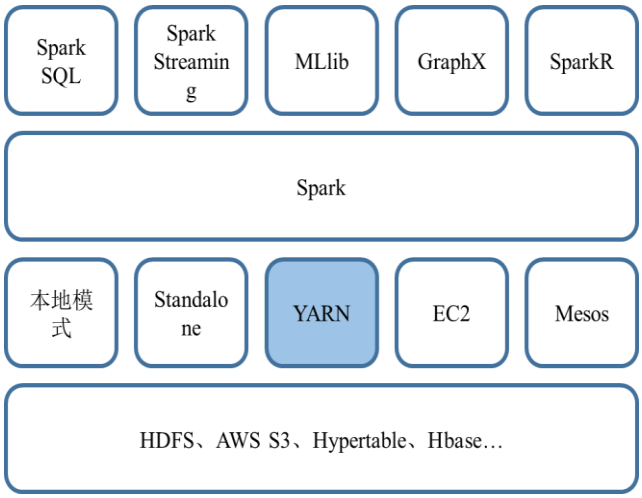
- Spark

Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎。2009 年，美国加州大学伯克利分校的 AMPLab 立足内存计算，从多迭代批量处理出发，兼顾数据仓库、流处理、机器学习和图计算等多种计算范式，正式将 Spark 作为研究项目，并于 2010 年进行了开源。(《Spark 核心技术与高级应用》，于俊等著)

Spark 最初的设计受到 MapReduce 的启发，但相对于传统 MapReduce 中间结果需写入磁盘且无法有效支持迭代的缺陷，Spark 通过有向无环图 DAG(Directed Acyclic Graph)、弹性分布式数据集 RDD（Resilient Distributed Dataset）可实现执行优化及内存计算，大幅提高了数据处理能力。Spark 在内存中的运行速度是 Hadoop MR 的 100 倍，磁盘中运行速度是其 10 倍，因此，对于处理大批量、静态且时间不敏感的任务时 Hadoop MR 性价比好，而对于流数据、低时延、迭代处理等计算要求则 Spark 表现更优。

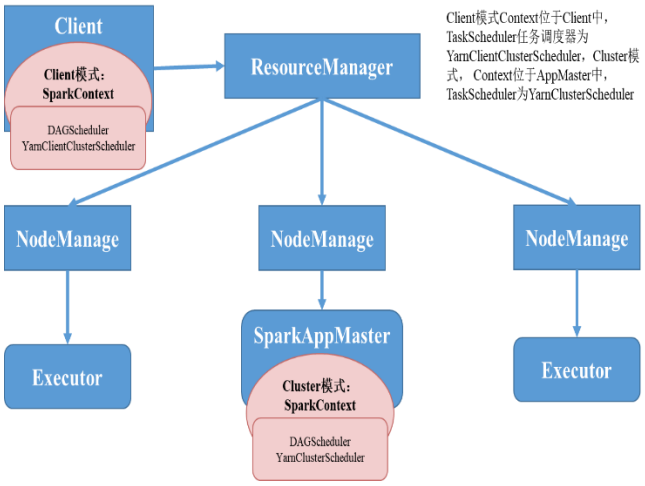
以 Spark 为核心的生态圈，最底层为分布式存储系统，如 HDFS、AWS S3、Hypertable 或其他格式；资源管理采用 Mesos、YARN 等集群资源管理，或 Spark 自带的独立运行模式及本地模式。同时，基于其基础平台 Spark 又扩展了一系列组件，主要包括支持交互式查询的 Spark SQL、实时数据处理的 Spark Streaming、机器学习的 MLlib、图计算的 GraphX、统计分析的 SparkR 等，各种程序组件并成软件栈与 Spark 核心 API 高度整合，可用来完成各种大数据运算，充分体现了 Spark“one stack to rule them all”的理念。

图表 24：Spark 框架构成



来源：《Spark 核心技术与高级应用》、国联证券研究所

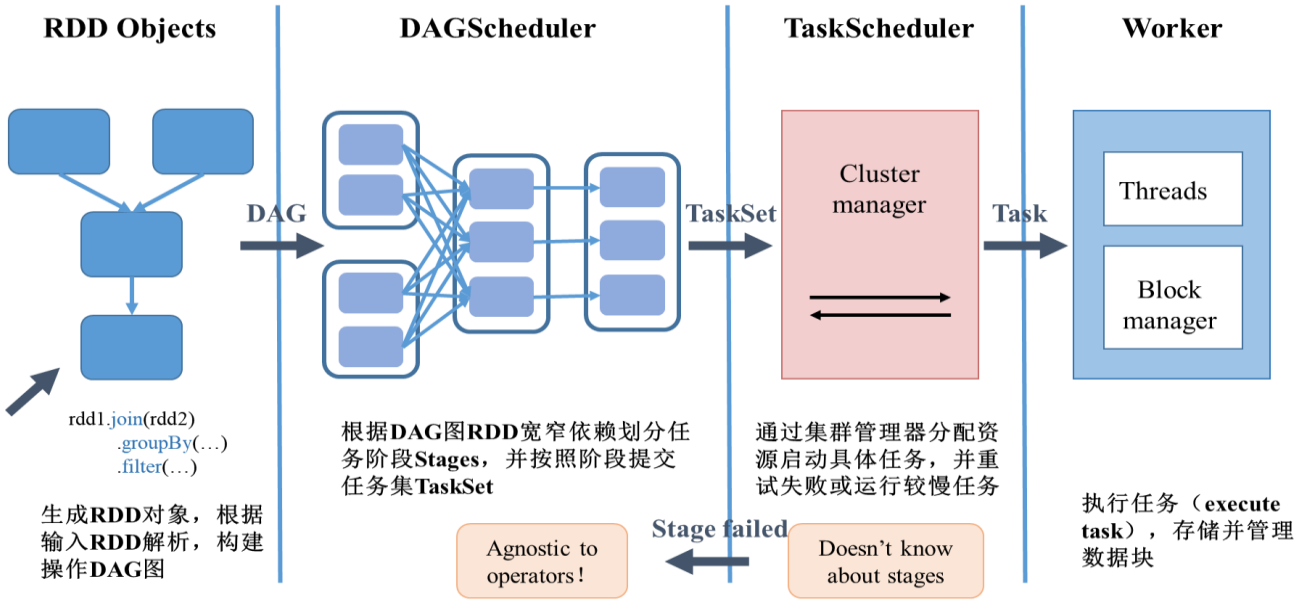
图表 25：基于 YARN 的 Spark 架构（类 MR-YARN）



来源：《Spark 核心技术与高级应用》、cnblogs、国联证券研究所

备注：YARN 模式分 Client 客户端及 Cluster 集群模式，YARN-Cluster 模式时 SparkAM 进行 SparkContext 初始化，类似 Standalone 模式，包含 DAGScheduler（负责根据 DAG 图划分 Stage，提交 TaskSet）和 YarnClusterScheduler（负责任务调度 TaskScheduler）。

图表 26：Spark 作业处理调度框架



来源：《Spark 核心技术与高级应用》、国联证券研究所

备注：作业调度过程包括生成 RDD 对象、构建 DAGScheduler、任务调度、作业执行等。创建 RDD 对象，然后通过转换（transformation）（lazy，只产生标记）生成新的 RDD，当遇到执行（action）操作时提交作业并进行计算：执行操作调用 runjob 方法，将代码作为 job 提交给 DAGScheduler，并绘制 DAG 图，然后 DAGScheduler 根据 DAG 图 RDD 依赖关系划分 Stage，对应每个 Stage 生成 TaskSet，将 TaskSet 交给 TaskScheduler，由 TaskScheduler 调动 Executor 进行任务的计算。

Spark SQL：是 Spark 的结构化数据处理模块，可以看做是分布式 SQL 查询引擎。

Spark Streaming：为 Spark 提供流处理能力。Spark 在设计之初也只用于批处理，为适应流处理模式，微批次（Micro-Batch）概念应运而生，原理是把一小段时间内的接入数据作为微批次来处理，极大地重用了核心模块（Spark Core），但延时比 Storm

等专用流处理框架高。

Spark MLlib：是 Spark 对常用的机器学习算法的实现库，包括相关的测试及数据生成器。MLlib 目前支持四种常见的机器学习问题：二元分类、回归、聚类、协同过滤，以及一个底层的梯度下降优化基础算法。

Spark GraphX：是 Spark 中用于图计算的组件，通过引入将有效信息放在顶点和边属性上的有向多重图，再加上日益增长的图算法和 builders 构造器集合，GraphX 为图计算、图挖掘提供了一站式解决方案，可有效处理社交网络关系网及语言建模等。

SparkR：是 AMPLab 发布的 R 开发包，为 Spark 提供了一个轻量级的前端。Spark 具备快速（fast）、可扩展（scalable）、交互（interactive）等特点，R 具备统计（statistics）、封装（packages）、绘制（plots）的优势，R 与 Spark 有效结合，解决了 R 语言中无法级联扩展的难题，也极大地丰富了 Spark 在机器学习方面能够使用的 Lib 库。

图表 27：Spark 特性

特性	内容
快速	利用 DAG 及 RDD，Spark 应用在内存中的运行速度是 Hadoop MapReduce 的 100 倍，磁盘运行速度是 10 倍
易用	Spark 支持 Scala、Java 及 Python 等编程语言，并提供了 80 多个高级运算符
通用	Spark 提供了大量的库，包括 SQL、DataFrames、MLlib、GraphX、Streaming，开发者可以在同一个应用程序中无缝组合使用这些库
支持多种资源管理器	Spark 支持 Hadoop YARN，Apache Mesos 及其自带的独立集群资源管理器等，也可以读取 HDFS、HBase、AWS S3 等多种数据源

来源：百度百科、国联证券研究所

随着设计的不断完善，Spark 已成为继 Hadoop MapReduce 之后，最具影响力的大数据框架之一。在快速查询、实时日志采集处理、业务推荐、定制广告、用户图计算等多个领域，Spark 都实现了较好的应用，IBM、Facebook、Intel、阿里巴巴、腾讯、网易、科大讯飞等都有实际业务运行其上，隐有企业应用王者之势。

图表 28：Spark 在各领域的应用

领域	应用
快速查询系统	基于日志数据的快速查询业务，利用 Spark 快速查询及内存表等优势，可以实现大部分数据的即时查询，且内存查询比 hadoop 快百倍
实时日志采集处理	利用 Spark Streaming 进行实时日志数据采集，快速迭代处理，并综合分析，满足线上系统分析要求
业务推荐	通过 Spark 将业务推荐系统的小时、天级别的模型训练转变为分钟级别，有效优化相关排名、个性化推荐以及热点点击分析等。
定制广告	在定制广告业务方面需要大数据做应用分析、效果分析、定向优化等，借助 Spark 快速迭代的优势，实现了在“数据实时采集、算法实时训练、系统实时预测”的全流程实时并行高维算法，支持上亿的请求量处理；模拟广告投放计算效率高、延迟小，同 MapReduce 相比延迟至少降低一个数量级
用户图计算	利用 GraphX 解决多种计算场景：基于度分布的中枢节点发现、基于最大连通图的社区发现、基于三角形计数的关系衡量、基于随机游走的用户属性传播等

来源：CSDN、国联证券研究所

3.2.2 数据处理算法：受益人工智能，神经网络算法关注度再次高涨

如果说一个稳定、可靠、高效的底层框架是数据处理的必要基础，那一系列优异的算法就是数据处理程序的设计范本。算法是解决特定问题求解步骤的描述，在计算机中表现为指令的有限序列，能够对一定规范的输入，在有限时间内获得所要求的输出，代表着用系统的方法描述解决问题的策略机制。

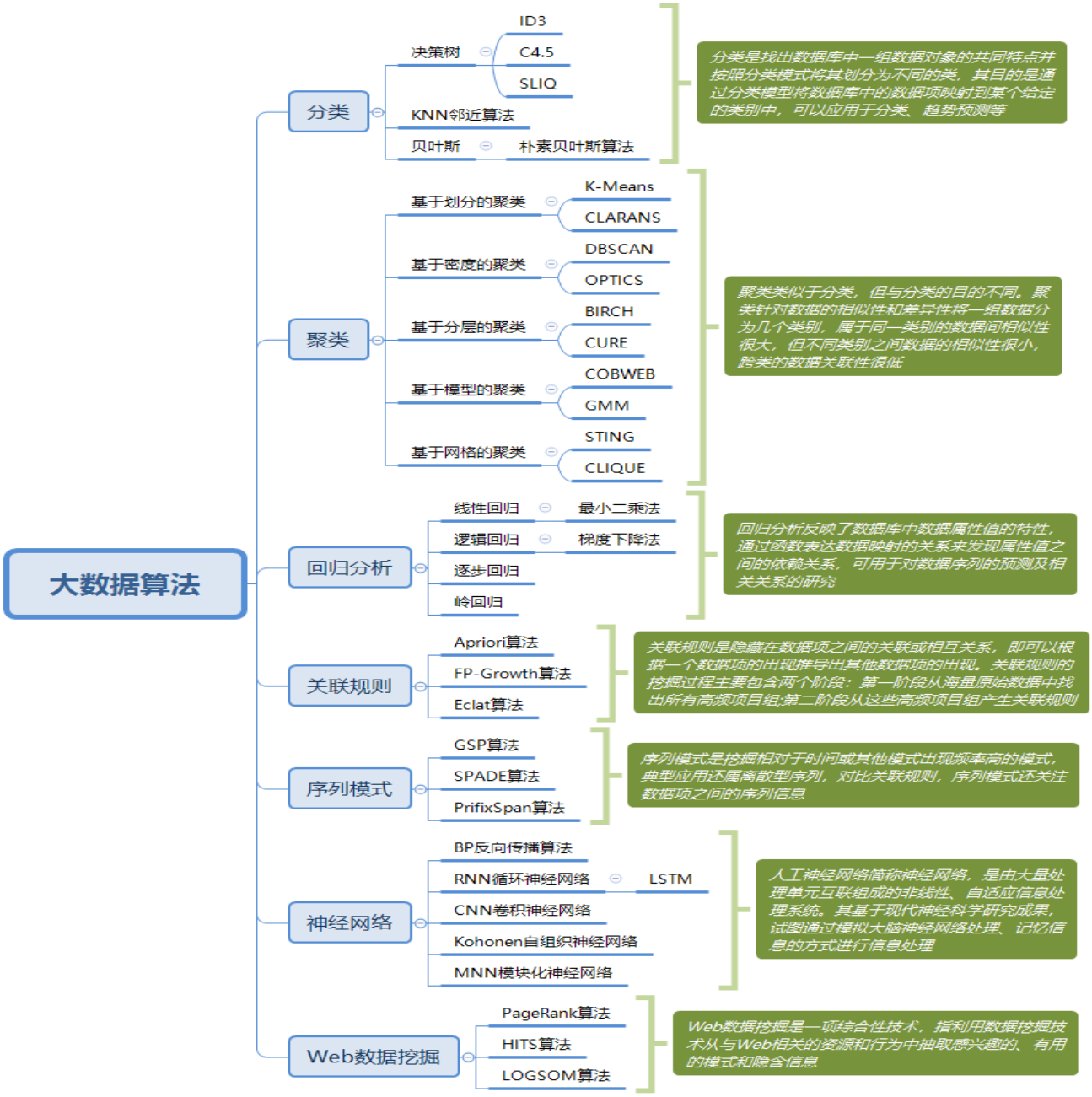
图表 29：评定算法优劣的依据

依据	内容
正确性	算法至少应该具有输入、输出和加工处理无歧义性、能正确反映问题的需求、能够得到问题的正确答案
可读性	算法可供人们阅读、理解和交流的容易程度
健壮性	当输入数据不合理时，算法也能做出相关处理，而不是产出异常或莫名其妙的结果，也称容错性
时间复杂度	执行算法所需要的计算时间
空间复杂度	执行算法需要消耗的内存空间

来源：百度百科、cnblogs、国联证券研究所

常用的数据处理算法包含分类、聚类、回归分析、关联规则、序列模式挖掘等。算法思想源远流长，发展到目前，可谓种类繁多，而受益于第三次人工智能浪潮，神经网络算法近来关注度再次高涨。

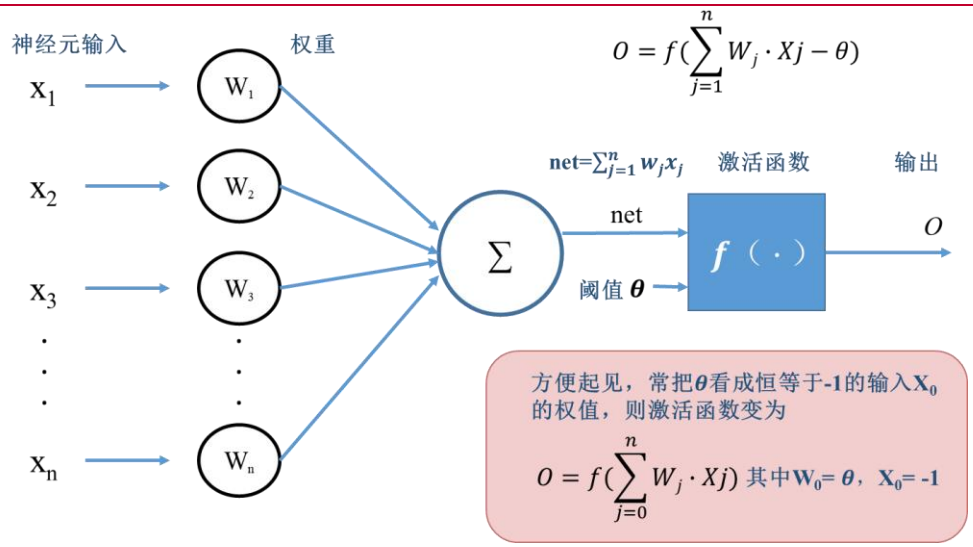
图表 30：大数据处理算法（非完全统计，由于神经网络算法近来关注度较高故单列）



来源：CSDN、cnblogs、国联证券研究所

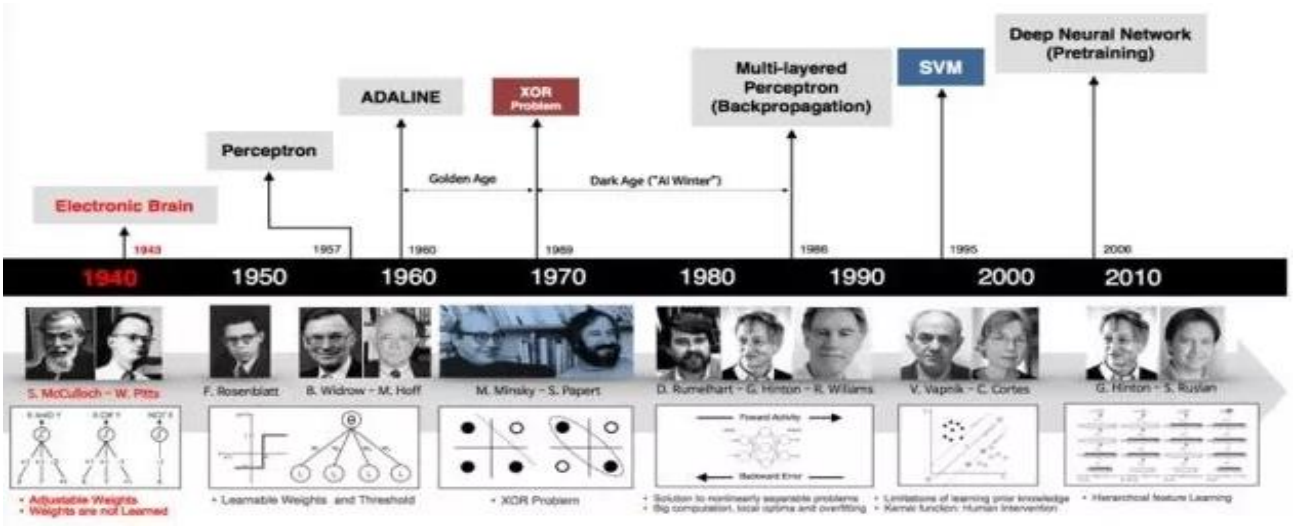
人工神经网络简称神经网络，是由大量处理单元互联组成的非线性、自适应信息处理系统。其基于现代神经科学研究成果，试图通过模拟大脑神经网络处理、记忆信息的方式进行信息处理。典型的神经网络模型有 BP 反向传播算法(back propagation)、RNN 循环神经网络（Recurrent Neural Network)、CNN 卷积神经网络（Convolutional Neural Network)、Kohonen 自组织神经网络等。

图表 31：神经网络处理单元模型（神经元）



来源：数据派 THU、docin、国联证券研究所

图表 32：神经网络算法发展历程



来源：数据派 THU、国联证券研究所

➢ BP 反向传播算法

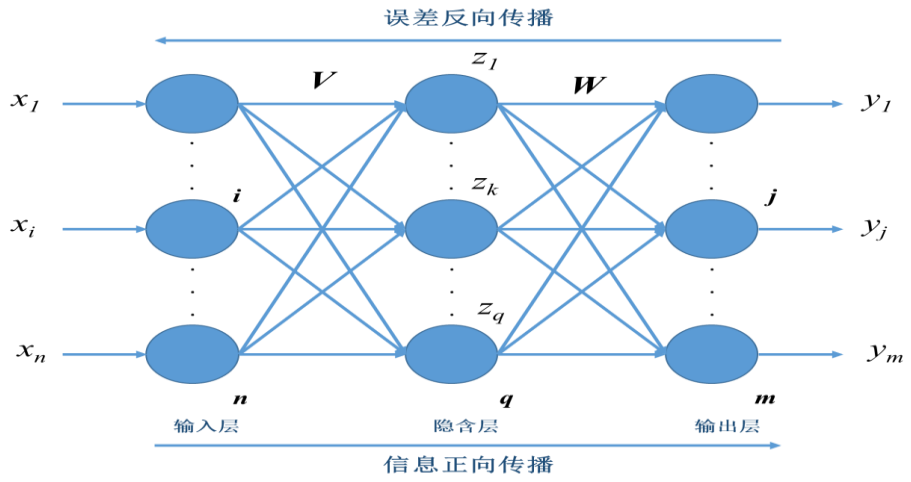
BP 算法是一种按误差逆传播方法训练的多层前馈网络，属迄今应用最成功的神经网络学习算法之一，大多神经网络使用 BP 算法进行训练。

BP 算法最早由 Werbos 于 1974 年提出，但当时并未受到应有的重视；1986 年 Rumelhart、McClelland 等学者出版《平行分布处理：认知的微观结构探索》一书，完整地提出了 BP 算法，系统地解决了多层网络中隐单元连接权的学习问题，并在数学上给出了完整的推导，助推了神经网络的第二次浪潮。

BP 算法的学习过程由信号的正向传播和误差的反向传播两个过程组成，属于有监督学习。其基本思想是将样本模式从网络的输入层输入，经隐含层逐层处理后传递至输出层，若输出层的实际输出与期望值一致则结束学习算法，若没有获得期望输出

结果则将误差进行反向传播；输出误差按原连接路径反转计算，在过程中根据误差信号不断调整各层神经元的权重及阈值，使误差达到最小。

图表 33：BP 算法结构图（3 层）



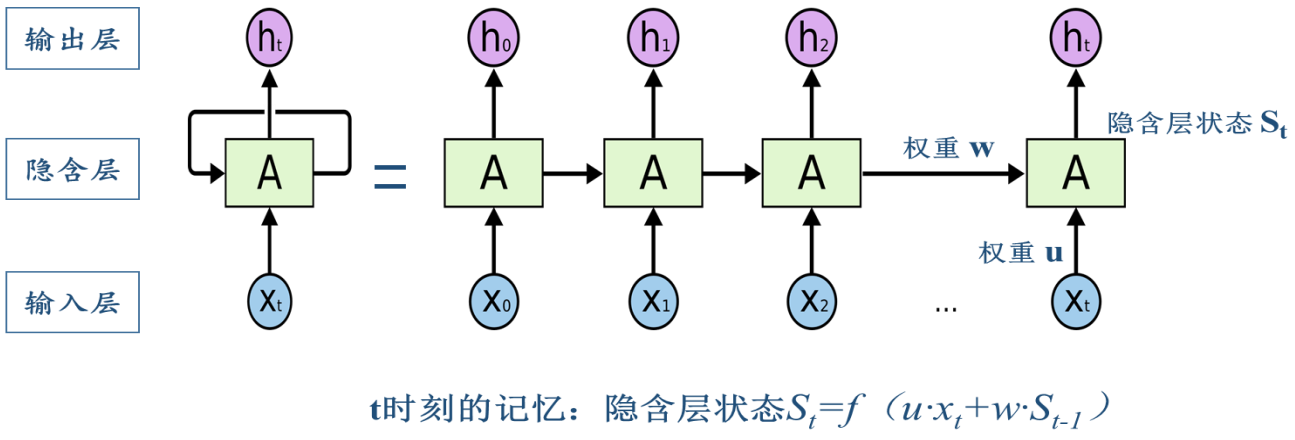
来源：CSDN、国联证券研究所

➢ RNN 循环神经网络

RNN 循环神经网络是一类适用于处理序列数据的神经网络模型，在语音识别、语言建模、翻译、图像标注等多个领域 RNN 都有很好的建树。

不同于基础前馈网络只在层与层之间建立权连接，RNN 通过自循环在同层级的神经元间也建立权连接，使得前面的信息得以记忆延续并应用于当前输出，用以处理存在序列关系的数据输入，实现信息的持久化。在 RNN 的计算过程中，序列数据前段信息通过隐含单元向后段传递，输入单元只与隐含单元相连，隐含单元不仅与输入、输出层连接还与前、后隐含单元连接，输出单元只接受隐藏单元的输入。

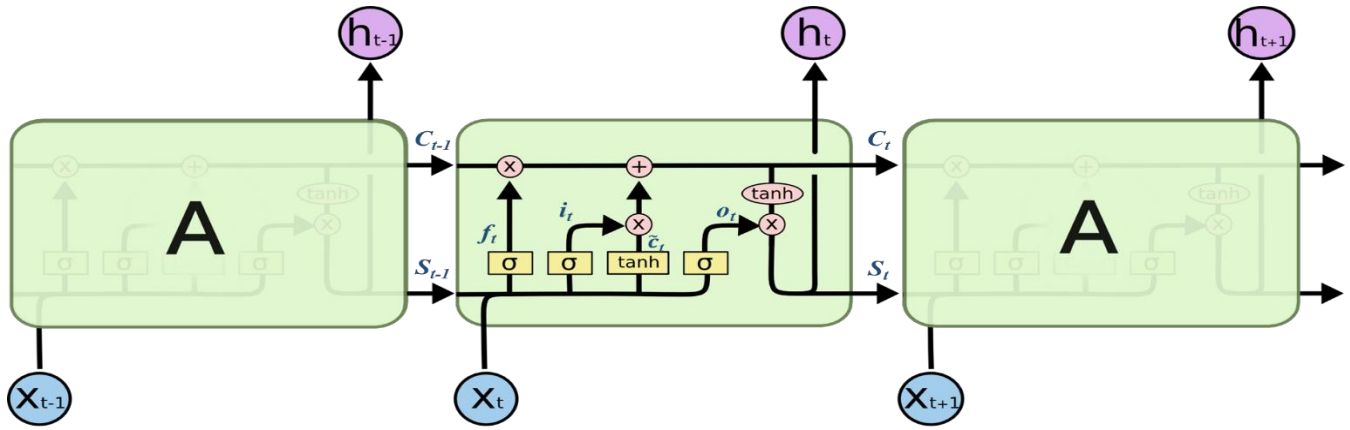
图表 34：RNN 循环展开结构



来源：github、CSDN、国联证券研究所

但由于梯度消失问题，RNN 存在无法记忆长程信息的弱点。1997 年，LSTM 长短期记忆网络(Long Short-Term Memory)应运而生，LSTM 由 Hochreiter、Schmidhuber 提出，并在近期被 Alex Graves 进行了改良和推广。LSTM 较 RNN 增加了 memory cell 单元，并通过遗忘门、输入门、输出门来删除、更新、输出 cell 中的数据内容，从而使得时间序列上的记忆信息可控，此举旨在克服误差回流问题，帮助 LSTM 能学习跨越 1000 步以上的时间间隔。

图表 35：LSTM 隐含单元结构



遗忘门$f_t=\sigma(W_f\cdot[S_{t-1},X_t]+b_f)$；输入门$i_t=\sigma(W_i\cdot[S_{t-1},X_t]+b_i)$；候选门$\tilde{c}_t=\tanh(W_c\cdot[S_{t-1},X_t]+b_c)$；输出门$o_t=\sigma(W_o\cdot[S_{t-1},X_t]+b_o)$；$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{c}_t$；$S_t=o_t\cdot\tanh(C_t)$；W为权重，b为阈值

来源：github、CSDN、数据派 THU、国联证券研究所

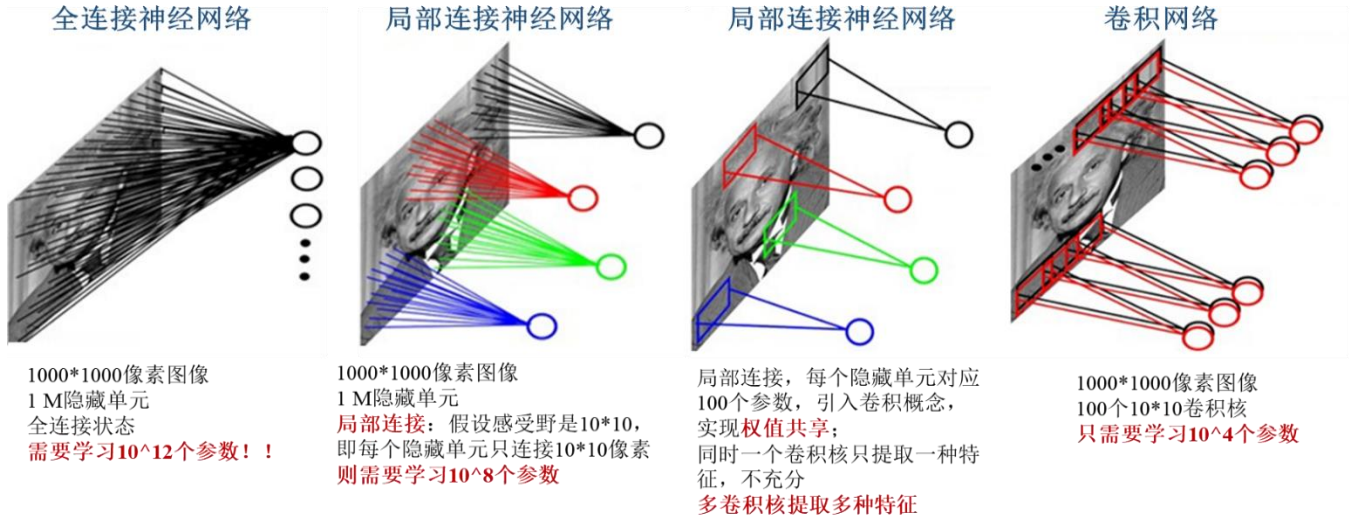
➢ CNN 卷积神经网络

CNN 卷积神经网络是为识别二维形状而特殊设计的一个多层感知器，这种网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性，多用于图像识别、人脸识别、视频分析、药物挖掘及游戏等。

CNN 的原理得益于 1962 年 Hubel、Wiesel 等提出的感受野[8]（receptive field）概念，Yann LeCun 最早将其应用于手写数字识别（1998 年），并被誉为“卷积神经网络之父”。CNN 模型强调的是卷积及池化过程：卷积是让卷积核在二维样本上按步长滑动，与对应位置数据计算乘积并求和，形成新矩阵，可以理解为通过过滤器（卷积核）提取样本局部特征形成特征图（feature map）；池化即下采样，是将局部区域中不同位置的特征进行聚合统计，主要起到降低数据维度的作用，可以保证特征的平移、旋转不变性，通常的方法有均值采样、最大值采样等。

[8] 1962 年，Hubel 等人通过对猫视觉皮层细胞研究发现，每一个视觉神经元只会处理一小块区域的视觉图像，即感受野概念。

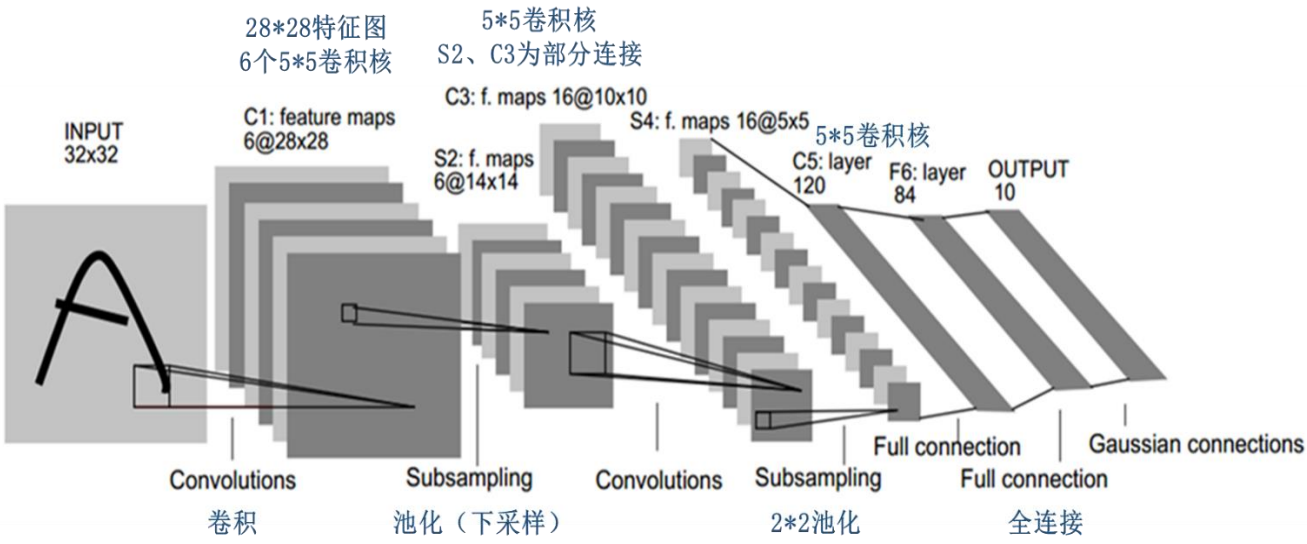
图表 36：卷积理念推演



来源：CSDN、国联证券研究所

CNN 这种局部感知、权重共享、池化的操作大大减少了参数，同时多卷积核实现了不同特征的提取，在正确识别的基础上降低了网络模型的复杂度，使之更类似于生物神经网络。

图表 37：CNN 经典结构（LeNet-5，Yann LeCun，1998）



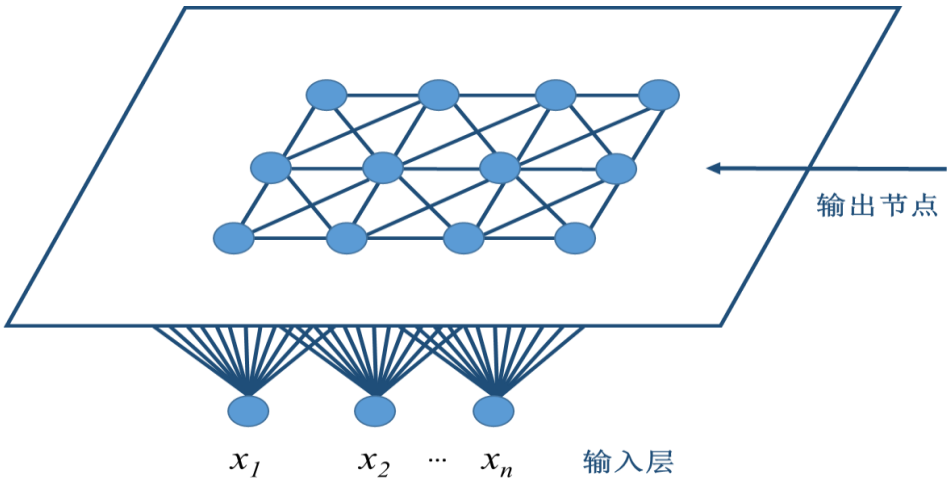
来源：CSDN、国联证券研究所

➢ Kohonen 自组织神经网络

自组织特征映射网络被称为 Kohonen 网，或 SOM（Self-organizing feature Map）网，通过自动寻找样本中的内在规律和本质属性，自组织、自适应地改变网络参数与结构，在模式识别、联想存储、样本分类、优化计算、机器人控制等领域中得到广泛应用。

1981 年芬兰学者 TeuvoKohonen 基于脑科学研究成果[9]提出 Kohonen 网。Kohonen 认为，神经网络在接受外界输入时，将会分成不同的区域，各区域对输入模式具有不同的响应特征，即不同的神经元以最佳方式响应不同性质的信号激励，从而在输出层形成一种拓扑意义上的映射图。此映射图中功能相同的神经元靠得较近，功能不同的神经元分得较开，自组织特征映射网络由此得名。

图表 38：Kohonen 网络基本结构（二维平面线阵）

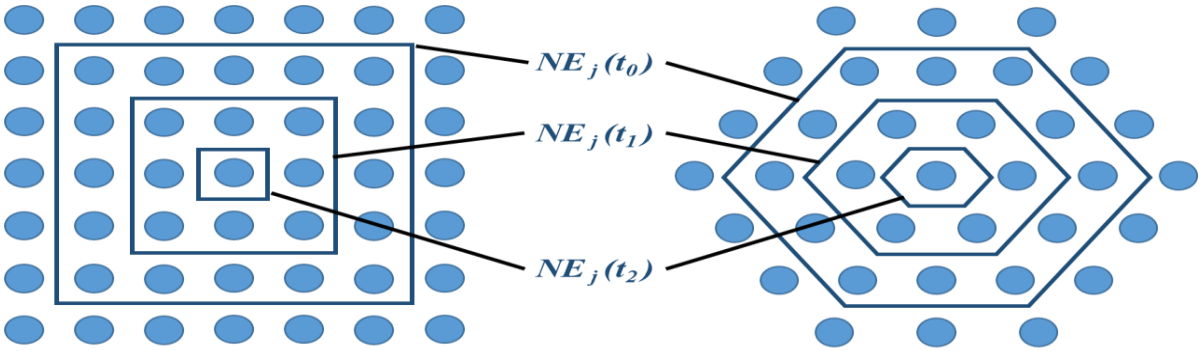


来源：《信息技术与信息化》、cnki、国联证券研究所

Kohonen 是一种竞争型网络，同一层神经元之间相互竞争，竞争胜利的神经元可修改与其相连的连接权值，胜利神经元某一半径邻域内的神经元不同程度得到兴奋，邻域外的则被抑制。并且，Kohonen 属于无监督自组织学习，在学习过程中只需提供学习样本，而无需提供理想的目标输出。Kohonen 天然具有降维的作用，适用于数据聚类等。

[9] 生物学研究表明，在人脑的感觉通道上，神经元的组织原理是有序排列的。当外界的特定时空信息输入时，大脑皮层的特定区域兴奋，并且类似的外界信息在对应的区域是连续映像的。如：生物视网膜中有许多特定的细胞对特定的图形较为敏感，当视网膜中有若干个接收单元同时受特定模式刺激时，就使大脑皮层中的特定神经元开始兴奋，输入模式接近，与之对应的兴奋神经元也接近；在听觉通道上，神经元在结构排列上与频率的关系亦十分密切，对于某个频率，特定的神经元具有最大的响应，位置相邻的神经元具有相近的频率特征，而远离的神经元具有的频率特征差别也较大。（CSDN、百度文库）

图表 39：领域示意图（可以是正方形或六角形等形状）



j是胜利神经元；NE_j是以j为中心不超过某一半径内的所有神经元集合；
$NE_j(t)$是t的函数，随着训练的进行，$NE_j(t)$的半径逐渐减小，最后剩下一个节点（或一组节点）
权值修正公式：$W_{ij}(t+1)=W_{ij}(t)+\eta(t)[X_i(t)-W_{ij}(t)]$；其中$\eta$为增益项，随时间变化逐渐下降到零，一般取$\eta(t)=1/t$或$\eta(t)=0.2(1-t/10000)$

来源：《信息技术与信息化》、cnki、国联证券研究所

3.3.数据应用：应用是完成产业商业化目标，实现价值的终点

应用为王，对大数据分析结果进行应用是完成产业商业化目标，实现价值的终点。经过近几年的发展，大数据应用已渗透政府、电信、金融、人力资源、医疗、物流、等多个行业，从产品角度而言，除传统的工具/产品化服务（精准营销、舆情监控等）外，整体式的解决方案亦愈加丰富。

图表 40：中国大数据应用领域企业

领域	企业
政府类	数字政通、美亚柏科、辰安科技、天源迪科、华宇软件、九次方大数据、国信优易等
电信类	天源迪科、东方国信、华为等
金融类	九次方大数据、WIND、同花顺等
人力资源类	智联招聘、猎聘、前程无忧等
医疗类	卫宁健康、创业软件、春雨医生、九安医疗、迪安诊断等
气象类	墨迹天气、华风气象等
环境及地理类	启迪桑德、雪迪龙、四维图新、高德等
生活类	小桔科技、无忧停车、电话邦等
物流类	Amazon、顺丰、联邦等
教育类	百度、新东方、华图、科大讯飞等
媒体类	浙报传媒、今日头条、华谊等

来源：贵阳大数据交易所、艾瑞咨询、国联证券研究所

我们认为应用市场的成熟程度与数据的完备性息息相关，当前，政府、BAT、运营商是数据源的主要拥有者，因此为这些领域服务的数据应用厂商拥有一定的先天优势，对比市场空间、政策倾向及惠及民生等方面，我们更为看好政务大数据及医疗大数据市场。而从产品形态看，整体解决方案商掌握多元技术、跨场景服务能力强、可解决客户的综合性需求，因此更容易树立标杆案例，灯塔效应明显。

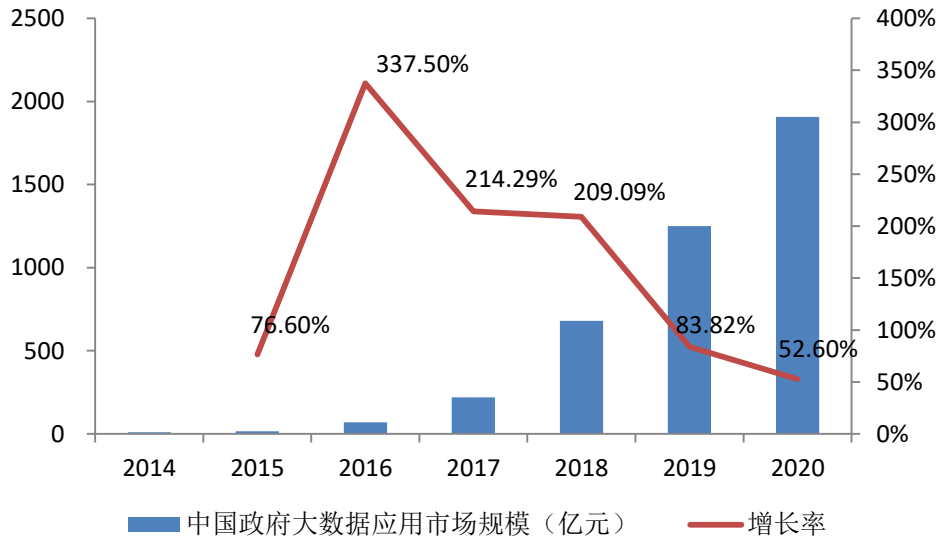
➢ 政府大数据

大数据是提升政府治理能力的重要方式之一。近年来，电子政务、政务系统等相关文件频发，尤其是 2017 年起“加快国务院部门和地方政府信息系统互联互通，形成全国统一政务服务平台”、“深入推进‘互联网+’行动和国家大数据战略”等要求陆续提出，为国内政府大数据建设提供了良好的政策环境。

政府部门信息化过程中积累了海量的政务数据，通过大数据应用不但可以改善对民众、企业的公共服务，也可以为政府管理决策辅以有效参考，帮助政府工作高效化、科学化。目前，政府大数据已渗透公安、税务、司法、金融、工商、海关、质监等多个部门，在公共安全领域如治安防控、情报研判、案情侦破等，交通管理领域如拥堵提醒、疏散管理、公交到站监测等，财税领域如逃税漏税分析、税改效果追踪、公共资源项目监管等，金融领域如企业征信、金融市场风险管控等，正做出越来越多的贡献。

根据贵阳大数据交易所统计，2014 年中国政府大数据应用市场规模为 9.06 亿元，2015 年，这一规模已快速增长至 16 亿元，预计今后几年，政府大数据应用市场规模仍将成倍增长，到 2020 年有望达到 1908 亿元。

图表 41：中国政府大数据应用市场规模



来源：贵阳大数据交易所、国联证券研究所

➢ 医疗大数据

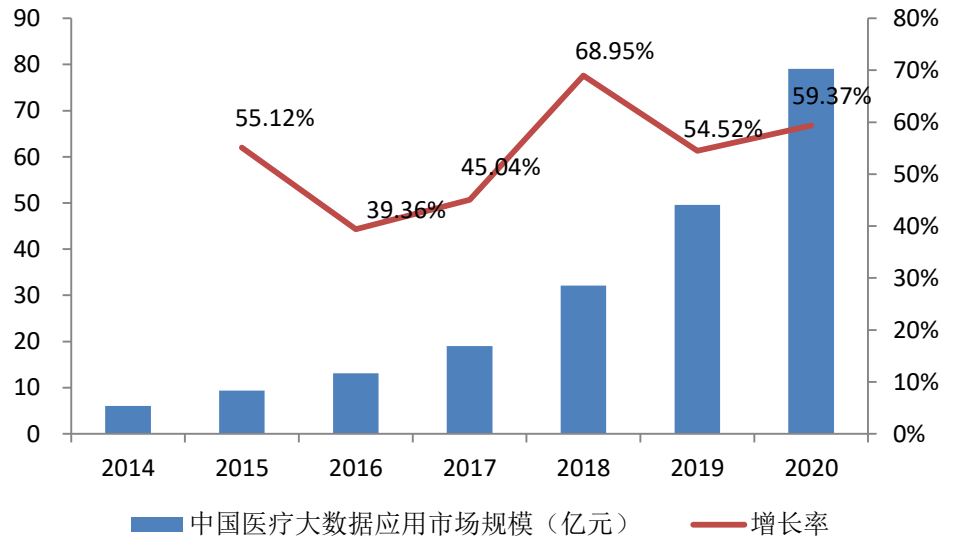
大数据是改善医疗矛盾的重要手段之一。如今，“效率较低的医疗体系、质量欠佳的医疗服务、看病难看病贵的就医现状”已成为民生焦点，大医院人满为患、社区医院无人问津、病人就诊手续繁琐等问题大都源于医疗信息不畅、医疗资源两极化、医疗监督体制不健全等。而随着互联网、大数据、云计算等信息技术快速发展，医疗市场的种种矛盾有望逐步得到解决。

就医疗大数据而言，其应用有效提高了医院长尾市场的信息流通，降低了广大受

众成本，能够使原有医疗服务体系更加完善、精准。对于医务人员，大数据可进行临床辅助决策、精准诊疗与个性化治疗、不良反应与差错分析提醒等；对于患者，大数据可助力自我健康管理、健康预测与预警、全生命周期健康档案建立等；对于管理者，大数据可提供精细化管理决策支持、感染爆发监控、疾病与疫情监测等；对于研究人员，大数据可服务其用药分析、药物研发等。

十二五“3521”工程[10]、十三五“开展健康中国云服务计划”、《新一代人工智能发展规划》、《关于促进“互联网+医疗健康”发展的意见》、《全国医院信息化建设标准与规范（试行）》等，都为中国医疗大数据应用的开展完善了信息建设基础。根据贵阳大数据交易所统计，2015 年中国医疗大数据应用市场规模为 9.4 亿元，预计随着应用领域不断增加，市场规模将不断扩张，到 2018 年形成 32 亿的规模，2020 年达到 79 亿。

图表 42：中国医疗大数据应用市场规模



来源：贵阳大数据交易所、国联证券研究所

4. 投资建议

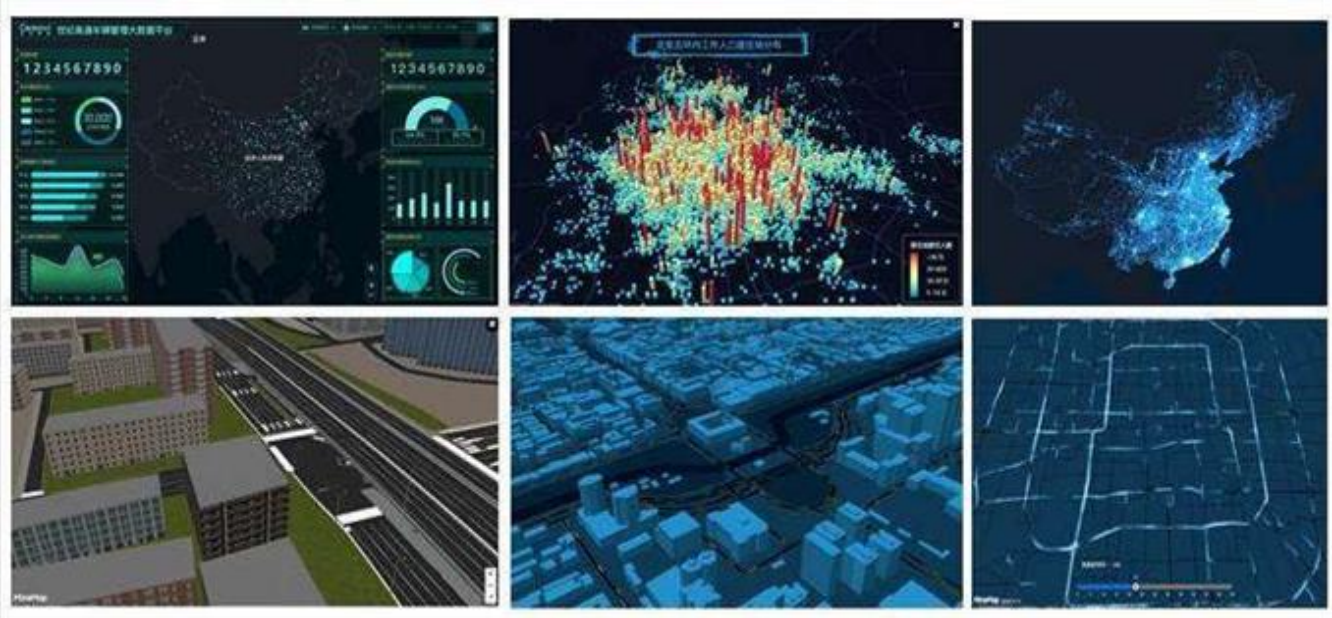
我们认为，数据是行业发展的源泉、存储是行业发展的支撑、安全是行业发展的保障，分析是行业发展的核心、应用是行业发展的价值实现。建议关注拥有位置领域入口资源的四维图新，布局芯片及 AI 服务器的中科曙光，以及掌握视频数据分析能力的海康威视，外加应用领域的智慧公安解决方案商美亚柏科、智慧医疗解决方案商创业软件等。

[10] “3521”工程包括 1）3 级卫生信息平台：建设国家级、省级和地市级 3 级卫生信息平台；2）5 项业务应用：加强公共卫生、医疗服务、新农合、基本药物制度、综合管理 5 项业务应用；3）2 个基础数据库：建设健康档案和电子病历 2 个基础数据库；4）1 个专用网络建设：建立一个卫生系统专用网络。

四维图新：

四维图新立足数字地图，经十余年的创新发展，已成为导航地图、导航软件、动态交通信息、乘用车和商用车车联网解决方案以及位置大数据服务领域的领导者。近年来，公司通过系列资源整合不断完善自身生态系统：收购 MapScape 成为全球领先的导航地图编译服务提供商，实现 NDS、RDF、GDF 等各种数据格式的编译转换，并面向宝马、戴姆勒等提供全球 NDS 数据编译服务；收购杰发科技布局 IVI 车载信息娱乐系统芯片、AMP 车载功率电子芯片，同时研发 MCU（BCM）车身控制芯片、TPMS 胎压监测芯片等，力争打破国外巨头垄断等等。未来，公司遵循“智能汽车大脑”战略，依靠高精度地图、高精度定位、算法、芯片、系统平台等优势，有望升级为自动驾驶解决方案提供商。

图表 43：四维图新位置大数据服务



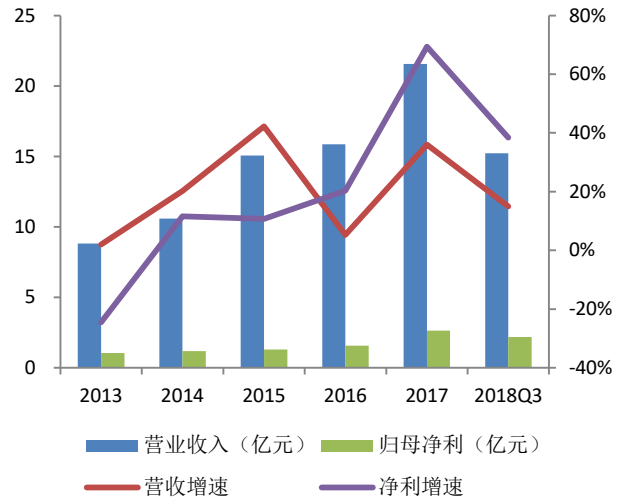
来源：公司公告、国联证券研究所

公司位置大数据服务主要基于海量交通出行大数据仓库、位置云服务平台及大数据生态系统，搭建集团级数据蜂巢系统，用最小的成本存储、定制、使用、交叉引用以及范式管理位置数据，形成 SaaS 平台输出能力，并面向政府机关、交警、公安、保险、互联网、运营商、物流、交通、气象等政企及行业用户，提供电子地图、位置分析研判及可视化处理、行业应用解决方案等，帮助用户获得位置大数据能力加成。目前，公司已与武大信息资源研究中心、公安部一所、中规院交通分院、苏州科达、上海电科、海信、浙江大华等多家行业领先机构达成战略合作。其创新应用产品 MineData 作为“数据+可视化+分析研判”的一站式位置大数据服务平台，已荣获 ITS Asia 2017“创新产品奖”，2018 年 6 月 MineData 2.0 版本发布，系统数据总量已经超过 4.7PB，数据日增量超过 3.3TB，助力公司不断提升产业赋能价值。

2017 年，公司实现营收 21.56 亿元，同比增长 36.03%；实现归母净利 2.65 亿元，

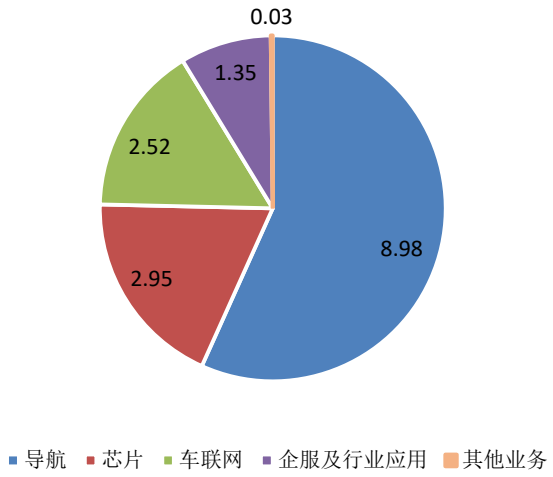
同比增长 69.38%。2018 年前三季度实现营收 15.23 亿元，同比增长 14.96%；实现归母净利 2.19 亿元，同比增长 38.38%。

图表 44：四维图新历年经营情况



来源：Wind、国联证券研究所

图表 45：四维图新分业务毛利情况（2017，亿元）



来源：Wind、国联证券研究所

中科曙光：

中科曙光实际控制人为中科院计算机所，是以国家“863”计划重大科研成果为基础组建的高新技术企业。公司自成立以来始终专注于高端计算机、存储、软件和云计算业务，已掌握大量核心技术，实现国内领先并达到国际先进水平。

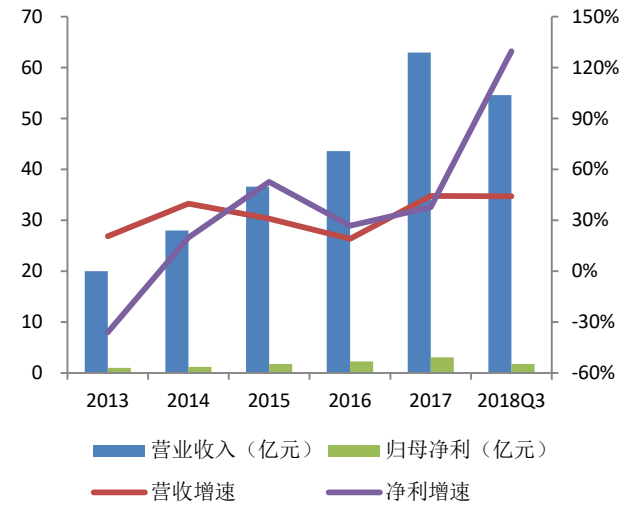
公司在通用芯片领域与 AMD 达成合作，通过参股公司天津海光获得 AMD 在中国地区的 Zen 架构授权，用于开发 X86 芯片；在 AI 芯片领域，与同属中科院的寒武纪达成合作，寒武纪是国内领先的 AI 芯片独角兽，2017 年两者合作推出了人工智能服务器“Phaneron”。公司高端计算机产品在《中国高性能计算机性能 TOP100 排行榜》连续多年获得数量份额第一，是国家突破“E 级超算”的一支重要力量，并于 2016 年获批承担国家重点研发计划的“E 级高性能计算机原型系统研制”项目。公司 ParaStor 存储产品在 HPC、石油地震、视频监控、人工智能等领域持续深耕细作，以业界最高的单节点 5GB/s 性能表现，为大规模 HPC 项目提供理想的数据共享方案。2017 年 2 月，公司与业界领先厂商 Promise 合资成立存储公司，为加快布局下一代统一架构存储产品创造有利条件。

而在云计算及大数据领域，公司于 2015 年提出“数据中国”战略，2016 年明确“数据中国加速计划”，2017 年启动“数据中国智能计划”，随着 Cloudview 云计算操作系统的不断升级完善，其建设城市云和行业云计算中心的能力显著提升。目前，公司已在全国 30 多个城市部署了城市云计算大数据中心，全面参与和支持地方政府的政务信息化和智慧城市建设。

2017 年，公司实现营业收入 62.94 亿元，同比增长 44.36%；实现归母净利润 3.09 亿元，同比增长 37.71%。2018 年前三季度实现营收 54.58 亿元，同比增长 44.06%；

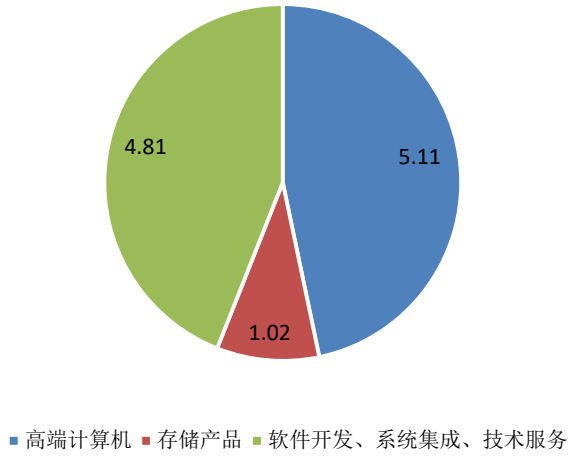
实现归母净利 1.74 亿元，同比增长 129.63%。

图表 46：中科曙光历年经营情况



来源：Wind、国联证券研究所

图表 47：中科曙光分业务毛利情况（2017，亿元）



来源：Wind、国联证券研究所

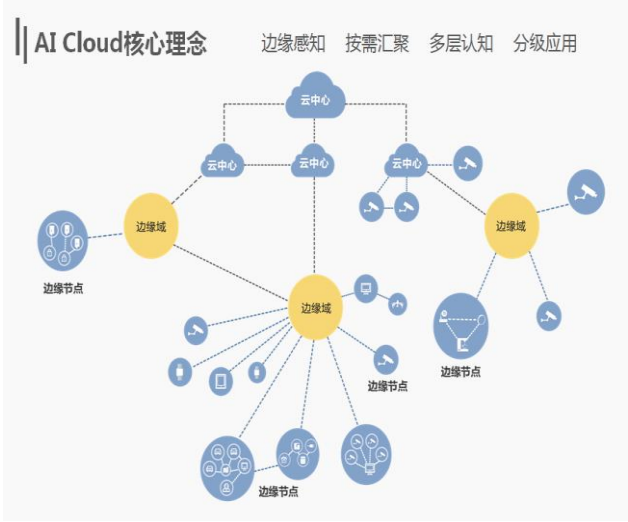
海康威视：

海康威视是全球最大的安防厂商，是视频监控数字化、网络高清化、智能化的见证者、践行者和重要推动者。根据 IHS 报告，海康威视连续 6 年蝉联视频监控行业全球第一，拥有全球视频监控市场份额的 21.4%。

公司以音视频压缩板卡起家，抓住 DVR、NVR 发展先机，并从后端产品拓展至前端摄像头领域，逐步发展成为涵盖整个视频监控行业安全和可视化管理需求的全系列产品及系统解决方案商，面向全球提供综合安防、智慧业务与大数据服务。

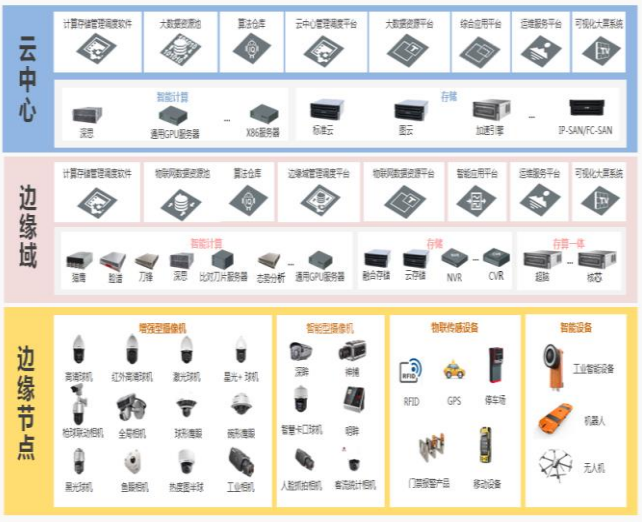
2017 年，海康开创性提出 AI Cloud 架构，通过边缘节点、边缘域、云中心三个层级，实现人工智能、大数据、云计算和终端设备的有机融合。边缘节点侧重多维感知、数据采集和前端智能处理；边缘域侧重感知数据汇聚、存储、处理和智能应用；云中心侧重业务数据融合及大数据多维分析应用。整体遵循“边缘感知、按需汇聚、多层认知、分级应用”的核心理念，赋能各行业应用的智能化转型。

图表 48：海康 AI Cloud 核心理念



来源：公司公告、国联证券研究所

图表 49：海康 AI Cloud 产品家族



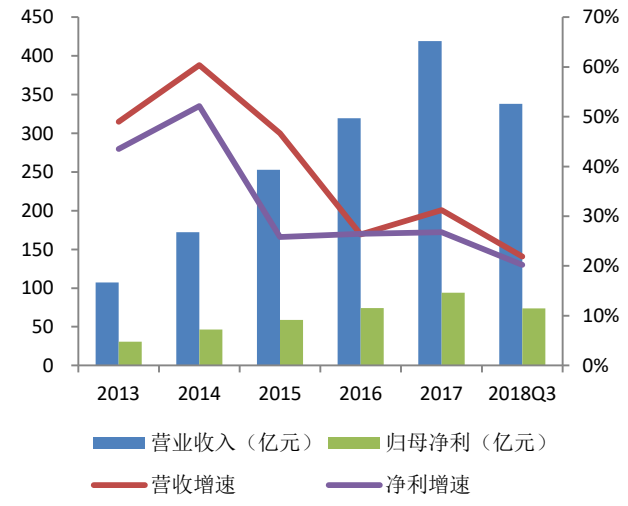
来源：公司公告、国联证券研究所

目前，海康“深眸”系列智能摄像机、“神捕”系列智能交通产品、“明眸”系列智能通道、“深思”系列智能服务器等代表产品，可实现人像比对、车牌识别、智能预警、卡口抓拍、客流分析、客群分析、人脸考勤、门禁管理、应急指挥等多种功能，在公共安全、公共服务、商业、金融等多个领域皆有建树。

除政企市场外，其基于民用兴起（家庭、小微商户等）而推出的“萤石”平台（2013年），截止 2017 年底已接入设备 2800 万，萤石云 APP 用户超过 2000 万，且营业收入突破 10 亿元并实现盈利，成为海康首个盈利的创新业务。

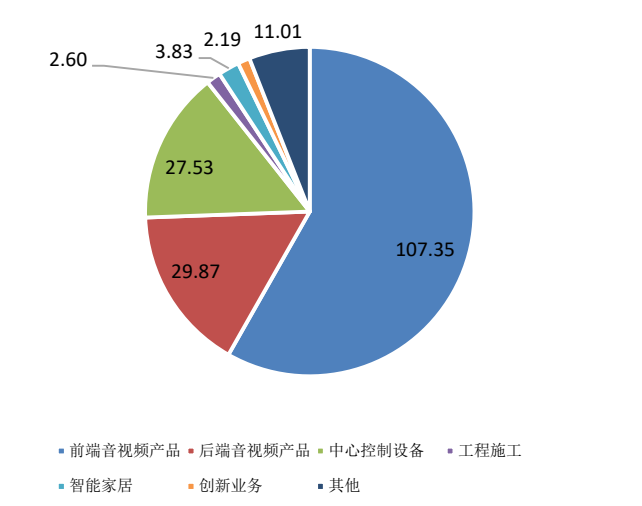
2017 年，公司实现营业收入 419.05 亿元，同比增长 31.22%；实现归母净利润 94.11 亿元，同比增长 26.77%。2018 年前三季度实现营收 338.03 亿元，同比增长 21.90%；实现归母净利 73.96 亿元，同比增长 20.20%。

图表 50：海康威视历年经营情况



来源：Wind、国联证券研究所

图表 51：海康威视分业务毛利情况（2017，亿元）



来源：Wind、国联证券研究所

美亚柏科：

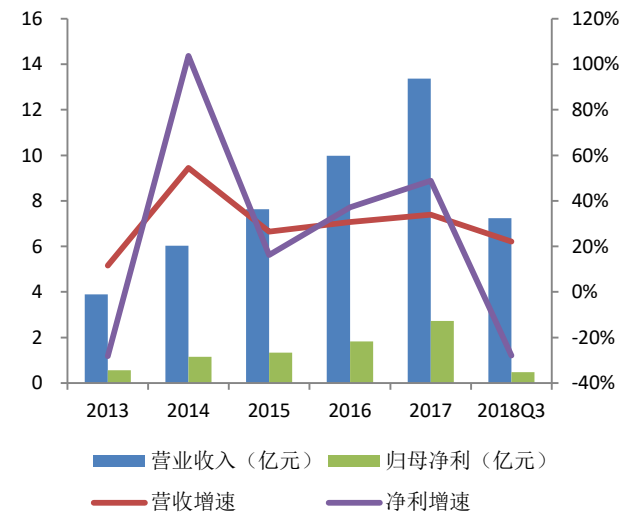
美亚柏科是国内电子数据取证领域龙头企业，网络空间安全及大数据信息化专家。多年来，围绕为司法机关打击犯罪、为行政执法部门实现社会治理，公司已形成电子数据取证产品、大数据信息化产品、网络空间安全产品、专项执法装备“四大产品”，以及存证云+、网络空间安全服务、数据服务和培训、技术支持增值服务“四大服务”。

公司以数字取证起家，除加大研发投入促进内生增长外，也积极利用上市公司融资平台围绕产业进行外延式布局。其取证产品由早期的计算机取证发展到移动设备取证、网络取证、云取证和取证智能化分析等，主营业务也由早期的电子数据取证拓展到大数据信息化、网络空间安全、便民惠民设备等。通过 1+1>2 的整合效益，公司纵向渠道下沉，由省部级向市级、区县级方向渗透；横向客户拓展，由网安警种逐步延伸到刑侦、经侦等其他警种以及监察委、食药监、海关、税务等其他领域，充分享受信息化时代红利。

同时，面对政府大数据浪潮，公司以取证设备为入口，在深刻理解各警种数据的基础上，形成了“数据理解→数据获取→数据应用/共享”的产业链布局，持续推动“公安大数据平台”建设。目前，公司大数据业务已拓展至市场监督管理、税务稽查、海关缉私等领域，形成了大数据+打击犯罪、大数据+社会安全/治理、大数据+铁路、大数据+税务、大数据+市场监管、大数据+质检、大数据+海关等行业解决方案，有望持续放量进一步增厚业绩。

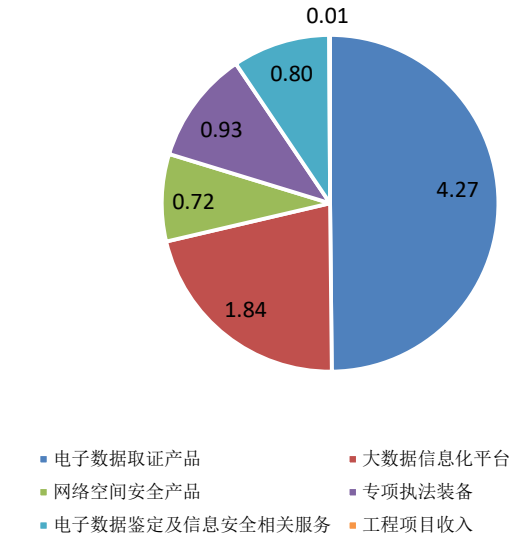
2017 年，公司实现营业收入 13.37 亿元，同比增长 33.94%；实现归母净利润 2.72 亿元，同比增长 48.78%。2018 年前三季度实现营收 7.24 亿元，同比增长 22.15%；实现归母净利 0.47 亿元，同比下滑 27.97%。（Q3 不达预期，主要是受国家组织机构调整影响，公司订单有所延缓，但短期承压不改长期向好逻辑）

图表 52：美亚柏科历年经营情况



来源：Wind、国联证券研究所

图表 53：美亚柏科分业务毛利情况（2017，亿元）



来源：Wind、国联证券研究所

创业软件：

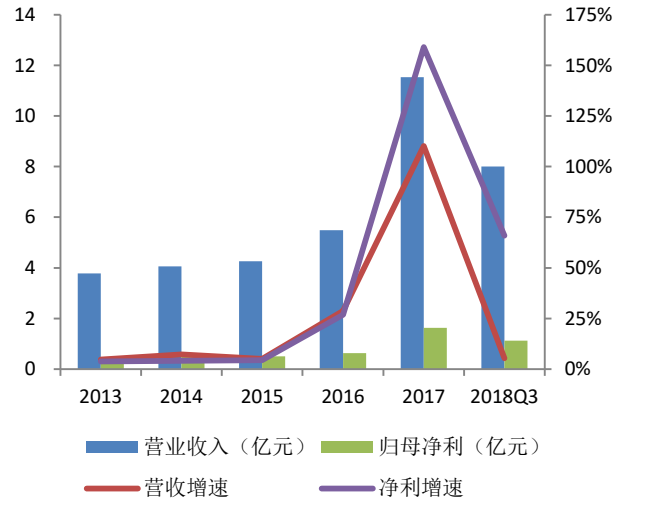
创业软件是国内较早进入医疗卫生信息化领域的软件企业之一。公司以智慧医疗、区域卫生、健康城市为主要发展方向，目前已累计实施了 10000 多个医疗卫生信息化建设项目，行业用户数量 6000 多家，公共卫生项目遍及全国 340 多个区县，积累超过 2.4 亿份居民健康档案。

18 年以来，《关于促进“互联网+医疗健康”发展的意见》、《全国医院信息化建设标准与规范（试行）》、《关于进一步推进以电子病历为核心的医疗机构信息化建设工作的通知》等政策频频，促使医疗信息化行业维持高景气度。公司作为国内智慧医疗第一梯队，将充分受益此次以临床信息系统为主导的医疗信息建设大潮，目前，公司实际在手订单约 7 亿左右，10 月更是中标淄博市中心医院信息化建设——5280 万元大单，预计全年订单有望持续保持较高增速。

在医疗大数据领域，公司 1）承接了南京浦口、浙江桐乡、广东珠海、福建龙岩、山西长治等地区数个涉及大数据应用订单，并与上海闵行卫计委、江阴市人民医院、白银市第一人民医院等 400 余家医疗卫生机构及卫生行政部门签署了大数据业务合作及实施意向协议。2）与蚂蚁金服、腾讯开展合作，试点医疗支付与支付宝应用结合，将腾讯觅影、智慧医院、人工智能、微信支付等产品融合到公司产品中。3）标杆平台——中山健康城市项目将于 12 月底完成建设，从 2019 年开始进入 10 年运营期，初步规划实施支付服务类、药品服务类、健康服务类、导流导购类、数据产品类等 9 大类运营服务内容，有望获得新增盈利点，同时中山模式复制拓展性较强，若实现全国推广，将打开新一轮增长空间。

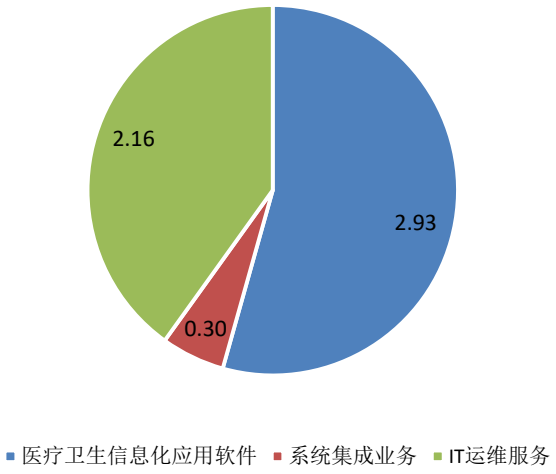
2017 年，公司实现营业收入 11.53 亿元，同比增长 110.15%；实现归母净利润 1.63 亿元，同比增长 159.10%。2018 年前三季度实现营收 8.00 亿元，同比增长 5.37%；实现归母净利 1.12 亿元，同比增长 65.94%。

图表 54：创业软件历年经营情况



来源：Wind、国联证券研究所

图表 55：创业软件分业务毛利情况（2017，亿元）



来源：Wind、国联证券研究所

5. 风险提示

1）技术发展遭遇瓶颈

2）政策推进有所延缓

3）订单落地低于预期

4）市场系统性风险

（感谢实习生练钊辰对资料搜集的贡献）

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

股票投资评级	强烈推荐	股票价格在未来 6 个月内超越大盘 20%以上
	推荐	股票价格在未来 6 个月内超越大盘 10%以上
	谨慎推荐	股票价格在未来 6 个月内超越大盘 5%以上
	观望	股票价格在未来 6 个月内相对大盘变动幅度为－10%～10%
	卖出	股票价格在未来 6 个月内相对大盘下跌 10%以上
行业投资评级	优异	行业指数在未来 6 个月内强于大盘
	中性	行业指数在未来 6 个月内与大盘持平
	落后	行业指数在未来 6 个月内弱于大盘

一般声明

特别声明

分公司机构销售联系方式

无锡

国联证券股份有限公司研究所

江苏省无锡市太湖新城金融一街 8 号国联金融大厦 9 层

电话：0510-82833337

传真：0510-82833217

上海

国联证券股份有限公司研究所

上海市浦东新区源深路 1088 号葛洲坝大厦 22F

电话：021-38991500

传真：021-38571373

地区	姓名	固定电话
北京	管峰	010-68790949-8007
上海	刘莉	021-38991500-831
深圳	薛靖韬	0755-82560810