# 软工综合训练新技术报告

## 一.大数据定义：

对于“大数据”（Big data）研究机构[Gartner](http://baike.baidu.com/view/1231927.htm" \t "_blank)给出了这样的定义。“大数据”是需要新处理模式才能具有更强的[决策力](http://baike.baidu.com/view/558688.htm" \t "_blank)、[洞察](http://baike.baidu.com/view/684004.htm)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](http://baike.baidu.com/view/42564.htm)。

[麦肯锡](http://baike.baidu.com/view/1013977.htm)全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用[分布式架构](http://baike.baidu.com/view/4792581.htm" \t "_blank)。它的特色在于对海量数据进行分布式数据挖掘，但它必须依托云计算的[分布式处理](http://baike.baidu.com/view/185352.htm)、[分布式数据库](http://baike.baidu.com/view/68389.htm)和[云存储](http://baike.baidu.com/view/2044736.htm)、[虚拟化技术](http://baike.baidu.com/view/13605.htm)。

随着[云时代](http://baike.baidu.com/view/2169141.htm" \t "_blank)的来临，大数据（[Big data](http://baike.baidu.com/view/7573514.htm)）也吸引了越来越多的关注。《[著云台](http://baike.baidu.com/view/8379937.htm" \t "_blank)》的分析师团队认为，大数据（Big data）通常用来形容一个公司创造的大量[非结构化数据](http://baike.baidu.com/view/2119114.htm" \t "_blank)和[半结构化数据](http://baike.baidu.com/view/888365.htm)，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和[云计算](http://baike.baidu.com/view/1316082.htm)联系到一起，因为实时的大型数据集分析需要像MapReduce一样的框架来向数十、数百或甚至数千的电脑分配工作。

大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模[并行处理](http://baike.baidu.com/view/494465.htm)（MPP）数据库、数据挖掘电网、[分布式文件系统](http://baike.baidu.com/view/771589.htm" \t "_blank)、[分布式数据库](http://baike.baidu.com/view/68389.htm)、[云计算平台](http://baike.baidu.com/view/2729362.htm)、互联网和可扩展的[存储系统](http://baike.baidu.com/view/51839.htm)。

最小的基本单位是bit，按顺序给出所有单位：[bit](http://baike.baidu.com/view/155429.htm" \t "_blank)、[Byte](http://baike.baidu.com/view/44243.htm" \t "_blank)、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。

它们按照进率1024（2的十次方）来计算：

1 Byte =8 bit

1 [KB](http://baike.baidu.com/subview/170925/4986980.htm) = 1,024 Bytes = 8192 bit

1 [MB](http://baike.baidu.com/subview/35572/5142531.htm) = 1,024 KB = 1,048,576 Bytes

1 [GB](http://baike.baidu.com/subview/50311/8909432.htm) = 1,024 MB = 1,048,576 KB

1 [TB](http://baike.baidu.com/subview/347017/5891845.htm) = 1,024 GB = 1,048,576 MB

1 [PB](http://baike.baidu.com/subview/344884/10542238.htm) = 1,024 TB = 1,048,576 GB

1 [EB](http://baike.baidu.com/subview/48988/10542260.htm) = 1,024 PB = 1,048,576 TB

1 ZB = 1,024 EB = 1,048,576 PB

1 [YB](http://baike.baidu.com/subview/478158/10542273.htm) = 1,024 ZB = 1,048,576 EB

1 [BB](http://baike.baidu.com/subview/86977/9552503.htm) = 1,024 YB = 1,048,576 ZB

1 [NB](http://baike.baidu.com/subview/3353/5725631.htm) = 1,024 BB = 1,048,576 YB

1 DB = 1,024 NB = 1,048,576 BB

全称：

1 Bit（比特） =Binary Digit

8 Bits = 1 Byte（字节）

1,000 Bytes = 1 Kilobyte

1,000Kilobytes = 1 Megabyte

1,000 Megabytes = 1 Gigabyte

1,000 Gigabytes = 1Terabyte

1,000 Terabytes = 1 Petabyte

1,000 Petabytes = 1 Exabyte

1,000Exabytes = 1 Zettabyte

1,000 Zettabytes = 1 Yottabyte

1,000 Yottabytes = 1Brontobyte

1,000 Brontobytes = 1 Geopbyte

## 二.大数据特征：

容量（Volume）：数据的大小决定所考虑的数据的价值的和潜在的信息；

种类（Variety）：数据类型的多样性；

速度（Velocity）：指获得数据的速度；

可变性（Variability）：妨碍了处理和有效地管理数据的过程。

真实性（Veracity）：数据的质量

复杂性（Complexity）：数据量巨大，来源多渠道

## 三. 大数据意义

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。

阿里巴巴创办人马云来台演讲中就提到，未来的时代将不是IT时代，而是DT的时代，DT就是Data Technology数据科技，显示大数据对于阿里巴巴集团来说举足轻重。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据是成为赢得竞争的关键。

大数据的价值体现在以下几个方面：1)对大量消费者提供产品或服务的企业可以利用大数据进行精准营销;2) 做小而美模式的中长尾企业可以利用大数据做服务转型;3) 面临互联网压力之下必须转型的传统企业需要与时俱进充分利用大数据的价值。

不过，“大数据”在经济发展中的巨大意义并不代表其能取代一切对于社会问题的理性思考，科学发展的逻辑不能被湮没在海量数据中。著名经济学家路德维希·冯·米塞斯曾提醒过：“就今日言，有很多人忙碌于资料之无益累积，以致对问题之说明与解决，丧失了其对特殊的经济意义的了解。”  这确实是需要警惕的。

## 四. 大数据结构

大数据就是[互联网](http://baike.baidu.com/view/6825.htm)发展到现今阶段的一种[表象](http://baike.baidu.com/view/39081.htm)或[特征](http://baike.baidu.com/view/1069886.htm)而已，没有必要神话它或对它保持敬畏之心，在以云计

算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

其次，想要系统的[认知](http://baike.baidu.com/view/69807.htm)大数据，必须要全面而细致的分解它，我着手从三个层面来展开：

第一层面是[理论](http://baike.baidu.com/view/96020.htm)，理论是认知的必经途径，也是被广泛认同和传播的基线。在这里从大数据的特征定义理解行业对大数据的整体描绘和定性；从对大数据价值的探讨来深入解析大数据的珍贵所在；洞悉大数据的发展趋势；从大数据[隐私](http://baike.baidu.com/view/350273.htm)这个特别而重要的视角审视人和数据之间的长久博弈。

第二层面是[技术](http://baike.baidu.com/view/45517.htm)，技术是大数据价值体现的手段和前进的基石。在这里分别从云计算、分布式处理技术、存储技术和感知技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。

第三层面是[实践](http://baike.baidu.com/view/51642.htm)，实践是大数据的最终价值体现。在这里分别从互联网的大数据，政府的大数据，企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

## 五. 大数据应用

洛杉矶警察局和加利福尼亚大学合作利用大数据预测犯罪的发生。

google流感趋势(Google Flu Trends)利用搜索关键词预测禽流感的散布。

统计学家内特.西尔弗(Nate Silver)利用大数据预测2012美国选举结果。

麻省理工学院利用手机定位数据和交通数据建立城市规划。

梅西百货的实时定价机制。根据需求和库存的情况，该公司基于SAS的系统对多达7300万种货品进行实时调价。

医疗行业早就遇到了海量数据和非结构化数据的挑战，而近年来很多国家都在积极推进医疗信息化发展，这使得很多医疗机构有资金来做大数据分析。

## 六.大数据趋势：

**趋势一：数据的资源化**

何为资源化，是指大数据成为企业和社会关注的重要战略资源，并已成为大家争相抢夺的新焦点。因而，企业必须要提前制定大数据营销战略计划，抢占市场先机。

**趋势二：与云计算的深度结合**

大数据离不开云处理，云处理为大数据提供了弹性可拓展的基础设备，是产生大数据的平台之一。自2013年开始，大数据技术已开始和云计算技术紧密结合，预计未来两者关系将更为密切。除此之外，物联网、移动互联网等新兴计算形态，也将一齐助力大数据革命，让大数据营销发挥出更大的影响力。

**趋势三：科学理论的突破**

随着大数据的快速发展，就像计算机和互联网一样，大数据很有可能是新一轮的技术革命。随之兴起的数据挖掘、机器学习和人工智能等相关技术，可能会改变数据世界里的很多算法和基础理论，实现科学技术上的突破。

**趋势四：数据科学和数据联盟的成立**

未来，数据科学将成为一门专门的学科，被越来越多的人所认知。各大高校将设立专门的数据科学类专业，也会催生一批与之相关的新的就业岗位。与此同时，基于数据这个基础平台，也将建立起跨领域的数据共享平台，之后，数据共享将扩展到企业层面，并且成为未来产业的核心一环。

**趋势五：数据泄露泛滥**

未来几年数据泄露事件的增长率也许会达到100%，除非数据在其源头就能够得到安全保障。可以说，在未来，每个财富500强企业都会面临数据攻击，无论他们是否已经做好安全防范。而所有企业，无论规模大小，都需要重新审视今天的安全定义。在财富500强企业中，超过50%将会设置首席信息安全官这一职位。企业需要从新的角度来确保自身以及客户数据，所有数据在创建之初便需要获得安全保障，而并非在数据保存的最后一个环节，仅仅加强后者的安全措施已被证明于事无补。

**趋势六：数据管理成为核心竞争力**

数据管理成为核心竞争力，直接影响财务表现。当“数据资产是企业核心资产”的概念深入人心之后，企业对于数据管理便有了更清晰的界定，将数据管理作为企业核心竞争力，持续发展，战略性规划与运用数据资产，成为企业数据管理的核心。数据资产管理效率与主营业务收入增长率、销售收入增长率显著正相关;此外，对于具有互联网思维的企业而言，数据资产竞争力所占比重为36.8%，数据资产的管理效果将直接影响企业的财务表现。

**趋势七：数据质量是BI(商业智能)成功的关键**

采用自助式商业智能工具进行大数据处理的企业将会脱颖而出。其中要面临的一个挑战是，很多数据源会带来大量低质量数据。想要成功，企业需要理解原始数据与数据分析之间的差距，从而消除低质量数据并通过BI获得更佳决策。

**趋势八：数据生态系统复合化程度加强**

大数据的世界不只是一个单一的、巨大的计算机网络，而是一个由大量活动构件与多元参与者元素所构成的生态系统，终端设备提供商、基础设施提供商、网络服务提供商、网络接入服务提供商、数据服务使能者、数据服务提供商、触点服务、数据服务零售商等等一系列的参与者共同构建的生态系统。而今，这样一套数据生态系统的基本雏形已然形成，接下来的发展将趋向于系统内部角色的细分，也就是市场的细分;系统机制的调整，也就是商业模式的创新;系统结构的调整，也就是竞争环境的调整等等，从而使得数据生态系统复合化程度逐渐增强。

## 七.国家政策

经李克强总理签批，2015年9月，国务院印发《促进大数据发展行动纲要》（以下简称《纲要》），系统部署大数据发展工作。

《纲要》明确，推动大数据发展和应用，在未来5至10年打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态。

《纲要》部署三方面主要任务。一要加快政府数据开放共享，推动资源整合，提升治理能力。大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商事服务便捷化，促进安全保障高效化，加快民生服务普惠化。二要推动产业创新发展，培育新兴业态，助力经济转型。发展大数据在工业、新兴产业、农业农村等行业领域应用，推动大数据发展与科研创新有机结合，推进基础研究和核心技术攻关，形成大数据产品体系，完善大数据产业链。三要强化安全保障，提高管理水平，促进健康发展。健全大数据安全保障体系，强化安全支撑。[15]

2015年9月18日贵州省启动我国首个大数据综合试验区的建设工作，力争通过3至5年的努力，将贵州大数据综合试验区建设成为全国数据汇聚应用新高地、综合治理示范区、产业发展聚集区、创业创新首选地、政策创新先行区。

围绕这一目标，贵州省将重点构建“三大体系”，重点打造“七大平台”，实施“十大工程”。

“三大体系”是指构建先行先试的政策法规体系、跨界融合的产业生态体系、防控一体的安全保障体系；“七大平台”则是指打造大数据示范平台、大数据集聚平台、大数据应用平台、大数据交易平台、大数据金融服务平台、大数据交流合作平台和大数据创业创新平台；“十大工程”即实施数据资源汇聚工程、政府数据共享开放工程、综合治理示范提升工程、大数据便民惠民工程、大数据三大业态培育工程、传统产业改造升级工程、信息基础设施提升工程、人才培养引进工程、大数据安全保障工程和大数据区域试点统筹发展工程。

此外，贵州省将计划通过综合试验区建设，探索大数据应用的创新模式，培育大数据交易新的做法，开展数据交易的市场试点，鼓励产业链上下游之间的数据交换，规范数据资源的交易行为，促进形成新的业态。

国家发展改革委有关专家表示，大数据综合试验区建设不是简单的建产业园、建数据中心、建云平台等，而是要充分依托已有的设施资源，把现有的利用好，把新建的规划好，避免造成空间资源的浪费和损失。探索大数据应用新的模式，围绕有数据、用数据、管数据，开展先行先试，更好地服务国家大数据发展战略。

## 八.大数据IT分析工具

大数据概念应用到[IT](http://baike.baidu.com/subview/30/16685180.htm)操作工具产生的数据中，大数据可以使[IT](http://baike.baidu.com/subview/30/16685180.htm)管理软件供应商解决大广泛的业务决策。[IT](http://baike.baidu.com/subview/30/16685180.htm)系统、应用和技术基础设施每天每秒都在产生数据。大数据非结构化或者结构数据都代表了‘所有用户的行为、服务级别、安全、风险、欺诈行为等更多操作’的绝对记录。

大数据分析的产生旨在于IT管理，企业可以将实时数据流分析和历史相关数据相结合，然后大数据分析并发现它们所需的模型。反过来，帮助预测和预防未来运行中断和性能问题。进一步来讲，他们可以利用大数据了解使用模型以及地理趋势，进而加深大数据对重要用户的洞察力。 他们也可以追踪和记录网络行为，大数据轻松地识别业务影响；随着对服务利用的深刻理解加快利润增长；同时跨多系统收集数据发展IT服务目录。

大数据分析的想法，尤其在[IT](http://baike.baidu.com/subview/30/16685180.htm)操作方面，大数据对于我们发明并没有什么作用，但是我们一直在其中。[Gartner](http://baike.baidu.com/view/1231927.htm)已经关注这个话题很多年了，基本上他们已经强调，如果[IT](http://baike.baidu.com/subview/30/16685180.htm)正在引进新鲜灵感，他们将会扔掉大数据老式方法开发一个新的[IT](http://baike.baidu.com/subview/30/16685180.htm)操作分析平台。