**项目一 大数据概述**

**知识目标**

* 熟悉数据的生命周期
* 了解数据战略
* 了解大数据分析
* 了解大数据的应用方向及产业情况

**能力目标**

* 掌握数据的概念
* 掌握数据的类型
* 掌握大数据概念
* 掌握大数据的特征及技术

**素质目标**

能掌握数据与大数据的概念，并准确把握七关联与区别；准确把握大数据的发展趋势。

**知识精讲**

**任务一 数据**

**一、数据的概念**

数据（data）指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。对数据的含义的理解要把握以下几点：

第一，数据是可识别的、抽象的符号。

第二，数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵。数据本身没有意义，数据只有对实体行为产生影响时才成为信息。因此，数据是信息的表现形式和载体，可以是符号、文字、数字、语音、图像、视频等。

第三，数据可以是连续的值，比如声音、图像，称为模拟数据。也可以是离散的，如符号、文字，称为数字数据。

第四，在计算机系统中，各种字母、数字符号的组合、语音、图形、图像等统称为数据，数据经过加工后就成为信息。在计算机系统中，数据是指所有能输入到计算机并被计算机程序处理的符号的介质的总称，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称。

**二、数据的类型**

**（一）按性质分为**

1.定位数据，如各种坐标数据；

2.定性数据，如表示事物属性的数据（居民地、河流、道路等）；

3.定量数据，反映事物数量特征的数据，如长度、面积、体积等几何量或重量、速度等物理量；

4.定时数据，反映事物时间特性的数据，如年、月、日、时、分、秒等。

**（二）按表现形式分为**

1.数字数据，如各种统计或量测数据；

2.模拟数据，由连续函数组成，是指在某个区间连续变化的物理量，又可以分为图形数据（如点、线、面）、符号数据、文字数据和图像数据等，如声音的大小和温度的变化等。

**（三）按记录方式分**

按记录方式分为地图、表格、影像、磁带、纸带等。

**（四）按数字化方式分**

按数字化方式分为矢量数据、格网数据等。

**【素质拓展】数据的语义特点**

数据的表现形式还不能完全表达其内容，需要经过解释，数据和关于数据的解释是不可分的。如130是一个数据，可以是某门课程的成绩，也可以是某个人的体重，还可以是某个班级的人数。数据的解释是指对数据含义的说明，数据的含义称为数据的语义。数据与其语义是不可分的。

**三、数据生命周期**

数据生命周期指的是数据从创建到销毁的整个过程，通常根据指定的策略将数据组织成各个不同的层，并基于那些关键条件自动地将数据从一个层移动到另一个层，作为一项规则，较新的数据和那些很可能被更加频繁访问的数据，应该存储在更快的，并且更昂贵的存储媒介上，而那些不是很重要的数据则存储在比较便宜的，稍微慢些的媒介上。

基于大数据环境下数据在组织机构业务中的流转情况，定义了数据生命周期的6个阶段，具体各阶段的定义如下:

1.数据采集∶指新的数据产生或现有数据内容发生显著改变或更新的阶段。对于组织机构而言，数据的采集既包含在组织机构内部系统中生成的数据也包含组织机构从外部采集的数据。

2.数据存储︰指非动态数据以任何数字格式进行物理存储的阶段。

3.数据处理∶指组织机构在内部针对动态数据进行的一系列活动的组合。

4.数据传输∶指数据在组织机构内部从一个实体通过网络流动到另一个实体的过程。

5.数据交换∶指数据经由组织机构内部与外部组织机构及个人交互过程中提供数据的阶段。

6.数据销毁∶指通过对数据及数据的存储介质通过相应的操作手段，使数据彻底丢失且无法通过任何手段恢复的过程。

特定的数据所经历的生命周期由实际的业务场景所决定，并非所有的数据都会完整的经历六个阶段。

**【素质拓展】数据资源使用许可与保密协议**

编号（ ）

**数据资源使用许可与保密协议**

管理方：

使用方：

数据用途及内容：

我单位承担xxx单位的《xxx项目》，需向xxx申请使用xxx数据，为规范xxx数据使用，保证数据应用安全，防止数据泄密，特签订如下协议。

一、使用方必须遵守以下使用协议

1、使用方从xxx获取的数据享有受限使用权，仅限于在xxx项目工作范围内使用，不得透露给任何第三方。

2、使用方必须在使用数据所形成的成果的显著位置注明该数据版权的所有者( xxx )。

3、使用方对许可使用的数据不拥有复制、传播、出版、翻译成外国语言等权利﹐不得以商业目的使用该数据或者开发和生产产品。数据的任何格式或者任何复制品视同原始数据。使用方可根据需要对数据内容进行必要的修改和对数据格式进行转换，但未经许可不得将修改、转换后的数据对外发布和提供，并须将修改、转换的情况及修改、转换的内容向管理方备案。

4、不得使用数据从事危害国家安全、社会公共利益和他人合法权益的活动。

5、若使用方违反本协议规定，管理方有权责令使用方停止使用共享数据并归还管理方，且将再复制的该数据及其衍生品全部删除。

6、在数据使用期限内，管理方有权对使用方数据成果使用

情况、数据存储设备管理情况、数据保密管理情况进行检查。如发现存在严重泄密倾向，将有权责令使用方停止使用共享数据，归还数据，将再复制的该数据及其衍生品全部删除。

7、使用方在数据使用期限（\_ 年 月 日至\_ 年 月 日)结束后须及时归还数据，将再复制的该数据及其衍生品全部删除。

二、使用方必须遵守以下保密协议

1、使用方必须按国家有关保密法律法规的要求，采取有效的保密措施，确保资料安全，严防丢失泄密。

2、使用xxx项目数据仅限用于申请使用的范围，不得挪作他用。发表论文、报告、讲话等涉及数据内容应书面告知管理方。

3、使用方在本单位内须严格数据使用管理，控制数据知悉范围，建立专人负责制度，制定领用管理台帐，告知使用人员保密要求，并与数据使用人员签订《保密承诺书》。

4、使用方必须设置数据专用计算机，专人负责，专机专用。数据专用计算机禁止连接互联网，禁止通过网络传输数据信息。不得将数据或衍生成果在互联网上登载。

三、违约责任

1、使用方使用xxx项目数据违反有关保密规定的，依照《中华人民共和国保密法》、《中华人民共和国测绘成果管理规定》等有关法律法规的规定处理。

2、使用方违反本协议规定的，管理方有权对因此造成的损失要求赔偿;构成犯罪的，由司法机关追究其刑事责任。

3、因使用方使用或保管数据不当，导致知识产权纠纷或失密事件，由使用方负全部法律责任。

四、本协议一式四份,管理方持三份,使用方持一份,具有相同的法律效力。

五、协议由双方法定代表或代理人签字后生效。

管理方:(盖章) 使用方:(盖章)

法人代表或代理人（签字) 法人代表或代理人（签字)

时间: \_ 年 月 日 时间: \_ 年 月 日

**四、数据战略**

随着数字经济在全球加速推进以及5G、人工智能、物联网等相关技术的快速发展，数据已成为影响全球竞争的关键战略性资源。只有获取和掌握更多的数据资源﹐才能在新一轮的全球话语权竞争中占据主导地位。目前，全球数据量在飞速增长，各国数据战略布局步伐加快。

1.美国联邦数据战略焦点从“技术”转移到“资源”

自2012年以来，美国极力推动大数据领域前沿核心技术的发展和科学工程领域的发明创造，致力打造有活力的数据创新生态。2019年美国白宫行政管理和预算办公室(OMB)发布《联邦数据战略与2020年行动计划》(以下简称“《联邦数据战略》”)，其核心目标是“将数据作为战略资源开发”。《联邦数据战略》确立了40项数据管理的具体实践目标,包括重视数据并促进共享、保护数据资源、有效使用数据资源三个层次。

2.欧盟数据战略致力于发展数据敏捷型经济体

数据已成为经济社会发展的重要命脉，欧盟致力于平衡数据流动和广泛使用,希望通过建立单一的数据市场，确保欧洲在未来的数据经济中占据领先地位。2020年2月19日，欧盟委员会公布了《欧盟数据战略》，提出了到2030年欧洲将成为世界上最具吸引力、最安全、最具活力的数据敏捷型经济体的愿景目标。即，在保持高度的隐私、安全和道德标准的前提下，充分发掘数据利用的价值造福经济社会，并确保每个人能从数字红利中受益。为推进欧盟数据一体化和提升欧盟国家的市场主体竞争力，《欧盟数据战略》提出了四大支柱性战略措施:一是构建跨部门治理框架；二是加强数据投入；三是提升数据素养；四是构建数据空间。

3.英国通过数据战略助力经济复苏

2020年9月9日，英国数字、文化、媒体和体育部(DCMS)发布《国家数据战略》，支持英国对数据的使用，设定五项“优先任务”，帮助该国经济从疫情中复苏。这五项任务包括:(1）释放数据的价值；(2）确保促进增长和可信的数据体制；(3）转变政府对数据的使用，以提高效率并改善公共服务；(4）确保数据所依赖的基础架构的安全性和韧性；(5)倡导国际数据流动。英国《国家数据战略》还包括设立政府首席数据官，改变政府当前的数据使用方式，以此提高效率并改善公共服务；通过立法提高智慧数据计划的参与度；在支持创新发展的同时致力于解决当前数据共享中存在的障碍等。

4.我国政府高度重视大数据的发展。自2014年以来，我国国家大数据战略的谋篇布局经历了四个不同阶段。

（1）预热阶段

2014年“大数据”一词首次写入政府工作报告，为我国大数据发展的政策环境搭建开始预热。从这一年起，“大数据”逐渐成为各级政府和社会各界的关注热点，中央政府开始提供积极的支持政策与适度宽松的发展环境，为大数据发展创造机遇。

（2）起步阶段

2015年国务院正式印发了《促进大数据发展行动纲要》(国发[2015]50号)，成为我国发展大数据的首部战略性指导文件,对包括大数据产业在内的大数据整体发展作出了部署，体现出国家层面对大数据发展的顶层设计和统筹布局。

（3）落地阶段

《十三五规划纲要》的公布标志着国家大数据战略的正式提出，彰显了中央对于大数据战略的重视。2016年工信部发布《大数据产业发展规划(2016-2020年)》，为大数据产业发展奠定了重要的基础。

（4）深化阶段

随着国内大数据迎来全面良好的发展态势，国家大数据战略也开始走向深化阶段。2017年党的十九大报告中提出推动大数据与实体经济深度融合，为大数据产业的未来发展指明方向。2019年3月，政府工作报告第六次提到“大数据”，并且有多项任务与大数据密切相关。2020年4月，中共中央、国务院发布《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”与土地、劳动力、资本、技术并称为五种要素，提出“加快培育数据要素市场”。5月18日，中央在《关于新时代加快完善社会主义市场经济体制的意见》中进一步提出加快培育发展数据要素市场。这标志着数据要素市场化配置上升为国家战略，将进一步完善我国现代化治理体系，有望对未来经济社会发展产生深远影响。

**任务二 大数据**

**一、大数据的概念**

“大数据”概念最早出现在1980年，由著名的未来学家阿尔文•托夫勒在其著作《第三次浪潮》中所提出。2009年美国互联网数据中心证实大数据时代的来临，而在今天，我们已经能充分感受到大数据的魅力和影响力。许多人看到“大数据”一词，会先入为主的认为大数据就是“大量数据”或者“强大的数据”。其实并非这样，“大”意为：“有用的”、“实用的”、“有价值的”。关于大数据的确切定义，不同组织从不同角度给出了不同的定义。

全球领先的管理咨询公司麦肯锡给出的大数据定义是：“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。”

著名研究机构高德纳咨询公司（Gartner）给出的定义是：“大数据是需要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来适应海量、高增长率和多样化的信息资产。”

全球最大的数据中心IDC则侧重从技术角度说明其概念：“大数据处理技术代表了新一代的技术架构，这种架构通过高速获取数据并对其进行分析和挖掘，从海量且形式各异的数据源中更有效地抽取出富含价值的信息。”

综合各种观点给出大数据的定义：大数据（big data）是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要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。

**【素质拓展】 数据与大数据的关联与区别**

|  |  |  |
| --- | --- | --- |
|  | 传统数据 | 大数据 |
| 数据格式 | 结构化数据 | 非结构化数据+结构化数据 |
| 存储模式 | 集中存储 | 分布式存储 |
| 计算平台 | 数据库查询平台有较好的安全机制 | 分布式计算处理平台几乎没有安全机制 |
| 复杂度 | 相对简单 | 由于异构性，导致复杂度增加 |
| 计算物理环境 | 以服务器为主，有向云上转移的趋势，有较清晰的边界 | 云是主要的承载物理平台，但仍有利用物理服务器，边界模糊 |
| 保护目标 | 机密性、完整性、可用性 | 机密性、完整性、可用性  同时要进一步考虑数据的真实性 |
| 数据库结构 | SQL | SQL+NOSQL |
| 软件栈 | C++为主 | Java为主 |
| 主流规模 | 1~10台 | 3~1000台，最高可支持上万台 |
| 包含的内容 | 集中存储、查询 | 存储、查询、计算、ETL、分布式应用程序协调服务 |

**二、大数据的特征**

大数据具有四个特征：第一是数据体量大（Volume）；第二是数据类型繁多（Variety）；第三是价值密度低（Value）；第四是处理速度快（Velocity）。业界将其归纳为4个“V”——Volume，Variety，Value，Velocity。

1、数据体量大（Volume）

大数据最明显的特点就是体量大。过去由于数据量不够，即使用了数据，依然不足以消除不确定性，因此数据的作用其实有限，很多人忽视它的重要性是必然的，数据的价值也就被严重低估。统计发现，非结构化数据占总数据量的80~90%，比结构化数据增长快10到50倍。据IDC预测，未来10年全球数据量将以大于40%的速度增长。

2、数据类型繁多（Variety）

大数据种类繁多，一般包括结构化、半结构化和非结构化等多类数据。这些数据在编码方式、数据格式、应用特征等多个方面存在差异性，多信息源并发形成大量的异构数据。不同结构的数据处理和分析方式也有所区别。

3、价值密度低（Value）

大数据价值密度的高低与数据总量大小成反比，单条数据本身并无太多价值，但庞大的数据量累积并隐藏了巨大的财富。其价值具备稀疏性、多样性和不确定性等特点。

4、处理速度快（Velocity）

数据的快速流动和处理是大数据区分于传统数据挖掘的显著特征。比如，涉及感知、传输、决策、控制开放式循环的大数据，对数据实时处理有着极高的要求，通过传统数据库查询方式得到的“当前结果”很可能已经没有价值。因此，大数据更强调实时分析而非批量式分析。

大数据中的数据量非常巨大，达到了PB级别。而且这庞大的数据之中，不仅仅包括结构化数据（如数字、符号等数据），还包括非结构化数据（如文本、图像、声音、视频等数据）。这使得大数据的存储，管理和处理很难利用传统的关系型数据库去完成。在大数据之中，有价值的信息往往深藏其中。这就需要对大数据的处理速度要非常快，才能短时间之内就能从大量的复杂数据之中获取到有价值的信息。在大数据的大量复杂的数据之中，通常不仅仅包含真实的数据，一些虚假的数据也混杂其中。这就需要在大数据的处理中将虚假的数据剔除，利用真实的数据来分析得出真实的结果。

**三、大数据的内涵理解**

大数据是一门技术，也是一种全新的商业模式，代表着一种思维方式。它是大规模数据的集合体，更是数据对象、集成技术、分析应用、商业模式、思维创新的统一体。

1.从对象角度来看，大数据是数据规模超出传统数据库处理能力的数据集合

大数据对象既可能是实际的、有限的数据集合，也可能是虚拟的、无限的数据集合。目前，数据的发展演进已由数据库时代走向大数据时代，数据量处于TB级，乃至PB级，甚至更高。但是，大数据并非大量数据简单、无意义的堆积，而是在数据之间存在或远或近、或直接或间接的关联性，具有分析挖掘的价值，并且数据集中储存和计算已经达到传统数据库软件无法处理的巨大数据量，具有非结构化数据无固定格式、变化多、并发高、增长速度快等特性。传统数据库研究讲究因果关系，强调的是数据精确性，而大数据研究则侧重于相关性，强调挖掘不同事物间的相关性，并以此作为各类判断的依据。此外，大数据使运算更依赖于数据而不是算法，较多的数据对于结果的影响要好于事先模型。

2.从技术角度来看，大数据是从海量数据中快速获得有价值信息的技术

大数据技术涉及数据采集、存储、管理、分析挖掘、可视化等技术及其集成。该技术可以从凌乱纷繁的数据背后找到更符合用户兴趣和习惯的产品和服务，并对产品和服务进行针对性的调整和优化。传统数据库软件在应对大数据多样化格式上较为吃力，其存储、计算也难以获得满意效果，因此并不适用于大数据分析，需要革新性的大数据技术来解决这些问题。现在常用的大数据技术包括：批量分布式并行计算Hadoop技术、实时分布式高吞吐高并发数据存取处理NoSQL技术、利用廉价服务器搭建高容错性并行计算架构技术等，涉及数据聚类、数据挖掘、分布式处理各领域。

3.从应用角度来看，大数据是对特定数据集合应用相关技术获得价值的行为

大数据有着旺盛的应用需求和广阔的使用前景，该技术可以释放商业价值，使数据更加透明，具有极强的行业应用需求特性。通过数据分析帮助企业了解不同市场之间的关联，发现新的产品和服务。企业可以将大数据分析技术用于在市场或行业内创造竞争优势，开拓新的商业机会。正由于与具体应用紧密联系，甚至是一对一的联系，才使得“应用”成为大数据不可或缺的内涵之一。

4.从商业模式角度来看，大数据是企业获得商业价值的业务创新方向

大数据资源与技术的工具化运用，推动大数据产业链形成，以大数据为中心的扩张引发行业的跨界与融合。大数据使得企业在价值主张、关键业务与流程、收益模式等方面发生转变，向着全数据模式演进，以利用数据价值为核心，新型商业模式不断涌现。企业在制定大数据业务战略时，需要分析自身业务基础和数据能力，选择适合的大数据商业模式。根据Bloomberg Venture 发布的大数据产业地图2.0版本，大数据产业可划分为6大类，共38种产品/商业模式，分别是大数据基础设施类、大数据分析类、大数据应用类、大数据数据源类、跨基础设施分析、开源项目。

5.从思维方式来看，大数据是从第三范式中分离出来的一种科研范式

科学研究的第一范式是实验归纳，第二范式是模型推演，第三范式是计算机仿真模拟，第四范式就是密集数据分析。图灵奖获得者吉姆•格雷（Jim Gray）基于e-Science的思路提出：大数据是科学研究的第四范式，即以大数据为基础的数据密集型科研。之所以将大数据科研从第三范式中分离出来，是因为其研究方式不同于基于数学模型的传统研究方式。PB级数据使得人们可以做到没有模型和假设就分析数据。将数据输入巨大的计算机机群中，只要有相互关系的数据，统计分析算法就可以发现传统科学方法发现不了的新模式、新知识，甚至新规律。科研第四范式不仅是科研方式的转变，也是人们思维方式的大变化。

**四、大数据分析**

大数据，表面上看就是大量复杂的数据，这些数据本身的价值并不高，但是对这些大量复杂的数据进行分析处理后，却能从中提炼出很有价值的信息。对大数据的分析，主要分为五个方面：可视化分析（AnalytIC Visualization）、数据挖掘算法（Date Mining Algorithms）、预测性分析能力（Predictive Analytic Capabilities）、语义引擎（Semantic Engines）和数据质量管理（Data Quality Management）。

可视化分析是普通消费者常常可以见到的一种大数据分析结果的表现形式，比如说百度制作的“百度地图春节人口迁徙大数据”就是典型的案例之一。可视化分析将大量复杂的数据自动转化成直观形象的图表，使其能够更加容易的被普通消费者所接受和理解。

数据挖掘算法是大数据分析的理论核心，其本质是一组根据算法事先定义好的数学公式，将收集到的数据作为参数变量带入其中，从而能够从大量复杂的数据中提取到有价值的信息。著名的“啤酒和尿布”的故事就是数据挖掘算法的经典案例。沃尔玛通过对啤酒和尿布购买数据的分析，挖掘出以前未知的两者间的联系，并利用这种联系，提升了商品的销量。亚马逊的推荐引擎和谷歌的广告系统都大量使用了数据挖掘算法。

预测性分析能力是大数据分析最重要的应用领域。从大量复杂的数据中挖掘出规律，建立起科学的事件模型，通过将新的数据带入模型，就可以预测未来的事件走向。预测性分析能力常常被应用在金融分析和科学研究领域，用于股票预测或气象预测等。

语义引擎是机器学习的成果之一。过去，计算机对用户输入内容的理解仅仅停留在字符阶段，不能很好的理解输入内容的意思，因此常常不能准确的了解用户的需求。通过对大量复杂的数据进行分析，让计算机从中自我学习，可以使计算机能够尽量精确的了解用户输入内容的意思，从而把握住用户的需求，提供更好的用户体验。苹果的Siri和谷歌的Google Now都采用了语义引擎。

数据质量管理是大数据在企业领域的重要应用。为了保证大数据分析结果的准确性，需要将大数据中不真实的数据剔除掉，保留最准确的数据。这就需要建立有效的数据质量管理系统，分析收集到的大量复杂的数据，挑选出真实有效的数据。

**任务三 大数据时代**

**一、大数据时代概述**

数据来源的极大丰富和数据体量的爆炸性增长促使大数据（big data）出现并得到广泛应用。

大数据正以前所未有的速度，颠覆人们探索世界的方法，驱动产业间的融合与分立。各领域新技术、新工艺、新材料的不断出现，引领着各种新思维和新变革的产生，改变着人们的工作、学习和生活。智能终端的快速普及、通信网络的升级换代、应用程序的丰富多彩、海量数据的深入分析使得移动互联网的发展正在逐步超过传统互联网。而云计算和物联网技术的出现带来了服务交付模式、商业应用模式、设备之间互联互通、处理规模与能力的创新与提高。由于这些新技术的不断发展和成熟，客观上为大数据的产生奠定了基础，从而揭开大数据时代的序幕。

大数据时代是建立在对互联网、物联网等渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和分发的信息时代。

无论从数据规模和结构，还是对社会生活和生产的影响来看，当下都已全面进入大数据时代。

**【素质拓展】 大数据的发展趋势**

****

**二、大数据时代下大数据的应用方向**

大数据成为时代发展一个必然的产物，大数据时代，一切可量化，一切可分析。大数据不仅意味着海量、多样、迅捷的数据处理，更是一种颠覆的思维方式、一项智能的基础设施、一场创新的技术变革。物联网、智慧城市、增强现实(AR)与虚拟现实(VR)、区块链技术、语音识别、人工智能、数字汇流是大数据未来应用的七大发展方向。

1.物联网

物联网是把所有物品通过信息传感设备与互联网连接起来，进行信息交换，即物物相息，以实现智能化识别和管理。

物联网是新一代信息技术的重要组成部分，也是“信息化”时代的重要发展阶段。

物联网的核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络；其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信，也就是物物相息。物联网用途广泛，遍及智能交通、环境保护、政府工作、公共安全、平安家居、智能消防、工业监测、环境监测、路灯照明管控、景观照明管控、楼宇照明管控、广场照明管控、老人护理、个人健康、花卉栽培、水系监测、食品溯源、敌情侦查和情报搜集等多个领域。

2.智慧城市

智慧城市就是运用信息和通信技术手段感测、分析、整合城市运行核心系统的各项关键信息；对包括民生、环保、公共安全、城市服务、工商业活动在内的各种需求做出智能响应。其实质是利用先进的信息技术，实现城市智慧式管理和运行，进而为城市中的人创造更美好的生活，促进城市的和谐、可持续成长。这项趋势的成败取决于数据量是否足够，这有赖于政府部门与民营企业的合作；

用途分为十大智慧体系，分别为：智慧物流体系、智慧制造体系、智慧贸易体系、智慧能源应用体系、智慧公共服务、智慧社会管理体系、智慧交通体系、智慧健康保障体系、智慧安居服务体系、智慧文化服务体系。

3.增强现实(AR)与虚拟现实(VR)

增强现实技术（Augmented Reality，简称 AR），是一种实时地计算摄影机影像的位置及角度并加上相应图像、视频、3D模型的技术，这种技术的目标是在屏幕上把虚拟世界套在现实世界并进行互动。

AR技术应用于广泛领域，诸如尖端武器、飞行器的研制与开发、数据模型的可视化、虚拟训练、娱乐与艺术等领域具有广泛的应用。由于其具有能够对真实环境进行增强显示输出的特性，在医疗研究与解剖训练、精密仪器制造和维修、军用飞机导航、工程设计和远程机器人控制等领域，具有比VR技术更加明显的优势。随着随身电子产品CPU运算能力的提升，预期增强现实的用途将会越来越广。

虚拟现实技术（Virtual Reality）是一种能够创建和体验虚拟世界的计算机仿真技术，它利用计算机生成一种交互式的三维动态视景，其实体行为的仿真系统能够使用户沉浸到该环境中。虚拟现实技术是一种可以创建和体验虚拟世界的计算机仿真系统，它利用计算机生成一种模拟环境；是一种多源信息融合的、交互式的三维动态视景和实体行为的系统仿真使用户沉浸到该环境中。

VR应用一开始以电玩为主，现在的应用却超越电玩，可以用来教学。VR已不仅仅被关注于计算机图像领域，它已涉及更广的领域，如电视会议、网络技术和分布计算技术，并向分布式虚拟现实发展。虚拟现实技术已成为新产品设计开发的重要手段。如地产漫游（在虚拟现实系统中自由行走、任意观看，冲击力强，能使客户获得身临其境的真实感受，促进了合同签约的速度）、网上看房（租售阶段用户通过互联网身临其境的了解项目的周边环境、空间布置、室内设计）等。

4.区块链技术

区块链技术，简称BT（Blockchain technology），也被称之为分布式账本技术，是一种互联网数据库技术，其特点是去中心化、公开透明，让每个人均可参与数据库记录。

区块链技术是指一种全民参与记账的方式。所有的系统背后都有一个数据库，你可以把数据库看成是就是一个大账本。

区块链有很多不同应用方式，美国几乎所有科技公司都在尝试如何应用，最常见的应用是比特币跟其他加密货币的交易。

5.语音识别技术

语音识别技术就是让机器通过识别和理解过程把语音信号转变为相应的文本或命令的高技术。与机器进行语音交流，让机器明白你说什么，这是人们长期以来梦寐以求的事情。中国物联网校企联盟形象得把语音识别比作为“机器的听觉系统”。语音识别技术主要包括特征提取技术、模式匹配准则及模型训练技术三个方面。

语音识别在移动终端上的应用最为火热，语音对话机器人、语音助手、互动工具等层出不穷。

目前，国外的应用一直以苹果的siri为龙头。而国内方面，盛大、捷通华声等系统都采用了最新的语音识别技术，市面上其他相关的产品也直接或间接嵌入了类似的技术。

预计未来10年内，语音识别技术将进入工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域。

6.人工智能(AI)

人工智能(Artificial Intelligence)是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。

用途范围︰机器翻译，智能控制，专家系统，机器人学，语言和图像理解，遗传编程机器人工厂，自动程序设计，航天应用，庞大的信息处理，储存与管理，执行化合生命体无法执行的或复杂或规模庞大的任务等。

7.数字汇流

数字汇流是对未来冲击最大的一项趋势,就是将上述六项趋势合并起来的效果。

比如,像是84亿个物联网设备,可用块链技术加强安全性；智慧城市通过物联网就能产生海量数据，这些数据需要由人工智能进行分析；虚拟现实和语音识别也需要通过人工智能不断学习,这些科技发展息息相关,相辅相成,所以数字汇流是最重要的趋势。

**三、大数据时代下大数据技术**

大数据技术起源于2000年前后互联网的高速发展。伴随着时代背景下数据特征的不断演变以及数据价值释放需求的不断增加，大数据技术已逐步演进针对大数据的多重数据特征，围绕数据存储、处理计算的基础技术，同配套的数据治理、数据分析应用、数据安全流通等助力数据价值释放的周边技术组合起来形成的整套技术生态。

如今，大数据技术的内涵伴随着大数据时代的发展产生了一定的演进和拓展，从基本的面向海量数据的存储、处理、分析等需求的核心技术延展到相关的管理、流通、安全等其他需求的周边技术，逐渐形成了一整套大数据技术体系，成为数据能力建设的基础设施。伴随着技术体系的完善，大数据技术开始向着降低成本、增强安全的方向发展。

1.大数据基础技术

大数据基础技术为应对大数据时代的多种数据特征而产生。大数据时代，数据量大、数据源异构多样、数据实效性高等特征催生了高效完成海量异构数据存储与计算的技术需求。在这样的需求下，面对迅速而庞大的数据量，传统关系型数据库单机的存储及计算性能有限，出现了规模并行化处理(Massively Parallel Processing，MPP )的分布式计算架构；面向海量网页内容及日志等非结构化数据,出现了基于Apache Hadoop和Spark生态体系的分布式批处理计算框架；面向对于时效性数据进行实时计算反馈的需求，出现了Apache Storm、Flink 和 Spark Streaming等分布式流处理计算框架。

2.数据管理类技术

数据管理类技术提升数据质量与可用性。在较为基本和急迫的数据存储、计算需求已在一定程度上得到满足后﹐如何将数据转化为价值成为了下一个最主要需求。最初，企业与组织内部的大量数据因缺乏有效的管理，普遍存在着数据质量低、获取难、整合不易、标准混乱等问题，使得数据后续的使用存在众多障碍。在此情况下，用于数据整合的数据集成技术，以及用于实现一系列数据资产管理职能的数据管理技术随之出现。

3.数据分析应用技术

数据分析应用技术发掘数据资源的内蕴价值。在拥有充足的存储计算能力以及高质量可用数据的情况下，如何将数据中蕴涵的价值充分挖掘并同相关的具体业务结合以实现数据的增值成为了关键。用以发掘数据价值的数据分析应用技术，包括简单统计分析与可视化展现技术，及以传统机器学习、基于深度神经网络的深度学习为基础的挖掘分析建模技术纷纷涌现，帮助用户发掘数据价值并进一步将分析结果和模型应用于实际业务场景中。

4.数据安全流通技术

数据安全流通技术助力安全合规的数据使用及共享。在数据价值的释放的同时，数据安全问题也愈加凸显，数据泄露、数据丢失、数据滥用等安全事件层出不穷，对国家、企业和个人用户造成了恶劣影响，如何应对大数据时代下严峻的数据安全威胁，在安全合规的前提下共享及使用数据成为了备受瞩目的问题。访问控制、身份识别、数据加密、数据脱敏等传统数据保护技术正积极向更加适应大数据场景的方向不断发展，同时，侧重于实现安全数据流通的隐私计算技术也成为了热点发展方向。

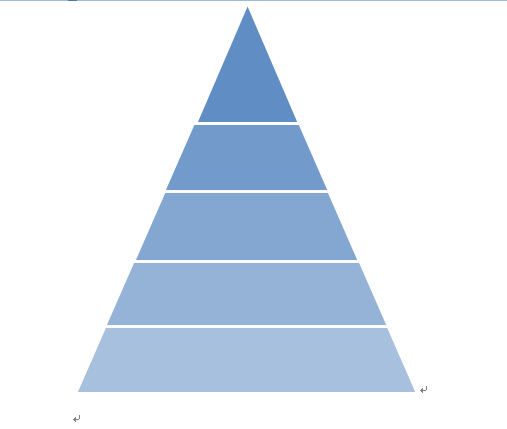
**【经典案例】通信大数据行程卡有效助力疫情防控**

2020年2月，在工业和信息化部领导下，中国信通院、中国电信、中国移动、中国联通共同推出“通信大数据行程卡”，并在国务院客户端微信小程序上线，为全国16亿手机用户免费提供14天内所到地市信息的查询服务。

通信大数据行程卡的技术原理是分析手机“信令数据”，获取用户设备所在位置信息。信令数据的采集、传输和处理过程自动化，有严格的安全隐私保障机制，不与其它个人信息进行匹配，查询结果实时可得且数据全国通用。行程卡APP2.0版本还引入了低功耗蓝牙技术(BLE)，为用户提供新冠肺炎密切接触者追踪提醒功能。截至2020年11月，累计查询量已超过42亿次。

**四、大数据时代下大数据产业**

理解大数据产业，首先要搞清楚大数据技术、大数据资源和大数据产业的关系。大数据技术是指采集获取汇聚处理数据的技术总称，包括数据的采集、数据预处理、分布式存储、数据库、数据仓库、机器学习、并行计算、可视化等；而大数据资源是指数据本身，是从资源利用的角度出发的，主要关心数据从哪里来、如何确权、如何治理、如何共享、如何交易流通、如何分析利用等问题。大数据产业则利用大数据技术作用于大数据资源，解决产业化落地问题。

****

**大数据软件**

**大数据服务**

**大数据硬件**

**大数据核心**

**产业**

**大数据**

**产业**

是企业购买的用于支撑企业内部大数据相关业务的硬件基础设施，包括服务器、网络设备和存储设备等。大数据系统-般使用通用硬件，技术上并无特殊性同地包含部分经过专门优化的一体机设备。

包括大数据分析公有云服务及大数据系统相关的咨询、设计和系统集成服务等。

指以Hadoop、Spark等技术架构为基础的新型分布式平台软件,以及基于这类平台软件的应用软件。

指大数据直接相关的软件.硬件及专业服务。大数据市场规模的计算口径,相应的包含大数据软件硬件及服务的直接产值，不含因大数据应用产生的附加值。

指以数据生产、采集存储、加工、分析.服务为主的相关经济活动，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。

图3 大数据产业趋势图

（资料来源：北大纵横管理咨询集团）

因此，大数据产业是以大数据技术为基础对数据生产、采集、储存、加工、分析、服务为主的相关经济活动，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。之所以能够形成大数据产业，主要原因是大数据是一种生产要素。在数字经济时代，数据如同农业经济时代和工业经济时代中的土地、劳动力、资本和石油一样成为关键生产要素。数据所蕴含的巨大创新价值，对于商业模式创新、产业数字化转型、经济高质量发展、治理能力现代化乃至重大科学发现都是必不可少的。

随着人类社会步入数据驱动的数字经济时代，数据要素进一步提升了全要素生产率。在数字社会，数据具有基础性战略资源和关键性生产要素的双重角色。一方面，有价值的数据资源是生产力的重要组成部分，是催生和推动众多数字经济新产业、新业态、新模式发展的基础。另一方面，数据区别于以往生产要素的突出特点是对其他要素资源的乘数作用，可以放大劳动力、资本等要素在社会各行业价值链流转中产生的价值。善用数据生产要素，解放和发展数字化生产力，有助于推动数字经济与实体经济深度融合，实现高质量发展。

“十三五”以来，我国大数据蓬勃发展，融合应用不断深化，数字经济量质提升，对经济社会的创新驱动、融合带动作用显著增强。工业和信息化部运行监测协调局发布的数据显示，2019年我国以云计算、大数据技术为基础的平台类运营技术服务收入2.2万亿元,其中，典型云服务和大数据服务收入达3284亿元，提供服务的企业达2977家3，大数据产业发展日益壮大[[1]](#footnote-2)。

**【素质拓展】大数据核心产业链**

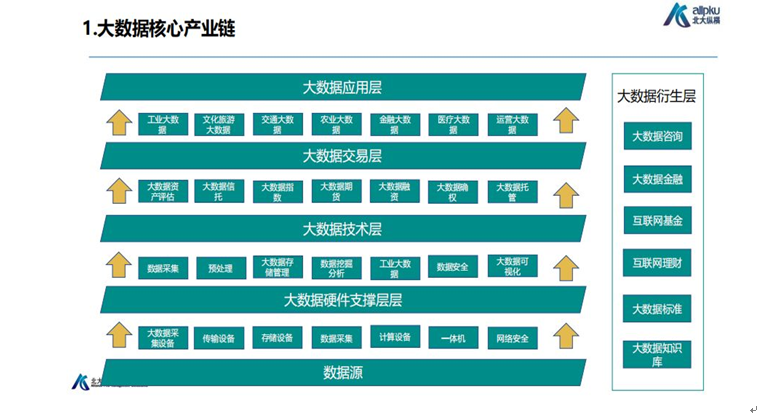


图2 大数据核心产业链

（资料来源：北大纵横管理咨询集团）

**实训练习**

**应知考核**

**一、单项选择题**

1.( )指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。

A.数据 B.数字 C.文字 D.信息

2.（ ）是信息的表现形式和载体，可以是符号、文字、数字、语音、图像、视频等。

A.数字 B.数字 C.数据 D.信息

3.（ ）指新的数据产生或现有数据内容发生显著改变或更新的阶段。

A.数据采集 B.数据存储 C.数据处理 D.数据传输

4.（ ）指非动态数据以任何数字格式进行物理存储的阶段。

A.数据采集 B.数据存储 C.数据处理 D.数据传输

5.（ ）指组织机构在内部针对动态数据进行的一系列活动的组合。

A.数据采集 B.数据存储 C.数据处理 D.数据传输

6.（ ）指数据在组织机构内部从一个实体通过网络流动到另一个实体的过程。

A.数据采集 B.数据存储 C.数据处理 D.数据传输

7.（ ）指数据经由组织机构内部与外部组织机构及个人交互过程中提供数据的阶段。

A.数据采集 B.数据交换 C.数据处理 D.数据传输

8.（ ）指通过对数据及数据的存储介质通过相应的操作手段，使数据彻底丢失且无法通过任何手段恢复的过程。

A.数据采集 B.数据交换 C.数据处理 D.数据销毁

9.大数据最明显的特点就是（ ）。

A.数据体量大 B.数据类型繁多 C.价值密度低 D.处理速度快

10.数据的（ ）是大数据区分于传统数据挖掘的显著特征。

A.数据体量大 B.数据类型繁多 C.价值密度低 D.处理速度快

11.（ ）是大数据分析最重要的应用领域。

A.可视化分析 B.语义引擎 C.预测性分析能力 D.数据质量管理

12.（ ）是机器学习的成果之一。

A.可视化分析 B.语义引擎 C.预测性分析能力 D.数据质量管理

**二、多项选择题**

1.数据按性质分为（ ）。

A.定位数据 B.定性数据 C.定量数据 D.定时数据

2.数据按按表现形式分为（ ）。

A.定位数据 B.数字数据 C.模拟数据 D.定时数据

3.数据按按记录方式可分（ ）。

A.表格 B.影像 C.磁带 D.纸带

4.数据按数字化方式分（ ）。

A.定位数据 B.定性数据 C. 矢量数据 D. 格网数据

5.《联邦数据战略》确立了40项数据管理的具体实践目标,包括（ ）三个层次。

A.重视数据并促进共享 B.保护数据资源 C.有效使用数据资源 D.推进数据的流动

6.《欧盟数据战略》提出了四大支柱性战略措施指的是（　　）。

A.构建跨部门治理框 B.加强数据投入 C.提升数据素养 D.构建数据空间

7.大数据的特征包括（ ）。

A.数据体量大 B.数据类型繁多 C.价值密度低 D.处理速度快

8.语音识别技术主要包括（ ）。

A.特征提取技术 B.模式匹配准则 C.模型训练技术 D.数据提取技术

**三、判断题**

1.数据本身没有意义，数据只有对实体行为产生影响时才成为信息。( )

2.数据只能是连续的值，不可以是离散的。（ ）

3.大数据（big data）是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要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。（ ）

4.大数据更强调批量式分析而非实时分析。（ ）

5.大数据对象既可能是实际的、有限的数据集合，也可能是虚拟的、无限的数据集合。

6.苹果的Siri和谷歌的Google Now都采用了语义引擎。

**应会考核**

1.阐述数据是类型。

2.阐述大数据的特征。

1. 数据来源于中国通信院《大数据白皮书（2020年）》 [↑](#footnote-ref-2)