数据分析

任课老师：

目录

项目一
大数据基础

任务1 大数据基础认知

任务2 数据库基础认知

什么是大数据

任务1大数据基础知识

一、数据与大数据

在目前高速发展的时代，科技发达、信息流通，大数据就是这个时代的产物。所谓数据，就是用来描述事物的符号或代码。在计算机系统中，各种数字、文字、字母、符号、图形、图像、音频等都被统称为数据，而数据通过一定的手段加工就得到了我们平时所说的信息。

在现实生活中，**数据无处不在**，例如，学生的学号、成绩、身份证号、快递单号等都包含了大量的数据。将数据整理成信息，可以分析出行、销售、生产等方向，从而达到最优组合。以天气为例，通过风速、湿度、云层的移动轨迹等数据进行分析，最终可以获得区域天气的相关信息，同时也可以较为精准地预测未来一定时间内的天气情况，这就是大数据时代最初的展现。所以，人们是通过数据来得到信息，从而认识世界的。

大数据 (big data或 mega data)，或称为巨量资料，指的是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产，是大的数据量与现代信息技术环境相结合涌现的结果。换而言之，它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。

大数据通常是以多元的形式从许多来源搜集庞大数据组，往往具有多样性。比如，在零售企业销售的数据库中，数据来源可能来自社交网络、电子商务网站、顾客来访记录等。例如天气的相关数据，要具有时效性，也就是及时性。除此之外，对于数据来说，最重要的一个特性就是准确性。

（一）大数据的起源

大数据源于互联网的发展。互联网运行产生了海量的信息数据，互联网的快速发展创造了大数据应用的规模化环境，互联网企塑也开发了处理软件，相对应地，大数据计算技术完美地解决了海量数据的收集、存储、计算和分析的问题。

大数据以多元形式产生。数据并非单纯地指人们在互联网中发布的信息。比如，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、湿度、温度乃至空气中化学物质的变化，同时也产生了海量的数据信息。

简单来说，当数据累积到一定数量时，通过数学模型进行建模分析，就是大数据分析的雏形。例如，在数学中，我们学到的方程就是线性回归模型的基础表现。以学生的学习成绩来说，假设影响成绩的因素有且只有学习的时长，那么可以将学习时长设为自变量X，成绩设为因变量Y。这样我们可以通过长期收集的数据推算出自变量的系数，得到一个线性方程，如图1-1所示。

如图中所展示的公式，代入不同的学习时长，就可以得到一个对应的学习成绩。通过这个简单的模型，我们可以知道，如果希望考试及格，需要分配多少时间来学习；如果希望成绩优秀(80分以上)，需要多少时间来学习。大数据时代，让我们可以更加合理地分配资源，获得更优的结果，最终推动了时代的发展。所以，未来将不再是IT时代，而是DT(Data Technology) 的时代，也就是数据科技所带来的新的发展。

（二）大数据的分类

数据根据其内容不同，　一般分为：

结构化数据、非结构化数据和半结构化数据三类。

1. 结构化数据

能够用数据或统一的结构加以表示的信息，称为结构化数据，如数字、符号等。传统的关系数据模型——行数据，存储于数据库，可用二维表结构表示，如图1-2所示。简单来说，结构化数据就是数据库，如企业ERP、 财务系统等。

员工ID	员工姓名	性别	部	Salary_In_lacs
2365	Rajesh Kulkarni	男	金融	650000
3398	Pratibha Josh	女	管理员	650000
7465	Shushil Roy	男	管理员	500000
7500	Shubhojit Das	男	金融	500000
7699	Priya Sane	女	金融	550000

图1-2 结构化数据

2. 非结构化数据

非结构化数据与结构化数据比对如图1-3所示。非结构化数据是指其字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库，用它不仅可以处理结构化数据(如数字、符号等信息)，而且更适合处理非结构化数据(如全文文本、图像、声音、网页、影视、超媒体等信息)。

图1-3　非结构化数据与结构化数据比对图

3. 半结构化数据

半结构化数据举例如图1-4所示。半结构化数据相对复杂，是介于完全结构化数据(如关系型数据库)和完全无结构的数据(如声音、图像文件等)之间的数据，XML、HTML文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

```
<person>
        <name>A</name>
        <age>13</age>
        <gender>female</gender>
</person>



<person>
        <name>B</name>
        <gender>male</gender>
</person>
```

图1–4半结构化数据举例

大数据的核心价值是**预测**，即在对海量数据进行存储和分析，运用数学算法预测事情发生的可能性。大数据已经渗透到了每一个行业和业务职能领域，逐渐成为重要的生产因素，而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。大数据是继云计算、物联网之后T 产业又一次颠覆性的技术变革。云计算为数据资产提供了技术支持手段，但数据才是真正有价值的资产。大数据技术的战略意义在于对数据进行专业化处理。没有互联网、云计算、物联网、移动终端与人工智能组合的环境，大数据毫无价值。

二、大数据发展之路

2005年，Hadoop 项目诞生，后因技术高效性，被Apache Software Foundation公司引入成为开元应用。

2008年年末，“大数据”得到部分美国知名计算机科学研究人员认可，《自然》杂志专刊提出Big Data概念。

2009年，印度、联合国、美国和欧洲一些领先研究机构进一步研究“大数据”，引起高潮。

2010年，肯尼斯·库克尔发表大数据专题报告，“大数据”词汇诞生。

2011年，大数据能力量现、内容得到丰富，得到进一步发展。

2012年，美国第一家大数据软件公司上市，联合国出台大数据白皮书，阿里巴巴全面推进“数据分享平台”战略，大数据价值得到进一步挖掘。

2015年，国务院正式印发《促进大数据发展行动纲要》，标志着大数据正式上升为国家战略。

2016年，大数据“十三五”规划出台，推动大数据在工业研发、制造、产业链全流程及服务业的发展。

2017年1月，工信部发布了《大数据产业发展规划(2016—2020年)》，进一步明确了促进我国大数据产业发展的主要任务、重大工程和保障措施。

2017年10月，中共十九大报告指出：加快建设制造强国，加快发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合。

（二）大数据与信息技术深度融合

大数据离不开云处理，云处理为大数据提供了可扩展的基础设备，是产生大数据的平台之一。自2013年开始，大数据技术已经和云计算技术深度融合。

物联网、云计算、移动互联网、车联网、手机、平板电脑、PC 以及遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式，包括网络日志、 RFID、 传感器网络、社会网络、社会数据、互联网文本和文件、互联网搜索索引、天文学、大气科学、生物地球学、军事侦察、医疗记录、大规模的电子商务等。

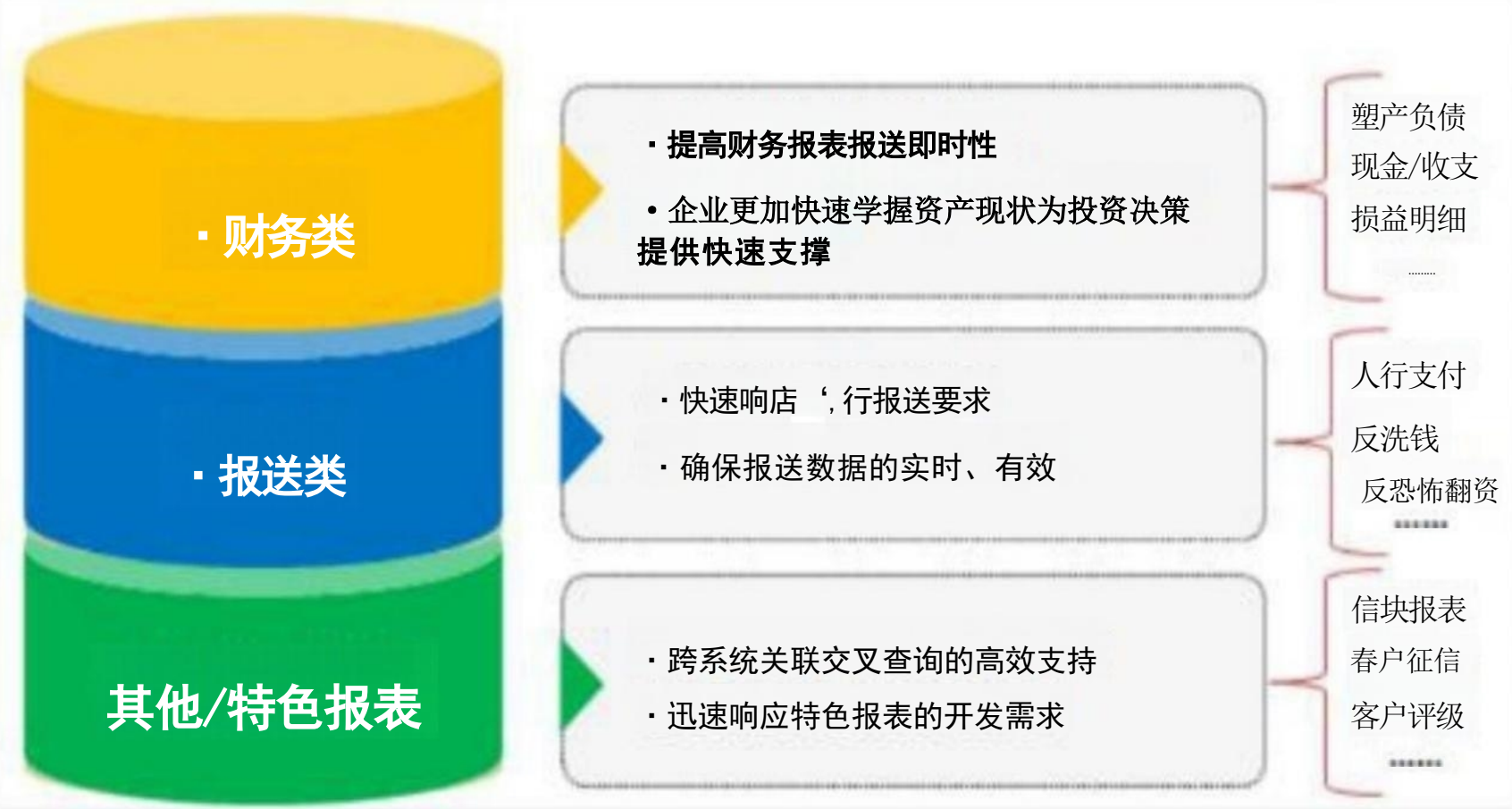

图1-5大数据的服务优势

(四)十大数据挖掘领域的经典算法

(1)C4.5 算法：是机器学习算法中的一种分类决策树算法。

(2)K-Means 算法：是一个聚类算法。

(3)Support Vector Machine(支持向量机):是一种监督式学习的方法。

(4)Apriori 算法：是一种最有影响的挖掘布尔关联规则频繁项集的算法。

(5)最大期望(EM) 算法：是在概率模型中寻找参数最大似然估计的算法。

(6)PageRank 算法：根据网站的外部链接和内部连接的数量和质量来衡量网站的价值。

(7)Adaboost 算法：是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。

(8)K 最近邻(KNN) 分类算法：如果一个样本在特征空间中，假定K个最相似(即特征空间中最近邻)的样本区间内，大多数属于某一类别，则该样本也属于这个类别。

(9)朴素贝叶斯模型：该模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。

(10)分类与回归树 (Classification and Regression Tree,CART): 在分类树下面有两个关键的思想， 一是关于递归地划分自变量空间的想法，二是用验证数据进行剪枝。

（五）常用的大数据分析工具

(1)DBE　财务大数据实践教学平台。 DBE 财务大数据实践教学平台是结合Python 数据获取、数据清洗、MYSQL数据存储、商业可视化分析软件、数据挖掘等多类大数据工具的综合软件，采用案例化教学模式呈现企业财务内部经营分析、外部投资决策实战应用场景等，是目前最为全面的分析软件之一。

(2)RapidMiner。RapidMiner　通过可视化程序进行操作，能够手动运作、分析和建模。它通过开源平台、机器学习和模型部署来提高数据工作效率。统一的数据科学平台可加速从数据准备到实现的分析工作流程，极大地提高技术人员的效率，是最易于使用的预测分析软件之一。

(3)Power BI。Microsoft Power BI同时提供本地和云服务。它最初是作为Excel插件引入的，不久Power BI凭借其强大的功能开始普及。目前，它被视为商业分析领域的软件领导者。它提供了数据可视化和BI功能，使用户可以轻松地以更低的成本实现快速、明智的决策，用户可协作并共享自定义的仪表板和交互式报告。

【任务要求】

（1）将从网站上采集的资产负债表、利润表和现金流量表上传至分析云，为下一步数据分析做准备，任务流程如图1-6所示。

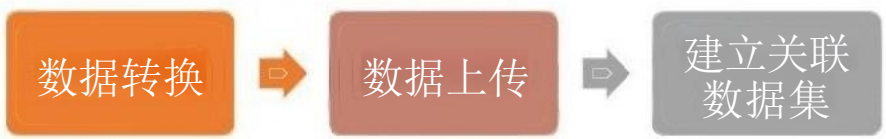

图1-6数据处理的流程

（2）建立关联数据集。

（3）进行可视化图表设计与呈现。

（4）构建数据管理驾驶舱。

【任务实施】分析云操作

小结

01 大数据的基本概念

02 数据的类型

03 数据处理流程

04 通过任务熟悉分析云平台

课后讨论

请每位同学查阅资料，就下面四个内容展开讨论：

1、 你所读的学院(专业)主要涉及的就业方向?

2、 你所读的学院(专业)可能涉及的大数据应用有哪些?

财务大数据分析

任课老师：